Psychological Test and Assessment Modeling, 2024.66:92-115

DOI: https://doi.org/10.2440/001-0014

Perception-based anticipation of social conflicts (ASK): IRT analysis of a new image-based method

Gebhard Sammer[1,2] & Annette Kroiß[1]

[1] Centre for Psychiatry, Justus Liebig University Giessen, Germany
[2] Department of Psychology, Justus Liebig University Giessen, Germany

Abstract*:*
People perceive everyday situations very individually and react to them depending on their personality, lifestyle, development or a psychological disorder. In this study, an image-based method was developed to capture the perception of social situations that represent potential social conflict situations. To date, there is no comparable psychometric instrument.

Based on what we know about social conflict, the construction of the items mainly included resource conflict and conflict of interest in various social situations. Eighteen images were presented as an online survey with a 7-point response scale. In addition, personality, anxiety, anger, alexithymia, and sociodemographic data were collected.

2074 participants were recruited via online portals, social media and mailing lists. Cases usable for analysis ($N = 1831$) included more women and more highly educated individuals.

The items and scale were tested for a Rasch model for polytomous responses with the partial credit model. After the number of response categories was reduced to 4, the item categories were ordered and fitted to the model. Person estimates showed a misfit of about 6% of cases. Model tests demonstrated the unidimensionality of the scale (Martin-Loef test) and revealed that some response categories were likely to deviate from the model (Wald-type test). The Anderson likelihood ratio test indicated model validity when smaller sample sizes were created and tested through resampling ($n = 300, 400, 500$). The reliability based on classical test theory was sufficient (McDonald's $\omega = 0.75$, $N = 1973$). However, the validity of the model is preliminary and needs to be tested empirically with a new data set without pooling of response categories.

Although the model fit statistics are preliminary due to the post-hoc pooling of response categories, it is proposed that the test be used to assess social cognition in the context of anticipating

Correspondence:

Gebhard Sammer, Centre for Psychiatry, Justus Liebig University Giessen, Klinikstrasse 36, 35392 Giessen, Germany; e-mail: gebhard.sammer@uni-giessen.de

social conflict (ASK), preferably for research questions and theory building on the perception of conflict-prone social situations.

Keywords: social cognition, social conflict, polytomous Rasch model, partial credit model, IRT

Introduction

So far, conflict management has focused on either sociological, educational (cf. Jonas, Otten, & Doosje, 2014; Bark, 2012; Schwarz & Siffert, 2010; Ahlbrecht, Bendiek, Meyers & Wagner, 2009) or intrapsychic or psychodynamic constructs (Horney, 1973). There is no uniform consensus on how to classify types of conflict (Bonacker, 2009). One encounters “hot” and “cold” conflicts, distribution, resource, value, goal or role conflicts (cf. Heigl, 2014). In daily practice, many groups of people (teachers, educators, doctors, police officers, psychologists, team leaders, etc.) are expected to adequately assess the general conflictual nature of situations. Existing instruments are largely based on the alexithymic concept or the theory of the mind construct (mentalization). They are constructed, for example, to assess emotion perception (MSCEIT, cf. Bracket & Salovey, 2006; Eyes-Test, cf. Baron-Cohen, Wheelwright, Hill, Raste & Plumb, 2001; CARAT, cf. Buck, 1979), empathy (MET, cf. Dziobek, Rogers, Fleck, Bahnemann, Heekeren, Wolf & Convit, 2008) or the recognition of an intention (Faux Pas Test, see Baron-Cohen, O’Riordan, Jones, Stone & Plaisted, 1999; TASIT, see McDonald, Bornhofen, Shum, Long, Saunders & Neulinger, 2006). Other tests that focus less on emotion or motive recognition, such as the Rosenzweig Picture Frustration Test (Coché & Meehan, 1979), are based on a psychodynamic construct and assess defense mechanisms or conflicts in a psychodynamic sense (cf. Corman, 1977).

Another approach to conflict relates to the perception of social interaction instead of intrapsychic mechanisms, whereby the type of social situation plays an important role. Systematic research into the interaction between situations and people has intensified in recent years (Rauthmann, Sherman & Funder, 2015; Rauthmann, 2016). The assumption that personality traits and other intrapsychic factors (motives, mood, attitudes, etc.) play a key role in the perception and assessment of situations seems undisputed (Fleeson, 2007; Flood, Hare & Wallis, 2001; Funder, 2016; Hogan, 2009; Johnson, 2009; Rauthmann, 2016; Berge & Raad, 2001). With the creation of a universal taxonomy of situations, Rauthmann, Gallardo-Pujol, Guillaume, Todd, Nave, Sherman, ... & Funder (2014) have created an instrument that makes it possible to describe every situation in a comparable manner based on its characteristics. Eight dimensions ("DIAMONDS") seem necessary and sufficient to describe a situation including Duty (something has to be fulfilled), Intellect (deep processing of information is required / desired), Adversity (someone is threatened), Pairing (a situation is sexually / romantically charged), Positivity (the situation brings fun / joy), Negativity (the situation can lead to negative feelings), Deception (one could mistrust someone), and Sociality (meaningful social interactions are required / desired) (Rauthmann et al., 2014, Rauthmann & & Sherman, 2015b). In the case of conflicts, it is still unclear which dimensions have to come together in order to classify a situation into a class of conflict situations (van Heck, 1984). This is made more difficult by the fact that a situation assessment is not only analytical and rational, as would be the case with an objective behavioral assessment, but also unconsciously and intuitively (Strack & Deutsch, 2004; Callenmark, Kjellin, Rönnqvist & Bölte, 2014).

One irritating problem with interpersonal conflicts is that there is no comprehensive, generally accepted, empirically based theory of interpersonal conflict or its perception, that gives advice to what for instance the external and internal determinants of the perception of interpersonal conflict are. The lack of such a theory makes the test construction difficult and prevents the verification of the construct validity. From our point of view, reading publications on interpersonal processes from various fields of application resulted in the following key points. A social conflict must be distinguished from intrapersonal conflict; it requires at least two people, relates to different conflict objects, such as diverging decision tendencies, identity-related action tendencies (roles) and diverging goal tendencies (Berkel, 2006), which in turn can have all conceivable different contents. Another important point seems to be that a conflict can be not only manifest but also latent. Latent conflicts here (in contrast to psychodynamic explanations) are understood to be those in which the potential for conflict, e.g. a wage difference, may not yet have been perceived by the interacting people (Rüttinger & Sauer, 2016). Conflicts go hand in hand with a high degree of emotionality and affectivity (Kreyenberg, 2005). The perception of a conflict depends on the perspective of the observer (Heigl, 2013; Thiel, 2021) and thus remains highly subjective.

A language-free and quickly administrable instrument for recording the perception of the conflictual nature of social situations is completely lacking. The availability of a short, language-free test would be important for use on people with current concentration difficulties, people with different languages, especially in an educational, clinical or forensic context. In this way, the relevant question could be systematically examined, for example whether people with schizophrenia, autism (Flood et al., 2011), depression, ADHD (Embregts & Van Nieuwenhuijzen, 2009), borderline personality disease or healthy people differ in their anticipation of social conflict situations. In addition, such a test could find practical application in many research and application areas in medicine, psychology or education. The aim of the current study was to develop such an instrument.

Methods

Making of the items

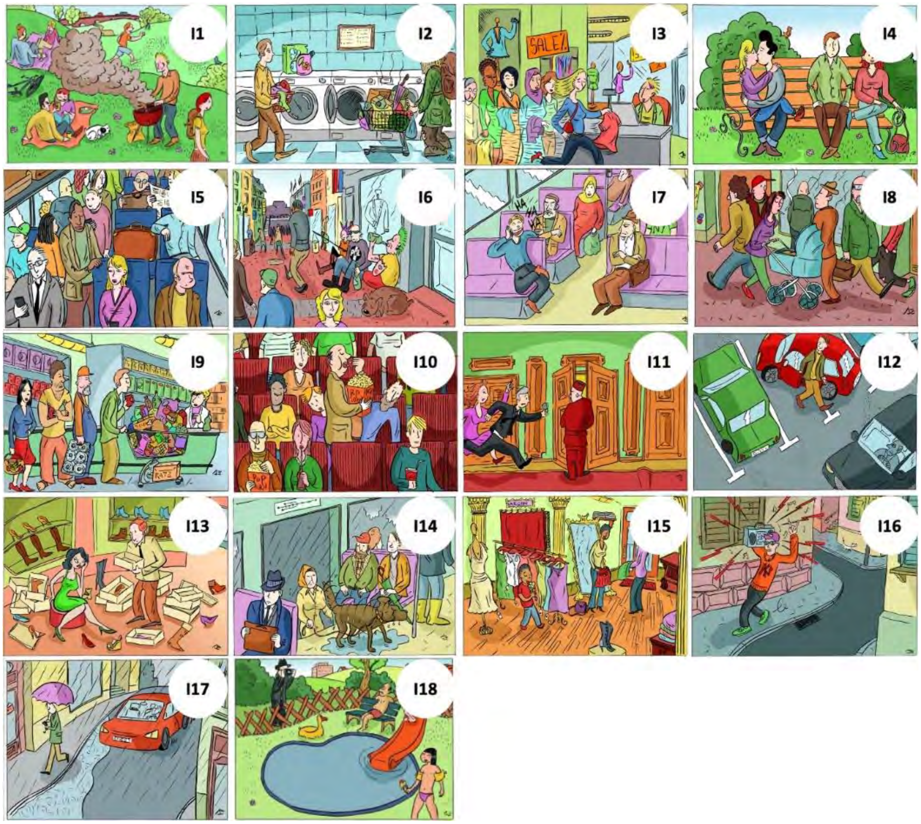
First, around 60 different everyday situations were outlined in writing, which should be evaluated with regard to possible conflicts. For example: “On a train platform. A train has arrived and is overcrowded, many people want to get off. The door is just opening. There are many people waiting to be let in to the left and right of the door. Plus, there’s a person standing right in front of the door that opens.”

Within the working group, this number of situations was reduced to 45 scenes based on convenience and difficulty (in terms of the availability of cues that make the assumption of a conflict easier or less obvious). In the next step, these verbally described situations were translated into cartoons by a professional illustrator according to our specifications. These items have been tested in advance. For the final test, items that showed ceiling and floor effects (score of 7 or 1) or had an overall item-test correlation ≤ 0.30 were removed. The number of items with corresponding social situations according to DIAMONDS has been adjusted.

Ultimately, 18 articles survived this process. The items are shown in Figure 1.

Figure 1

The items that make up the test used in this study. The items were presented in color.



Test application

The items were presented to the study participants via the online survey tool "Lime Survey" (https://www.limesurvey.org) from the Justus Liebig University of Giessen. For each scene, a 7-point rating scale was used to indicate whether or not there would be conflict in the scene.

In addition, the test subjects were given psychometric questionnaires on potentially relevant aspects of conflict perception. These included the areas of anxiety, fear, anger, intention recognition (theory of mind), and attachment type (theory). These results are reported elsewhere. The sociodemographic variables recorded were gender, age, marital status, education, employment and finances.

Sample

The sample was collected in multiple rounds using online survey tools. The only difference between each survey was that various additional tests were carried out to check aspects of validity. This kept the overall duration of the surveys short and improved participant compliance.

No representative sample was taken, but rather an “ex post facto” strategy was pursued. Participants were excluded if they did not complete the survey or if all items of the ASK test were not answered. In addition, the data was checked for quality criteria for online surveys.

No a priori sample size planning was performed for this IRT analysis study as it was originally aimed at a standard CTT analysis. There are increasing studies on sample size estimation for IRT models; Alexandrowicz & Draxler (2015) on the Rasch model with CML and Bootstrap; Dai, Vo, Kehinde, He, Xue, Demir and Wang (2021) on the graded response model and the generalized partial credit model; Draxler (2010) on Rasch models; Smith, Rush, Falowfield, Welikova & Sharpe (2008) on t-tests; Tekle, Gudicha & Vermunt (2016) on bootstrap LRT for LCM; and Zimmer (2023) on testing linear hypotheses in item response theory (IRT). In summary, there appears to be no general method for estimating sample size. Dai et al. (2021) argued that AIC, BIC or LL may not be meaningful when the sample size is less than 300 and the number of items is less than 5. However, the problem is also discussed that irrelevant model deviations could become significant when the samples are large and the tests are overpowered (Alexandrowicz & Draxler, 2015). Based on this, the expected test power for the likelihood ratio test (LRT) and the Wald-type test was estimated downwards for the sample sizes N = 300, 500 and 1000. The estimation used tcl::post_hocPCM() of the recently released R package *tcl* (Draxler & Kurz, 2023) with the study data, random group splits and bootstrapping (rep = 1000) for N = 300, 500 and 1000. The results showed for the LR-test for N = 300 a power of p = .93 [.791 < CI95 < .996], for N = 500 a power of p =.95 [.85 < CI95 < .997], and for N = 1000 a power of p =.97 [.88 < CI95 < .998]. The results for the Wald-type test showed for N = 300 a power of p = .92 [.77 < CI95 < .992), for N = 500 a power of p = .95 [.83 < CI95 < .996], and for N = 1000 a power of p = .96 [.87 < CI95 < .997]. This was taken as an indication that the model tests may be overpowered given the sample size of the final data set of N = 1831.

Statistics

Psychometric characteristics were assessed using item response theory (IRT) using polytomous response models. Main parameters of classical test theory (CTT) were also estimated.

Classical Test Theory

The internal consistency of the test was estimated as McDonald's Omega. Average item correlation and item-test correlation are reported. JAMOVI (The jamovi project (2023), *jamovi* (Version 2.3) [Computer Software] retrieved from https://www.jamovi.org) was used for calculations.

Item Response Theory

The test consisted of 7 ordered response categories (translated from german: “definitely not – probably not – not very likely – maybe – probably – very likely – definitely”) for each of the 18 items. The partial credit model was used for item and test analysis.

Preliminary analysis revealed that adjacent response categories did not follow the ordinal model. Andrich (2013) found that such an anomaly in response thresholds is due to substantive rather than statistical reasons. Accordingly, the problem with the 7-point rating scale may be that some of the category labels did not evoke the same meaning among respondents (e.g. “may be” – “probably”). This problem was solved by subsequently reducing the number of answer categories from {0-1-2-3-4-5-6} to {0-[1,2]-[3,4]-[5,6]}. The Partial Credit Model (PCM) was chosen because it proved to be superior to the Rating Scale Model (RSM) using mirt::anova(rpcm.MIRT,pcm.MIRT,gpcm.MIRT) of the R package mirt (Chalmers, 2012). Anova showed that (log-likelihood_rsm = 66857.85) > (log-likelihood_pcm = 66685.77) > (log-likelihood_gpcm = 66688.19) with p = 0.

The analysis consisted of an estimation of the parameters of the PCM, an item analysis (thresholds, location parameters, fit statistics, item information), a person analysis (ability, fit statistics, person separation reliability, misfit estimation), and model testing for Rasch conformity using the Anderson Likelihood Ratio Test (LRT) and the Martin-Loef Likelihood Ratio Test (MLoefT). The Wald-type test was calculated for the analysis of differential item functioning. LRT and Wald-type test were calculated for groups with higher and lower values (median split), gender (female, male), and age (median split), respectively. PCM and all associated analyzes were calculated using eRM (Mair & Hatzinger, 2007) and standard R functions. The graphical model test (GMT) for polytomous response models was performed using the GMX library (Alexandrowicz, 2022).

Results

Sample Description

A total of 2074 people took part in the study. The *mean age* was 35.0 years (*SD* = 13.4; *range* 14–85; *skewness* = 0.96; *kurtosis* = 0.14). Of these, 1416 (68.3%) participants reported their gender as female, 636 (30.7 %) as male, 8 (0.4 %) as non-binary and 4 (0.2 %) as both. 10 (0.5 %) did not select any of these categories, instead indicating "other".

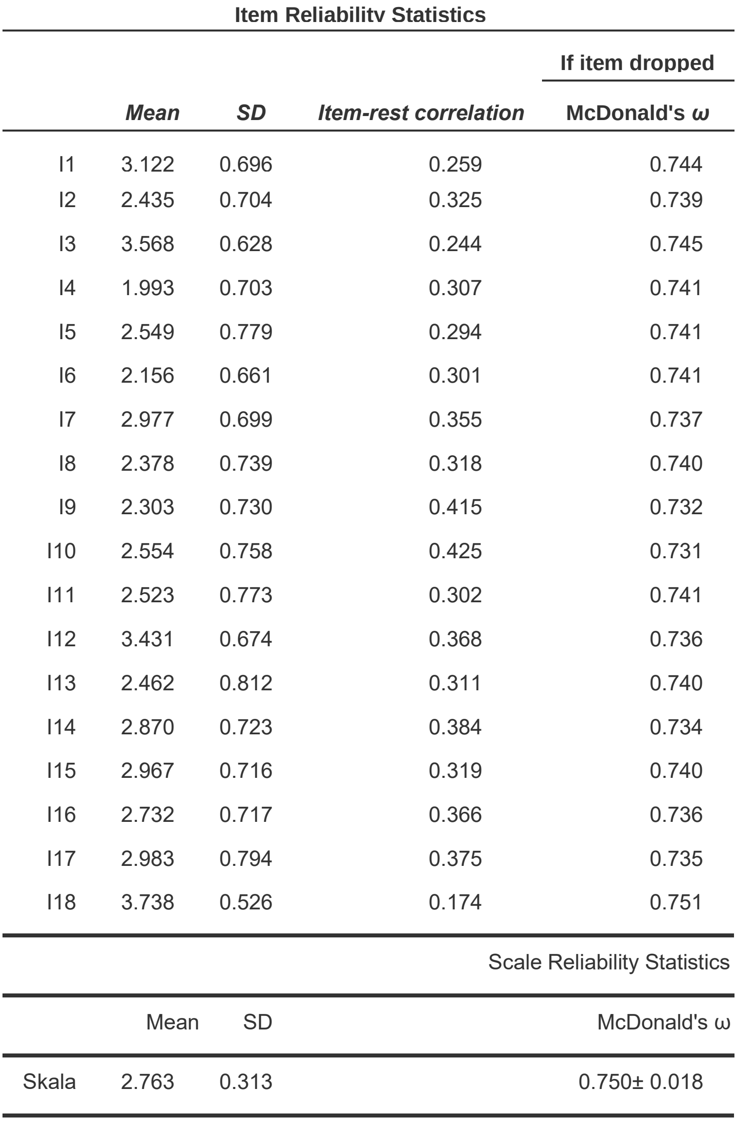
Some participants had to be excluded from the analysis due to missing data. As a result, the final data set for analysis consisted of 1831 cases (1255 women, 576 men; *mean age* = 34.9, *SD* = 13.3). Marital status (single: 54.4 %, married or living together 39.3 %, separated, divorced, widowed 5.46 %, other 0.84 %), education (high school diploma 45.3 %, university degree 39.96 %; secondary school diploma 11.2 %, other 3.6 %), professional activity (education 45.7 %, employed 45.9 %, retired 3.1 %, other 5.3 %) and finances (income < 50 % of all incomes 66, 9 %, > 50 % of all incomes 33.1 %).

CTT-based analysis

Item and scale reliability statistics were calculated for the entire data set based on classical test measurement error theory. The results are listed in Table 1. The raw values of the items ranged between 1.99 (Item 4) and 3.74 (Item 18). The standard deviation of the item scores was between 0.53 (item 18) and 0.81 (item 13). The average inter-item correlation was *aiC* = 0.15 ± 0.12. The item-rest correlations were between *r* = 0.17 (item 18) and *r* = 0.43 (item 10). The internal consistency reliability measure for the scale was calculated as McDonald's $\omega = 0.75 \pm 0.018$.

Table 1

Classical test theory (CTT) item and scale statistics



Item Reliability Statistics

				If item dropped
	Mean	***SD***	***Item-rest correlation***	**McDonald's *ω***
I1	3.122	0.696	0.259	0.744
I2	2.435	0.704	0.325	0.739
I3	3.568	0.628	0.244	0.745
I4	1.993	0.703	0.307	0.741
I5	2.549	0.779	0.294	0.741
I6	2.156	0.661	0.301	0.741
I7	2.977	0.699	0.355	0.737
I8	2.378	0.739	0.318	0.740
I9	2.303	0.730	0.415	0.732
I10	2.554	0.758	0.425	0.731
I11	2.523	0.773	0.302	0.741
I12	3.431	0.674	0.368	0.736
I13	2.462	0.812	0.311	0.740
I14	2.870	0.723	0.384	0.734
I15	2.967	0.716	0.319	0.740
I16	2.732	0.717	0.366	0.736
I17	2.983	0.794	0.375	0.735
I18	3.738	0.526	0.174	0.751

Scale Reliability Statistics

	Mean	SD	McDonald's ω
Skala	2.763	0.313	0.750± 0.018

N = 1860; Range 1-4

IRT-based analysis

In summary, the informative responses of 1831 participants to an 18-item image perception test were analyzed. Originally, seven response categories were recategorized into four categories to obtain ordered responses. The aim of the study was to examine the item responses for evidence of conformity to the polytomous Rasch model. The Partial Credit Model (Masters, 1982) to estimate the model statistics was used as implemented in the R package Extended Rasch Models ("eRm") by Mair, Hatzinger & Maier, 2020. PCM was run with "eRm::PCM(X = data)". The resulting conditional log_likelihood (CML) was -27381.68 (97 iterations, 53 parameters).

Item Analysis

The step difficulty parameters (eta) of the item category are listed in Table 2. The parameters of the categories with 0 answers were set to 0, as was the first category of the first item. The category difficulties of items 3, 12 and 14 are not ordered. The confidence intervals overlap across categories in all items. The location and thresholds estimates for each item can also be found in Table 2. It can be seen that the thresholds are ordered as expected for the PCM ($\tau 1 \leq \tau 2 \leq \tau 3$ for each item). This and the category probability curves (Figure 2) indicate that the categories are in correspondence with the latent dimension (higher category have higher locations). Such distinct categories describe a meaningful range of locations on the logit scale that represent the construct under consideration. However, since the order found is based on the subsequent merging of categories, this conclusion remains rather tentative.

Table 2

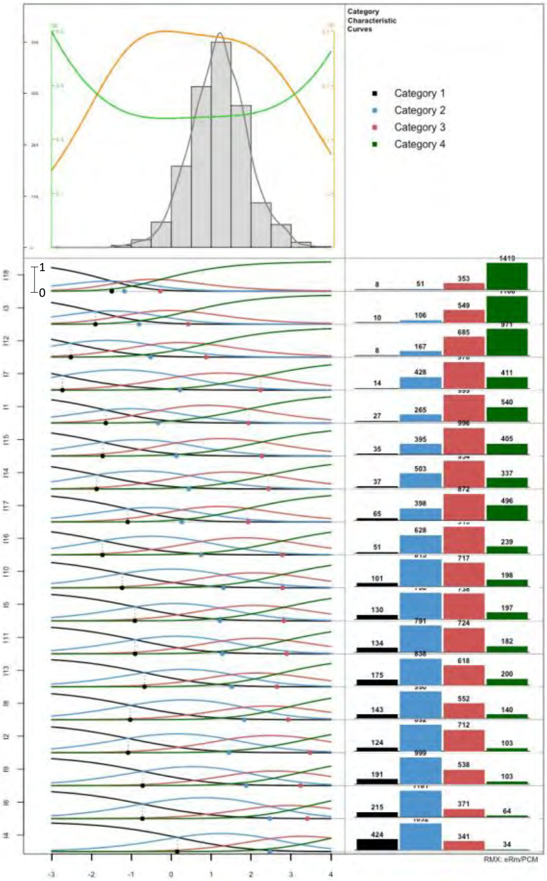
Estimates of the item-category easiness parameters (eta), standard errors and 95% confidence intervals. Last four columns contain the item location (loc) and threshold parameter estimates τ for each item.

Resp.Cat.	1				2				3				Thres		
	eta	*SE*	*CI-*	*CI+*	*Est.*	*SE*	*CI-*	*CI+*	*Est.*	*SE*	*CI-*	*CI+*	*loc*	*τ1*	*τ*
I1	-	-	-	-	-1,973	0,193	-2,352	-1,594	-0,039	0,198	-0,427	0,348	-0.013	-1.642	-0.3
I2	-1,085	0,098	-1,276	-0,894	0,358	0,107	0,149	0,566	3,837	0,150	3,544	4,131	1.279	-1.085	1.4
I3	-1,902	0,326	-2,540	-1,263	-2,709	0,310	-3,316	-2,101	-2,287	0,304	-2,883	-1,692	-0.762	-1.902	-0.8
I4	0,148	0,062	0,025	0,270	2,613	0,088	2,440	2,787	6,623	0,193	6,244	7,002	2.208	0.148	2.4
I5	-0,913	0,097	-1,103	-0,723	0,306	0,105	0,101	0,511	3,123	0,133	2,863	3,383	1.041	-0.913	1.2
I6	-0,725	0,077	-0,877	-0,574	1,740	0,098	1,548	1,932	5,147	0,161	4,832	5,462	1.716	-0.725	2.4
I7	-2,727	0,266	-3,249	-2,205	-2,511	0,260	-3,021	-2,001	-0,279	0,261	-0,791	0,233	-0.093	-2.728	0.2
I8	-1,030	0,091	-1,209	-0,850	0,795	0,104	0,592	0,998	3,725	0,138	3,453	3,996	1.242	-1.030	1.8
I9	-0,718	0,082	-0,879	-0,558	1,159	0,096	0,970	1,347	4,397	0,142	4,118	4,676	1.466	-0.718	1.8
I10	-1,233	0,107	-1,443	-1,024	0,074	0,114	-0,149	0,298	2,858	0,139	2,585	3,131	0.953	-1.234	1.3
I11	-0,907	0,095	-1,094	-0,720	0,371	0,104	0,168	0,575	3,258	0,133	2,997	3,520	1.086	-0.907	1.2
I12	-2,517	0,355	-3,213	-1,822	-3,038	0,342	-3,709	-2,367	-2,168	0,336	-2,825	-1,510	-0.723	-2.517	-0.5
I13	-0,672	0,086	-0,840	-0,504	0,840	0,097	0,650	1,031	3,488	0,126	3,241	3,735	1.163	-0.672	1.5
I14	-1,871	0,169	-2,202	-1,540	-1,432	0,168	-1,761	-1,104	1,006	0,178	0,658	1,355	0.335	-1.871	0.4
I15	-1,721	0,175	-2,064	-1,378	-1,599	0,172	-1,936	-1,262	0,668	0,180	0,315	1,021	0.223	-1.721	0.1
I16	-1,725	0,145	-2,009	-1,441	-0,977	0,146	-1,264	-0,690	1,807	0,163	1,488	2,127	0.602	-1.725	0.7
I17	-1,095	0,134	-1,358	-0,831	-0,831	0,134	-1,093	-0,569	1,090	0,145	0,806	1,374	0.363	-1.095	0.2
I18	-1,493	0,375	-2,228	-0,758	-2,668	0,348	-3,351	-1,986	-2,944	0,339	-3,609	-2,279	-0.981	-1.493	-1.1

Infit and outfit statistics for items were calculated as the squared difference (residuals) between metrics and model-based expectations (MSQ). Both indicated being productive for measurement (Bond & Fox, 2007, [0.8 < MSQ < 1.2]; often 0.5 < MSQ < 1.5) as they did not exhibit over- or under-fitting (Figure 3). This means that the stochastic information components correspond to what would be expected if the model were valid.

Figure 2

PIccc plot (according to Kabic & Alexandrowicz, 2023) of the PCM model parameters estimated with eRm. The upper left panel shows the person parameters (histogram and density curve), the test information function TIF (orange) as an additive function of the item information functions and the standard error SE (green). The test information function indicates the range of ability scores at which the test performs best. The lower left panel shows the item characteristics and thresholds for each category of each item. The items are ordered by level of difficulty. The lower right panel shows the response frequencies for each category of each item.

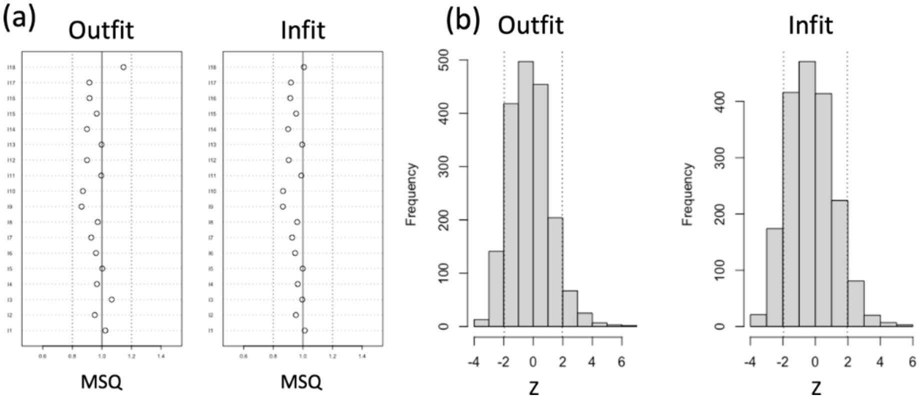


Person Analysis

Maximum likelihood estimation of the person parameters was performed. Figure 2 shows the distribution of the person parameters together with the item- and category difficulties as a person-item map. As can be seen from the test information function, the test differentiated more strongly towards the more difficult items, i.e. the perception of pictures of social situations, where it was more difficult to perceive a potential social conflict in the scene. eRm::PersonMisfit was used to count the number of people who did not fit (deviation from predicted response pattern in terms of chi-square based *Z*-score > 1.96). The results showed a proportion of 6.2 % (i.e. 112 out of 1831 individuals) of misfitting individuals in the sample (Figure 3b). This is not a large percentage considering that everyone was able to take the test without restrictions.

Figure 3

(a) Outfit and Infit Statistics (MSQ) for each item. Values between [0.8 < MSQ < 1.2] indicate agreement with model expectations. (b) Person outfit and infit statistics. 6.2 percent of people showed z > |1.96| (vertical dashed lines) and therefore did not fit the model.



The separation reliability (SepRel) was calculated using eRm::SepRel. The result was *SepRel* = 0.75. This function calculates the proportion of person variance that is not due to error. The concept of person separation reliability is more similar to reliability indices such as Cronbach's α. However, CTT and IRT reliability concepts differ fundamentally, as does the calculation of separation reliability in other program packages. In summary, it can be said that the separation reliability for the test under consideration is sufficiently high.

Table 3

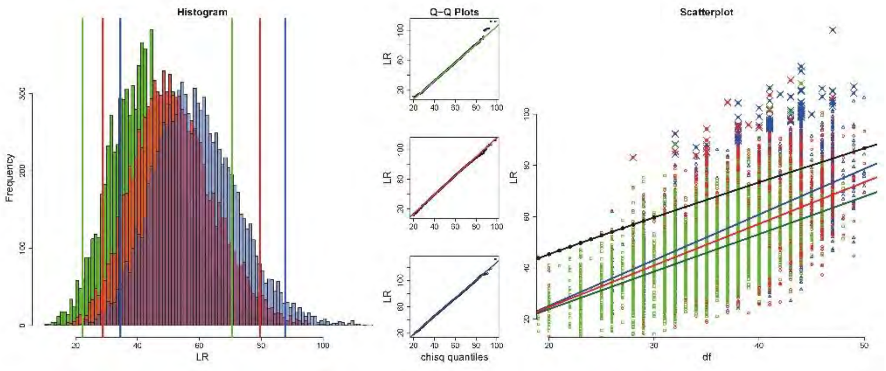
Results for the Andersen Likelihood Ratio Model Test.

sample	groups	c^2	*df*	*p*
total	score	132.9	47	.00
total	gender	114.7	53	.00
total	age	236.9	53	.00
females	score	104.0	44	.00
males	score	63.9	44	.03

The LR tests listed in Table 3 clearly failed for the total sample (*N* = 1831) and the female sample. In the smaller male sample, the p for the LR test was higher. As discussed in the Sample section, the study's larger sample size may have resulted in "significant" model tests due to small, potentially irrelevant model deviations (see graphical model test below in the Results section). Therefore, resampling was performed using a delete-d jackknife approach. To calculate the LRTs, random samples of *N* = 300, 400 or 500 were drawn from the total sample in three runs (cp. power calculation in the “Sample” section). For each sample size, 10,000 LRTs were calculated. Figure 4 shows the results of this procedure. The distribution of LRs fit the chi-square distribution well, the majority of LRs conformed to the null hypothesis. This is taken as an indication that the LR test likely overestimated the difference between the models due to the sample size.

Figure 4

Results of Jackknife – d estimation of LR-Test. Results of the LRT for random samples of size n = 300 (green), 400 (red) and 500 (blue) people. The left panel shows the histogram of the LRT values. Vertical colored lines represent 95% CI. The middle panel shows Q-Q plots. The right panel shows a scatterplot of LRT values versus df. Regression lines are displayed in the corresponding color. The black line shows the critical chi-square values at p < .05. Crosses indicate LRT values with p < .05, corrected for multiple testing (Bonferroni).



The Martin-Loef test (MLT) was calculated to examine unidimensionality by testing whether two sets of items correspond to a Rasch scale. The MLT was computed using the median split criterion and two MLTs with additional, more content-oriented criteria. We grouped items showing fewer than four people in the social scene versus ≥4 people. Another criterion for classifying the objects were images that show social scenes inside a room compared to outside scenes. The results for the MLT with split-criterion median item difficulty were *log-likelihood* = -27381.68, *LR* value = 499.224, *df* = 728, *p* = 1. Item groups were formed by group 1 (items: 1, 3, 7, 12, 14, 15, 16, 17, 18), *log-likelihood*$_{\textit{group 1}}$ = -11110.48, *log-likelihood*$_{\textit{group 2}}$ = -12702.26. The results with split-criterion "*number of persons in the scene*" were *log-likelihood* = -27381.68, *LR* = 310.867, *df* = 719, p = 1 (Item group ≥ *4 persons* [1, 3, 5, 6, 7, 8, 9, 10, 14, 15], *log-likelihood*$_{\geq 4}$ = -13860.58, *log-likelihood*$_{<4}$ = -10336.17). The results with split-criterion "*inside:outside*" were *log-likelihood* = -27381.68, *LR* = 323.434, *df* = 719, p = 1 (Item group *outside* [1, 4, 6, 8, 12, 16, 17, 18], *log-likelihood*$_{\text{outside}}$ = -9866.93, *log-likelihood*$_{\textit{inside}}$ = -14316.3). The results of the MLTs therefore support the hypothesis that the test captures the construct to be measured one-dimensionally.

The Wald-type test (WT) is a common tool among IRT-based methods for detecting differential item functioning (DIF). One of the basic assumptions of IRT models is the invariance of the item parameters. DIF represents a violation of this assumption and occurs when the item parameters have different values in different groups of subjects. The WT compares estimates of item or category parameters obtained from different groups. The Wald-type test was calculated for the same group splittings as the LRT. The results can be found in Table 4.

Table 4

The Wald-type test for the group splittings score, gender, and age. z and p values are shown for the response categories of each item.

Split. Criterion	SCORE Median (lower:higher)						Gender (female:male)						AGE Median (younger:older)					
Resp. Category	1		2		3		1		2		3		1		2		3	
	z	*p*	*z*	*p*	*z*	*p*	*z*	*p*	*z*	*p*	*z*	*p*	*z*	*p*	*z*	*p*	*z*	*p*
I1	1.60	0.11	**2.29**	**0.02**	1.77	0.08	-0.74	0.46	-0.94	0.35	-1.79	0.07	1.46	0.15	1.00	0.32	-0.08	0.94
I2	0.45	0.65	-0.65	0.52	-0.70	0.48	0.10	0.92	-0.29	0.78	0.03	0.97	-0.72	0.47	0.04	0.97	0.46	0.64
I3	0.10	0.92	0.63	0.53	0.44	0.66	0.73	0.46	0.52	0.61	-0.04	0.97	0.39	0.70	1.22	0.22	1.86	0.06
I4	0.04	0.97	-1.27	0.21	1.05	0.30	1.52	0.13	0.96	0.34	0.58	0.56	-0.38	0.70	-0.54	0.60	-0.39	0.70
I5	**2.59**	**0.01**	1.54	0.12	0.88	0.38	**-2.10**	**0.04**	-1.94	0.05	**-2.96**	**0.00**	1.63	0.10	0.82	0.42	0.78	0.43
I6	-0.87	0.38	-0.78	0.43	-0.02	0.98	0.61	0.55	-1.36	0.18	-1.01	0.31	-1.07	0.28	-1.55	0.12	-1.29	0.20
I7	-0.05	0.96	-0.24	0.81	-0.69	0.49	1.75	0.08	1.83	0.07	1.55	0.12	-1.08	0.28	-1.21	0.23	-0.98	0.33
I8	-1.77	0.08	-1.82	0.07	-0.88	0.38	-0.07	0.94	0.09	0.93	-1.01	0.31	1.60	0.11	2.03	0.04	1.64	0.10
I9	-0.40	0.69	**-2.23**	**0.03**	**-2.45**	**0.01**	0.72	0.47	**2.23**	**0.03**	0.41	0.68	0.01	0.99	0.10	0.92	-0.55	0.58
I10	-1.63	0.10	**-2.57**	**0.01**	**-3.32**	**0.00**	0.27	0.79	0.03	0.99	-0.91	0.36	-0.80	0.42	-1.60	0.11	**-3.87**	**0.00**
I11	-0.81	0.42	-0.91	0.36	-0.67	0.51	1.09	0.28	-0.36	0.72	-0.84	0.40	0.13	0.90	0.45	0.65	-0.75	0.46
I12							0.02	0.99	0.34	0.74	-0.03	0.97	0.20	0.84	0.19	0.85	0.24	0.81
I13	-1.05	0.29	-1.03	0.30	0.01	0.99	0.80	0.43	0.03	0.98	-0.93	0.35	2.06	0.04	**3.09**	**0.00**	**2.44**	**0.02**
I14	1.52	0.13	0.43	0.67	-0.14	0.89	0.99	0.32	0.70	0.49	0.14	0.89	-1.39	0.16	**-2.88**	**0.00**	**-3.99**	**0.00**
I15	0.32	0.75	-0.31	0.76	-0.26	0.79	0.87	0.39	1.70	0.09	1.68	0.09	-0.49	0.69	-0.66	0.51	-1.89	0.06
I16	-0.80	0.43	-1.58	0.11	**-2.32**	**0.02**	-0.67	0.51	-1.32	0.19	-2.31	0.02	1.20	0.23	1.86	0.06	0.86	0.39
I17	1.07	0.29	-0.13	0.90	-0.75	0.46	1.94	0.05	0.75	0.45	-1.00	0.32	-0.69	0.49	-1.45	0.15	-1.95	0.05
I18	1.94	0.05	**3.03**	**0.00**	**3.19**	**0.00**	-0.99	0.32	-0.34	0.74	-0.71	0.48	-0.33	0.74	0.50	0.63	1.27	0.20

Note: For item 12, the Wald-type test for the differential item function could not be calculated for item 12 under the median split criterion because there were not the same number of categories in both groups. To avoid such a problem, eRm suggests using a different splitting criterion. Results with |z| > 2 are printed in bold.

The graphical model test was performed according to the LRT with group split *score* using the GMX library (Alexandrowicz, 2022). The model test was plotted for thresholds and betas, 95% confidence ellipses are shown (Figure 5). The correlation between the two sets was r > .93, the regression line differed only slightly from the diagonal line. As the Wald-type test of the differential item function shows, most response categories fit the model, but some response categories deviate.

To illustrate this difference in more detail, the confidence intervals for the groups' category thresholds were plotted for the easiest (18), most difficult (4), and middle (17) items of the item set. The results are shown in Figure 6. There appears to be a general problem with Category 1 in item 18, as it has a large confidence interval that spans both other categories. The category appears to vary in age group distribution. While item 4 appears to be almost perfect, item 17, the item with medium difficulty, shows an influence of gender in the highest response category.

Figure 5

Graphical model test. Grouped by score. The solid diagonal line is the 45° line, the dashed line is the regression line. The correlation coefficient of the two sets of estimates can be found at the top left. Ellipses represent 95% CI. Those that do not intersect the 45° line differ between the two subgroups. Alpha inflation has not been taken into account in this presentation. The scaling is different in both diagrams.

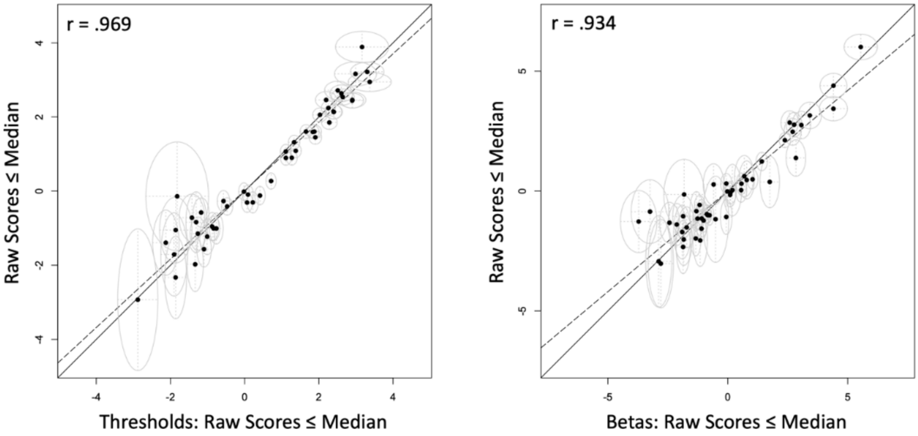
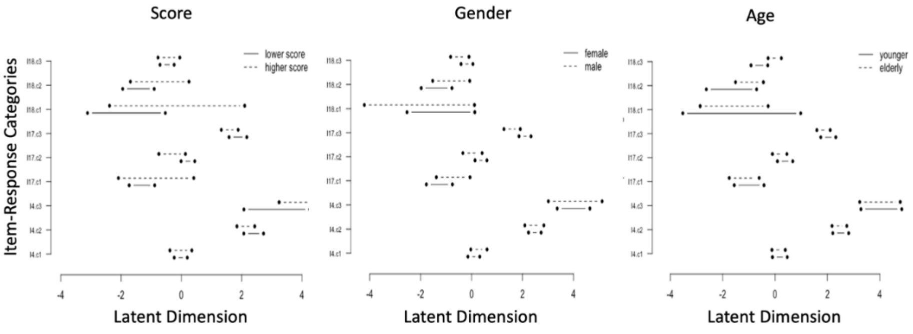


Figure 6

Differential Item Functioning with eRm::plotDIF() to plot each response category (y-axis) of items 18 (easiest item), 17 (medium level of difficulty) and 4 (most difficult item). Charts are displayed for all group splittings used for LRT (score, gender, age). 95% confidence intervals are Bonferroni corrected.



Discussion

A new picture-based method was developed and evaluated that serves to capture the anticipation of social conflicts based on perception (ASK). The focus of this work was on the analysis of the assessment quality with regard to item analysis and model conformity.

The importance of developing such a scale is that there are usually only a few tests to measure social cognition and these are particularly language-based, i.e., they require good language skills. To our knowledge, the proposed test is the only one that focuses on social conflict. Such a test is advantageous for clinical purposes, since the perception of social situations is an important prerequisite for social functioning, which is most often impaired in people with psychiatric illnesses. No less important, especially from a theoretical point of view, such a test can be very useful for the development of models of "social conflicts".

The proposed test consists of 18 items rated on a 7-point rating scale regarding the perceived likelihood that a conflict will occur. An initial preliminary analysis showed that the number of categories was not optimal. For analysis, the answers were reduced to four categories. For this new test, a comprehensive item analysis was carried out based on item response theory (IRT).

The CTT item and test analysis was also calculated. Classical Test Theory (CTT) analysis showed adequate reliability. The scale mean was 2.76 points (possible range 1-4 points). The lowest item value was 1.99 points, which shows that the data was slightly skewed to the right, i.e. more answers were given in the higher categories ("It is easier to say that there will be a conflict"). Item-residual correlations were moderately high and all items contributed uniformly to reliability. In summary, it can be said that the requirements for the CTT are met.

The partial credit model was chosen for the IRT analysis assuming equal item discrimination functions. Model tests showed that the partial credit model was superior to the rating scale model. It is likely that a generalized partial credit model would have resulted in a better model fit, but we were interested in testing a Rasch model rather than a model that allows for different discrimination between items or categories. Nevertheless, the parameters of the generalized partial credit model and the partial credit model were estimated using *mirt* (Chalmers, 2012). However, in our opinion, the results do not allow a clear statement to be made as to which items or response categories ultimately contribute to the slightly better model fit, so the PCM was retained.

The category characteristics and the associated data showed ordered categories, the item fit statistics showed no significantly different items in either the infit or the outfit statistics.

The test information function showed that the test can cover a range between approximately -2 and 3 of the latent dimension. The scale therefore tends to cover the area with a slightly higher level of difficulty in stating that a conflict will arise in the perceived social situation. With the exception of two items, the individual items also captured a broader range of the latent dimension.

Person fitting was successful for approximately 94 % of participants, while model fit parameters were above or below average for the remaining 6 %. It should be noted that this was an ex post facto sample consisting of potentially very diverse participants.

The next step is to analyze which people fall out of the tested model. Of interest, for example, is whether people with altered social cognition not only show a quantitative shift towards a reduced or increased assessment of the potential for conflict in a social situation, but also show a changed reaction pattern. For example, one might expect certain types of conflict to occur more frequently in certain social situations. This, in turn, could interact with various personal characteristics and lead to more conflictual perceptions in certain social situations. The situations were therefore qualified in advance using the DIAMONDS concept (Rauthman et al., 2014). However, this is reported elsewhere.

Three model tests were carried out with different model test approaches. While the Anderson likelihood ratio test indicated differences in item parameter estimates for groups with high vs. low scores, female vs. male, and older vs. younger participants, therefore rejecting the scale as Rasch model compliant, the scale survived the Martin-Löf-test very well. The ML test supports the hypothesis that the test captures a one-dimensional construct. This discrepancy motivated a more in-depth study of the scale.

The graphical model test for the LRT showed few and rather small differences in the estimated category parameters between the splitting groups. The correlation between the parameters of both groups was also high (> .90). Therefore, we tested by resampling whether the small p-values of the LRT tests could also be due to the larger sample size. The distribution of resampling estimates matched the theoretical distribution very well. The confidence intervals and number of p-values likely to indicate “significant” LRT were very small. It appears (see graphical model test) that the model deviations are rather small and are due to specific group differences in some response categories. Therefore, the results of the Andersen LR model test cannot be assessed clearly and conclusively.

The *Wald-type* model test was calculated to test the differential item or category function effect for the same group splittings as for the LRT. The results showed that some few response categories differed between the compared groups (score 6/54; gender 4/54; age 4/54). However, there was no consistent pattern or entire items showing differences between groups.

Limitations

Three key limitations should be discussed here. One concerns the sample (1), one concerns the problem of pooling the original response categories (2), and the third limitation concerns the weakness of the psychological theory of social conflict (3).

Anyone could take part in the study, but not everyone takes part in such a study. Experience has shown that it is mainly younger women (2 thirds) and one third younger men with higher educational qualifications who take part in studies, unless there is active recruitment within other population groups. However, post-ex facto study designs are useful when post-hoc classification of participants' characteristics is possible.

Advantages and disadvantages of combining adjacent answer categories or dichotizing answers have been discussed as "Disordered threshold Controversy" (Adams, Wu & Wilson, 2012; Andersen, 1977; Andrich, 2009, 2013; and others). Concluding, the key message is that such an approach should be followed by an empirical test of the new/changed response model to avoid circular conclusions. Since we currently have no new data for four response categories (a corresponding study has already been started), the current model is currently considered preliminary.

Another important point is that there currently appears to be no comprehensive theory of social conflict. This makes test construction more difficult from a theoretical perspective and, for example, does not allow for an assessment of construct validity. However, analysis of the data in this study shows that there is most likely a unidimensional construct underlying this. This is encouraging and stimulates the search for better theoretical support for the apriority hypothesis of this test construction venture.

Conclusion

In summary, the results of the item analysis and the model test overall show that the test is consistent with a one-dimensional Rasch model. However, some answer categories do not meet the requirements of the model brilliantly, but on the other hand they do not fail completely either. Two reasons could be considered to explain the discrepancies. First, some response categories have larger confidence intervals, and in some group splits used for model testing, there were also differences in the estimates of some response categories between groups. This shows that it is worthwhile to further investigate such effects, focusing on the characteristics of the respondents. Furthermore, the analysis approach could be extended to a generalized partial credit model that allows for unequal item discrimination functions.

In its current form, and assuming that the pooled response categories used in this analysis can be validated, the test could continue to be used to assess social cognition in the context of anticipating social conflict (ASK), preferably for research questions on the perception of conflict-prone social situations.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch Rating Model and the Disordered Threshold Controversy. Educational and Psychological Measurement, 72(4), 547-573. https://doi.org/10.1177/0013164411432166

Alexandrowicz, R. W., & Draxler, C. (2015). Testing the Rasch model with the conditional likelihood ratio test: sample size requirements and bootstrap algorithms. Journal of Statistical Distributions and Applications, 3, 1-25.

Ahlbrecht, K., Bendiek, A., Meyers, R., & Wagner, S. (2009). Konflikte: Definitionen, Erscheinungsformen und Ursachen–Versuch einer Typologie. In *Konfliktregelung und Friedenssicherung im internationalen System* (pp. 23-51). VS Verlag für Sozialwissenschaften.

Alexandrowicz RW (2022). GMX: Extended Graphical Model Checks: A Versatile Replacement of the plotGOF () Function of eRm. Psychological Test and Assessment Modeling. 64(3). 214-224.)

Andersen, E.B. Sufficient statistics and latent trait models. Psychometrika 42, 69–81 (1977). https://doi.org/10.1007/BF02293746

Andrich, D. (2009) Understanding the Polytomous Rasch model Understanding the response structure and process in the polytomous Rasch model. In Nering, M. L., & Ostini, R. (eds.) (2010). Handbook of Polytomous Item Response Theory Models, VI, pp 123-152.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any “threshold disorder controversy”. Educational and Psychological Measurement, 73(1), 78-124.

Bark, S. (2012). *Zur Produktivität sozialer Konflikte*. Springer-Verlag.

Baron-Cohen, S., O’Riordan, M., Jones, R., Stone, V., & Plaisted, K. (1999). A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*(5), 407-418.

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, 42(2), 241-251.

Berge, M. T. & Raad, B. D. (2001). The construction of a joint taxonomy of traits and situations. *European Journal of Personality*, *15*(4), 253-276.

Berkel, B. (2006). *Konflikt als Motor europäischer Öffentlichkeit: Eine Inhaltsanalyse von Tageszeitungen in Deutschland, Frankreich, Großbritannien und Österreich*. Springer-Verlag.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates

Bonacker, T. (2009). Konflikttheorien. In *Handbuch Soziologische Theorien* (pp. 179-197). VS Verlag für Sozialwissenschaften.

Brackett, M. A., & Salovey, P. (2006). Measuring emotional intelligence with the Mayer-Salovery-Caruso Emotional Intelligence Test (MSCEIT). *Psicothema*, *18*(Suplemento), 34-41.

Buck, R. (1979). Measuring individual differences in the nonverbal communication of affect: The slide-viewing paradigm. *Human Communication Research*, *6*(1), 47-57.

Callenmark, B., Kjellin, L., Rönnqvist, L., & Bölte, S. (2014). Explicit versus implicit social cognition testing in autism spectrum disorder. *Autism*, 1362361313492393.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Coché, E., & Meehan, J. (1979). Factor and cluster analyses with the Rosenzweig Picture Frustration Study. *Journal of personality assessment*, *43*(1), 39-44.

Corman, L. (1977). *Der Schwarzfuss-Test (SF-Test)*. Ernst Reinhardt.

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021, September). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. In Frontiers in Education(Vol. 6, p. 721963). Frontiers Media SA.

Draxler, C. (2010). Sample size determination for Rasch model tests. Psychometrika, 75, 708-724.

Draxler, C., Kurz, A. (2023). Testing in Conditional Likelihood Context. Cran R Package (https://cran.r-project.org/web/packages/tcl/index.html)

Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T., & Convit, A. (2008). Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET). *Journal of autism and developmental disorders*, *38*, 464-473.

Embregts, P. J. C. M., & Van Nieuwenhuijzen, M. (2009). Social information processing in boys with autistic spectrum disorder and mild to borderline intellectual disabilities. *Journal of Intellectual Disability Research*, *53*(11), 922-931.

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of personality*, *75*(4), 825-862.

Funder, D. C. (2016). Taking situations seriously: The situation construal model and the Riverside Situational Q-Sort. *Current Directions in Psychological Science*, *25*(3), 203-208.

Flood, A. M., Hare, D. J., & Wallis, P. (2011). An investigation into social information processing in young people with Asperger syndrome. *Autism*, *15*(5), 601-624.

Heigl, N. J. (2013). *Konflikte verstehen und steuern*. Springer-Verlag.

Hogan, R. (2009). Much ado about nothing: The person–situation debate. *Journal of Research in Personality*, *43*(2), 249.

Horney, K. (1973). *Unsere inneren Konflikte*. Kindler Taschenbücher.

Johnson, J. A. (2009). Wrong and right questions about persons and situations. *Journal of Research in Personality*, *43*(2), 251-252.

Jonas, K. J., Otten, M., & Doosje, B. (2014). Humiliation in conflict: Underlying processes and effects on human thought and behavior. In *Social Conflict within and between Groups* (pp. 37-54). Psychology Press.

Kabic, M., & Alexandrowicz, R. W. (2023). RMX/PIccc: An Extended Person–Item Map and a Unified IRT Output for eRm, psychotools, ltm, mirt, and TAM. *Psych,* 5(3), 948-965.

Kreyenberg, J. (2005). Transactional analysis in organizations as a systemic constructivist approach. *Transactional Analysis Journal*, *35*(4), 300-310.

McDonald, S., Bornhofen, C., Shum, D., Long, E., Saunders, C., & Neulinger, K. (2006). Reliability and validity of The Awareness of Social Inference Test (TASIT): a clinical test of social perception. *Disability and rehabilitation*, *28*(24), 1529-1542.

Mair P, Hatzinger R (2007). "Extended Rasch modeling: The eRm package for the application of IRT models in R." *Journal of Statistical Software*. 20. doi:10.18637/jss.v020.i09.)

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., ... & Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677.

Rauthmann, J. F. (2016). Motivational factors in the perception of psychological situation characteristics. *Social and Personality Psychology Compass*, *10*(2), 92-108.

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations. *European Journal of Personality*, *29*(3), 363-381.

Rauthmann, J. F., & Sherman, R. A. (2015b). Measuring the Situational Eight DIAMONDS Characteristics of Situations: An Optimization of the RSQ-8 to the S8*.

Robitzsch A, Kiefer T, Wu M (2022). *TAM: Test Analysis Modules*. R package version 4.1-4. https://CRAN.R-project.org/package=TAM

Rüttinger, B., & Sauer, J. (2016). *Konflikt und Konfliktlösen: Kritische Situationen erkennen und bewältigen*. Springer-Verlag.

Schwarz, B., & Siffert, A. (2010). Die Bedrohlichkeit elterlicher Konflikte aus Sicht der Kinder: Eine deutsche Fassung der Skala Threat aus der Children's Perception of Interparental Conflict Scale. *Diagnostica*, *56*(4), 222-229.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. BMC medical research methodology, 8, 1-11.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and social psychology review*, *8*(3), 220-247.

Steinfeld YT & Kiefer T (2015). Statistical Power Simulation for Testing the Rasch Model. Version 0.1-2. Date 2015-09-28

Tekle, F. B., Gudicha, D. W., & Vermunt, J. K. (2016). Power analysis for the bootstrap likelihood ratio test for the number of classes in latent class models. Advances in Data Analysis and Classification, 10, 209-224.

Thiel, K. (2021). Organisation, Motivation und Konflikte in der Freiwilligenarbeit. *Eine organisationspsychologische Analyse freiwilligen Engagements in Non-Profit-Organisationen. Katholische Universität Eichstätt-Ingolstadt: Dissertation.*

Van Heck, G. L. (1984). The construction of a general taxonomy of situations. *Personality psychology in Europe: Theoretical and empirical developments*, *1*, 149-164.

Zimmer, F. (2023). New Methods for Power Analysis and Sample Size Planning (Doctoral dissertation, University of Zurich).

Acknowledgement: Our special thanks go to the artist Andreas Eickenroth (https://andreas-eikenroth.de), who, in excellent collaboration, created these wonderful, timeless images that served as test objects. Thanks also go to the former master's students Bonnie Röhrig, Lena Fuß, Mareike Ludwig, Stine Ramsl and Sabrina Wirth for their help with data collection.