Qual Quant (2015) 49:2169–2185
DOI 10.1007/s11135-014-0100-1

Maximum Likelihood Difference Scaling versus Ordinal Difference Scaling of emotion intensity: a comparison

Martin Junge · Rainer Reisenzein

Published online: 21 September 2014

Abstract We compare the utility of Maximum Likelihood Difference Scaling (MLDS) and Ordinal Difference Scaling (ODS) for the measurement of emotion intensity. MLDS and ODS are both nonmetric probabilistic scaling methods based on difference measurement; however, MLDS uses quadruple comparisons (comparison of pairs of stimulus pairs) as input data, whereas ODS uses graded pair comparison judgments, where participants indicate the size of the difference between two stimuli on an ordinal response scale. In two studies using different kinds of emotional stimuli (disgust-inducing pictures, and descriptions of situations eliciting relief), quadruple comparisons and graded pair comparisons of the stimuli were collected and submitted to MLDS and ODS respectively. The scaling solutions were compared in terms of the reliability of the estimated scale values and their correlations to direct ratings of emotion intensity. The findings of both studies suggest that ODS performs at least as well as MLDS on these criteria. In addition, in most cases, good to high agreement between the scale values estimated by the two methods was found. Hence, ODS may be used as an economical alternative to MLDS for the difference scaling of emotional stimuli.

Keywords Graded pair comparisons · Quadruple comparisons · Maximum likelihood difference scaling · Ordinal difference scaling · Emotion intensity · Emotion measurement

1 Introduction

The most widely used methods of emotion measurement are based on self-reports of emotional experience (see e.g., Mauss and Robinson 2009; Pekrun and Bühner 2014). Among these, direct scaling methods (e.g., Stevens 1975; Torgerson 1958), typically in the form of category rating scales (e.g., "How happy do you feel right now on a scale from 0 = *not at all happy* to 10 = *extremely happy*?"), are by far the most common. Although ratings have several advantages (e.g., economy, ease of use), they also have serious drawbacks, including

M. Junge (✉) · R. Reisenzein
Institute of Psychology, University of Greifswald, Franz-Mehring-Str. 47, 17487 Greifswald, Germany
e-mail: martin.junge@uni-greifswald.de

limited resolution, limited reliability, and a measurement level that is only ordinal, or at best somewhere in between ordinal and interval (see e.g., Krantz et al. 1971; O'Brien 1985). To overcome the limitations of direct emotion intensity ratings, Junge and Reisenzein (2013) proposed to apply indirect scaling methods, developed in psychophysics, to the measurement of emotion intensity. In support of their proposal, they showed that a recent probabilistic scaling method based on difference measurement, *Maximum Likelihood Difference Scaling* (MLDS; e.g., Maloney and Yang 2003; Knoblauch and Maloney 2008, 2012) yielded highly reliable measurements of emotion intensity, that also fitted theoretical models of the determinants of emotion intensity much better than direct intensity ratings did.

The aim of the studies reported in this article is to compare MLDS of emotion intensity to a closely related indirect scaling method, *Ordinal Difference Scaling* (ODS; Boschman 2001; see also Agresti 1992). Like MLDS, ODS is a nonmetric, probabilistic, unidimensional scaling method based on difference measurement. However, whereas MLDS uses comparisons of pairs of stimulus pairs (quadruple judgments, QCs) as input data, ODS uses graded pair comparisons (GPCs), where participants indicate the size of the difference between two stimuli on an ordinal response scale. Paralleling this difference in scaling tasks, MLDS and ODS differ with respect to the underlying scaling models (see Sect. 1.1).

Compared to MLDS, ODS has a number of potential advantages for emotion measurement, in particular greater economy and the absence of the need to know the rank order of the stimuli on the judgment dimension (for details, see Sect. 1.1.3). Given these advantages, it is important to know whether the results of ODS are comparable to those of MLDS. The aim of the two studies reported in this article was to investigate this question empirically, using two different kinds of emotional stimuli: disgust-inducing pictures, and descriptions of situations inducing relief. In both studies, QCs and GPCs of the stimuli were collected and submitted to MLDS and ODS respectively, and the scaling solutions were compared in terms of the reliability of the estimated scale values, their correlations to direct emotion intensity ratings, and their agreement with each other. Before presenting the studies, the scaling methods are described in more detail and briefly compared.

1.1 Maximum Likelihood Difference Scaling versus Ordinal Difference Scaling

1.1.1 Maximum Likelihood Difference Scaling

Following up on seminal work by Schneider and co-workers on difference scaling (e.g., Schneider 1980; Schneider et al. 1974), Maloney and Yang (2003) proposed Maximum Likelihood Difference Scaling (MLDS) as a new unidimensional scaling method with several desirable properties (see also Knoblauch and Maloney 2008, 2012; Kingdom and Prins 2010): It is suited for the scaling of suprathreshold stimulus differences, requires only binary comparative judgments as input data, is probabilistic (i.e., explicitly takes judgment errors into account), is robust to violations of its distributional assumptions, allows missing data, yields stable solutions if only a fraction of the possible stimulus comparisons are scaled, and is grounded in an axiomatic measurement model (the difference measurement model; Krantz et al. 1971). Although MLDS has so far mainly been used for the scaling of sensory features such as image quality (e.g., Charrier et al. 2007; Maloney and Yang 2003; Menkovski et al. 2011), it has also been found suitable for the measurement of emotion intensity (Junge and Reisenzein 2013).

As its name indicates, MLDS is a method for the scaling of difference judgments. The input data to MLDS are 0/1 dominance judgments as in classical Thurstonian pair comparison scaling (Thurstone 1927; see Böckenholt 2003); however, different from the classical pair comparison task, comparisons are made between pairs of stimuli (a, b) and (c, d) rather than between single stimuli, and the participant's task is to judge which difference on the to-be measured dimension is larger: that between a and b, or that between c and d. To illustrate this quadruple comparison (QC) task[1], in Study 1 participants were presented with pairs of disgusting pictures (a, b) and (c, d) and indicated which of the two pairs differed more strongly in regard to the intensity of elicited disgust.

The statistical model underlying MLDS can be regarded as a miniature psychological theory of the judgment processes that underlie the participant's responses in the QC task. The model can be summarized by two equations. The first of these describes the relation between the scale values of the stimuli presented in a trial of the QC task, ψ_a, ψ_b, ψ_c, ψ_d (in Study 1, the intensities of disgust elicited by the stimuli) and the internal decision variable $\Delta_{ab,cd}$ that represents the result of comparing these stimuli, and is the basis of the participant's overt response $R_{ab,cd}$. The second equation maps the decision variable into the overt response.

$$\Delta_{ab,cd} = |\psi_d - \psi_c| - |\psi_b - \psi_a| + \varepsilon \quad \text{with } \varepsilon \sim \mathrm{N}(0, \sigma^2) \tag{1}$$

$$R_{ab,cd} = 1 \quad \text{if } \Delta_{ab,cd} > 0; \text{ else } R_{ab,cd} = 0 \tag{2}$$

Equation 1 implies that the participant in a QC task first (implicitly) computes the difference between the scale values of the members of the stimulus pairs (a, b) and (c, d), and then computes the difference between the (absolute) differences of these intervals, to determine which of them is larger. The internal judgment processes—including the initial mental representation of the stimuli and, in the case of emotion measurement, the elicitation of emotional reactions by the stimuli—are assumed to be biased by independent random error stemming from a normal distribution with constant variance σ^2. Equation 2 implies that, if the judgment error were zero, the difference between stimuli c and d would be judged as greater than that between a and b (i. e., $R_{ab,cd} = 1$) whenever the difference between the scale values of c and d is greater than that between a and b; however, due to the presence of error, the wrong response will occasionally be given, and this will occur more frequently, the more similar in size the compared intervals are. The aim of MLDS scaling is to estimate, from the observable responses $R_{ab,cd}$ (the 0/1 dominance judgments of the stimulus pairs (a, b) and (c, d)), the latent scale values of the stimuli assumed to underlie these responses. This is achieved by collecting responses to a sufficiently large number of quadruples (see Sect. 1.1.3), and then fitting the statistical model described by Eqs. 1 and 2 to these data using maximum likelihood estimation. As explained in Sect. 2.1.4, if the stimuli are arranged in increasing order, the MLDS model becomes a special case of binary probit regression (see Hardin and Hilbe 2007), that can be fitted using available software for generalized linear models.

1.1.2 Ordinal Difference Scaling

Whereas MLDS is appropriate for the scaling of comparisons of differences (standardly, QC judgments),ODS is tailored to the scaling of ordinal difference judgments (Boschman 2001),

[1] MLDS can also be used to scale triad judgments, where participants compare two stimuli to a third and judge which of them is closer to the target on the judgment dimension (see Devinck and Knoblauch 2012; Knoblauch and Maloney 2008, 2012). Triad judgments can be conceptualized as a special case of QCs with overlapping stimulus pairs, i.e. (a, b) is compared to (a, c). They allow to reduce the number of necessary stimulus comparisons relative to QCs.

or graded pair comparisons (Bechtel 1967; De Beuckelaer et al. 2013). The GPC task is like the classical Thurstonian (Thurstone 1927) pair comparison task in that only pairs of stimuli are judged; however, different from the classical pair comparison task, the participants are asked to state not only which of the two stimuli has the larger value on the judgment dimension, but also how much the stimuli differ from each other, using an ordered category response scale (i.e., the reported differences are assumed to have an ordinal scale level only). To illustrate, in Study 1 the participants were presented with pairs of disgusting pictures and were asked to indicate which picture was more disgusting, and how much more disgusting it was, using a response scale with 12 ordered categories from "the left picture is extremely more disgusting than the right" to "the right picture is extremely more disgusting than the left". The statistical model underlying ODS can be described by the following two equations:

$$\Delta_{a,b} = \psi_a - \psi_b + \varepsilon \ \text{ with } \ \varepsilon \sim \ N\,(0, \sigma^2) \tag{3}$$

$$R_{a,b} = j \quad \text{if } \theta_{j-1} < \Delta_{a,b} \le \theta_j, \text{ with } j = 1, \ldots, \ J$$
$$\text{and } -\infty = \theta_0 < \theta_1 < \cdots < \theta_{J-1} < \theta_J = +\infty \tag{4}$$

ψ_a and ψ_b are the scale values of the two stimuli a and b compared in a trial of the GPC task, and $\Delta_{a,b}$ is the internal decision variable on which the overt response $R_{a,b}$ is based. In addition, the ODS model contains $\theta_1, \ldots, \theta_{J-1}$ unknown thresholds separating the response categories, which, like the scale values, must be estimated. Equation 3 assumes that the participant in a GPC task (implicitly) computes the difference between the scale values of the two presented stimuli, and that the judgment process—including the initial mental representation of the stimuli and the elicitation of emotional responses—is biased by independent random error stemming from a normal distribution with constant variance σ^2. Equation 4 implies that if the judgment error were zero, the decision variable $\Delta_{a,b}$ (which in our case represents the perceived difference between the intensities of the emotions elicited by stimuli a and b) would be mapped into category j of the response scale consisting of J ordered categories, whenever $\Delta_{a,b}$ lies between the thresholds θ_{j-1} and θ_j that mark the boundaries of j on the latent continuum; however, due to the presence of error, the wrong response category will occasionally be chosen, and this will happen more frequently, the closer the stimuli are on the judgment dimension. The aim of ODS scaling is to estimate, from the observable responses $R_{a,b}$ (the ordinal difference judgments of stimuli a and b), the latent scale values of the stimuli assumed to underlie these responses.

As just described, the ODS model is a special case of the ordered (or cumulative) probit model (McKelvey and Zavoina 1975; Greene and Hensher 2010), which can be obtained in a straightforward manner from applying the ordered probit model to graded pair comparisons (Agresti 1992). Boschman's (2001) version of ODS is a restricted version of the general ODS model, in which the threshold parameters are constrained to be symmetric around the middle threshold (in case of an even number of response categories), or around the two middle thresholds (in case of an odd number of response categories). This is done to take account of the fact that the response scale in GPC tasks typically consists of two symmetric halves with identically labeled categories plus, sometimes, a middle category of "no difference".

1.1.3 A comparison of MLDS and ODS for the measurement of emotion intensity

MLDS and ODS are in many ways similar: Both are unidimensional, nonmetric, probabilistic scaling methods based on difference measurement, and both are suitable for the scaling of suprathreshold stimulus differences. What is more, the cognitive processes assumed in the scaling models of MLDS and ODS overlap: As a comparison of Eqs. 1 and 3 reveals, the first step of the QC judgment process—computing the differences between the scale values of stimuli (a, b) and (c, d)—corresponds to two instances of the first step of the GPC judgment process. Therefore, the GPC task could be regarded as an abbreviated version of the QC task, in which participants are asked to verbalize the results of the first step of the QC judgment process, assuming that they can report these results on at least an ordinal scale; whereas the second step of the QC judgment process (determining which of the two intervals is larger) is omitted.

Compared to MLDS, ODS has two potential advantages for the measurement of emotion intensity. First, ODS is more economical than MLDS, as it needs fewer input data. This is a direct consequence of the fact that the scaling task associated with ODS is based on the comparison of pairs, whereas that associated with MLDS is based on the comparison of pairs of pairs. Although the number of QCs required for MLDS can be greatly reduced by means of random or systematic subsampling, or a combination of both (Maloney and Yang 2003; see also Burton 2003), the differences to GPCs remain substantial. For example, Maloney and Yang (2003, see also Knoblauch and Maloney 2008, 2012) propose to use only quadruples (a, b, c, d) with nonoverlapping stimuli and increasing values on the to-be-measured psychological dimension (i.e., stimuli with scale values $\psi_a < \psi_b < \psi_c < \psi_d$) (as argued by Knoblauch and Maloney, this subsample prevents the possible use of judgment heuristics that circumvent the comparison of intervals). This amounts to 210 QCs for 10 stimuli (our Study 1), 495 for 12 stimuli (Study 2), and 1365 for 15 stimuli. Furthermore, simulation studies by Maloney and Yang (2003) suggest that reliable scale values can still be obtained with a random subsample of the systematically selected QCs; for 10–15 objects, about 30 % seem to be sufficient. This would mean that the number of QCs required for MLDS scaling (10, 12, 15) stimuli is (70, 165, 455). Although these numbers are manageable in basic research, they may be regarded as prohibitive in applied contexts. Furthermore, researchers are often interested in measuring several different psychological magnitudes to relate them in a theory (e.g., Junge and Reisenzein 2013), which means that separate scalings need to be made for each dimension of interest. At least in these cases, ODS recommends itself as an economical alternative: The complete set of GPCs for (10, 12, 15) objects comprises just (45, 66, 105) comparisons; and as in the case of MLDS, these can be further reduced by random or systematic subsampling.

In addition to its economy, a second consideration that speaks in favor of ODS specifically for the scaling of emotional stimuli is that, different from MLDS, ODS does not presuppose knowledge of the rank order of the stimuli on the judgment dimension (see Sect. 2.1.6 for an explanation why this rank order is needed for MLDS). For sensory dimensions such as loudness or image quality, the stimulus rank order is often evident because it is monotonically related to a physical stimulus dimension, and can therefore be specified a priori by the experimenter (Knoblauch and Maloney 2012). In contrast, the rank order of the intensities of an emotion elicited by a set of stimuli is usually not evident and can vary considerably between participants. Therefore, if emotional stimuli are scaled using MLDS, their rank order of intensity must usually be empirically estimated for each participant. If this rank order is only used to fit the MLDS model, it can be estimated from the QCs (see Sect. 2.1.6); however, if it is also used to select the quadruples for the scaling task (Maloney and Yang 2003), it must be separately estimated, for example by a preceding rank-ordering or rating task.

Although economy and the lack of need to know the stimulus order recommend GPC-based ODS for the scaling of emotional stimuli, these advantages could come at the cost of a lower quality of the scaling solution. Two considerations might lead one to expect that MLDS yields more precise estimates of scale values than ODS. First, it might be held that the "greater/smaller" judgments used as input to MLDS are easier to make, and therefore more reliable, than the graded difference judgments required for ODS. Second, it could be argued that MLDS should yield more precise estimates of scale values than ODS because it is based on a larger number of judgments.

However, on second thought, both arguments appear doubtful. As to the first argument, QCs are certainly less complex than GPCs with regard to the final stage of the judgment process (Eqs. 2 vs. 4), but they are more complex than GPCs with respect to the computations assumed to result in the internal decision variable (Eqs. 1 vs. 3). It can be argued that if the complete judgment process is considered, the cognitive requirements of the two tasks are actually quite similar: In the GPC task, participants are in each trial asked to estimate a single interval and rank-order it relative to a set of verbally described comparison intervals (e.g., "small difference", "medium difference", "large difference"); in the QC task, they are required to estimate two intervals, compare them to each other, and judge which is larger. As to the second argument, the advantage of MLDS of a larger number of input data could be balanced by the fact that a single QC judgment contains much less information than a single GPC judgment.

The two scaling studies reported in this article were conducted to provide an empirical answer to the question of how ODS compares to MLDS for the measurement of emotion intensity. In Study 1, we scaled the intensity of disgust experiences induced by pictures and in Study 2 the intensity of (imagined) relief experienced in hypothetical scenarios. In Sects. 2 and 3, the two studies are described; in Sect. 4, we summarize the results and present some proposals for future research.

2 Scaling the intensity of disgust experiences

2.1 Method

2.1.1 Participants

The participants of Study 1 were seven students and three non-students, seven of them females, with a mean age of $M = 25.1$ ($SD = 7.8$). They were paid 8 Euros per hour.

2.1.2 Materials

To induce disgust, we used 10 pictures representing major categories of disgust-eliciting objects (see Curtis and Biran 2001) and a reasonably wide range of disgust intensities. The pictures showed a moldy piece of bread and a rotten animal carcass (from the disgust category "decay and spoiled food"; Curtis and Biran 2001); a cockroach, a spider, and maggots (from the category "particular living creatures"); a toilet with feces, vomit, and a purulent finger (from the category "bodily excretions and body parts"); garbage-polluted water, and an oil-contaminated swan. The last two pictures, which elicited only mild disgust, were included to increase the range of disgust intensities. All pictures were $300 * 360$ pixels in size and were presented on notebooks with a screen resolution of $1280 * 800$ pixels.

2.1.3 Procedure

The participants performed three scaling tasks: direct scalings (ratings), graded pair comparisons (GPCs) and quadruple comparisons (QCs). All scaling tasks were programmed using DMDX (Forster and Forster 2003). Each participant performed the ratings, GPCs and QCs in this order two times, with a three days interval. The choice of this task order was motivated by the consideration that the task order should not disadvantage the QCs (we assumed that performing the rating and GPC tasks first would, if anything, facilitate the subsequent QCs). Participants received a notebook with the pre-installed scaling experiments and were instructed how to start them. They completed the tasks at home at their own leisure within a period of two weeks.

Direct scaling task (Ratings) In the rating task, the ten pictures were separately presented in an individual random order to the participants, who rated how disgusting they found each picture on an 11-point rating scale ranging from 0 = "not at all disgusting" to 10 = "extremely disgusting". Responses were entered by pressing labeled keys (0–10) on the keyboard.

Indirect scaling task I (GPCs) Following the rating task, the participants made all possible $(10 * 9)/2 = 45$ GPCs of the 10 pictures. The comparisons were presented in a different random order to each participant; furthermore, in half of the comparisons involving a given picture, it was presented on the left side of the screen and in the other half, on the right side. For each pair, participants indicated which picture was more disgusting and how much more disgusting it was. Answers were given on a bipolar 12-category response scale ranging from "The left picture is extremely more disgusting than the right" to "The right picture is extremely more disgusting than the left". The intermediate scale points on both halves of the scale were labeled "very much more", "much more", "more", "a little more", and "just barely more". The response scale was positioned below the pictures in such a way that its left half extended below the left picture and its right half below the right picture. Answers were entered by pressing appropriately labeled keys on the keyboard. An "equally intense" answer was disallowed to encourage the participants to discriminate even small intensity differences (see Böckenholt 2001). We assumed that if the participants could not detect a difference, their responses would be determined by guessing.

Indirect scaling task II (QCs) Following the GPCs, the participants made all possible QCs of the 10 pictures, i.e. the $(45 * 44)/2 = 990$ comparisons of the 45 picture pairs used in the GPC task.[2] Because of the large number of QCs, they were randomly divided into three separate blocks of 330 that were to be judged in separate sessions. Within each block, trials were individually randomized. In each trial, two pairs of disgust-eliciting pictures were presented on the left and right side of the screen, respectively, with the pictures belonging to a pair shown one above the other. The participants were asked to indicate which of the two picture pairs differed more with respect to the intensity of elicited disgust. Answers were entered by pressing the left or right arrow key. The position of the pictures was balanced across trials such that each picture pair appeared equally often on the left and right side of the screen, and each picture within a pair appeared equally often in the top and the bottom position.

2.1.4 Estimation of the scale values

The scaling models were separately fitted to the data of each participant at each of the two measurement points. The input data to MLDS were all possible QCs except those with overlapping stimulus pairs, such as (a, b; a, c). This left 630 of the 990 QCs. We also scaled the subset of 210 QCs with increasing scale values, as proposed by Maloney and Yang (2003); however, because the reliabilities of these scalings and their correlations to the direct emotion

[2] Although we only used a subset of the possible QCs for MLDS, we collected the complete set for the purpose of additional analyses (not reported in this article).

intensity ratings were on average somewhat lower than the scalings of the complete set of nonoverlapping QCs, these results are not reported in detail. For ODS, the complete set of GPCs (45) was used as input.

The parameters of the MLDS model were initially estimated using the *glm* function of R (R Core Team 2014). The binomial distribution family with a probit link was specified (Knoblauch and Maloney 2008, 2012). This specification implies a normal error distribution for the latent decision variable, as assumed in the MLDS model (Eq. 1). The ODS model was initially estimated using the R package *ordinal* (Christensen 2013), which allows to fit a variety of cumulative link models with and without constraints on the thresholds. A cumulative probit model was specified (Eqs. 3 and 4). However, in several cases estimation problems caused by separation were encountered; to overcome these, we switched to biased-reducing estimation methods (see Sect. 2.1.5).

2.1.5 Coping with separation

A technical problem that can occur when trying to fit binary regression and cumulative link models, especially with sparse data, is the occurrence of complete or quasi-complete separation (e.g., Albert and Anderson 1984; Agresti 2010; Allison 2008; Kosmidis 2014; see also, Knoblauch and Maloney 2008, 2012, for the case of MLDS). In binary regression models (MLDS), complete separation is present if one or a combination of several of the predictors permit a perfect separation of the sample responses into "0" and "1"; or in other words, if they allow a perfect prediction of the response. In cumulative link models (ODS), complete separation is present if separation exists for each of the possible collapsings of the ordinal response to a binary response (Agresti 2010 p. 64). Quasi-complete separation is present if predictability of the response is near-perfect. Although perfect or near-perfect predictability of the criterion variable would seem to be desirable, it has the undesired side-effect that unique maximum likelihood estimates of the coefficients of the responsible predictor variables do not exist. Typical indications of the occurrence of complete or quasi-complete separation are the presence of extreme values and very high standard errors for one or more parameter estimates. Some statistical programs also report that the estimation algorithm failed to converge or give some other warning message (see Allison 2008).

Although separation can occur in both MLDS and ODS, it is more likely in ODS because of the smaller number of input data. Inspection of the MLDS solutions suggested that no case of separation occurred if the full set of nonoverlapping quadruples (630) was used as input, whereas separation occurred for three of the ten participants at one or both measurement points if the subset of quadruples with increasing scale values (210) was used. Inspection of the ODS solutions suggested the occurrence of separation for four participants at one or both measurement points.

To overcome this problem, we reestimated the MLDS model using the bias-reducing maximum likelihood estimation method proposed by Firth (1993, see also Kosmidis and Firth 2009), which has been implemented, for binary regression models, in the R package *brglm* (Kosmidis 2013). Likewise, we reestimated the ODS model using bias-reducing estimation for the cumulative probit model, implemented in the R function *bpolr* (Kosmidis 2014).[3] Note that the bias-reducing estimation method provides not only an effective solution to the problem of separation (Heinze and Schemper 2002), but also reduces the bias inherent in the standard maximum likelihood estimates of regression parameters (Firth 1993). In

[3] Thanks are due to Ioannis Kosmidis, who kindly made an updated version of *bpolr* available to us.

cases not affected by separation, the scale values estimated by the bias-reducing method correlated near-perfectly (all $r > .999$) with the original maximum likelihood estimates, but were, as expected, numerically somewhat smaller. In cases affected by separation, the bias-reducing estimation method effectively reduced extreme parameter values obtained by the standard maximum likelihood estimation method. Even in these cases, however, the correlations between the original and bias-reduced scale values remained very high (all $r > .95$).

2.1.6 Determining the rank order of the stimuli

To fit the MLDS model using binary regression software such as *glm* and *brglm* in R, it is necessary to specify to the program the rank order of the stimuli on the judgment dimension (Knoblauch and Maloney 2008). The reason is that the MLDS model is nonlinear (owing to the occurrence of the absolute value function in Eq. 1). However, if the rank order of the stimuli on the judgment dimension is known, the quadruples can be reordered in such a way that $\psi_a < \psi_b$ and $\psi_c < \psi_d$ before the data are scaled (the response is recoded as necessary). As a result, the differences $(\psi_d - \psi_c)$ and $(\psi_b - \psi_a)$ in Eq. 1 are always positive, and the absolute signs in the equation can be dropped (see Knoblauch and Maloney 2008, 2012). The MLDS model then becomes a standard generalized linear model.

As mentioned in Sect. 1.1.3, the rank order of the intensities of an emotion elicited by a set of stimuli is usually not known a priori. To address this problem, we estimated the stimulus rank order from the available data, trying several different methods: (a) nonmetric multidimensional scaling (NMDS) restricted to one dimension (Kruskal 1964; Cox and Cox 2001); (b) a partial enumeration method, the "branch-and-bound" method proposed by Brusco and Stahl (2005), as implemented in the R package *seriation* (Hahsler et al. 2008); and (c) a linear scaling of the GPC judgments based on an additive functional measurement model (AFM; Anderson 1970, see Junge and Reisenzein 2013). The first two methods can be used to estimate the rank order implicit in GPCs as well as (after suitable data transformations) in QCs, whereas the third method is only applicable to GPCs.[4] We then fitted MLDS models to the QC judgments, using the rank orders of the stimuli suggested by the different estimation methods, and compared the fit of the models using the Akaike Information Criterion (AIC_c; Sugiura 1978; Hurvich and Tsai 1991).

We found that the AFM-based rank order yielded consistently better fits (lower AIC_c values) than the rank orders estimated by NMDS and the partial enumeration method. In addition, the retest correlations of the estimated MLDS scale values were highest for the AFM-based rank order. Therefore, we used the AFM results to specify the rank order of the stimuli to MLDS.[5]

2.2 Results

The mean disgust ratings for the 10 pictures, aggregated across the two measurement points, ranged from $M = 1.25$ ($SD = 2.05$) for the least disgusting picture to $M = 7.65$ ($SD = 2.28$) for the most disgusting picture. A two-way within-subjects analysis of variance (ANOVA)

[4] Details of the estimation procedures are available from the first author on request.

[5] It might be objected that using the GPC-based AFM rank order of stimulus intensities for the estimation of the MLDS model inflates the agreement between the ODS and the MLDS solution, because the latter is constrained by the rank order of the stimuli implicit in the GPC judgments. However, the finding that the AFM-based rank order yielded the best MLDS fits speaks against this possibility and suggests, rather, that the GPC judgments contained more reliable information about the stimulus order than the QCs.

Table 1 Comparison of MLDS and ODS scale values, Study 1

Participant no.	Fit (AIC_c)				Reliability			Agreement		Correlation with ratings			
	t_1		t_2		r_{t_1,t_2}			$r_{MLDS,ODS}$		t_1		t_2	
	MLDS	ODS	MLDS	ODS	Rating	MLDS	ODS	t_1	t_2	MLDS	ODS	MLDS	ODS
1	352.86	125.83	258.35	87.08	.85	.99	.91	.94	.98	.91	.92	.97	.96
2	272.96	136.57	315.73	114.59	.71	.99	.98	.96	.98	.59	.73	.94	.94
3	470.70	145.84	479.79	101.08	.73	.79	.95	.24	.48	.18	.89	.32	.95
4	286.41	103.67	282.11	93.72	.89	.96	.95	−.18	−.22	−.40	.82	−.45	.95
5	647.86	131.43	489.67	128.26	.74	.37	.75	.75	.23	.40	.74	−.07	.85
6	294.54	131.66	250.46	90.57	.86	.98	.93	.93	.99	.93	.79	.98	.97
7	561.75	129.41	625.45	173.04	.68	.99	.94	.92	.97	.60	.62	.86	.93
8	524.88	127.93	576.79	194.39	.83	.99	.86	.99	.88	.92	.94	.95	.95
9	537.19	144.01	523.42	154.42	.83	.99	.97	.96	.92	.82	.74	.97	.96
10	626.08	149.19	535.32	119.61	.84	.61	.89	.73	.83	.43	.87	.87	.71
M	457.52	132.55	433.71	125.68	.79	.87	.91	.72	.70	.54	.81	.63	.92
SD	144.31	12.97	142.24	36.98	.07	.21	.07	.39	.41	.42	.10	.52	.08

MLDS Maximum Likelihood Difference Scaling, ODS Ordinal Difference Scaling, t_1 1st measurement, t_2 2nd measurement

with factors stimuli and measurement occasion revealed a significant main effect of stimuli, $F(9, 81) = 12.25$, $p < .001$, but no significant main effect of measurement occasion, $F(1, 9) < 1$, and no significant interaction between stimuli and occasion, $F(9, 81) = 1.47$, $p = .18$. This suggests that the disgust feelings elicited by each picture were of similar intensity at the two measurement points.

2.2.1 Fit of the MLDS and ODS models

As a preliminary datum, we report the fit of MLDS and ODS models to their respective input data. Model fit was evaluated using AIC_c. The fit values obtained for the individual participants at the two measurement points are shown in Table 1 (columns 1 to 4). Lower AIC_c values reflect a better fit. If the MLDS models were estimated from the 210 quadruples with increasing scale values instead of the 630 quadruples with nonoverlapping elements, the AIC_c fit values ($M = 133.9$) were practically identical to those of the ODS model ($M = 129.12$); however, as mentioned, the retest reliabilities of the MLDS scalings and their correlations to the ratings decreased in several cases. It may be concluded that, if each model is fitted to the data appropriate for it, similar overall fit values can be achieved.

2.2.2 Retest reliabilities

As an estimate of the reliabilities of the indirect scalings and ratings, we computed the correlations between the respective variables at the two measurement points. These retest reliabilities are shown in Table 1 (columns 5 to 7). As expected, for most participants, the reliabilities of the indirect scalings ($M_r = .87$ for MLDS and .91 for ODS) were substantially higher than those of the ratings ($M_r = .79$), whereas the reliabilities of the ODS and MLDS scalings were comparable.

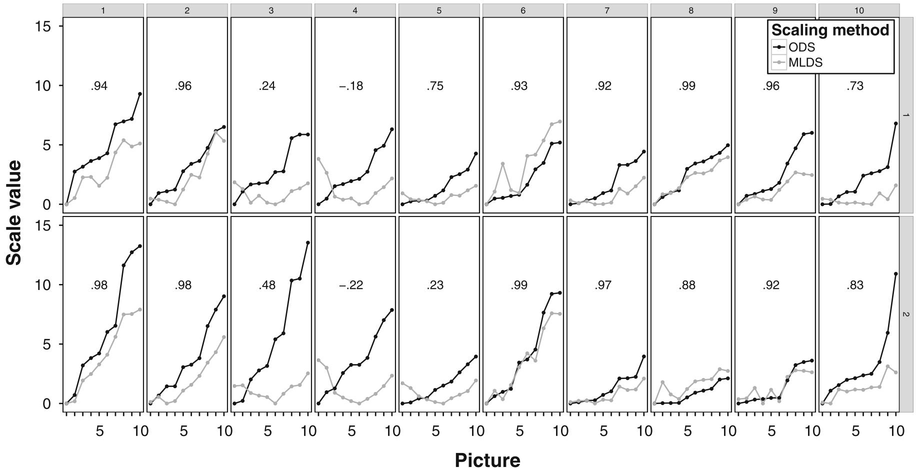

Fig. 1 Individual MLDS and ODS scale values estimated from QCs (*grey*) and GPCs (*black*) respectively, separately for the first (*top*) and second (*bottom*) measurement point. To facilitate comparison, stimuli have been ordered by their ODS scale values, and the scale values have been transformed to have zero as the minimum. Correlations between the scale values are also shown

2.2.3 Agreement of the MLDS and ODS scale values

If GPC-based ODS and QC-based MLDS are equivalent scaling methods for the stimuli used in Study 1, then the estimated scale values should be highly (linearly) correlated. Figure 1 shows graphs of these scale values and the corresponding correlations, separately for each participant and measurement point (see also Table 1, columns 8 and 9). As can be seen, in most cases, the correlations were from good to high. Low (in two cases even negative) between-method correlations were obtained for two participants (No. 3 and 4), and a third (No. 5) had a low correlation at the second measurement point. The latter case can be partly explained by the low reliability of the MLDS scaling. In contrast, for participants No. 3 and 4, the inter-method correlation was low despite high reliabilities of both the MLDS and ODS scale values.

2.2.4 Correlations between indirect scale values and ratings

If one accepts that the indirect scaling procedures and the direct ratings used in our study are different methods for measuring the same latent variable (the intensity of disgust), their correlation can be interpreted as an index of the validity of the indirect scalings (although due to the limited reliability of the variables, in particular the ratings, this correlation cannot be expected to be perfect). As can be seen from Table 1 (columns 10–13), the correlations between the indirect scale values and the direct ratings of disgust intensity were higher if the indirect scale values were estimated using ODS ($M_r = .81$ for the first and .92 for the second measurement point) than if they were estimating using MLDS ($M_r = .54$ and .63). This difference remained stable after correcting for possible outliers by computing the 10 % trimmed mean.

2.2.5 MLDS scaling of the GPCs

Junge and Reisenzein (2013) previously used MLDS for the measurement of emotion intensity and found the method to be superior to direct ratings of emotion intensity. However, different from the present studies, the QCs required as input to MLDS were not obtained in a QC task, but were analytically derived from GPCs by assuming (see e.g., Roberts 1979) that (a, b) > (c, d) if the graded comparison of a and b has a higher rank than that of c and d. A possible advantage of scaling GPCs with MLDS rather than ODS is that the MLDS model contains no thresholds and therefore has much fewer parameters than the ODS model, which can help to avoid estimation problems in ODS. However, it can be argued that MLDS is not entirely appropriate for GPC-derived QCs, because they are not statistically independent (as each pair of stimuli occurring in the derived QCs is judged only once in the GPC task). To check whether MLDS of GPC-based QCs nonetheless yields similar scale values as the theoretically more appropriate ODS scaling of the GPCs, we expanded the GPCs to QCs and submitted the latter to MLDS. Again only the 630 QCs with nonoverlapping stimuli were considered. Of these, quadruples for which the intervals (a, b) and (c, d) were equal according to the GPC judgments (21.8 %) had to be excluded because MLDS (in its current form) does not allow for equal responses.

We found that the MLDS scalings of GPC-derived QCs correlated very strongly with the ODS scalings of the original GPCs ($M_r = .99$, $SD = .02$). This finding suggests that the obtained differences between MLDS (of directly made QCs) and ODS reported in Table 1 and Fig. 1, are primarily due to differences in the input data.

2.3 Discussion

In Study 1, the intensities of disgust feelings evoked by ten pictures were estimated from QCs using MLDS, and from GPCs using ODS. The scaling solutions were compared with regard to the retest reliability of the estimated scale values and their correlation with direct scalings (ratings). The findings suggest that GPC-based ODS performs at least as well as QC-based MLDS on these criteria. In addition, in most cases, good to high agreement between the scale values estimated by MLDS and ODS was found. Finally, we found that MLDS of GPC-derived QCs yields nearly the same scale values as ODS of the original GPCs.

3 Scaling the intensity of relief experiences

3.1 Method

3.1.1 Participants

The participants of Study 2 were five males and five females from the same subject pool as in Study 1, with a mean age of 25.8 years ($SD = 7.8$). They were paid 8 Euros per hour.

3.1.2 Materials

The participants judged brief (one- or two sentence) descriptions of 12 hypothetical situations that were found to induce (imagined) relief feelings from low to high intensity in a previous study (unpublished). The scenarios depicted common relief-inducing situations of student life, such as “You make it for the lecture just in time”, “You finally found an apartment for rent”, or “A friend of yours has stopped smoking”. The scenarios were presented on the computer monitor in small text boxes.

3.1.3 Procedure

As in Study 1, the participants worked on three tasks: direct scalings (ratings), GPCs and QCs. The tasks were completed (in this order) two times, within a three days interval.

Direct scaling task (Ratings) In the rating task, the 12 relief scenarios were separately presented to the participants in random order. For each scenario, they rated how relieved they would feel in the described situation if it were real, using an 11-point rating scale ranging from 0 = “not at all relieved” to 10 = “extremely relieved”.

Indirect scaling task I (GPCs) Following the direct ratings of relief intensity, the participants made all possible $12 * 11/2 = 66$ GPCs of the 12 scenarios. The stimulus pairs were presented in a different random order to each participant. For each pair, the participants indicated which situation would be more relieving, and how much more. Answers were given on a bipolar 12-category response scale ranging from “The left situation would be extremely more relieving than the right” to “The right situation would be extremely more relieving than the left”. Intermediate scale points were labeled analogous to Study 1.

Indirect scaling task II (QCs) The complete set of QCs for 12 objects comprises $66 * 65/2 = 2145$ comparisons, a number beyond what can be reasonably demanded of a participant. To reduce the number of QCs, we first omitted quadruples comprised of pairs with nonoverlapping stimuli (cf. Study 1). From the remaining 1485 comparisons, we drew a semi-random sample of 1/3 with the restriction that each object appeared equally often (165 times) in the comparisons. This resulted in 495 QCs, the same number that would be obtained using the selection method proposed by Maloney and Yang (2003), i.e. if only QCs with increasing scale values were included (see Sect. 1.1.3, this method could not be used in our case because the rank order of the stimulus intensities was not known a priori). Apart from this difference, the QC judgment procedure was analogous to Study 1: In each trial, the participants were presented with two pairs of situations, and judged which pair differed more with respect to the intensity of relief elicited by the described events.

3.1.4 Estimation of the scale vvalues

As in Study 1, MLDS and ODS scale values were estimated using bias-reducing maximum likelihood estimation (Kosmidis 2013, 2014), and linear scaling (AFM) of the GPCs was used to specify the rank order of the stimuli to MLDS.

3.2 Results

The mean relief ratings for the 12 scenarios, aggregated across the two measurement points, ranged from $M = .75$ ($SD = 2.17$) to $M = 8.4$ ($SD = 1.96$). A two-way within subjects ANOVA with factors stimuli and measurement occasion revealed a significant main effect of stimuli, $F(11, 99) = 18.02$, $p < .001$, but no significant main effect of measurement occasion, $F(1, 9) = 3.03$, $p = .12$, and no significant interaction between stimuli and occasion, $F(11, 99) < 1$. This suggests that the relief feelings elicited by each scenario were of similar intensity at the two measurement points.

Table 2 Comparison of MLDS and ODS Scale Values, Study 2

Participant no.	Fit (AIC_c)				Reliability			Agreement		Correlation with ratings			
	t_1		t_2		r_{t_1,t_2}			$r_{MLDS,ODS}$		t_1		t_2	
	MLDS	ODS	MLDS	ODS	Rating	MLDS	ODS	t_1	t_2	MLDS	ODS	MLDS	ODS
1	246.59	154.02	179.97	109.88	.98	.98	.97	.98	.98	.98	.96	.99	.98
2	318.44	180.65	238.15	148.52	.83	.95	.90	.93	.96	.76	.83	.91	.98
3	291.24	187.47	262.72	151.99	.87	.15	.95	.95	.24	.90	.94	−.12	.85
4	342.72	205.01	234.65	180.78	.77	.96	.96	.96	.99	.88	.90	.96	.97
5	279.82	151.90	238.33	136.35	.69	.99	.98	.97	.99	.66	.66	.96	.99
6	424.91	178.12	170.30	203.98	.71	.96	.91	.75	.92	.71	.85	.97	.94
7	266.10	266.61	548.84	212.33	.37	.40	.58	.18	−.11	−.02	.53	−.30	.78
8	429.98	170.64	357.33	219.58	.49	.74	.76	.73	.26	.66	.88	.09	.87
9	279.66	164.09	126.87	139.75	.66	.77	.77	.82	.99	.48	.84	.97	.97
10	237.05	227.61	595.84	189.42	.72	.49	.73	.88	.90	.86	.95	.85	.91
M	311.65	188.61	295.30	169.26	.71	.74	.85	.82	.71	.69	.83	.63	.92
SD	68.45	35.77	158.82	37.03	.18	.30	.13	.24	.41	.29	.14	.52	.07

MLDS Maximum Likelihood Difference Scaling, ODS Ordinal Difference Scaling, t_1 1st measurement, t_2 2nd measurement

3.2.1 Fit of the MLDS and ODS models

The fit values (AIC_c.) of the MLDS and ODS models to their respective input data (Table 2, columns 1 to 4) were similar to those obtained in Study 1.

3.2.2 Retest reliabilities

As can be seen from comparing Table 2 to Table 1, the retest reliabilities of the indirect scalings and ratings were on average somewhat lower than those obtained in Study 1. This may have been due to the nature of the stimuli (written scenarios), which were more complex, and therefore probably more difficult to rate and compare than the pictures used in Study 1. However, as in Study 1, the reliabilities of the indirect scalings ($M_r = .74$ for MLDS and $M_r = .85$ for ODS) were higher than those of the ratings, whereas the reliabilities of the indirect scalings were in most cases similar (exceptions are participants No. 3 and 10).

3.2.3 Agreement of the MLDS and ODS scale values

As shown in Table 2 (columns 8 and 9), the correlations between the MLDS and ODS scale values were in most cases from good to high: For one participant (No. 7) the correlations were low at both measurement occasions; for two more participants (No. 3 and 8), the correlation was low at the second measurement point. In the case of participants No. 3 and 7, the low correlations can at least in part be attributed to low reliability of one of the scalings.

3.2.4 Correlations between indirect scale values and ratings

As can be seen from Table 2 (columns 10 to 13), the correlations between the indirect and the direct scalings (ratings) of relief intensity were on average higher for ODS ($M_r = .83$ for the first and .92 for the second measurement) than for MLDS ($M_{\mathrm{r}} = .69$ and .63, respectively), and this difference remained stable after correcting for possible outliers by computing the 10 % trimmed mean.

3.2.5 MLDS scaling of the GPCs

As in Study 1, the GPCs were also submitted to MLDS, after first expanding them to QCs. Again, the obtained scale values correlated highly with the ODS scalings of the original GPCs ($M_{\mathrm{r}} = .98$, $SD = .02$).

3.3 Discussion

In Study 2, we estimated the intensity of relief feelings in 12 hypothetical scenarios from QCs using MLDS, and from GPCs using ODS. The results corroborated those of Study 1: The retest reliabilities of the two indirect scalings were similar for most participants, their correlation to the direct ratings were in most cases as high or higher for ODS, and there was from good to high agreement of the obtained solutions for most participants. We also replicated the finding of Study 1 that MLDS of GPC-derived QCs yields nearly the same scale values as ODS of the original GPCs.

4 General discussion

In two scaling studies of emotion intensity, we compared MLDS to ODS with regard to the retest reliability of the estimated scale values and their correlation with direct ratings. The results suggest that, at least for the stimuli used in our studies, GPC-based ODS is equivalent, and certainly not inferior, to QC-based MLDS on these criteria. In addition, in most cases, good to high agreement between the scale values estimated by MLDS and ODS was found.

Because ODS requires much fewer input data than MLDS and does not presuppose knowledge of the stimulus order on the judgment dimension, these findings recommend GPC-based ODS as an economical alternative to MLDS for the measurement of emotion intensity. Its combination of precision and economy could make ODS particularly attractive to researchers as an alternative to direct ratings and other direct scaling methods (e.g., Stevens 1975) for the measurement of emotions, as well as the measurement of other psychological magnitudes beyond the domain of classical psychophysics, such as preferences, attitudes, and personality judgments. It may also be noted that several other scaling models can be applied to GPCs, including nonmetric multidimensional scaling (restricted to one dimension) and additive functional measurement (AFM; see Junge and Reisenzein 2013).

4.1 Future research

Scaling models such as MLDS and ODS can be regarded as miniature psychological theories of the judgment processes involved in particular scaling tasks. To further evaluate these models, it would therefore be reasonable to investigate the cognitive processes that underlie the associated judgment tasks in more detail. Process tracing methods could be helpful for this purpose. For example, eye-tracking (e.g., Schulte-Mecklenbeck et al. 2011) could be used to test hypotheses about how participants actually make quadruple judgments. Furthermore, with respect specifically to the measurement of emotion intensity, it would be desirable to relate the process assumptions of scaling models to more general models of emotional introspection (i.e., the introspection of emotions; e.g., Robinson and Clore 2002). The proposed research could provide a deeper explanation of the advantages of indirect scaling methods for the measurement of emotion intensity and beyond that, may aid the development of new self-report methods based on theoretical models of emotional introspection.

References

Agresti, A.: Analysis of ordinal paired comparison data. J. Royal Stat. Soc. Series C **41**, 287–297 (1992)

Agresti, A.: Analysis of ordinal categorical data. Wiley, Hoboken (2010)

Albert, A., Anderson, J.A.: On the existence of maximum likelihood estimates in logistic regression models. Biometrika **71**, 1–10 (1984)

Allison, P.: Convergence failures in logistic regression. SAS global forum (2008). http://www2.sas.com/proceedings/forum2008/360-2008.pdf

Anderson, N.H.: Functional measurement and psychophysical judgment. Psychol. Rev. **77**, 153–170 (1970)

Bechtel, G.G.: The analysis of variance and pairwise scaling. Psychometrika **32**, 47–65 (1967)

Böckenholt, U.: Thresholds and intransitivities in pairwise judgments: a multilevel analysis. J. Educ. Behav. Stat. **26**, 269–282 (2001)

Böckenholt, U.: Thurstonian-based analyses: past, present, and future utilities. Psychometrika **71**, 615–629 (2003)

Boschman, M.C.: DifScal: a tool for analyzing difference ratings on an ordinal category scale. Behav. Res. Methods Instr. Comput. **33**, 10–20 (2001)

Brusco, M., Stahl, S.: Branch-and-bound applications in combinatorial data analysis. Springer, New York (2005)

Burton, M.L.: Too many questions? the uses of incomplete cyclic designs for paired comparisons. Field Methods **15**, 115–130 (2003)

Charrier, C., Maloney, L.T., Cherifi, H., Knoblauch, K.: Maximum likelihood difference scaling of image quality in compression-degraded images. J. Opt. Soc. America **24**, 3418–3426 (2007)

Christensen, R.H.B.: ordinal–regression models for ordinal data (2013). R package version 2013.9-30 http://www.cran.r-project.org/package=ordinal/

Cox, T.F., Cox, M.A.A.: Multidimensional scaling. Chapman & Hall, Boca Raton (2001)

Curtis, V., Biran, A.: Dirt, disgust and disease: is hygiene in our genes? Perspect. Biol. Med. **44**, 17–31 (2001)

De Beuckelaer, A., Kampen, J.K., Van Trijp, H.C.: An empirical assessment of the cross-national measurement validity of graded paired comparisons. Qual. Quant. **47**, 1063–1076 (2013)

Devinck, F., Knoblauch, K.: A common signal detection model accounts for both perception and discrimination of the watercolor effect. J. Vision **12**, 1–14 (2012)

Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika **80**, 27–38 (1993)

Forster, K.I., Forster, J.C.: DMDX: a windows display program with millisecond accuracy. Behav. Res. Methods Instr. Comput. **35**, 116–124 (2003)

Greene, W.H., Hensher, D.A.: Modeling ordered choices: a primer. Cambridge University Press, Cambridge (2010)

Hahsler, M., Hornik, K., Buchta, C.: Getting things in order: an introduction to the R package seriation. J. Stat. Softw. **25**, 1–34 (2008)

Hardin, J., Hilbe, J.M.: Generalized linear models and extensions. A Stata Press publication. Taylor & Francis, College Station (2007)

Heinze, G., Schemper, M.: A solution to the problem of separation in logistic regression. Stat. Med. **21**, 2409–2419 (2002)

Hurvich, C.M., Tsai, C.L.: Bias of the corrected AIC criterion for underfitted regression and time series models. Biometrika **78**, 499–509 (1991)

Junge, M., Reisenzein, R.: Indirect scaling methods for testing quantitative emotion theories. Cognit. Emot. **27**, 1247–1275 (2013)

Kingdom, F.A.A., Prins, N.: Psychophysics: a practical introduction. Elsevier, London (2010)

Knoblauch, K., Maloney, L.T.: MLDS: maximum likelihood difference scaling in R. J. Stat. Softw. **25**, 1–28 (2008)

Knoblauch, K., Maloney, L.T.: Modeling psychophysical data in R. Springer, New York (2012)

Kosmidis, I.: brglm: bias reduction in binary-response generalized linear models (2013). http://www.ucl.ac.uk/ucakiko/software.html. R package version 0.5–9

Kosmidis, I.: Improved estimation in cumulative link models. J. Royal Stat. Soc. **76**, 169–196 (2014)

Kosmidis, I., Firth, D.: Bias reduction in exponential family nonlinear models. Biometrika **96**, 793–804 (2009)
Krantz, D., Luce, R., Suppes, P., Tversky, A.: Foundations of measurement, Vol. 1: additive and polynomial representations. Academic Press, New York (1971)
Kruskal, J.B.: Nonmetric multidimensional scaling: a numerical method. Psychometrika **29**, 115–129 (1964)
Maloney, L.T., Yang, J.N.: Maximum likelihood difference scaling. J. Vision **3**, 573–585 (2003)
Mauss, I.B., Robinson, M.D.: Measures of emotion: a review. Cognit. Emot. **23**, 209–237 (2009)
McKelvey, R., Zavoina, W.: A statistical model for the analysis of ordinal level dependent variables. J. Math. Sociol. **4**, 103–120 (1975)
Menkovski, V., Exarchakos, G., Liotta, A.: The value of relative quality in video delivery. J. Mobile Multimed. **7**, 151–162 (2011)
O'Brien, R.: The relationship between ordinal measures and their underlying values: why all the disagreement? Qual. Quant. **19**, 265–277 (1985)
Pekrun, R., Bühner, M.: Self-report measures of academic emotions. In: Pekrun, R., Linnenbrink-Garcia, L. (eds.) International handbook of emotions in education, pp. 561–579. Taylor & Francis, New York (2014)
R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2014). http://www.R-project.org/
Roberts, F.S.: Measurement theory: with applications to decisionmaking, utility, and the social sciences. Addison-Wesley, Reading (1979)
Robinson, M.D., Clore, G.L.: Episodic and semantic knowledge in emotional self-report: evidence for two judgment processes. J. Personal. Soc. Psychol. **83**, 198–215 (2002)
Schneider, B.: Individual loudness functions determined from direct comparisons of sensory intervals. Percept. Psychophys. **28**, 493–503 (1980)
Schneider, B., Parker, S., Stein, D.: The measurement of loudness using direct comparisons of sensory intervals. J. Math. Psychol. **11**, 259–273 (1974)
Schulte-Mecklenbeck, M., Kühberger, A., Ranyard, R.: The role of process data in the development and testing of process models of judgment and decision making. Judgm. Dec. Making **6**, 733–739 (2011)
Stevens, S.S.: Psychophysics. Wiley, New York (1975)
Sugiura, N.: Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Stat. **7**, 13–26 (1978)
Thurstone, L.L.: A law of comparative judgment. Psychol. Rev. **34**, 273–286 (1927)
Torgerson, W.S.: Theory and methods of scaling. Wiley, New York (1958)