




Analyzing the Factor Structure of the Toronto Empathy Questionnaire: Dimensionality, Reliability, Validity, Measurement Invariance and One-Year Stability of the German Version

Tobias Janelt, Tobias Altmann, R. Nathan Spreng & Marcus Roth


To cite this article: Tobias Janelt, Tobias Altmann, R. Nathan Spreng & Marcus Roth (2024) Analyzing the Factor Structure of the Toronto Empathy Questionnaire: Dimensionality, Reliability, Validity, Measurement Invariance and One-Year Stability of the German Version, Journal of Personality Assessment, 106:2, 230-241, DOI: [10.1080/00223891.2023.2224873](https://doi.org/10.1080/00223891.2023.2224873)


To link to this article: <https://doi.org/10.1080/00223891.2023.2224873>

 View supplementary material 

 Published online: 03 Jul 2023.

 Submit your article to this journal 

 Article views: 440





 View related articles 

 View Crossmark data 

 Citing articles: 1 View citing articles 



Analyzing the Factor Structure of the Toronto Empathy Questionnaire: Dimensionality, Reliability, Validity, Measurement Invariance and One-Year Stability of the German Version

Tobias Janelt¹ , Tobias Altmann¹ , R. Nathan Spreng²  and Marcus Roth¹ 

¹Department of Psychology, University of Duisburg-Essen, Essen, Germany; ²Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada

ABSTRACT

In the face of heterogeneity in the measurement of empathy, the *Toronto Empathy Questionnaire* (TEQ; Spreng et al., *Journal of Personality Assessment*, 91(1), 62–71 (2009)) was developed as a brief unidimensional tool by statistically forming a consensus from existing measures of the construct. The present study aimed to (1) validate a German version of the TEQ, and (2) contribute empirical evidence to the ongoing debate regarding a singular versus multidimensional factor structure of the TEQ. One cross-sectional and two longitudinal studies were performed, with a total of 1,075 participants. Our initial exploratory factor analyses suggested either a one- or a two-factor structure (with the two-factors clustering straight and reverse-scored items); the two-factor model outperformed the one-factor model using confirmatory factor analyses. However, after negated items were replaced by positively reworded alternatives, both models fit the data equally well. A comparison of the correlation patterns with numerous external measures indicated that a second factor of the TEQ is a methodological artifact of item wording. Finally, a unidimensional TEQ scale showed sufficient internal consistency, two-week test-retest reliability, one-year stability, as well as convergent and discriminant validity with measures of empathy, emotion recognition, emotion regulation, altruism, social desirability, and the Big Five personality traits.

ARTICLE HISTORY

Received 16 December 2022
Accepted 31 May 2023

The concept of empathy plays an essential role for human communication and social interaction. Originally referring to Titchener's (1909) translation of the German word *Einfühlung* (literally *feeling into*), empathy advanced to a construct of enormous scientific attention (for a review see Bošnjaković & Radionov, 2018). Nevertheless, a consensus regarding its conceptualization remains absent (Hall & Schwartz, 2019).

Most authors share a basic conceptual dualism of the empathy construct, namely the distinction between an affective and a cognitive component (Hall & Schwartz, 2019). *Affective empathy* implies experiencing emotions that are similar, or parallel, to one's perception of another person's emotions so that one's emotions are more congruent with the other person's situation than with one's own situation (Hoffman, 2000). *Cognitive empathy* can be defined as the intellectual or imaginative apprehension of another's emotional state.

Many authors (e.g., Baron-Cohen & Wheelwright, 2004; Bošnjaković & Radionov, 2018) associate—or even equate—cognitive empathy with the concept of theory of mind, i.e., the “ability to think about the contents of other minds” (Lawrence et al., 2004, p. 911).

Unsurprisingly, the conceptual heterogeneity of the empathy construct is mirrored in a multiplicity of instruments for *assessing* these diverse constructs, most of them self-report measures (Hall & Schwartz, 2019). However generous the offer of these various suggestions may appear, the psychometric quality of

most available measures is rather insufficient: As recently reviewed by Lima and Osório (2021), most instruments did not reach a consensus concerning dimensionality, while evidence for convergent validity is likewise limited in most cases. Obviously, there is still a glaring research gap regarding the psychometrically sound measurement of empathy *via* self-report. In order to address this gap, the present study aims to contribute evidence to clarifying the dimensionality, as well as convergent validity and reliability, of a specific instrument, namely the *Toronto Empathy Questionnaire* (Spreng et al., 2009).

The Toronto empathy questionnaire

In light of the conflicting and heterogeneous measurement approaches occurring in the empathy field, comparison between studies remains difficult (Gerdes et al., 2010). To address this issue, Spreng et al. (2009) presented a factor-analytic approach, examining what all of the competing empathy measures' items had in common, to create a new and parsimonious tool that captured empathy at the widest range, which they called the *Toronto Empathy Questionnaire* (TEQ). Since its development in 2009, the TEQ has received rising attention by the scientific community with over 500 empirical papers (collected *via* PsycINFO and Google Scholar) using it in one of at least nine language versions.

Scale development of the TEQ

The TEQ was constructed by conducting an exploratory factor analysis on items pooled from several self-report measures of empathy, including the Interpersonal Reactivity Index (IRI; Davis, 1980), Hogan's (1969) Empathy Scale, the Questionnaire Measure of Emotional Empathy (Mehrabian & Epstein, 1972), the Balanced Emotional Empathy Scale (Mehrabian, 2000), and others (see Spreng et al., 2009, for details). Items were reworded to consistently assess frequency of behavior and used a five-point Likert scale. Participant responses were run in an exploratory factor analysis that forced all items to load onto a single factor. Subsequently, items were selected based on their psychometric properties, resulting in the final scale containing eight positively and eight negatively worded items. Item contents capture empathy as a primarily affective phenomenon, e.g., "When someone else is feeling excited, I tend to get excited too." (Item 01).

Factor structure of the TEQ

The initial validation study of the TEQ demonstrated its clear unidimensional structure in two independent samples (Spreng et al., 2009) using exploratory factor analyses (EFA). This one-factor model was, according to fit indices in confirmatory factor analyses (CFA), replicated by three subsequent studies, which adapted the TEQ into other languages: Czech (Novak et al., 2021), Turkish (Totan et al., 2012) and Greek (Kourmoussi et al., 2017).

In contrast, Chiorri (2016) found two correlated factors for the Italian version of the TEQ, clustering straight (= positively worded/scored) and reverse (= negatively worded/scored) items, which he labeled as *empathy* and *callousness*, respectively. The two-factor model provided better model fit than a one-factor model in a replication sample. Other language validation studies proposed a three-dimensional structure for the TEQ (Ursoniu et al., 2021, Romanian; Xu et al., 2020, Chinese; Yeo & Kim, 2021, Korean). However, these studies' results also allow different conclusions: The CFA by Xu et al. (2020) demonstrated the superiority of the three-factor model over the one- and two-factor model although it might be questionable whether a factor containing only two items (08 and 09) can represent a meaningful facet of empathy. The other two factors clustered the (remaining) straight and reverse items, respectively, with the only exception of item 02. Ursoniu et al. (2021) used EFA and interpreted the scree plot to suggest a three-factor structure (however, visual examination also supports a one factor structure, contrary to the authors, conclusions). Moreover, the fit indices of the three-factor model tested within CFA appear, according to our interpretation, rather unsatisfactory. Nevertheless, item assignments appear interesting: Except items 01, 08 and 09, all positively worded items were assigned to one factor, while all negatively worded items were clustered by the other two factors. Yeo and Kim (2021) used CFA to demonstrate the superiority of a three-factor model over the one-factor model (with the latter showing poor fit), but did not test the fit of other models, such as a two-factor model.

Taken together, literature concerning the TEQ's factor structure is inconsistent: While four studies propose a

one-factor structure, four studies propose multidimensional solutions for the TEQ. Even taking into account the different languages, this inconsistency represents a serious issue, since the instrument's construction and scoring is based on a clear unidimensional structure (Spreng et al., 2009).

Hence, two questions arise: At first, what could (besides language/cultural context) be responsible for this inconsistency? Secondly, what could be done to regain a consistent factor structure? A recurring pattern in the multidimensional structure is the separation of straight and reverse items. The question arises whether this pattern either points out that both TEQ item types (straight and reverse) capture two substantially different constructs or merely reflects a methodological artifact of item wording. Chiorri (2016) addressed this question by building two subscales of the TEQ, containing the eight straight and eight reverse items, respectively. Comparing the correlation patterns of both subscales with several (overall 16) external scales only revealed a significant difference between both subscales for the associations with *one* measure (Emotional Intelligence Scale; Schutte et al., 1998), questioning the justifiability of a model with two distinct constructs. However—even if the reverse items produce a methodological artifact—does this phenomenon represent a crucial obstacle for a clear one-factor structure?

This question was addressed by a recent validation study (Novak et al., 2021), in which the authors positively reformulated negated TEQ items. EFA showed that reworded items had higher factor loadings and communalities than the originally negated ones and increased the scale's internal consistency. CFA indicated that the rather poor fit of the one-factor model found for the scale including negated items turned into an excellent fit when considering the reworded alternatives instead. Since merely the item direction (but not item content) was changed by the authors, these results suggest a methodological artifact of item wording. However, as promising as these findings may look, they are limited to one study and one cultural context (Czech). Replication studies are required to test the generalizability of these findings (e.g., for other language environments).

Convergent and discriminant validity

Several studies demonstrated strong positive associations between the TEQ and the IRI subscale *empathic concern* (EC) and moderate to strong positive associations between the TEQ and the IRI subscale *perspective taking* (PT; e.g., Baldner & McGinley, 2014). This pattern is consistent with the idea that the TEQ might, on the one hand, tap empathy as a primarily emotional process, but, on the other hand, still captures empathy on a broad level, which includes the cognitive dimension (Spreng et al., 2009).

Associations between the TEQ and several other self-report empathy measures have been reported. For instance, the TEQ was found to correlate positively with the Empathy Quotient (Baron-Cohen & Wheelwright, 2004) and the Basic Empathy Scale (BES; Jolliffe & Farrington, 2006), as also shown by, e.g., Baldner and McGinley (2014), the Jefferson Scale of Physician Empathy (Hojat et al., 2001), as shown by Leloirain et al. (2013), as well as with constructs

related to empathy, such as Theory of mind (Kaviani & Kinman, 2017), compassion (Novak et al., 2021), emotional intelligence (Chiorri, 2016) and altruism (Mulet et al., 2022).

Spreng et al. (2009) also reported positive associations with task-based measures of interpersonal perception, namely the Reading the Mind in the Eyes Test-Revised (Baron-Cohen et al., 2001). Studies found negative associations between the TEQ and psychopathy (e.g., Luckhurst et al., 2017). Several studies analyzed the correlation pattern between the Big Five personality traits and the TEQ (e.g., Chiorri, 2016; Stewart et al., 2019), which can be summarized as following: Mostly there were no or weak associations with Extraversion, Conscientiousness, Neuroticism and Openness, while Agreeableness showed significant weak to strong positive associations.

The present study

The present study pursues two major goals. First, the present research aims to validate a German translation of the TEQ (TEQ-D) to enable the use of the instrument for the German-speaking community. Second, the present research aims to contribute to the cross-lingual analysis of the TEQ which has been inconsistent regarding the instrument's overall factor structure.

In Study 1, the dimensionality of a first version of the TEQ-D will be analyzed. Study 2 will examine its one-year longitudinal construct stability across four measurement occasions and the association with a task-based measure of emotion recognition. In addition to further psychometric and factor analyses, Study 3 will specifically focus on convergent and discriminant validity of the TEQ-D, investigate its two-week test-retest-reliability as well as the influence of item negations on model fit when using positively reworded items.

Study 1

Study 1 describes the translation of the TEQ and analyzes the factor structure of the translated version (TEQ-D) *via* exploratory as well as confirmatory methods using existing data already collected by the authors. Within CFA, the fit of the originally proposed (Spreng et al., 2009) one-factor model as well as the fit of a model with two correlated factors clustering straight and reverse items, respectively (Chiorri, 2016), was tested. Based on the previous literature, we expected to find superior fit of the two-factor model compared to the one-factor model.

Methods

Participants and data collection

Participants were recruited *via* flyers, e-mails and postings in social media, mainly distributed to students at the university of Duisburg-Essen (Germany). The data were collected *via* computers and in a controlled laboratory setting in groups of up to twelve persons. A total of 745 subjects (76.9% female) aged between 18 and 69 years ($M=26.1$, $SD=8.7$) completely participated (incomplete responses were deleted). 86.4% were students, 13.4% were employed,

0.1% were unemployed. 91.9% were highly educated (A-levels or higher), while 7.2% reported "O-levels" and 0.8% "high school" as their highest level of education. The analysis was preregistered in the Open Science Framework [<https://osf.io/yvh7j>].

Measure

The original TEQ was independently translated into German by two researchers, both native speakers of the German language and fluent in the English language. Differences were discussed, resulting in a provisional German version. The latter was translated back into the English language by a professional translator (English native speaker). Then another professional translator (German native speaker) compared the back translated TEQ with the original TEQ. Differences were discussed and the provisional German version was slightly adjusted. The resulting TEQ-D is (analogous to the original TEQ) a 16-item empathy questionnaire containing eight straight and eight reverse items as described above. The German items can be found in the [supplementary material](#). Answers are given on a five-point Likert scale (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, 5 = *always*).

Data analysis

The sample was randomly split in half. In one half, an EFA was performed. A CFA was performed on the other half of the sample. The EFA was conducted with the R package *psych* (Revelle, 2020), using principle axis factoring method and, oriented to previous studies which compared a one- and two-factor model for the TEQ (Chiorri, 2016; Novak et al., 2021), oblique rotation (Oblimin). Parallel analysis (Horn, 1965) was performed using both principle component analysis (PCA) and principle axis factor analysis, based on 1000 bootstrapped samples (following Buja & Eyuboglu, 1992). CFA was performed with the R package *lavaan* (Rosseel, 2012) using the Maximum Likelihood estimator. Differences between information criteria were computed and interpreted according to Burnham and Anderson (2004) and Raftery (1995).

Results

EFA

The factorability of data was good (Kaiser-Meyer-Olkin: .82). The initial PCA yielded five components with an eigenvalue > 1 (eigenvalues: 3.91, 1.39, 1.17, 1.08, 1.02) explaining 54% of the variance in the data. The scree plot suggested the extraction of one component (see supplement 2), consistent with Velicer's (1976) Minimum average partial test (average squared correlations of 0, 1, 2, 3 and 4 components, respectively: .042, .011, .015, .021, .029). In contrast, Parallel analysis (see also supplement 2) suggested the extraction of two components.

Subsequently, items were forced to load onto either one or two factors. The models explained 20% and 24% of variance in the data, respectively. The two factors were negatively correlated, $r=-.59$. As can be seen in [Table 1](#), the two factors clustered straight and reverse items, respectively, and could conceptually be labeled as *empathy* and *callousness*.

Table 1. Factor loadings and communalities for the one- and two-factor solution of the preliminary TEQ-D (EFA).

Item	One-factor		Two-factor		
	λ_{total}	h^2	λ_{empathy}	$\lambda_{\text{callousness}}$	h^2
01	.20	.0416	.24	.02	.0500
03	.57	.3286	.56	−0.06	.3581
05	.61	.3711	.35	−0.33	.3694
06	.65	.4178	.59	−0.12	.4371
08	.43	.1849	.59	.14	.2725
09	.36	.1305	.41	.02	.1547
13	.60	.3637	.57	−0.09	.3852
16	.61	.3752	.70	.04	.4600
02	−0.08	.0062	−0.04	.05	.0062
04	−0.43	.1884	−0.12	.39	.2193
07	−0.42	.1741	−0.02	.49	.2508
10	−0.25	.0610	.06	.37	.1127
11	−0.30	.0908	.04	.41	.1513
12	−0.47	.2255	−0.09	.47	.2830
14	−0.16	.0259	< .01	.20	.0380
15	−0.46	.2088	.03	.59	.3368

Notes. λ : factor loading, h^2 : communality. In bold the higher absolute value of factor loading of each item in the two-factor solution. Items are sorted by their wording/scoring (first eight positive, last eight negative). The preliminary TEQ-D includes negated items.

CFA

The factorability of data was good (Kaiser-Meyer-Olkin: .85). Three models were specified. The first model corresponds to the originally proposed one-factor model (Spreng et al., 2009). The second model corresponds to the two-factor model suggested by Chiorri (2016) with two correlated factors represented by straight and reverse scored items. The third model entails a general factor represented by all items and a *reverse item method factor* represented by all reverse scored items (RMF model; correlation between factors constrained to zero). Figure 1 depicts the three models.

As shown in Table 2, the two-factor model (correlation between factors: $r = -.714$) and the RMF model demonstrated reasonable absolute fit to the data, while the one-factor model did not. According to χ^2 differences, the one-factor model was outperformed by both the two-factor model, $\Delta\chi^2$ ($\Delta df = 1$) = 55.201, $p < .001$, and the RMF model, $\Delta\chi^2$ ($\Delta df = 8$) = 83.304, $p < .001$. Applying the Satorra-Bentler χ^2 correction (Satorra & Bentler, 2010) due to a violation of the multivariate normality assumption did not produce results leading to different conclusions.

ΔAIC indicated better fit of the RMF model compared to both the one- as well as the two-factor model. ΔBIC , however, indicated preference of the two-factor model compared to the other two models. When using the sample-size adjusted BIC (ssBIC) following Henson et al. (2007), $\Delta ssBIC$ preferred the RMF model compared to both the one- and the two-factor model.

Internal consistency

The internal consistency was $\alpha = .75$, $\omega = .81$ for the total scale; $\alpha = .76$, $\omega = .82$ for the straight item subscale (SIS); and $\alpha = .55$, $\omega = .66$ for the reverse item subscale (RIS).

Discussion

Considering the inconsistency of EFA and CFA results, Study 1 could not satisfyingly clarify the dimensionality of the TEQ-D.

One might consider the mathematically inherent characteristic of BIC—compared to AIC—to more strongly prefer less complex models (due to a higher penalty term for parameter count). A relatively high BIC value of the RMF model thus might not seem very surprising. Apart from that, there is evidence (Henson et al., 2007), suggesting the ssBIC is superior to other information criteria (e.g., BIC) in latent variable estimation, arguing for the RMF model. Altogether, CFA results suggest, on the one hand, very clearly that a one-factor solution was not sufficient for explaining the TEQ-D data, while, on the other hand, providing some interesting clues (e.g., AIC, ssBIC) that the second factor *could be* of methodological nature. This issue is investigated further in Study 3.

Study 2

Study 2 data had been collected by the authors before Study 1 was conducted. Therefore, Study 2 examines the preliminary version of the TEQ, as introduced in Study 1 (for the final version: See Study 3). Since Study 1 results were indecisive with regard to the superiority of the one- or the two-factor model, analyses of Study 2 are run using the total TEQ-D scale as well as the straight items subscale (SIS) and the reverse items subscale (RIS).

The TEQ-D aimed to measure trait empathy. Consequently, the TEQ-D score is expected to show an at least mid-term stability (i.e., stable associations during longer intervals than the frequently assessed “two weeks apart”). Therefore, Study 2 tests the stability of empathy measured by the TEQ-D in intervals of approximately 3.5, 7.0 and 10.5 months. The instrument’s convergent validity was also assessed, by correlating it with a measure of emotion recognition, the short form of the Geneva Emotion Recognition Test (GERT-S; Schlegel & Scherer, 2016). According to initial reports (Spreng et al., 2009) of positive associations between the TEQ and task-based measures of interpersonal perception (Reading the Mind in the Eyes Test-Revised; Baron-Cohen et al., 2001), a positive association between the TEQ-D and the GERT-S was predicted.

Methods

Participants and data collection

The data were collected within a larger training evaluation project, which focused on psychological strain in social professions and in health care providers (Deckers et al., 2021). For the present study, only the untreated control group from the research project is considered. The data collection process consisted of four measurement occasions with three to four months between assessments: Assessment t1 assessed at the beginning, assessment t2 assessed 3–4 months after t1, assessment t3 assessed 6–8 months after t1, and assessment t4 assessed 9–12 months after t1.

The TEQ-D was applied at all four occasions. 135 persons (82.2% female) aged between 20 and 61 years ($M = 39.3$, $SD = 11.4$) participated on all occasions. All participants were employed. 70.4% indicated their highest level of education as A-levels or higher, while 29.6% reported “O-levels” as their highest level of education. 131 subjects answered the GERT-S at

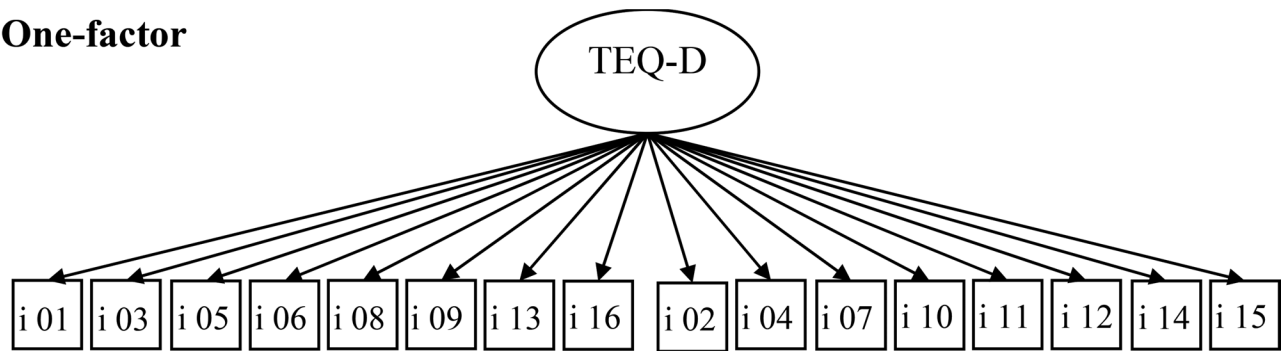
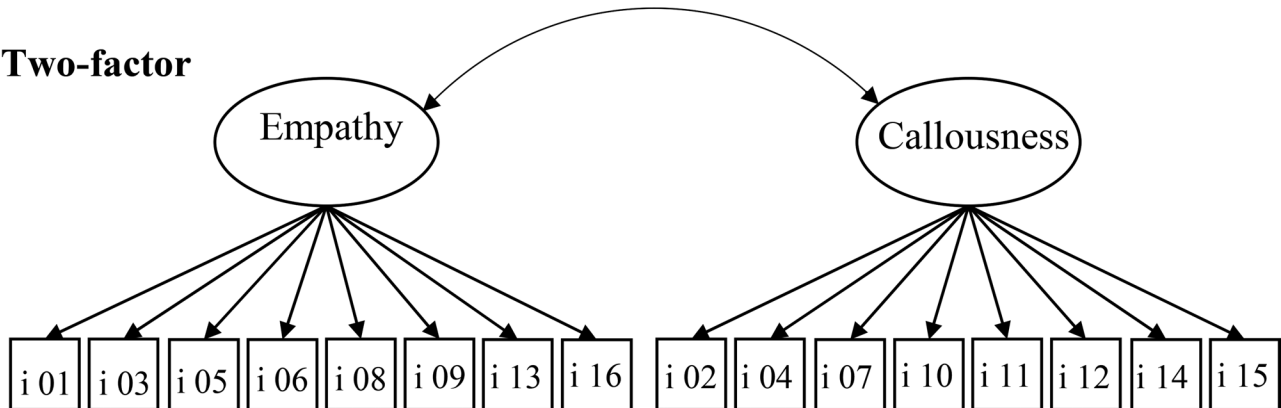
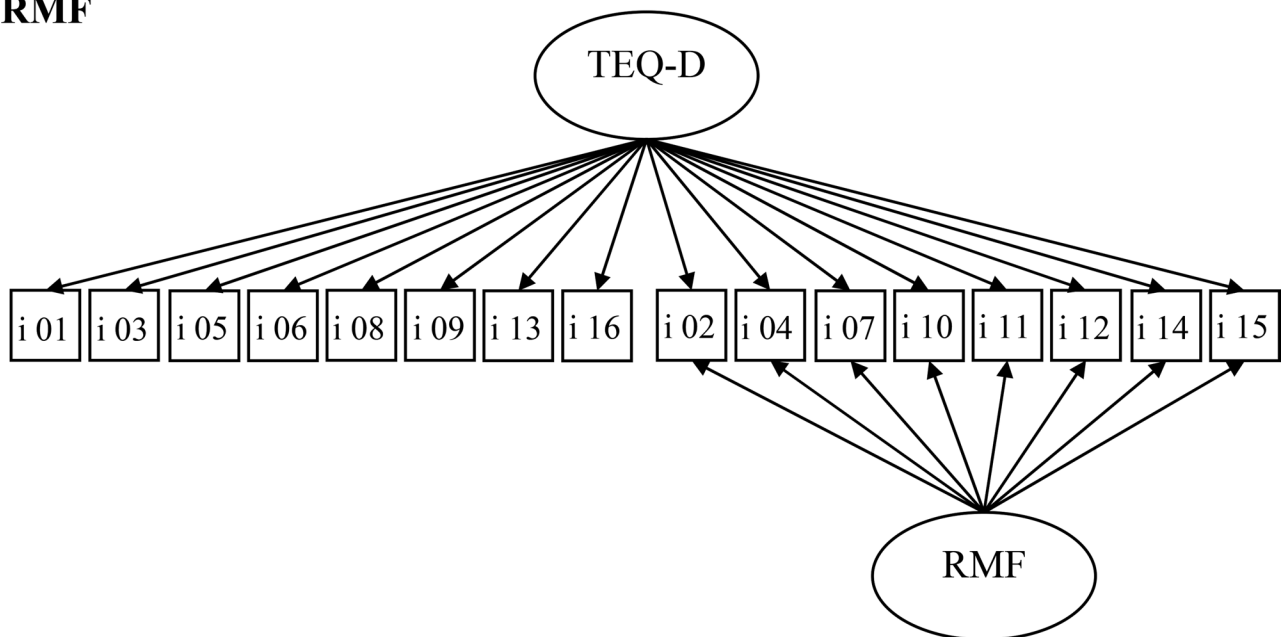
One-factor**Two-factor****RMF**

Figure 1. Measurement models for the TEQ-D.

the second measurement occasion in addition to the TEQ-D. The analysis was preregistered [<https://osf.io/wp98a>].

Measures

TEQ-D. See Method section of Study 1 for a description of the TEQ-D.

GERT-S. The GERT-S (Schlegel & Scherer, 2016) is a performance test aimed at measuring the capability of recognizing other peoples' emotions. It includes 42 short video clips

with sound, in which actors/actresses express 14 different emotions. After each clip, participants are asked to indicate which of the 14 emotions the person in the video expressed. Responses are coded dichotomously (correct, incorrect). The internal consistency of the GERT-S in the current sample was $\alpha = .73$. Moreover, the split-half reliability was computed using the odd-even method. This method implies assigning every second item to the first half of the test and all other items to the second half and then correlating both

Table 2. Fit indices of measurement models for the preliminary TEQ-D (CFA).

	One-factor	Two-factor	RMF
χ^2	278.732 (235.803)	223.531 (192.802)	195.428 (170.763)
df	104	103	96
RMSEA	.067 (.058)	.056 (.048)	.053 (.046)
CFI	.842 (.840)	.891 (.891)	.910 (.909)
TLI	.818 (.815)	.873 (.873)	.888 (.886)
AIC	14656.943	14603.742	14589.638
Δ AIC	67.305	14.104	–
BIC	14782.347	14733.065	14746.394
Δ BIC	49.282	–	13.329
ssBIC	14680.821	14628.366	14619.486
Δ ssBIC	61.335	8.88	–

Note. Satorra-Bentler corrected test statistics in parentheses.

halves. The split-half reliability of the GERT-S was .59 (95% CI [.46, .69]), $p < .001$.

Results

Stability

TEQ-D total scores were significantly (all $ps < .001$) and highly correlated across the intervals of three to four months, $r_{t1-t2} = .74$, $r_{t2-t3} = .70$, $r_{t3-t4} = .75$, as well as the intervals of six to eight months, $r_{t1-t3} = .69$, $r_{t2-t4} = .78$, and the interval of nine to twelve months, $r_{t1-t4} = .68$. The correlations corresponding to the SIS and the RIS across all intervals ranged between .68-.74 and .54-.68, respectively (all $ps < .001$).

Association with GERT-S

There was neither a significant association between the GERT-S and the total TEQ-D, $r = .02$, $p = .399$, nor the SIS, $r = .09$, $p = .148$, nor the RIS, $r = .06$, $p = .257$.

Internal consistency

Internal consistencies of the TEQ-D total scale at the four measurement occasions ranged between .59 and .69 for Cronbach's alpha and .70 to .78 for McDonald's omega. When separating straight (SIS) and negatively scored items (RIS) into two scales, results were similar for the SIS but lower for the RIS: SIS: $\alpha = .64$ -.74, $\omega = .71$ -.81; RIS: $\alpha = .34$ -.42, $\omega = .46$ -.60.

Discussion

As predicted, Study 2 demonstrated a high stability for periods up to one year, indicating that the TEQ-D captures a stable personality construct (instead of a short-term state). Contrary to prediction, the TEQ-D was not associated with the GERT-S, a performance-based measure of emotion recognition. Study 2 provided additional evidence for poor internal consistency of the RIS, which represents a serious issue for the approach of aggregating the reverse items into a separate callousness score.

Study 3

While CFA results of Study 1 clearly showed that the one-factor model was outperformed by a two-dimensional structure for the preliminary TEQ-D, Study 3 aimed to

clarify whether this second factor represents a truly dissociable construct or a methodological artifact of item wording.

Another popular personality questionnaire, the Rosenberg Self-Esteem scale (RSES; Rosenberg, 1965), has been the object of the same question (e.g., Tomas & Oliver, 1999). A proposal for answering this question was provided by Greenberger et al. (2003), who demonstrated that changing the wording of the RSES, so that all items are written in a consistent direction, substantially improved the fit of the one-factor model. This rewording of negatively scored items to the positive provided strong evidence that the two-factor solution was a methodological artifact of item wording.

This approach of reformulating negatively worded items was recently pursued for the TEQ by Novak et al. (2021), which yielded a substantial improvement in fit for a one-factor model and increased internal consistency. Following Greenberger et al. (2003) in general and Novak et al. (2021) in particular, negated items of the preliminary TEQ-D were reformulated for the present study. We predicted a positive impact in fit for a one-factor model as well as greater internal consistency for the total scale. However, to avoid the possible objection that “re-wording of the scale items (...) essentially changes the construct being measured by these items” (Greenberger et al., 2003, p. 1252), we focus on reformulating *literally negated* items which include words like *not*, instead of rephrasing all negatively scored items (see Method section below for details).

Additionally, Study 3 addresses convergent and discriminant validation of the TEQ-D: Generally, we predicted positive associations between the TEQ-D and other empathy measures (based on previous studies, see Introduction), namely the IRI (Davis, 1980; Hypothesis 1) and the BES (Jolliffe & Farrington, 2006; Hypothesis 2) as well as the related construct of emotional contagion (measured by the Emotional Contagion Scale; ECS; Doherty, 1997; Hypothesis 3). However, since the TEQ has been shown to tap the affective component of empathy (e.g., Spreng et al., 2009), a stronger association with the IRI subscale EC compared to the other three IRI subscales was expected (Hypotheses 1a–c). Similarly, we hypothesized a stronger association of the TEQ-D with the BES subscale *affective empathy* (AE) compared to the cognitive BES subscale (Hypothesis 2a).

Moreover, a positive correlation between the TEQ-D and emotion regulation (Hypothesis 4) was assumed, with respect to a proposed “common conceptual ground” of the latter construct and empathy (as summarized by Morawetz et al., 2022, p. 1). However, this association was expected to be weaker than the associations between the TEQ-D score and the IRI subscale EC, the BES subscale AE and the ECS due to the higher theoretical overlap of the latter measures with the TEQ-D (Hypotheses 4a–c).

In line with previous studies (see introduction), we additionally hypothesized a positive association between TEQ-D and altruism (Hypothesis 5), but also expected this association to be weaker than TEQ-D's associations with the IRI subscale EC, the BES subscale AE and the ECS (Hypotheses 5a–c). We predicted the TEQ-D to be associated with the Big Five personality trait agreeableness, but not with other traits (Hypotheses 6a–e). Finally, a positive association

between the TEQ-D and social desirability was predicted (see, e.g., Chiorri, 2016).

Test-retest reliability of the TEQ-D was assessed, as well as the measurement invariance, between two measurement occasions separated by two weeks.

Methods

Participants and data collection

The data were collected at two measurement occasions, 14 days apart, *via* online surveys. Participants were recruited *via* flyers, e-mails and postings in social media, mainly addressed to students at the University of Duisburg-Essen (Germany). Participants could either receive partial course credit for their participation at the first measurement occasion or participate in a raffle. For the (optional) participation at the second measurement occasion, subjects were able to participate in another raffle. A total of 195 subjects (80.5% female) aged between 18 and 62 years ($M=23.0$, $SD=6.6$; one missing value) participated at the first measurement occasion. 93.3% were students, 4.6% were employed, 1.0% were unemployed and 1.0% were retired. 98.5% were highly educated (A-levels or higher), while 1.5% reported “O-levels” as their highest level of education. 101 participants returned for a second assessment. Study 3 was preregistered before data collection [<https://osf.io/wujc2>].

Measures

At the first measurement occasion, all self-report measures described below were collected. For the second assessment, only the preliminary TEQ-D was administered.

TEQ-D. The TEQ-D (as described in the Method section of Study 1) was administered. Following Novak et al. (2021), negated items were positively reformulated by eliminating/changing the negating words: For instance, the word *not* was eliminated within Item 12 (“I am not really interested in how other people feel”). A total of five negated items (02, 04, 10, 12, 14) were detected within the instrument and reworded.

These five positively reworded items were administered in the assessment in addition to the original wording, so that the preliminary scale (i.e., including the originally negated items) as well as the reworded scale (by replacing the five negated items with the reworded alternatives) could be examined.

To avoid biased responding, other self-report measures were placed between the preliminary 16 TEQ-D items and the additional five reworded items. To avoid sequence effects, participants were randomly counterbalanced to either answer the preliminary scale at the beginning and the five reworded items at the end of the survey or, *vice versa*, answer the reworded scale at the beginning and the five originally negated items at the end of the survey. Since the retest-analysis only focused on the TEQ-D, no other measures were applied at the second measurement occasion. Thus, to avoid biased responding, only the preliminary TEQ-D scale was presented at the second measurement occasion.

IRI. The IRI (Davis, 1980) consists of the four subscales *perspective taking (PT)*, *empathic concern (EC)*, *fantasy (FT)*

and *personal distress (PD)*, each containing seven items. The PT and EC subscales capture the cognitive and affective component of empathy, respectively. The FT subscale captures the propensity to identify with fictional characters, whereas the PD subscale measures the tendency to feel uneasy and anxious while recognizing the negative experiences of others. In the present study, the German version developed by Paulus (2009) was used, which comprises 16 (positively scored) items. Answers are given on a five-point Likert scale ranging from 1 = *Never* to 5 = *always*.

BES. The BES is a self-report empathy measure, originally developed by Jolliffe and Farrington (2006), containing a cognitive and an affective empathy subscale (CE and AE, respectively). The items address the four emotions *anger*, *fear*, *sadness* and *joy*. Responses are given on a five-point Likert scale ranging from 1 = *I don't agree* to 5 = *I fully agree*. In the current study, the German adaptation containing 12 items (Heynen et al., 2016) was administered.

ECS. The ECS is a 15-item questionnaire developed by Doherty (1997), assessing the construct of emotional contagion, meaning the “susceptibility to others’ emotions resulting from afferent feedback generated by mimicry” (Doherty, 1997, p. 131). Responses are given on a four-point Likert scale ranging from 1 = *never* to 4 = *always*. In the present study, a German adaptation (Falkenberg, 2005) was administered.

Emotion regulation questionnaire. The Emotion Regulation Questionnaire (ERQ) is a ten-item self-report instrument, originally developed by Gross and John (2003). It captures two emotion regulation strategies: (1) *suppression*, or describing a process of repressing one’s behavior and expression due to emotional experiences, and (2) *reappraisal*, a cognitive reinterpretation of a pending emotional situation (Abler & Kessler, 2009). Responses were given on a seven-point Likert scale ranging from 1 = *strongly disagree* to 7 = *strongly agree*. In the present study, a German version (Abler & Kessler, 2009) was administered.

Facets of altruistic behaviors scale. The Facets of Altruistic Behaviors Scale (FAB; Windmann et al., 2021) is a questionnaire measuring different facets of altruistic behavioral traits. For economic reasons, only the five-item subscale *help giving (HG)* was applied in the present study. This subscale captures the propensity of “*sharing* one’s resources with needy or deserving others”. Responses are given on a six-point Likert scale ranging from 1 = *strongly disagree* to 6 = *strongly agree*.

NEO-five-factor inventory-30. The 30-item short version (NEO-FFI-30) of the German adaption (Borkenau & Ostendorf, 1993) of the NEO-five-factor inventory (Costa & McCrae, 1989) was developed by Körner et al. (2008) and assesses the Big Five personality traits. Responses are given on a five-point Likert scale ranging from 1 = *Doesn't apply* to 5 = *Applies*.

Short scale social desirability-gamma. The Short Scale Social Desirability-Gamma (KSE-G; Kemper et al., 2012) is a brief tool for measuring the tendency to respond in a social desirable manner to self-report measures. It includes the two subscales *exaggeration of positive qualities (PQ+)* and *understatement of negative qualities (NQ-)*, each consisting of three items. Responses are given on a five-point Likert scale ranging from 1 = *Doesn't apply* to 5 = *Applies*. The KSE-G items were presented interspersed within the NEO-FFI-30.

Data analysis

CFA. CFA was conducted and interpreted pursuant to Study 1.

Correlational analyses. Product moment correlations were computed. $Z_{contrast}$ tests were run for comparing correlations using the R package *cocor* (Diedenhofen & Musch, 2015). To avoid Type-I-error inflation, p values were adjusted.

Measurement invariance. Multiple-group confirmatory factor analyses were performed to assess measurement invariance across the two measurement occasions two weeks apart, using the R package *lavaan* plus the R package *semTools* (Jorgensen et al., 2021). Pursuant to Hirschfeld and von Brachel (2014), the following four types of invariance were examined: *Configural invariance* describes a gross factor structure equivalence across groups, *weak invariance* implies equal factor loadings across groups, while *strong invariance* also implicates equality of intercepts. Finally, *strict invariance* also implies the equality of residual variances across groups. Besides χ^2 difference tests, the difference (ΔCFI) between CFI values of each model and the less restricted model were computed, applying the often used (Hirschfeld & von Brachel, 2014) cutpoint of $\Delta CFI < .01$ to test for weak, strong and strict invariance.

Results

Internal consistency

The internal consistency of the TEQ-D total scale increased by reformulating negated items from $\alpha = .74$, $\omega = .79$ to $\alpha = .79$, $\omega = .84$.

CFA

The factorability of data was good (Kaiser-Meyer-Olkin: .78 for the preliminary scale and .84 for the reworded scale). The one-factor model yielded a higher absolute and incremental fit within the reworded scale compared to the original scale (see Table 3), but this improvement in fit was limited. χ^2 difference tests comparing the one- and the two-factor model indicated that the two-factor model had a significantly better fit than the one-factor model in the preliminary scale, $\Delta\chi^2 (\Delta df=1) = 16.87$, $p < .001$, but there was no substantial difference between both models in the reworded scale, $\Delta\chi^2 (\Delta df=1) = 0.719$, $p = .419$. This interpretation did not change by applying the Satorra-Bentler χ^2 correction (used due to a violation of multivariate normality) and is also confirmed by the computed fit indices, including information criteria (see Table 3).

Correlational analyses for clarifying factor structure

Since the one- and the two-factor solutions yielded a similar fit within the reworded TEQ-D, correlational analyses were run to determine whether both subscales suggested by the two-factor model had a differential pattern of correlations with the external measures (using the reworded TEQ-D data). The magnitude of the two subscales' associations with the external measures did not differ. Not one single comparison was significant ($p < .05$) after applying the Bonferroni

Table 3. Fit indices of the one- and two-factor model for the preliminary and reworded TEQ-D.

	Preliminary scale		Reworded scale	
	One-factor	Two-factor	One-factor	Two-factor
χ^2	176.95 (154.44)	160.09 (139.61)	174.78 (155.60)	174.061 (154.92)
df	104	103	104	103
RMSEA	.060 (.050)	.053 (.043)	.059 (.050)	.059 (.051)
CFI	.838 (.857)	.874 (.896)	.887 (.889)	.887 (.888)
TLI	.813 (.835)	.853 (.879)	.870 (.871)	.868 (.869)
AIC	7393.436	7378.566	6832.017	6833.301
ΔAIC	14.870	—	—	1.284
BIC	7498.172	7486.575	6936.753	6941.310
ΔBIC	11.597	—	—	4.557
ss BIC	7396.801	7382.036	6835.382	6836.771
$\Delta ssBIC$	14.765	—	—	1.389

Note. Satorra-Bentler corrected test statistic in parentheses.

Table 4. Internal consistencies of external measures, correlations between TEQ-D scales and other measures and comparisons between correlations of TEQ-D subscales with other measures.

Measure - scale	α	ω	TEQ-D Scale			
			Final	SIS	RIS	p
IRI - EC	.58	.63	.66***	.63***	.59***	>.999
IRI - PT	.79	.84	.30***	.28***	.26***	>.999
IRI - FT	.68	.79	.43***	.38***	.41***	>.999
IRI - PD	.80	.82	.02	.05	−0.01	>.999
BES - AE	.64	.79	.61***	.62***	.50***	.068
BES - CE	.78	.85	.53***	.54***	.43***	.289
ECS - total	.73	.79	.53***	.52***	.45***	>.999
ERQ - Rea	.82	.89	.13	.16	.08	>.999
ERQ - Sup	.74	.79	−0.33***	−0.25***	−0.35***	.697
FAB - HG	.71	.75	.40***	.40***	.34***	>.999
NEO-FFI-30 - E	.75	.86	.29***	.31***	.21**	.731
NEO-FFI-30 - A	.69	.80	.52***	.47***	.49***	>.999
NEO-FFI-30 - C	.79	.85	.36***	.36***	.29***	>.999
NEO-FFI-30 - N	.86	.91	−0.07	−0.09	−0.04	>.999
NEO-FFI-30 - O	.82	.86	.30***	.29***	.25***	>.999
KSE-G - PQ+	.47	.48	.22*	.25***	.15	.595
KSE-G - NQ-	.55	.60	.32***	.27***	.31***	>.999

correction (see Table 4). To ensure the interpretability of findings, the (less conservative) corrections proposed by Holm (1979), Hochberg (1988) and Benjamini and Hochberg (1995) were additionally applied. However, none of these alternative methods changed the interpretation of any comparison. Overall, the positively worded first-factor “sub-scale” could not be dissociated from the negatively worded second factor “sub-scale” in the patterns of convergent validity. This indicates that the original TEQ likely measures a singular construct. Two factor solutions for the TEQ are therefore likely driven by the methodological artifact of item wording.

In rewording negation statements, but not their content, the psychometrics of the modified TEQ improved. For this reason, the reworded scale was selected for the final TEQ-D scale, providing a unidimensional total score. The factor loadings, communalities, means, standard deviations, item-total correlations and response probabilities of the final TEQ-D scale are displayed in Table 5.

Correlational analyses for testing convergent and discriminant validity

Convergent and discriminant associations were examined using the final TEQ-D scale. Results of the main Hypotheses (1–5, 6a–e and 7) are displayed in Table 4. The TEQ-D was

Table 5. Psychometric properties of the final TEQ-D.

Item	λ	h^2	M	SD	r_{tt}	P
01	.28	.0770	3.08	0.80	.31	.62
02	.04	.0018	2.61	0.86	.06	.52
03	.70	.4857	4.44	0.68	.71	.89
04	.50	.2534	4.15	0.76	.50	.83
05	.55	.3027	4.33	0.74	.54	.87
06	.50	.2521	3.77	0.87	.52	.75
07	.31	.0961	4.26	0.82	.29	.85
08	.37	.1346	3.94	0.69	.39	.79
09	.31	.0978	3.41	0.72	.31	.68
10	.41	.1669	3.42	0.85	.39	.68
11	.30	.0872	3.61	1.09	.31	.72
12	.68	.4637	4.14	0.73	.68	.83
13	.63	.3933	3.88	0.77	.62	.78
14	.78	.6056	4.25	0.71	.76	.85
15	.39	.1489	4.44	0.84	.41	.89
16	.59	.3445	4.15	0.73	.58	.83

Notes. λ : factor loading, h^2 : communality, r_{tt} : part-whole corrected item-total correlation, P : response probability. Negated items were reworded.

strongly positively associated with the affective empathy subscales of the IRI (EC) and the BES (AE), also strongly positively with the ECS, moderately to strongly positively with altruism (FAB-HG) and moderately positively with social desirability (KSE-G), while it was not significantly positively associated with the ERQ. The TEQ-D did not significantly correlate with Neuroticism, was strongly positively associated with Agreeableness, while it was also moderately positively associated with Extraversion, Conscientiousness and Openness.

Except for the tests corresponding to the comparative Hypotheses 2a ($p = .075$) and 5c ($p = .051$), all tests comparing correlations between external measures and the final TEQ-D were significant, all $ps < .001$ (except hypothesis 5b, $p = .002$). Applying the Bonferroni correction did not produce any result leading to divergent conclusions (hypothesis 5b: $p_{Bon} = .026$).

Retest analyses

Test-retest reliability. The association of the originally worded TEQ-D between the two assessments was $r = .68$, $p < .001$.

Measurement invariance. The originally worded TEQ-D one-factor baseline model had suboptimal fit to the data, $\chi^2_{\text{Satorra-Bentler}}(104) = 174.55$, $p < .001$, $RMSEA_{SB} = .058$, $CFI_{SB} = .790$. The configural model did not show an acceptable fit either, $\chi^2_{SB}(208) = 259.57$, $p = .009$, $RMSEA_{SB} = .050$, $CFI_{SB} = .830$. However, both the (Satorra-Bentler corrected) χ^2 difference tests ($ps > .819$ for all three comparisons) as well as the differences between CFI values did not indicate a decrease in model fit for the weak ($CFI_{SB} = .864$), strong ($CFI_{SB} = .880$) and strict ($CFI_{SB} = .902$) invariance models. Therefore, it can be assumed that the originally worded TEQ-D showed strict invariance across the two measurement occasions.

Discussion

This study demonstrated that, as predicted, the fit of the one-factor structure as well as the internal consistency of the TEQ-D increased by modifying the negatively worded items through positively reworded alternatives. Further, rewording negative to positive items eliminated the superiority in fit for a two-factor model. While CFA could not show the

superiority of the one-factor model after item rewording, correlational analyses revealed that the subscales suggested by the two-factor model did not show dissociable correlation patterns with various external measures. The idea that the two subscales (including only the positively and only the negatively scored items, respectively) capture two distinct constructs is not tenable. In conclusion, the TEQ-D should be viewed as a unidimensional scale.

The final TEQ-D showed promising convergent validity, while evidence for discriminant validity was mixed. Its psychometric properties and internal consistency were found to be sufficient. Analyses of the original scale suggested satisfactory test-retest reliability as well as strict measurement invariance between measurement occasions.

General discussion

The present study conducted a psychometric analysis of the TEQ-D. Factor analyses, at first, suggested a two-dimensionality of the instrument, which was not supported by comparisons in their associations with external measures. After item negations were eliminated, the final TEQ-D showed unidimensionality, satisfactory convergent validity and internal consistency as well as sound psychometric properties.

Factor structure

Studies 1 and 3 examined the factor structure of the TEQ-D using EFA and CFA. EFA of Study 1 suggested that the number of factors in the preliminary TEQ-D data was either one or two. CFA demonstrated that the two-factor solution suited the preliminary TEQ-D data better than the one-factor model. However, a model with one substantial factor and a reverse item method factor (RMF model) yielded an even better fit concerning most criteria.

CFA results of Chiorri (2016) comparing 15 different factor models for the TEQ seem—from a certain perspective—quite consistent with our results: If we ignore (justified due to problems of interpretability) as well the models with correlated uniquenesses as a model with one factor and three method factors, the RMF model likewise showed the best fit.

Study 3 replicated the superiority of the two-factor model over the one-factor model using CFA for the preliminary

TEQ-D data, but this superiority vanishes after eliminating literal negations in item wording. Correlational analyses using the reworded TEQ-D data clearly demonstrated that the factors of the two-factor model were not empirically distinguishable and are therefore indicative of a methodological artifact produced by item negations. The improvement in fit of the one-factor structure as well as the increase in internal consistency after rewording negated items replicated the results of Novak et al. (2021).

The general issue of negated items is obviously not limited to the TEQ, as it has been demonstrated for other questionnaires before, such as the RSES (see, e.g., Tomas & Oliver, 1999) and the IRI (e.g., Paulus, 2009). An underlying mechanism may represent the potential duality of expressing agreement with a negated item, which does not equally occur for positively formulated items, as summarized by Moosbrugger and Kelava (2020). Other potential causes of biased responding to negated items are participants' inattention as well as a higher verification difficulty of negated items, meaning that verifying those items require greater cognitive resources (Swain et al., 2008).

Convergent and discriminant validity

The present study failed to demonstrate an association between the TEQ and a performance-based measure of interpersonal perception (in the present study: GERT-S), contrary to previous observations (Spreng et al., 2009). The absence or weakness of associations between behavioral-based measures of interpersonal perception and self-reported trait empathy are not unusual (Hall & Schwartz, 2019). This once again reflects the issue of heterogeneity the empathy field suffers from and the fact that behavioral-based measures rarely correlate with self-report (e.g., Melchers et al., 2015). The (final) TEQ-D was consistent with previous studies with positive associations with affective and cognitive empathy subscales of the IRI, as well as the BES and the ECS.

Moreover and in line with, e.g., Mulet et al. (2022), the final TEQ-D significantly correlated with altruism. However, this relationship was, against theoretical expectations, not substantially lower than final TEQ-D's relationship with the BES subscale AE as well as the ECS. This issue might be affected by some items' content (e.g., 05, 13, 16), which is obviously not limited to a non-behavioral, emotional empathic response, but also includes prosocial behavior, e.g., "I enjoy making other people feel better" (Item 05). It might, from a theoretical perspective, seem questionable whether such items are reasonably placed within an *empathy* questionnaire. Nevertheless, it should be noted again that the TEQ was statistically developed out of existing empathy measures and thereby—invariably—reflects underlying problems in the field (e.g., discrepancies between conceptual and operational definitions; Hall & Schwartz, 2019).

Retest analyses

The stability of empathy measured by the preliminary TEQ-D can be estimated as remarkably high. The present

study is, to our knowledge, the first which analyzed the associations between TEQ scores over periods lasting several months or even up to one year. Additionally, the preliminary TEQ-D scale showed sufficient two-week test-retest reliability, thereby replicating results of previous TEQ validation studies (e.g., Totan et al., 2012). Moreover, the present study was the first to analyze measurement invariance of the TEQ between measurement occasions. Although results suggested strict invariance, these conclusions have to be treated with caution: First, the baseline model did not fit the data well—which though appears not surprising since invariance was investigated within the preliminary scale. Secondly, the fit of the configural variance model was not sufficient—which should, however, not surprise either, since a good fit of the baseline model is crucial for assessing configural invariance (Hirschfeld & von Brachel, 2014).

Limitations and future directions

The approach of eliminating negations—regardless of its psychometric benefit—inevitably implicates the elimination of inverse scored items, which are though important to control for acquiescence bias (e.g., Ray, 1983). Nevertheless, the final TEQ-D still includes three inverse scored items (07, 11, 15)—so called *polar opposites* (e.g., Weijters & Baumgartner, 2012), meaning that these do not require a negation for being reverse coded (e.g., "I become irritated when someone cries"; Item 11). Since the present study demonstrated that item negations can produce methodological artifacts, it seems advisable for future research to rely on polar opposites when selecting reverse coded items for the development of new personality questionnaires.

Another limitation of the present study is the fact that the reworded items were only administered in one sample. Future studies will need to cross-validate the unidimensional structure of the final TEQ-D. Finally, an issue of any so far published study examining the validity of the TEQ is the simultaneous translation into another language. Future studies need to analyze the instrument's dimensionality and the impact of item (re)wording without unintended confounds of cultural linguistic effects.

Conclusions

A two-factor structure can be interpreted as a methodological artifact driven by item wording and should not challenge the practice of computing a total score for the TEQ(-D). Overall, the TEQ-D demonstrated satisfactory internal consistency, one-year stability, test-retest reliability as well as convergent and discriminant validity. We recommend that future studies attempt to replicate these results and test the generalizability of findings (e.g., across less educated samples).

Acknowledgments

We thank Victoria Schönefeld for contributing to the translation. The BMBF was not involved in study design, data collection, data analysis, data interpretation, and in the writing of this report.

Data availability statement

The data that support the findings of this study can be retrieved from the OSF (<https://osf.io/qz7y3/>).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Grant 02L14A150 from the Federal Ministry of Education and Research (BMBF).

ORCID

Tobias Janelt  <http://orcid.org/0000-0002-3138-5818>
 Tobias Altmann  <http://orcid.org/0000-0001-7294-7808>
 R. Nathan Spreng  <http://orcid.org/0000-0003-1530-8916>
 Marcus Roth  <http://orcid.org/0000-0002-5676-8137>

References

- Abler, B., & Kessler, H. (2009). Emotion regulation questionnaire—Eine deutschsprachige Fassung des ERQ von Gross und John. *Diagnostica*, 55(3), 144–152. <https://doi.org/10.1026/0012-1924.55.3.144>
- Baldner, C., & McGinley, J. J. (2014). Correlational and exploratory factor analyses (EFA) of commonly used empathy questionnaires: New insights. *Motivation and Emotion*, 38(5), 727–744. <https://doi.org/10.1007/s11031-014-9417-2>
- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34(2), 163–175. <https://doi.org/10.1023/B:JADD.0000022607.19833.00>
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “reading the mind in the eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 42(2), 241–251. <https://doi.org/10.1111/1469-7610.00715>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <http://www.jstor.org/stable/2346101> <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae: Handanweisung*. Hogrefe. <https://pub.uni-bielefeld.de/record/1902849>
- Bošnjaković, J., & Radionov, T. (2018). Empathy: Concepts, theories and neuroscientific basis. *Alcoholism and Psychiatry Research: Journal on Psychiatric Research and Addictions*, 54(2), 123–150. <https://doi.org/10.20471/dec.2018.54.02.04>
- Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4), 509–540. https://doi.org/10.1207/s15327906mbr2704_2
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2), 261–304. <https://doi.org/10.1177/0049124104268644>
- Chiorri, C. (2016). Competing factor structures for the Toronto Empathy Questionnaire. In D. F. Watt & J. Panksepp (Eds.), *Psychology and neurobiology of empathy* (pp. 399–432). Nova Science.
- Costa, P. T., & McCrae, R. R. (1989). *The NEO PI/FFI manual supplement*. Psychological Assessment Resources.
- Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, 10, 85.
- Deckers, M., Schönefeld, V., Altmann, T., & Roth, M. (2021). Forschungsergebnisse zur Wirksamkeit des empCARE-Konzeptes. *empCARE* (pp. 149–174). Springer.
- Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *Plos One*, 10(3), e0121945. <https://doi.org/10.1371/journal.pone.0121945>
- Doherty, R. W. (1997). The emotional contagion scale: A measure of individual differences. *Journal of Nonverbal Behavior*, 21(2), 131–154. <https://doi.org/10.1023/A:1024956003661>
- Falkenberg, I. (2005). *Wahrnehmung und Expression von Emotionen durch Mimik: Eine Untersuchung über emotionale Ansteckung bei Gesunden und Patienten mit Schizophrenie*. Universität Tübingen.
- Gerdes, K. E., Segal, E. A., & Lietz, C. A. (2010). Conceptualising and measuring empathy. *British Journal of Social Work*, 40(7), 2326–2343. <https://doi.org/10.1093/bjsw/bcq048>
- Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences*, 35(6), 1241–1254. [https://doi.org/10.1016/S0191-8869\(02\)00331-8](https://doi.org/10.1016/S0191-8869(02)00331-8)
- Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2), 348–362. <https://doi.org/10.1037/0022-3514.85.2.348>
- Hall, J. A., & Schwartz, R. (2019). Empathy present and future. *The Journal of Social Psychology*, 159(3), 225–243. <https://doi.org/10.1080/00224545.2018.1477442>
- Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(2), 202–226. <https://doi.org/10.1080/10705510709336744>
- Heynen, E. J. E., Van der Helm, G. H. P., Stams, G. J. J. M., & Korebrits, A. M. (2016). Measuring empathy in a German youth prison: A validation of the German version of the Basic Empathy Scale (BES) in a sample of incarcerated juvenile offenders. *Journal of Forensic Psychology Practice*, 16(5), 336–346. <https://doi.org/10.1080/15228932.2016.1219217>
- Hirschfeld, G., & von Brachel, R. (2014). Improving Multiple-Group confirmatory factor analysis in R—A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation*, 19(1), 7.
- Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), 800–802. <https://doi.org/10.1093/biomet/75.4.800>
- Hoffman, M. L. (2000). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.
- Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology*, 33(3), 307–316. <https://doi.org/10.1037/h0027580>
- Hojat, M., Mangione, S., Gonnella, J. S., Nasca, T., Veloski, J. J., & Kane, G. (2001). Empathy in medical education and patient care. *Academic Medicine: Journal of the Association of American Medical Colleges*, 76(7), 669–669. <https://doi.org/10.1097/00001888-200107000-00001>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <http://www.jstor.org/stable/4615733>
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>
- Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence*, 29(4), 589–611. <https://doi.org/10.1016/j.adolescence.2005.08.010>
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y. (2021). Useful tools for structural equation modeling. R package version 0.5-5. <https://CRAN.R-project.org/package=semTools>
- Kaviani, H., & Kinman, G. (2017). Relationships between psychosocial characteristics and democratic values in Iranians: A cross-cultural study. *Journal of Social Sciences*, 3(1), 12–22.
- Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2012). *Eine Kurzsкала zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: Die Kurzsкала Soziale Erwünschtheit-Gamma*

- (KSE-G) (GESIS-Working Papers, 2012/25). GESIS - Leibniz-Institut für Sozialwissenschaften.
- Körner, A., Geyer, M., Roth, M., Drapeau, M., Schmutzer, G., Albani, C., Schumann, S., & Brähler, E. (2008). Persönlichkeitsdiagnostik mit dem neo-fünf-faktoren-inventar: Die 30-item-kurzversion (neo-ffi-30). *Psychotherapie, Psychosomatik, Medizinische Psychologie*, 58(6), 238–245. <https://doi.org/10.1055/s-2007-986199>
- Kourmoussi, N., Amanaki, E., Tzavara, C., Merakou, K., Barbouni, A., & Koutras, V. (2017). The Toronto Empathy Questionnaire: Reliability and validity in a nationwide sample of Greek teachers. *Social Sciences*, 6(2), 62. <https://doi.org/10.3390/socsci6020062>
- Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: Reliability and validity of the Empathy Quotient. *Psychological Medicine*, 34(5), 911–919. <https://doi.org/10.1017/s0033291703001624>
- Lelorain, S., Sultan, S., Zenasni, F., Catu-Pinault, A., Jaury, P., Boujut, E., & Rigal, L. (2013). Empathic concern and professional characteristics associated with clinical empathy in French general practitioners. *The European Journal of General Practice*, 19(1), 23–28. <https://doi.org/10.3109/13814788.2012.709842>
- Lima, F. F. d., & Osório, F. d. L. (2021). Empathy: Assessment instruments and psychometric quality—A systematic literature review with a meta-analysis of the past ten years. *Frontiers in Psychology*, 12, 781346. <https://doi.org/10.3389/fpsyg.2021.781346>
- Luckhurst, C., Hatfield, E., & Gelvin-Smith, C. (2017). Capacity for empathy and emotional contagion in those with psychopathic personalities. *Interpersona: An International Journal on Personal Relationships*, 11(1), 70–91. <https://doi.org/10.5964/ijpr.v11i1.247>
- Mehrabian, A. (2000). Manual for the Balanced Emotional Empathy Scale (BEES). Unpublished manuscript. Available from Albert Mehrabian, 1130 Alta Mesa Road, Monterey, CA 93940.
- Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality*, 40(4), 525–543. <https://doi.org/10.1111/j.1467-6494.1972.tb00078.x>
- Melchers, M., Montag, C., Markett, S., & Reuter, M. (2015). Assessment of empathy via self-report and behavioural paradigms: Data on convergent and discriminant validity. *Cognitive Neuropsychiatry*, 20(2), 157–171. <https://doi.org/10.1080/13546805.2014.991781>
- Moosbrugger, H., & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion* (3. Aufl. 2020 ed.). Springer. <https://doi.org/10.1007/978-3-662-61532-4>
- Morawetz, C., Berboth, S., Kohn, N., Jackson, P. L., & Jauniaux, J. (2022). Reappraisal and empathic perspective-taking—More alike than meets the eyes. *NeuroImage*, 255, 119194. <https://doi.org/10.1016/j.neuroimage.2022.119194>
- Mulet, M., Vuillemin, Q., Lachaux, J., Trousselard, M., & Ferrer, M.-H. (2022). Perceived stress, personality traits, and state of victims' consciousness: Impact on tourniquet application time and effectiveness. *Military Medicine*, 187(9–10), e1216–e1224. <https://doi.org/10.1093/milmed/usab092>
- Novak, L., Malinakova, K., Mikoska, P., van Dijk, J. P., Dechterenko, F., Ptacek, R., & Tavel, P. (2021). Psychometric analysis of the Czech version of the Toronto Empathy Questionnaire. *International Journal of Environmental Research and Public Health*, 18(10), 5343. <https://doi.org/10.3390/ijerph18105343>
- Paulus, C. (2009). *Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index*. <https://doi.org/10.23668/psycharchives.9249>
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–163. <https://doi.org/10.2307/271063>
- Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *The Journal of Social Psychology*, 121(1), 81–96. <https://doi.org/10.1080/00224545.1983.9924470>
- Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research*. <https://CRAN.R-project.org/package=psych>
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.
- Rossee, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, 75(2), 243–248. <https://doi.org/10.1007/s11336-009-9135-y>
- Schlegel, K., & Scherer, K. R. (2016). Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation. *Behavior Research Methods*, 48(4), 1383–1392. <https://doi.org/10.3758/s13428-015-0646-4>
- Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., & Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and Individual Differences*, 25(2), 167–177. [https://doi.org/10.1016/S0191-8869\(98\)00001-4](https://doi.org/10.1016/S0191-8869(98)00001-4)
- Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, 91(1), 62–71. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2775495/pdf/nihms-74226.pdf> <https://doi.org/10.1080/00223890802484381>
- Stewart, C., Lawrence, S., & Burg, M. A. (2019). Exploring the Relationship of personality characteristics and spirituality to empathy: Does spirituality add to our understanding? *Journal of Religion & Spirituality in Social Work: Social Thought*, 38(1), 3–20. <https://doi.org/10.1080/15426432.2018.1548953>
- Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed likert items. *Journal of Marketing Research*, 45(1), 116–131. <https://doi.org/10.1509/jmkr.45.1.116>
- Titchener, E. (1909). *Elementary psychology of the thought process*. Macmillan.
- Tomas, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 84–98. <https://doi.org/10.1080/10705519909540120>
- Totan, T., Dogan, T., & Sapmaz, F. (2012). The Toronto Empathy Questionnaire: Evaluation of psychometric properties among Turkish University Students. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, 12(46), 179–198.
- Ursoniu, S., Serban, C. L., Giurgi-Oncu, C., Ravis, I. A., Bucur, A., Bredicean, A.-C., & Papava, I. (2021). Validation of the Romanian Version of the Toronto Empathy Questionnaire (TEQ) among undergraduate medical students. *International Journal of Environmental Research and Public Health*, 18(24), 12871. https://mdpi-res.com/d_attachment/ijerph/ijerph-18-12871/article_deploy/ijerph-18-12871-v2.pdf <https://doi.org/10.3390/ijerph182412871>
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321–327. <https://doi.org/10.1007/BF02293557>
- Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, 49(5), 737–747. <https://doi.org/10.1509/jmr.11.0368>
- Windmann, S., Binder, L., & Schultze, M. (2021). Constructing the Facets of Altruistic Behaviors (FAB) Scale. *Social Psychology*, 52(5), 299–313. <https://doi.org/10.1027/1864-9335/a000460>
- Xu, R. H., Wong, E. L. Y., Lu, S. Y. J., Zhou, L. M., Chang, J. H., & Wang, D. (2020). Validation of the Toronto Empathy Questionnaire (TEQ) among medical students in China: Analyses using three psychometric methods. *Frontiers in Psychology*, 11, 810. <https://doi.org/10.3389/fpsyg.2020.00810>
- Yeo, S., & Kim, K. J. (2021). A validation study of the Korean version of the Toronto empathy questionnaire for the measurement of medical students' empathy. *BMC Medical Education*, 21(1), 119. <https://doi.org/10.1186/s12909-021-02561-7>