COGNITION AND EMOTION, 2013
Vol. 27, No. 7, 1247–1275, http://dx.doi.org/10.1080/02699931.2013.782267

Routledge
Taylor & Francis Group

Indirect scaling methods for testing quantitative emotion theories

Martin Junge and Rainer Reisenzein

Institute of Psychology, University of Greifswald, Greifswald, Germany

Two studies investigated the utility of indirect scaling methods, based on graded pair comparisons, for the testing of quantitative emotion theories. In Study 1, we measured the intensity of relief and disappointment caused by lottery outcomes, and in Study 2, the intensity of disgust evoked by pictures, using both direct intensity ratings and graded pair comparisons. The stimuli were systematically constructed to reflect variables expected to influence the intensity of the emotions according to theoretical models of relief/disappointment and disgust, respectively. Two probabilistic scaling methods were used to estimate scale values from the pair comparison judgements: Additive functional measurement (AFM) and maximum likelihood difference scaling (MLDS). The emotion models were fitted to the direct and indirect intensity measurements using nonlinear regression (Study 1) and analysis of variance (Study 2). Both studies found substantially improved fits of the emotion models for the indirectly determined emotion intensities, with their advantage being evident particularly at the level of individual participants. The results suggest that indirect scaling methods yield more precise measurements of emotion intensity than rating scales and thereby provide stronger tests of emotion theories in general and quantitative emotion theories in particular.

Keywords: Emotion intensity; Emotion measurement; Direct scaling methods; Indirect scaling methods; Rating scales; Graded pair comparisons; Maximum likelihood difference scaling; Quantitative emotion theories.

Perhaps the most salient property of emotions, apart from their quality (happiness, anger, fear, etc.), is their intensity (e.g., Frijda, Ortony, Sonnemans, & Clore, 1992; Reisenzein, 1994): Each emotion quality can be exemplified in different grades or degrees, ranging from just noticeable to highly intense. However, whereas many emotion theories try to account for the qualitative differentiation of emotions (e.g., Lazarus, 1991; Ortony, Clore, & Collins, 1988; Scherer, 2001), there are only few explicit attempts to model the intensity of emotions (see, e.g., Gratch, Marsella, Wang, & Stankovic, 2009; Reisenzein, 2009a, for reviews; and the introduction to Studies 1 and 2 below).

One reason for why most emotion theories have remained on the qualitative level may be the problems associated with *testing* quantitative theories. To test a quantitative emotion theory, emotion intensity must be measured with sufficient precision to allow the verification or rejection of the

Correspondence should be addressed to: Martin Junge, Institute of Psychology, University of Greifswald, Franz-Mehring-Strasse 47, D-17487 Greifswald, Germany. E-mail: martin.junge@uni-greifswald.de

We thank Siegfried Macho for his helpful comments on a previous version of the manuscript.

quantitative laws proposed by the theory (see Study 1 for examples). Behavioural emotion indicators (e.g., physiological reactions or facial expressions) are of limited usefulness for this purpose because they are not differentiated (emotion-specific) enough, and because they are not strongly enough associated with emotional experience (see e.g., Mauss & Robinson, 2009; Reisenzein, Studtmann, & Horstmann, 2013).[1] Patterns of brain activity corresponding to specific emotions may ultimately provide objective and precise measures of both the quality and intensity of emotions; but such measures still need to be developed. This leaves self-reports of emotional experience as the most specific and sensitive measures of emotion currently available (see also Reisenzein & Junge, 2012). However, despite their virtues and widespread use in emotion research, self-reports of emotional experiences are often regarded as too imprecise to be suitable for testing quantitative emotion theories.

Criticisms of self-reports of emotions are usually targeted at their most frequently used incarnation: direct intensity ratings of emotion on quality-plus-intensity scales (e.g., "How happy are you right now?" from 0 = *Not at all* to 10 = *Extremely*). The basic criticism is this: Even granting that emotional experiences are continuous magnitudes with metric structure (Michell, 1990), it is unlikely that ratings preserve enough information about these magnitudes to be useable for testing quantitative theories. Suitable measurements must meet two basic requirements (e.g., O'Brien, 1985): (i) they must be on a metric (i.e., an interval or even a ratio) scale level; and (ii) they must be reasonably free of (other) measurement errors, both random and systematic (with the latter including errors due to categorisation, or a loss of resolution; see O'Brien, 1985). According to their critics, rating scales are unlikely to fulfil these requirements. With regard to scale level, it is widely held that rating scales are only ordinal, or at best somewhere in between ordinal and interval (e.g., Krantz, Luce, Suppes & Tversky 1971). There is also little doubt that ratings have only limited resolution and contain a considerable amount of random noise. Although random measurement error can be reduced by using repeated measurements or multiple indicators (e.g., Bagozzi, 1980; Bollen & Noble, 2011), in practice there are limits to what can be achieved this way. The measurement problems of rating scales become particularly salient if one wants to test emotion theories at the level of individual participants rather than at the group (mean) level only. However, given that emotion theories describe mental processes of individuals, this is what one should try to do (e.g., Reisenzein, 2000; for a more general discussion of this issue see, e.g., Cohen, Sanborn, & Shiffrin, 2008).

Because of the measurement problems associated with category ratings and other direct intensity judgements (e.g., magnitude scaling; Stevens, 1975), there have long been attempts to develop alternative scaling methods that yield more precise measurements of the intensity of subjective experiences. The basic idea behind these alternative scaling methods is to estimate subjective intensities from judgements that demand less of the participants than direct intensity ratings do, and that they are (therefore) able to make reliably. The most frequently proposed kind of simpler judgements are ordinal comparisons of intensities (e.g., in the case of emotion: "the intensity of relief elicited by event a is greater than that elicited by event b"). From these data, the underlying absolute intensities of the experiences caused by the stimuli are then estimated with the help of scaling models, which are really miniature models of the judgement processes thought to underlie the pair comparison responses (see below for more detail). Because the intensities of experiences (the scale values) are thus indirectly determined rather than directly reported by the subject, these scaling methods are often called *indirect* (e.g., Borg & Staufenbiel, 2007).

[1] We conceptualise emotions as mental states that are subjectively experienced as feelings and that manifest themselves in self-reports, expressive behaviours, and actions. Readers who prefer a multi-component view of emotion (according to which expressive behaviours and actions are components rather than indicators of emotions) should read our term "emotion" as referring to the feeling component of the multi-component state.

Although indirect scaling methods have much to recommend them (see, e.g., Böckenholt, 2004; Maydeu-Olivares & Böckenholt, 2008), they have been rarely used in emotion research. One reason for this may be emotion researchers' lack of familiarity with indirect scaling methods, particularly their recent developments. Another reason may be the belief that these methods, although useful for the measurement of sensations, cannot be used to measure emotional experiences. Perhaps the most important reason, however, is that emotion researchers have not been convinced that indirect scaling methods provide an increase in precision that justifies the greater effort of data collection and analysis required by these methods. Indeed, there seems to be no study in which direct and indirect scalings of emotion intensity have been compared in terms of their usefulness for testing substantive hypotheses. The aim of the present article is to fill this research lacuna. In two studies, we investigated to which degree indirect scaling methods can improve the testing of emotion theories compared to direct ratings of emotion intensity, as commonly used in emotion research. In Study 1, we compared the usefulness of direct ratings and indirect scalings of relief and disappointment about lottery outcomes for the testing of quantitative belief–desire models of these emotions. In Study 2, we compared the utility of direct ratings and indirect scalings of the intensity of disgust induced by pictures to test a semi-quantitative model of disgust intensity.

Scaling methods

We compared direct emotion intensity ratings obtained using numerically labelled rating scales with scale values obtained by means of two indirect scaling methods, both of which are (in our application) based on graded pair comparisons. Graded pair comparisons (Bechtel, 1967) or difference ratings (Boschman, 2001) are a variant of the well-known pair comparison method (e.g., Borg & Staufenbiel, 2007; Torgerson, 1958). They differ from the standard pair comparison task in that the participants judge not only which of the two stimuli in a pair is greater than the other on a specified judgement dimension, but also how much greater it is. In other words, participants judge the degree of the difference between the two compared stimuli on the judgement dimension. To illustrate, the participants in Study 1 were presented with pairs of disappointing lottery outcomes and were asked to indicate, for each pair, which of the two outcomes was more disappointing, and how much more disappointing it was (from "*Just barely*" to "*Extremely*").

Although comparatively rarely used (see De Beuckelaer, Kampen, & Van Trijp, 2013; Oishi, Schimmack, Diener, & Suh, 1998, for examples)—and never before, to our knowledge, to measure the intensity of specific emotions—graded pair comparisons have much to recommend them. Different from direct ratings, graded pair comparisons do not presuppose that people can accurately judge absolute intensities of experiences; only that they can judge intensity differences. Since the beginnings of psychophysical measurement, it has been argued that even though people may be unable to accurately report the absolute intensities of experiences, they are able to judge intensity differences (e.g., Titchener, 1905; see Michell, 2006). Furthermore, even the difference judgements are assumed to be only on an ordinal scale level in one of the two scaling models we used; hence, ultimately only ordinal comparison is required, as with binary pair comparison judgements (Thurstone, 1927). However, different from binary pair comparisons (see Thurstone, 1927; Torgerson, 1958), graded pair comparisons can also be used for the scaling of clear suprathreshold intensity differences.

To derive the emotion intensities from the graded pair comparison judgements, we fitted two different probabilistic scaling models to the data: an additive functional measurement model (AFM) and the maximum likelihood difference scaling model (MLDS). Both models can be regarded as descendants of the well-known Thurstonian scaling model (Thurstone, 1927; see, e.g., Böckenholt, 2003; Borg & Staufenbiel, 2007). In particular, in agreement with Thurstone (1927), both models assume that the comparative judgements provided by the participants are based on differences in latent scale values that are perturbed by random error. The most important difference between the two models is this: AFM presupposes that the graded

difference judgements are on a metric scale, whereas MLDS requires them to be ordinal only.

Additive functional measurement (AFM). The statistical model underlying AFM scaling can be regarded as a special version of the additive functional measurement model proposed by Anderson (1970, 1982) because it can be obtained from applying this model to graded pair comparisons (Boschman, 2001; see also, Bechtel, 1967; Bechtel & O'Connor, 1979). Alternatively, AFM can be regarded as a unidimensional version of probabilistic metric multidimensional scaling (e.g., MacKay & Zinnes, 1986). The AFM model assumes that the response function that maps the internal states (in our case, the perceived differences between the intensities of compared emotion stimuli or objects) into overt judgements, is linear and more precisely proportional. This implies that the difference judgements, like the latent scale values, are interpreted as numerical responses on a ratio scale. Accordingly, the graded response R_{ab} to a pair of objects (a,b); (e.g., in Study 1, the difference in the intensity of disappointment evoked by two lottery outcomes a and b, coded on a numerical response scale from -6 to $+6$) is proportional to the latent decision variable Δ_{ab}, which is the perceived difference in emotion intensities; i.e., $R_{ab} = \alpha\Delta_{ab}$. The decision variable Δ_{ab}, in turn, is assumed to be internally computed as the difference between the scale values ψ_a, ψ_b of the compared stimuli plus a random error term ε that summarises diverse kinds of randomly fluctuating components of the judgement process: The stimuli do not always evoke the same intensity of emotion when presented repeatedly; the computation of the difference may be imprecise, and so on. Following Thurstone's (1927) classical case V model, the errors are assumed to be independent and normally distributed with mean 0 and constant variance δ^2. The AFM model can be succinctly summarised by two equations, the first of which describes how the decision variable is computed, and the second, how it is mapped into an overt graded response:

$$\Delta_{ab} = \psi_b - \psi_a + \varepsilon \text{ with } \varepsilon \sim N(0, \delta^2) \quad (1)$$

$$R_{ab} = \alpha\Delta_{ab}. \quad (2)$$

The scale values ψ_a, $\psi_b, \ldots,$ ψ_n and the error variance δ^2 are unknown parameters of the model that must be estimated from the data (the graded pair comparisons). Hence, as with all probabilistic scaling methods, AFM scaling (the process of determining the scale values of the stimuli) involves fitting a probabilistic judgement model to the data.

Maximum likelihood difference scaling (MLDS). As mentioned, the AFM model assumes that the graded difference ratings are on a metric scale level. Although this assumption (that judgements of *intensity differences* are on a metric scale) is weaker and intuitively more plausible (Titchener, 1905) than the assumption that absolute intensity ratings are metric, it is far from uncontroversial. To address this concern, we included a second recently developed scaling method that requires only ordinal difference judgements: MLDS (Knoblauch & Maloney, 2008; Maloney & Yang, 2003). Different from the AFM model, the response variable in MLDS is dichotomous, and the response function that maps the decision variable into the overt response is a threshold function. MLDS can be regarded as a probabilistic unidimensional version of nonmetric multidimensional scaling (Maloney & Yang, 2003, p. 582).[2]

[2] Another probabilistic scaling model suitable for graded pair comparisons is the *cumulative probit model* (e.g., Boschman, 2001; Greene & Hensher, 2010). The cumulative probit model (also called *ordinal regression*) is similar to the AFM model in that it uses the original graded comparison judgements as input, and similar to the MLDS model in that it requires only ordinal data. Its main disadvantage is that it has many more parameters than MLDS because, in addition to the m scale values and the error variance, $m-1$ threshold parameters that mark the category boundaries need to be estimated. This can lead to estimation problems, particularly at the individual level. In addition, we prefer MLDS because it is based on a representational measurement model (Krantz et al., 1971; see also the general discussion). However, it should be noted that Boschman (2001) found that the scale values obtained with the cumulative probit model were highly similar to those obtained with the AFM model.

The input data for MLDS are 0/1 dominance judgements as in classical Thurstonian pair comparison scaling; however, comparisons are made between pairs of objects (a,b) and (c,d) rather than between single objects, and the participant's task is to judge which difference is larger: the difference between a and b or the difference between c and d. In our studies, the MLDS input data represent, for example, whether the difference in the intensity of disappointment caused by two lottery outcomes a and b is greater or smaller than the difference in disappointment caused by outcomes c and d. In previous applications of MLDS, the input data were obtained by actually presenting participants with (a subset of) the possible quadruples (ab,cd) and asking them to judge whether the difference between a and b is greater or less than that between c and d (e.g., Maloney & Yang, 2003). However, it is also possible to use the data from a graded pair comparison task as input to MLDS.[3] To do so, the graded pair comparisons are transformed into dominance judgements for quadruples as follows (see e.g., Roberts, 1979, p. 135): For each quadruple (each pair of stimulus pairs; ab,cd), ab > cd if the graded comparison (judged difference) of a and b has a higher rank than the difference of c and d. For example, assume that the disappointment about lottery outcome b is judged as "a little more intense" (difference score 2) than the disappointment about outcome a, whereas the disappointment about d is judged as "much more intense" (difference score 4) than the disappointment about c. Because the judged difference between c and d is greater than that between a and b, the quadruple (ab,cd) is assigned the dominance score 1. By contrast, if the judged difference between a and b is smaller than that between c and d (ab,cd), is assigned the dominance score 0. Because only the ranks of the graded pair comparison judgements are used to decide which difference is larger, MLDS requires that these judgements be on only an ordinal scale.

Analogous to the AFM model, the MLDS judgement model can be summarised by two equations, the first of which describes how the decision variable is computed from the scale values of the stimuli, and the second how it is mapped into an overt dominance judgement.

$$\Delta_{ab,cd} = |\psi_d - \psi_c| - |\psi_b - \psi_a| + \varepsilon \text{ with } \varepsilon \sim N(0, \delta^2) \quad (3)$$

$$R_{ab,cd} = 1 \text{ if } \Delta_{ab,cd} > 0; \text{ else } R_{ab,cd} = 0. \quad (4)$$

According to the psychological law expressed by Equation 3, the participant in a quadruple judgement task implicitly computes the (absolute) difference between the two members of each stimulus pair (e.g., how much more or less disappointing lottery outcome a is compared to b, and c compared to d), and then again computes the difference between these intervals, $|\psi_d - \psi_c| - |\psi_b - \psi_a|$, to determine which of them is larger. Furthermore, as in the AFM model, the internal judgement process is assumed to be contaminated by independent random error stemming from a normal distribution with variance δ^2. It should be noted, however, that according to simulation studies by Maloney and Yang (2003), MLDS is remarkably robust against misspecifications of the error distribution as well as violations of the assumption of independent errors. Analogous to the AFM model, the error component ε can be thought of as summarising all sources of random error in the judgement process: The stimuli may not always elicit the same intensity of emotion, there may be random errors when computing the intervals, and so on. Note that because in our application of the method, the 0/1 responses to the stimulus pairs were deduced from the graded pair comparison judgements, random errors associated with the computation of the differences between the two intervals *ab* and *cd* were excluded; hence one possible source of error in direct quadruple judgements was eliminated. Equation 4 implies

[3] In a recent study (Junge & Reisenzein, 2013a), we compared the MLDS solutions obtained from quadruple comparisons of emotional stimuli with those derived from graded pair comparisons of the same stimuli. For most participants, high agreements of the solutions were found.

that, if the error were zero, the difference between c and d would be judged as greater than that between a and b (i.e., $R_{ab,cd} = 1$) whenever $= |\psi_d - \psi_c| > |\psi_b - \psi_a|$; however, due to the presence of error, the other response ($R_{ab,cd} = 0$) will occasionally be given, particularly if the two intervals are of similar size.

STUDY 1: INTENSITY OF RELIEF AND DISAPPOINTMENT EXPERIENCES

For a small set of emotions, quantitative theories—theories that explicitly seek to explain not only the quality but also the intensity of emotions—have been proposed by philosophers (e.g., Davis, 1981), economists (e.g., Bell, 1985; Loomes & Sudgen, 1986), and psychologists (e.g., Anderson, 1989; Mellers, Schwartz, Ho, & Ritov, 1997). Although they use different terminology (e.g., strength of desire is referred to as utility or value, and strength of belief as subjective probability or expectancy), most of these theories can be regarded as variants of a common underlying emotion theory, the belief–desire theory of emotion (e.g., Davis, 1981; Green, 1992; Reisenzein, 2009a). Integrating this and other previous research, Reisenzein (2009a) proposed a simple quantitative model of belief–desire theory that includes intensity laws for happiness and unhappiness, relief and disappointment, hope and fear, and surprise. As formulated in Reisenzein (2009a), the model is restricted to situations in which a single desired or undesired event can occur (hope, fear) and then either occurs or does not occur (relief, disappointment, surprise). In Study 1, we tested the proposed intensity laws for relief and disappointment, using two-outcome lotteries where either a gain (vs. no gain) or a loss (vs. no loss) could occur.

The disappointment model assumes: (i) disappointment occurs if the person believes that a desired outcome p (in Study 1, this was a monetary gain) might occur, but then comes to believe not-p (i.e., that p did not or will not occur); and (ii) the intensity of disappointment about not-p is a strictly increasing function, f, of the product of the strength of the person's prior belief that p would occur, b(p), and the strength of her/his (positive) desire for p, d(p).[4] We will assume here that f is the identity function. Then, the proposed intensity model for disappointment simplifies to:

$$\text{disappointment(not-p)} = b(p) \times d(p) \text{ if } d(p) > 0; \text{ else } 0. \quad (5)$$

Analogously, the relief model assumes: (i) relief is experienced if the person believes that an outcome p may happen to which s/he is averse or that s/he wishes would not occur (in Study 1, a monetary loss), but then comes to believe not-p; and (ii) the intensity of relief about not-p is determined by the product of belief strength and the degree of undesiredness of, or aversion toward p:

$$\text{relief(not-p)} = |b(p) \times d(p)| \text{ if } d(p) < 0; \text{ else } 0. \quad (6)$$

The absolute value of $b(p) \times d(p)$ is used in Equation 6 only to make the intensity of relief a positive number. Note that the emotion models described by Equations 5 and 6 are in agreement with Mellers et al.'s (1997) theory of the determinants of *positive affect* (relief) and *negative affect* (disappointment) when this theory is restricted to two-outcome lotteries with a zero outcome.

Both the emotions and their proximate mental causes (beliefs and desires) are not intersubjectively observable but must be inferred. In our

[4] In keeping with the common representation of subjective probability and utility, we assume that b(p), the belief strength, is represented by real numbers from [0,1], with 1 denoting certainty that p, 0.5 maximal uncertainty, and 0 certainty that not-p; and that d(p) represents the direction and strength of the desire for p, with values >0 denoting positive desire, 0 indifference, and values <0 denoting negative desire, or an aversion to p.

study, the emotion intensities were measured using direct ratings and indirect scaling methods, whereas the belief and desire strengths were not measured in that way, but were experimentally manipulated by presenting lotteries with different outcome probabilities and payoffs. The values of b(p) and d(p) corresponding to these probabilities and payoffs were (implicitly) estimated during the process of fitting the emotion models, simultaneously with the parameters of these models (Equations 1 and 2), by specifying plausible "psychophysical" functions[5] for the mapping of the objective outcome probabilities into degrees of belief and the monetary payoffs into degrees of desire (see Tversky & Kahneman, 1992, for an analogous method in the domain of decision making; and Mellers et al., 1997, for a related approach). Based on prior research, we assume that both input functions are power functions. More precisely, using m_p for the size of the monetary gain or loss associated with outcome p, and $prob_p$ for the objective probability of p, we assume: $d(p) = \lambda m_p^{\alpha}$ for gains and $\lambda'|m_p|^{\beta}$ for losses; and $b(p) = \delta\, prob_p^{\gamma}$; where α, β and γ reflect the curvature of the power functions and λ, λ', and δ their steepness. For d(p), these assumptions amount to adopting the familiar power function for utility typically assumed and obtained in decision-making studies (e.g., Tversky & Kahneman, 1992; see Fox & Poldrack, 2009; Stott, 2006). For monetary gains and losses, utilities are typically negatively accelerated power functions (i.e., the exponents α and β are between 0 and 1; Fox & Poldrack, 2009; Galanter, 1990). For the mapping of objective probabilities into subjective quantities, decision theorists typically prefer more complicated, inverted S-shaped functions (e.g., Gonzalez & Wu, 1999; Prelec, 1998; Stott, 2006). However, it is at present not clear that these probability weighting functions are also descriptive of the process of emotion generation. Previous studies on relief, disappointment and other emotions by the authors (Reisenzein & Junge, 2006; Reisenzein & Junge, 2013a) found that power functions provided a good fit to the data, as did a study by Gratch et al. (2009). At minimum, then, a power function seems to be a good approximation to the true "emotion weighting" function (see Footnote 5).

The actual emotion models that we estimated are shown in Equations 7 and 8. These equations were obtained by plugging the power-function predictors for d(p) and b(p) into Equations 1 and 2 and combining the two steepness parameters λ, λ' and δ into single parameters $\lambda\delta$, $\lambda'\delta$. For convenience, these combined parameters are here also labelled λ, λ':

$$\text{disappointment(not-p)} = \lambda m_p^{\alpha} \times prob_p^{\gamma} \text{ if } m_p > 0;\ \text{else } 0. \quad (7)$$

$$\text{relief(not-p)} = |\lambda'|m_p|^{\beta} \times prob_p^{\gamma}| \text{if } m_p < 0;\ \text{else } 0. \quad (8)$$

The parameters of the emotion models were estimated using nonlinear regression (e.g., Bates & Watts, 1988; Ritz & Streibig, 2008). Figure 1 shows the disappointment and relief functions graphically for the range of outcomes and the probabilities used in Study 1, using parameters close to the median of the individual parameter values obtained in the study.

Method

Participants

Participants were 39 students (six males and 33 females) from different majors at the University of Greifswald, with a mean age of 22.3 years ($SD = 4.8$) who responded to a posting on the

[5] In the case of belief, it is actually not quite correct to speak of a "psychophysical" function because b(p) also includes (and in our view, even mainly reflects) the weight attached to perceived probabilities in the emotion-generating process. The case here is parallel to that of the decision-weighting function proposed in prospect theory (Fox & Poldrack, 2009; Kahneman & Tversky, 1979). Analogous to the decision weights of prospect theory, we propose speaking of "emotion weights" (Junge & Reisenzein, 2010). However, important as this clarification (or amendment) of the belief–desire theory of emotion is on a theoretical level, it makes no difference to the empirical tests of the emotion models reported here.

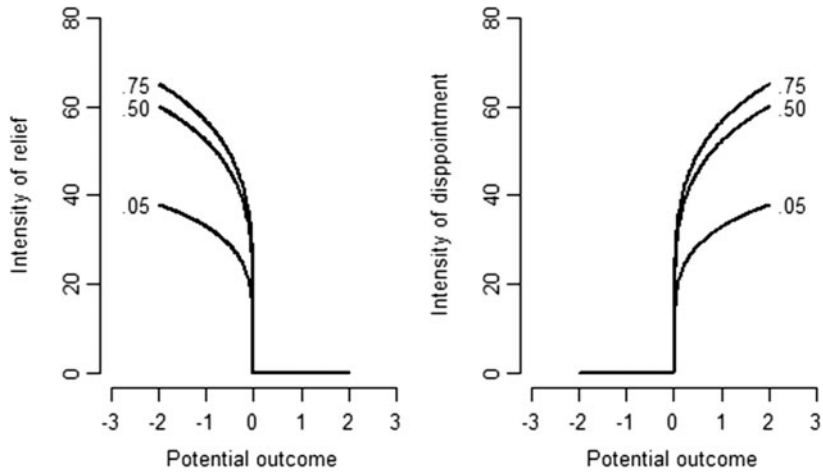

Figure 1. *Graphs of the relief and disappointment models for the outcome range and the probabilities used in Study 1. The parameters used in these examples are $\lambda = 60$, $\alpha = .20$, and $\gamma = .20$ for the disappointment model and $\lambda' = 60$, $\beta = 0.20$, and $\gamma = .20$ for the relief model.*

internet student message board. The study was described as dealing with the subjective experience of gambling. Potential participants were also informed that they could win a few Euros in a laboratory lottery game.

Materials

To induce disappointment and relief, we used a lottery paradigm similar to that used by Mellers et al. (1997). The experiment was created using WEXTOR, a free online web experiment generator (Reips & Neuhaus, 2002) and consisted of a set of HTML pages into which FLASH-based "money wheels" (wheels of fortune) were embedded. Fifty-one different money wheels were used. For each trial, the participant could either win or lose money, or else neither win nor lose. Outcome probabilities were manipulated by varying the size of the gain or loss sector of the wheel. The wheel was divided into 20 equal sectors, each of which therefore represented a probability of .05. Gain sectors were coloured in green, loss sectors in red, and the zero outcome sector in grey. The amount at stake in a trial was indicated by placing pictures of the corresponding coins at the centre of the wheel.

Design

The lotteries were constructed according to a two-factor design with factors Potential gain or loss ($-2, -0.50, -0.10, 0.10, 0.50$, and 2 Euros) and Outcome probability (.05, .50, .75). We focused on the 18 lotteries with zero outcomes because these were the occasions where relief and disappointment were primarily expected to occur.[6] However, to keep up the appearance of a real lottery, we also included 15 trials with nonzero outcomes. Furthermore, to estimate the reliability of the direct emotion rating and to increase its reliability by averaging, all zero-outcome lotteries were presented twice.

Procedure

The experiment was conducted in a single session that took about 60 minutes. Participants were tested individually.

[6] Losses can also cause disappointment, and gains can cause relief, under certain circumstances: namely, when participants focus on avoiding the possible loss or on missing the possible gain. However, we have found that without prompting (e.g., by explicitly asking participants how disappointed they feel about failing to avoid a possible loss), only a subset of the participants spontaneously construe the lottery outcomes in this way.

Direct scaling task (ratings). Participants were told that, to create a realistic gaming experience, they would participate in a series of lotteries in which they could win or lose money. In each trial, they could win or lose up to 2 Euros, but their net outcome would at worst be set to 0 Euros at the end of the experiment. In each trial, the participants were first presented with a money wheel and were asked to consider it carefully to gain a clear understanding about the chances for winning or losing that it offered. They then set the wheel into motion by clicking the "start" button. To support the impression that the outcomes were random, the time until the wheel came to a standstill varied randomly between 3.2 and 4 s. The outcome of the lottery was additionally communicated to the participants by a sentence that appeared below the wheel (e.g., "You won 2 Euros" vs. "You won nothing" in the case of the gain lotteries and "You lost 2 Euros" vs. "You lost nothing" in the case of the loss lotteries). Finally, the participants indicated how disappointed and relieved they felt about the outcome by moving a slider along rating scales ranging from 0 (*Not at all* [*disappointed/relieved*]) to 100 (*Extremely* [*disappointed/relieved*]). To allow for finely graded ratings, the currently selected scale value was continuously displayed in numerical format (0–100) immediately above the midpoint of the scale.

Indirect scaling task (graded pair comparisons). The second part of the experiment, also programmed in WEXTOR, was a graded pair comparison task (e.g., Bechtel & O'Connor, 1979; Boschman, 2001). The participants were informed that they would now be presented with all possible pairings of missed gains (disappointment) and avoided losses (relief) from the first part of the experiment. Because there were nine lotteries of each type, the participants judged $(9 \times 8)/2 = 36$ pairs of lotteries of each type. Half of the participants worked on the relief lotteries first and the other half on the disappointment lotteries. Within each block, the comparisons were presented in a random order. In each trial, the two money wheels that had to be compared were shown side by side on the screen. In four of the eight comparisons involving a given lottery, it was presented on the left side of the screen and in the other four, on the right side. Participants were asked again to imagine that they were participating in a real lottery even though no actual money was at stake. To aid their imagination, they were asked to spin the left money wheel, wait until it stopped, and then do the same for the right wheel. To save time, the wheels revolved more quickly than in the first part of the experiment and stopped after about 2 s. Subsequently, the participants indicated which of the two disappointing [relief-inducing] outcomes would have caused stronger disappointment [relief] if they had played for real money, and how much more disappointment [relief] it would have elicited. Answers were given on a bipolar rating scale (without numerical labels) ranging from *The left outcome is extremely much more disappointing* [*relieving*] to *The right outcome is extremely much more disappointing* [*relieving*]. Intermediate scale points on both sides were labelled *very much more*, *much more*, *more*, *a little more*, and *just barely more*. An answer of *equally intense* was not allowed to encourage participants to discriminate even small intensity differences. There is some evidence (e.g., Gridgeman, 1959) that discrimination may deteriorate when a tie category is included, possibly because its presence suggests that indeed not all of the items are distinguishable (Böckenholt, 2001). We assume that if participants cannot detect a difference, their responses are determined by guessing. The response scale was placed below the lottery wheels such that its right half extended below the right wheel and its left half below the left wheel.

Estimation of scale values

As detailed by Critchlow and Fligner (1991), many scaling models for pair comparisons can be conceptualised as generalised linear models (GLIMs; see, e.g., Hardin & Hilbe, 2007), as a consequence of which standard GLIM software can be used to estimate the model parameters (the scale values and the error variance), using maximum likelihood. This is also true for the AFM model, although for complete pairwise comparison

data, estimation is also possible using ordinary least squares (in this case, the maximum likelihood solution corresponds to the OLS solution; see Hardin & Hilbe, 2007). We preferred maximum likelihood estimation because we wanted to use the same estimation method as in MLDS, and because it allows the estimation of model parameters from subsets of the data. This feature was exploited to compute the reliability of the scale values by separately estimating the scale values from the two halves of a random split of the pair comparison judgements.

The AFM and MLDS scaling models were fitted separately to the data of each participant using the glm function of R (R Development Core Team, 2011). For the AFM model, the details were as follows: The nine stimuli (lotteries) were coded as dummy variables with values 1 and −1 whenever the stimuli were included in a comparison, and as 0 otherwise (Critchlow & Fligner, 1991). These dummy variables were used to predict the graded pair comparison judgements, coded as −6 to +6. In agreement with Equations 1 and 2, the Gaussian distribution family with an identity link was specified. One dummy variable was excluded to allow identification of the model (Critchlow & Fligner, 1991); the scale value for this stimulus was set to 0.

For the MLDS model, the graded pair comparisons were first transformed into quadruples (ab,cd). For n stimuli (objects), there are $m = n(n-1)/2$ pairs of objects, which in turn can be combined to $m(m-1)/2$ object pairs. In Study 1, there were nine stimuli (lotteries) for each of the two emotions (disappointment and relief) and hence 36 stimulus pairs ab (those used in the graded pair comparison task) and 630 quadruples (ab,cd) that could be used as input to MLDS. However, because MLDS has been found to perform well even if only a fraction of the possible quadruples are used (Maloney & Yang, 2003), we decided to include only the 378 quadruples that did not contain the same stimulus twice, such as (ab,ac). Of these, we had to eliminate quadruples for which the judged differences ab and cd were equal in size (23.4%) because MLDS does not allow for "equal" responses. As in the case of the AFM model, the parameters of the MLDS model were estimated with the glm function of R, but this time the binomial distribution family with a probit link was specified (Knoblauch & Maloney, 2008). This specification implies a normal error distribution for the latent decision variable, as assumed in the MLDS model (Equation 3). To fit the MLDS model using standard GLIM software such as glm, it is in addition necessary to order the predictor variables (stimuli) in the model matrix according to their scale values (Knoblauch & Maloney, 2008). Hence one needs to know the rank order of the stimulus values. When simple sensory stimuli are scaled, this rank order is often evident and can therefore be specified by the experimenter (Knoblauch & Maloney, 2008). By contrast, in the case of the emotion stimuli used in our studies, the rank order of intensities is not so clearly evident. To circumvent this problem, we used the participants' AFM scale values to specify the rank order of the stimuli.[7]

Results

Reliabilities

The reliabilities and intercorrelations of the direct and indirect intensity measurements are shown in Table 1 (note that these are individual-level statistics). For the ratings, the index of reliability is the retest correlation computed from the two (first and second) presentations of the zero-outcome lotteries; for the MLDS and AFM scale values, it is the split-half reliability, computed as the correlation between the scale values estimated from two randomly determined halves of the graded pair comparisons (AFM) or quadruples (MLDS). The Spearman–Brown corrected reliabilities are shown in parentheses. The corrected values best reflect the reliability of the intensity measurements actually used in the subsequent analyses (the scale values derived from the complete set of pair comparisons, and the means of the

[7] We also tried nonmetric multidimensional scaling (restricted to one dimension) for this purpose, but obtained inferior results.

Table 1. *Reliabilities and intercorrelations of the ratings and indirect scalings of emotion intensity, Study 1*

		M	*SD*	*Min*	*Max*
Reliabilities[a]					
	Relief				
	Ratings	.75 (.82)	.28 (.27)	.00 (.00)[b]	.99 (.99)
	AFM	.84 (.90)	.17 (.14)	.12 (.22)	.98 (.99)
	MLDS	.96 (.98)	.06 (.03)	.70 (.83)	.99 (.99)
	Disappointment				
	Ratings	.75 (.83)	.25 (.23)	.00 (.00)[b]	.98 (.99)
	AFM	.85 (.91)	.12 (.09)	.32 (.48)	.98 (.99)
	MLDS	.98 (.99)	.04 (.02)	.76 (.86)	.99 (.99)
Correlations					
	Relief				
	Rating–AFM	.69	.33	−.67	.97
	Rating–MLDS	.64	.33	−.36	.93
	AFM–MLDS	.94	.16	.01	.99
	Disappointment				
	Rating–AFM	.71	.36	−.48	.96
	Rating–MLDS	.67	.40	−.41	.97
	AFM–MLDS	.95	.10	.43	.99

Notes: [a] The computation of the reliabilities is explained in the text. Numbers in parentheses are Spearman–Brown corrected reliabilities.
[b] For one participant, the correlations between the two relief ratings as well as the two disappointment ratings were negative; for another participant, the correlation between the relief ratings was negative. For these participants, the reliability of the respective rating was defined as zero.

two ratings, respectively). As can be seen from Table 1, for both relief and disappointment, the average reliabilities of the scale values obtained via AFM scaling and MLDS were consistently higher (Spearman–Brown corrected .90/.91 for AFM and .98/.99 for MLDS) than those of the ratings (.82/.83).[8] Furthermore, the AFM and MLDS scale values were highly correlated for both relief (average $r = .94$) and disappointment (.95), whereas their correlations with the direct ratings were much lower (AFM: .69 and .71; MLDS: .64 and .67).

Quantitative model fits

The quantitative models of relief and disappointment (Equations 7 and 8) were estimated separately for each participant and type of intensity measurement (rating, AFM, MLDS) using the nonlinear regression function nls of R (R Development Core Team, 2011). Only the measurements for negative unobtained outcomes were used to fit the relief model, and only the measurements for positive unobtained outcomes were used to fit the disappointment model, because only these measurements were available from the indirect scaling methods. However, an examination of the *ratings* of relief and disappointment revealed that, as predicted by the emotion models, unobtained positive outcomes elicited virtually no relief and unobtained negative outcomes elicited virtually no disappointment (94% zero judgements). Hence, one may assume that the "else 0" parts of the emotion models (Equations 7 and 8) are correct.

The indirect scale values were first linearly transformed, separately for each participant, in such a way that their minimum and maximum corresponded to the minimum and maximum of

[8] In addition, the reliabilities of the MLDS scalings were higher (.98/.99) than those of the AFM scalings (.90/.91). We attribute this to the fact that the split-half scale values whose correlation was used as the index of reliability were estimated from many more, as well as more diverse data points, in the case of MLDS than in the case of AFM. As a result, the MLDS estimates were more stable.

the participant's ratings. This was done to obtain an estimate of the zero point (scale origin) of the indirect scales, as pair comparison data per se provide no information about the zero point (see Böckenholt, 2004; Guilford, 1954), but this information is needed to obtain correct estimates of the parameters of the emotion models. To ensure that the estimated power functions were monotonically increasing (or at least non-decreasing), their exponents were constrained to be ≥ 0. R^2, the squared correlation between actual and predicted values, was used as the index of model fit (see Zheng & Agresti, 2000, for the advantages of R^2). For one participant, R^2 could not be computed for the ratings and the AFM scale values because the estimated exponents of the power functions were zero; for another participant, the same problem occurred for the AFM and MLDS scale values. The data of these participants were excluded from the statistical analyses involving model fit.

Relief. The distributions of the individual fit values obtained for the relief and disappointment models are shown in Figure 2. For relief, the median fit value (the median was used because of the skewed distributions) was $\text{Mdn}(R^2) = .76$ (MAD, the median of absolute deviations from the median = .19) for the direct ratings, .97 (MAD = .02) for the AFM scale values, and .95 (MAD = .05) for the MLDS scale values. Wilcoxon signed-rank tests revealed that the latter two model fits were significantly higher than those obtained with the direct ratings, $W = 693$, $p < .001$ for the AFM, and $W = 660$, $p < .001$ for MLDS. In addition, the fit obtained with the AFM scale values was significantly higher than that obtained with the MLDS scale values, $W = 636$, $p < .001$. Compared to the ratings, the variance explained by the relief model was, on average, 27% and 24% higher when using the AFM or MLDS scale values, respectively. Furthermore, in contrast to the ratings, a close fit was obtained for most participants, leaving on average but 6% and 9% of the variance unexplained. Inspection of the distribution of the fit values (Figure 2) revealed that with the direct ratings, 13% of the participants attained R^2 values > .90. By contrast, when the AFM scale values were used to fit the relief model, 92% of the participants had $R^2 > .90$, and with the MLDS scale values, 68% had $R^2 > .90$.

Disappointment. Highly similar results were obtained for disappointment. The median model fit was $\text{Mdn}(R^2) = .86$ (MAD = .11) for the direct

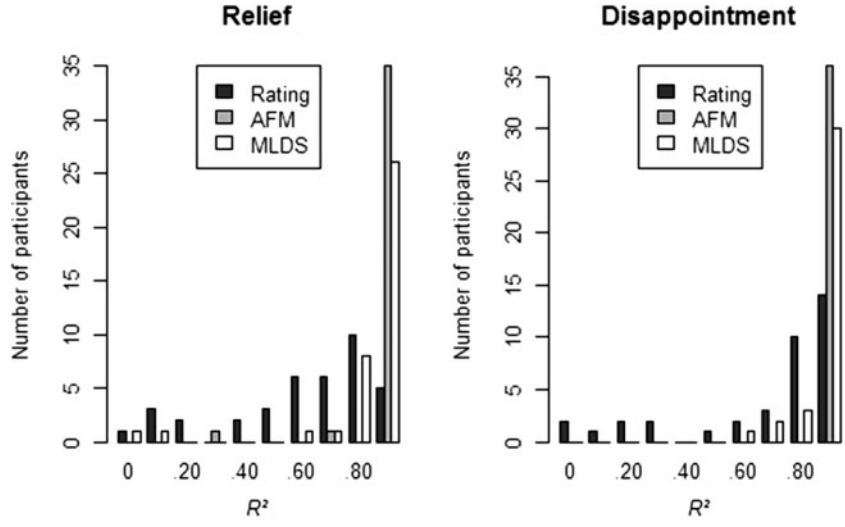

Figure 2. *Histograms of the individual model fits (R^2) for relief and disappointment.*

ratings, .96 (MAD = .02) for the AFM scale values, and .94 (MAD = .03) for the MLDS scale values. Again the differences between the direct and indirect scaling methods, and between AFM scaling and MLDS, were highly significant ($ps <$ .001). For the direct ratings, 38% of the participants had fit values > .90. By contrast, 95% had $R^2 > .90$ if the AFM scale values were used and 77% if the MLDS scale values were used.

Figure 3 shows the data and fitted values for relief of two participants representative of two main subgroups found in the data. For the first group, both the ratings and the AFM and MLDS scale values showed good fit to the relief model (Figure 3, top row). The second group comprised participants whose ratings did not fit the relief model but whose indirect scale values fit the model well (Figure 3, bottom row). A third group, comprising only two participants, showed poor model fit for both the ratings and the indirect scale values. In no case did the ratings but not the indirect scale values fit the model. Hence, as judged by model fit, the indirect scaling methods improved the quality of emotion measurement for many participants but never resulted in a deterioration. Parallel findings were obtained for disappointment.

Effects of aggregation

Although the indirect scaling methods were superior to the direct ratings in terms of model fit on the individual level, ratings might achieve

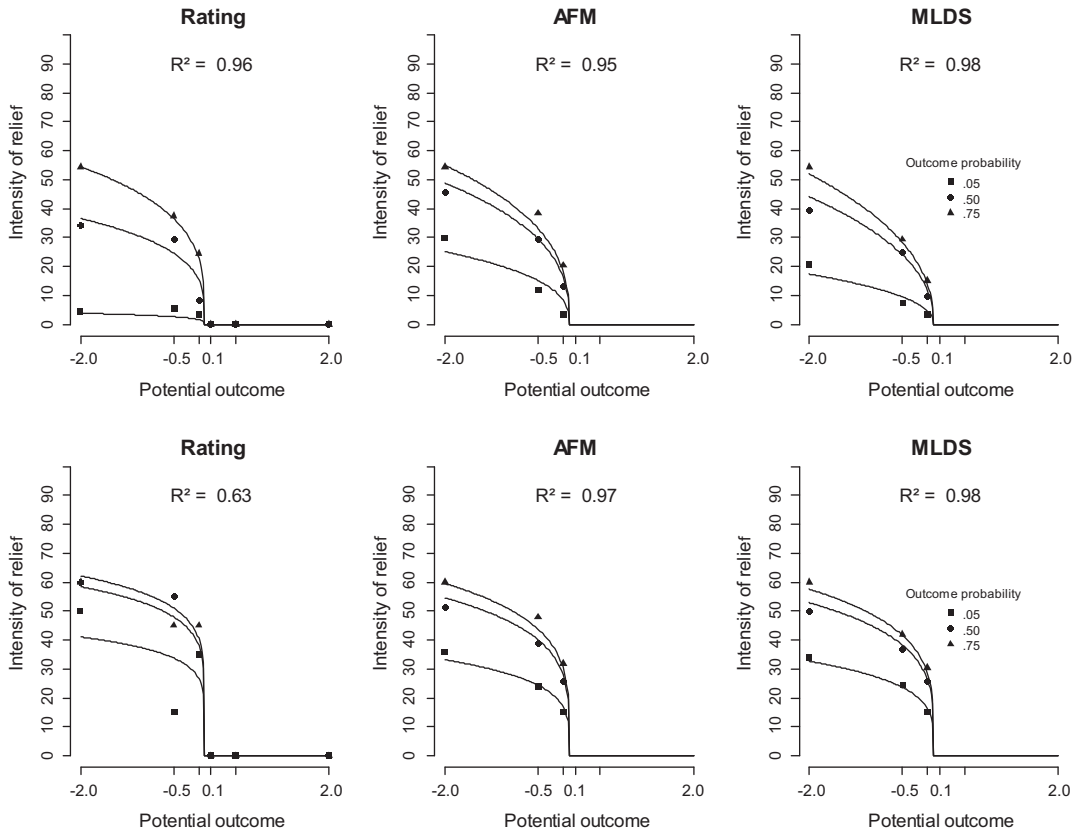

Figure 3. *Data and model fits for two selected participants, Study 1. From left to right: Direct ratings of relief, AFM scalings and MLDS scalings. For each measurement, the scale values are plotted against the potential outcome, with different symbols for each level of outcome probability. The lines show the predictions of the fitted relief models.*

comparable fit if the data are aggregated across participants to reduce measurement error. This data analytic strategy was previously used by Mellers et al. (1997). To explore this possibility, the relief and disappointment models were fitted to the means of the emotion intensity measurements in the different Outcome × Probability conditions. Indeed, we found that the fit values obtained for the mean ratings ($R^2 = .96$ for relief and $R^2 = .99$ for disappointment) were comparable to those obtained for the mean AFM scale values ($R^2 = .98$ for relief and $R^2 = .98$ for disappointment) and better than those obtained for the mean MLDS scale values (.95 for relief and .95 for disappointment). This finding may at first sight appear to be reassuring, because it seems to suggest that direct ratings of emotion intensity, although much noisier than indirectly determined scale values, do not lead to systematic distortions or information loss and can therefore be used for testing quantitative theories of emotion on the mean level. However, even a very high R^2 value can hide theoretically important local deviations from a theoretical model (e.g., Anderson, 1982; Birnbaum, 2011). Indeed, as reported below, information about one model component—the bilinear interaction between belief and desire for relief—was lost in the ratings (but not the indirect scalings) during aggregation. On a more general note, aggregation presupposes homogeneity of individuals in terms of their conformity to the theoretical model, which should not simply be presupposed, but should be checked by fitting the model to individual data. Nevertheless, our data suggest that aggregation biases may be less severe in the case of subjective ratings of emotion intensity than in some other areas of research (see Cohen et al., 2008).

Test of the Belief × Desire interaction

A central assumption of the belief–desire models of relief and disappointment is that belief and desire combine multiplicatively to determine emotion intensity. This assumption can be tested separately (i.e., without making assumptions about the form of the input functions that map the monetary outcome and its probability into desire and belief strengths, respectively) by using analysis of variance (ANOVA) techniques (Anderson, 1982). In addition, it is advisable to test this assumption separately because, as mentioned, even a high overall fit of a quantitative model can hide theoretically important local deviations of the data from the model. If the assumption of a multiplicative combination is correct, a two-way ANOVA with factors Potential outcome and Probability of outcome should reveal a significant interaction effect, and the interaction should be concentrated in the bilinear component (i.e., the linear × linear part of the interaction; see Keppel, 1991), reflecting a fan-shaped pattern of means in an Outcome × Probability plot (Anderson, 1982). As recently discussed by Nagengast et al. (2011) in the context of testing expectancy × value models of action, predicted multiplicative effects are often difficult to detect using ratings due to the presence of error. Hence, the empirical detection of (possibly weak) theoretically predicted interactions is another occasion for indirect scaling methods to prove their value.

We conducted separate two-way (3 × 3) ANOVAs with Potential loss (for relief) or Potential gain (for disappointment) and Outcome probability as within-subjects factors for the three measurements (rating, AFM, MLDS) and used generalised eta squared (η_g^2; Olejnik & Algina, 2003) as the measure of effect size, as recommended by Bakeman (2005) especially for repeated-measurement designs. Tests of the bilinear trend component of the interaction were conducted using the estimated marginal means of the two factors to construct contrast coefficients for the whole group.

Relief. For relief, the ANOVAs revealed significant main effects and interactions for all three kinds of measurement, ps $< .05$ or better. However, the obtained effects were consistently stronger for the indirect scaling methods: For Potential outcome, effect sizes were $\eta_g^2 = .19$ (ratings) versus .34 (AFM) and .33 (MLDS); for Outcome probability, they were .16, .36, and .31; for the interaction, they were .010, .024, and .051. More importantly, the bilinear interaction component

was very small and not significant for the ratings, $F(1, 152) = 2.03$, $p = .16$, $\eta_g^2 = .001$; whereas it was significant for the AFM scale values and accounted for nearly all of the interaction variance, $F(1, 152) = 109.31$, $p < .001$, $\eta_g^2 = .024$. Likewise, the bilinear interaction component was significant for the MLDS scale values and accounted for nearly all of the interaction variance, $F(1, 152) = 104.69$, $p < .001$, $\eta_g^2 = .046$.

To answer the question of how well the predicted bilinear interaction showed up at the level of individual participants, we tested its significance for the individual data using Tukey's (1949) test for additivity (see Anderson, 1982, p. 176). Note that because the small number of data points (i.e., nine) included in these tests, it is difficult to reach conventional levels of significance even if a bilinear interaction exists. Nevertheless, Tukey's test reached significance ($p < .05$) in 17 (44%) of 39 cases for the AFM scaling and in 21 (54%) for MLDS, whereas it was only significant in six (15.4%) cases for the rating. McNemar's test, an analogue to the dependent t-test for binary variables (Bortz, Lienert, & Boehnke, 2008), revealed that the percentage of model-consistent participants increased significantly from the rating to AFM, $\chi^2(1, N = 39) = 5.88$, $p < .05$, and to MLDS, $\chi^2(1, N = 39) = 9.33$, $p < .01$. As an alternative, we compared the variance explained by the bilinear interaction component for the direct and indirect scalings at the individual level (partial η^2) using dependent t-tests. Confirming the results of the Tukey test approach, the explained variance was significantly higher for the AFM scale values ($M = 0.61$, $SD = 0.32$) and the MLDS scale values ($M = 0.65$, $SD = 0.34$) than for the ratings ($M = 0.35$, $SD = 0.32$), ts$(38) > 3.52$, ps $< .01$.

Disappointment. For disappointment, the ANOVAs revealed significant main effects and interactions for all three kinds of measurement, ps $< .05$ or better; but with one exception, the effects were stronger for the indirect scalings: For Potential outcome, $\eta_g^2 = .14$ (ratings) versus .28 (AFM) and .25 (MLDS); for Outcome probability, $\eta_g^2 = .20$, .38, and .35; for the interaction, $\eta_g^2 = .03$, .02, and .04. Different from the results for relief, the bilinear interaction component was also significant for the ratings, $F(1, 152) = 35.48$, $p < .001$, $\eta_g^2 = .02$. For the AFM scale values, the bilinear interaction component was not only significant, but accounted for nearly all of the interaction variance, $F(1, 152) = 103.14$, $p < .001$, $\eta_g^2 = .02$. For the MLDS scale values, the bilinear interaction component was also significant and accounted for more than three quarters of the interaction variance, $F(1, 152) = 87.99$, $p < .001$, $\eta_g^2 = .03$.

Tukey's tests of additivity was significant (ps $< .05$) for seven of the 39 (18%) participants with the direct rating, for 16 (41%) with AFM scaling, and for 14 (35.9%) with MLDS. The increase in the percentage of model-consistent participants was marginally significant for the AFM scaling, McNemar $\chi^2(1, N = 39) = 3.37$, $p = .07$, but not significant for MLDS, $\chi^2(1, N = 39) = 2.40$, $p = .12$. However, the variance explained by the bilinear interaction component was significantly higher for both the AFM scalings ($M = 0.62$, $SD = 0.32$), $t(38) = 2.35$, $p < .05$, and the MLDS scalings ($M = 0.61$, $SD = 0.31$), $t(38)$, $p < .05$, than for the ratings ($M = 0.46$, $SD = 0.29$).

Discussion

Emotion intensities estimated using indirect scaling methods resulted in substantially improved fits of quantitative models of relief and disappointment. This was true for both indirect scaling methods used (AFM and MLDS). The superiority of the indirect scaling methods was manifest in higher reliabilities, higher sensitivities to interaction effects, higher effect sizes, and increased numbers of model-consistent participants. The advantages of the indirect scaling methods were particularly pronounced at the level of individuals: The median fit index increased from $R^2 = .76$ (rating) to .97/.95 (AFM/MLDS) for relief and from .86 to .94/.95 for disappointment; furthermore, whereas the predicted Belief × Desire interaction was significant for 15% (relief) and 18% (disappointment) of the participants when the direct ratings were used, the respective percentages

increased to 44/41% for the AFM scale values and to 54/36% for the MLDS scale values. Visual inspection of the data revealed that even in the cases with non-significant interactions, the pattern of means obtained using the indirect scaling methods was often close to the predicted fan shape (altogether, about 70% of the participants showed this pattern). Therefore, a significant bilinear interaction might conceivably have been found for the majority of the participants if more, or more extreme, payoffs and probabilities had been used.

The advantage of the indirect scaling methods was also apparent at the group level of analysis, where it revealed itself in increased effect sizes in the ANOVAs. Furthermore, the predicted bilinear (belief × desire) interaction for relief was detected with the indirect scale values, but not with the direct ratings.

It might be argued that our comparison of direct and indirect scaling methods is not "fair" to the ratings because the AFM and MLDS scale values were derived from twice as many data points (36 graded pair comparisons) than the direct scale values (two ratings of nine lotteries = 18 ratings). However, this objection would miss the aim of our studies. Our aim was not to compare two types of intensity judgements (absolute vs. comparative) *in isolation* (i.e., while keeping all other factors constant), but *in their typical implementation*. That is, we wanted to compare the typical form of direct ratings (the standard method for measuring emotion intensity), which involves only one or at best very few ratings of an object, with the standard form of graded pair comparisons, which involves judging the full set of object pairs. Seen from this perspective, the greater number of judgements required by the indirect scaling method is the price that has to be paid for the gain in measurement precision. Whether this is a price worth paying will be discussed in the general discussion, where we will also consider in more detail the reasons for the better performance of the indirect scaling methods.

Finally, Study 1 found that AFM was superior to MLDS in terms of model fit. This finding, too, will be addressed in the general discussion.

STUDY 2: MEASURING THE INTENSITY OF DISGUST EXPERIENCES

In Study 2, we examined whether the indirect scaling methods used in Study 1 can be used to a similar advantage for the measurement of disgust—an emotion that, according to several authors, is much less dependent on cognitions (beliefs) than are relief and disappointment. For example, Royzman and Sabini (2001) have argued that, in contrast to typical emotions such as joy or disappointment, no plausible abstract appraisal pattern or "core relational theme" (Lazarus, 1991) has yet been identified for disgust, suggesting that disgust might not depend on appraisals. Similarly, Reisenzein (2009b; 2010) proposed that, different from paradigmatic emotions, disgust (or at least the subform of disgust called "core disgust" by Rozin, Haidt, and McCauley, 2008) is a "sensory emotion" because it does not presuppose beliefs and desires about its objects, but is directly elicited by certain sensory features of the objects. More precisely, building on the evolutionary theory of disgust as an evolved disease-avoidance mechanism (Curtis & Biran, 2001; Nesse & Williams, 1995), Reisenzein (2010) proposed that disgust is elicited by sensory properties of objects that were predictive of contamination in evolution. Elaborating on this idea, a semi-quantitative model of disgust was proposed according to which the intensity of disgust is a monotonically increasing function of (i) the similarity of the eliciting object, in terms of its sensory features, to evolutionary disgust prototypes and (ii) the apparent physical closeness of the person to the disgusting object.[9]

[9] The model is called semi-quantitative because it does not specify the form of the proposed function beyond declaring it to be increasing; however, a more precise specification seems within reach (e.g., Tversky, 1977).

To test this hypothesis, Reisenzein (2010; Reisenzein & Junge, 2013b) studied the effects of an experimental manipulation of pictures of disgusting objects on experienced disgust. To manipulate the similarity of the objects to the evolutionary prototype, the colour and form of the objects were changed, whereas the apparent closeness to the objects was manipulated by varying the picture (and hence object) size. In line with predictions, it was found, among other things, that changing the colour of disgusting objects from natural to unnatural, and their size from large to small, reduced feelings of disgust. However, although these effects were statistically significant at the group level, they were not consistently found at the level of individuals. We propose that this is to a large degree due to random and systematic errors involved in the measurement of disgust feelings with rating scales. Accordingly, in Study 2, we compared direct ratings of disgust and scale values derived from graded pair comparisons with respect to their ability to confirm the predictions of the sensory theory of disgust. As in the previous studies, the colour (natural vs. unnatural) and size (large vs. small) of disgusting pictures were experimentally varied. We predicted that both variables would contribute to the intensity of experienced disgust elicited by an object, with their joint effect being either additive or superadditive (a superadditive effect would be predicted if one assumes that the apparent closeness to a disgusting object potentiates the effects of colour). Hence, the disgust model tested in Study 2 was:

$$\text{disgust}_{ij} = \mu + \text{colour}_i + \text{size}_j + \text{cs}_{ij} + \varepsilon_{ij} \text{ with } \varepsilon_{ij} \sim N(0, \delta^2). \quad (9)$$

Here, disgust_{ij} is the intensity of disgust caused by an object with colour level i and size j, μ is the mean disgust elicited by the stimuli, colour_i and size_j are the effects of colour level i and size level j, cs_{ij} is the (possibly zero) superadditive effect, and ε_{ij} is random error assumed to be normally and independently distributed with mean 0 and variance δ^2. This model can be tested using ANOVA methods (see method).

Method

Participants

Twenty female and four male psychology students participated. Most were between 18 and 24 years old; two were over 30 years. They received course credit for participation.

Materials and design

Disgust was induced by means of pictures similar to those used in previous studies (Reisenzein, 2010; Reisenzein & Junge, 2013b). Four pictures representing different major categories of disgust-inducing objects (e.g., Curtis & Biran, 2001) were used. They showed, respectively: a toilet with faeces, maggots, a purulent finger, and a plate of viscous liquid resembling bodily fluids (this picture was taken from Curtis, Aunger, & Rabie, 2004). Each picture was varied on two dimensions predicted to influence disgust intensity: colour (natural colour vs. unnatural colour) and size (large vs. small) as a proxy for apparent distance to the object. Four variants of each picture were constructed according to the 2 (Colour) $\times$ 2 (Size) design. Each participant rated all of the resulting $4 \times 4 = 16$ pictures. Pictures were presented on a 19-inch colour monitor with a screen resolution of $1{,}024 \times 768$ pixels. The large picture was 300×360 pixels and the small picture was 125×150 pixels. The unnaturally coloured versions of the pictures were obtained by tinting the originals in different neon colours (blue, green, purple, and violet) using picture-editing software.

Procedure

Like Experiment 1, the study comprised a graded pair comparison task and a direct scaling task. Participants were tested in small groups of two to four in a laboratory room equipped with several computer work places separated by room dividers. To further reduce possible acoustic interference, the participants wore sound-dampening headphones.

The two scaling tasks took 15–20 minutes to complete.

Indirect scaling task (graded pair comparisons). In Experiment 2, the graded pair comparison judgements were made before the direct ratings to see whether extended experience and ensuing familiarity with the stimuli would improve the quality of the direct intensity ratings. All possible $(16 \times 15)/2 = 120$ pairwise combinations of the 16 pictures were presented in an individually randomised order to each participant using DMDX (Forster & Forster, 2003). For each picture pair, participants indicated which picture was more disgusting and how much more. Analogous to Study 1, answers were given on a bipolar rating scale ranging from *The left picture is extremely much more disgusting* to *The right picture is extremely much more disgusting*. Intermediate scale points on both sides were labelled *Very much more*, *Much more*, *More*, *A little more*, and *Just barely more*. In half of the comparisons involving a picture, it was presented on the left side of the screen and in the other half, on the right side. It may be noted that there is little adaptation to disgusting pictures across repeated presentations, at least in the short run. Therefore, genuine disgust feelings were evoked by the stimuli during the pair comparisons. By contrast, in Study 1, participants had to compare simulated or recalled emotions.

Direct scaling task (ratings). The 16 pictures were presented in a separate random order to each participant together with an intensity rating scale ranging from 0 (*No disgust at all*) to 10 (*Very strong disgust*). Participants rated the intensity of disgust elicited by each picture by pressing a labelled button on the keyboard.

Results

Reliabilities

Table 2 shows the reliabilities and intercorrelations of the disgust measurements. The reliabilities of the indirect scalings (AFM, MLDS) are split-half reliabilities computed as in Study 1. The reliability of the disgust rating was estimated from the data of a different study in which a set of disgusting pictures, including three of those used in the present study, were first rated on a 10-point disgust rating scale and later scaled using a combined rating/ranking task where participants had to place the pictures alongside a 100 cm disgust scale; the reported reliability is the correlation between these two judgements. The Spearman–Brown correction is not meaningful in this case because only a single rating of each picture was available for the subsequent analyses.

Group level analysis

To facilitate the comparison of the different disgust measures, the AFM and MLDS scale values of each participant were transformed into the range of the participant's disgust ratings, analogous to Study 1. We first analysed the data

Table 2. *Reliabilities and intercorrelations of the ratings and indirect scalings of disgust intensity, Study 2*

		M	*SD*	*Min*	*Max*
Reliabilities[a]					
	Rating	.76 (—)	.21 (—)	.00 (—)[b]	.99 (—)
	AFM	.87 (.93)	.08 (.05)	.64 (.78)	.95 (.97)
	MLDS	.99 (.99)	.01 (.01)	.97 (.98)	.99 (.99)
Correlations					
	Rating–AFM	.82	.19	.14	.96
	Rating–MLDS	.76	.22	−.02	.96
	AFM–MLDS	.91	.11	.49	.99

Notes: [a] The computation of the reliabilities is explained in the text. Numbers in parentheses are Spearman–Brown corrected reliabilities. For the ratings, the Spearman–Brown correction is not meaningful because only a single rating was available for each picture. [b] For one participant, the correlation between the rating and the scale values obtained from the combined rating/ranking task was negative. For this participant, the reliability of the rating was defined as zero.

at the group level to see whether an advantage of the scalings was already apparent at this level. The means of disgust in the four cells of the 2 (Colour) × 2 (Size) design are plotted separately for the four disgust objects in Figure 4. A preliminary three-way ANOVA with Type of object, Picture size, and Colour as within-subjects factors revealed that Object Type interacted significantly with Colour for all three kinds of measurement, Fs(3, 69) > 8, ps < .001, and also with size, Fs(3, 69) > 5, ps < .05. These interactions were mainly due to the fact that the plate with "bodily fluids" elicited much less disgust than the other objects, and the effects of colour and size were also reduced for this object (see Figure 4). We therefore computed separate 2 × 2 ANOVAs for the four disgust pictures.

For maggots and purulent finger, both main effects of Colour and Size and the interaction were significant for all three disgust measurements, Fs(1, 23) > at least 4.4, ps at least < .05. For toilet and plate with bodily fluids, all effects

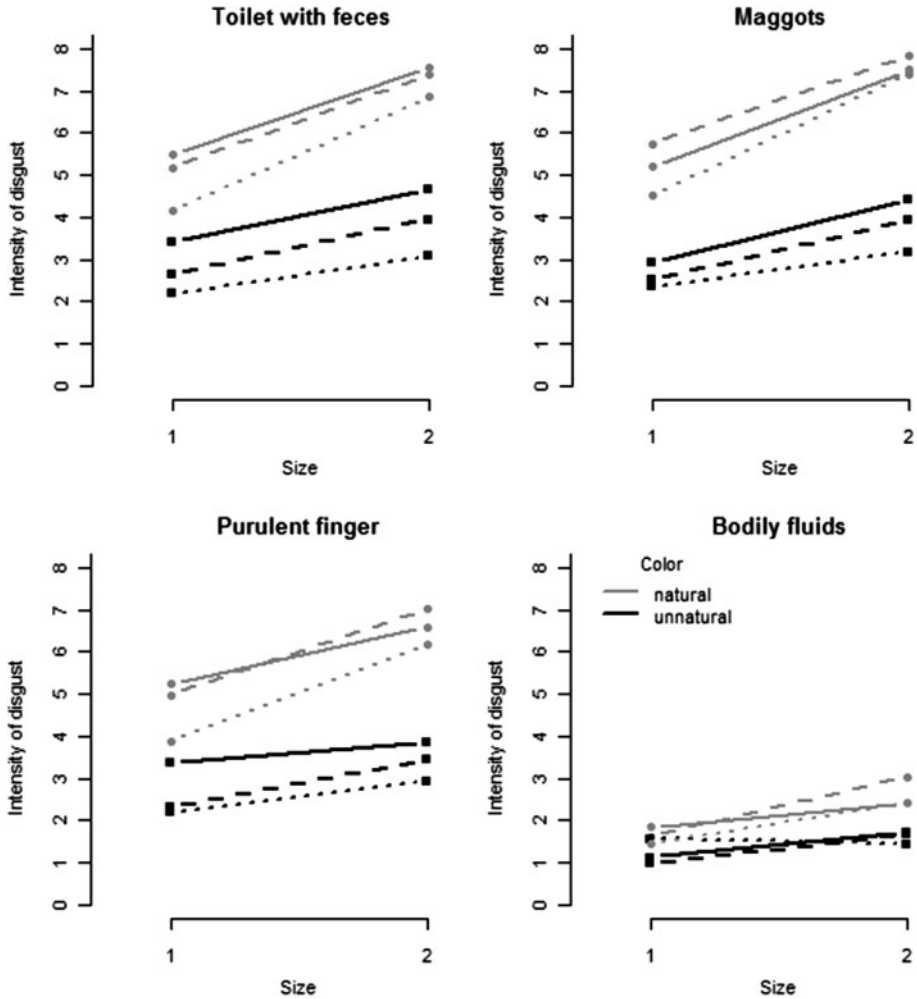

Figure 4. *Means of the disgust measurements in the four experimental conditions of Study 2, shown separately for the different objects. Mean disgust intensities are plotted against picture size, with a separate line for naturally and unnaturally coloured pictures. Solid lines are the direct disgust ratings, dashed lines are the AFM scale values, and dotted lines are the MLDS scale values.*

were significant, Fs(1, 23) > 6, ps < .05, with the exception of the interaction effect for the ratings, which was marginally significant for toilet, $F(1, 23) = 3.93$, $p = .06$, and not significant for plate with bodily fluids, $F < 1$. In addition, no significant effect of colour was obtained for the MLDS scale values for plate with bodily fluids, $F(1, 23) = 2.56$, $p = .12$. The form of the mean Colour × Size interaction agreed in all cases with the superadditive model (Figure 4).

Figure 5 shows the effect sizes (η_g^2, generalised η^2). As can be seen, for toilet, maggots, and purulent finger, both main effects and the interaction increased from the ratings to the MLDS and AFM scale values. Likewise, for plate with bodily fluids, all effects increased from the ratings to the AFM scale values. However, only the interaction effect increased from the ratings to the MLDS scale values, whereas the main effects of picture size and colour decreased.

Individual level analysis

To compare the ratings to the AFM and MLDS scale values at the level of individuals, we computed the number of participants whose response patterns conformed to the ordinal predictions of the disgust model. According to the model, picture size and colour should have an additive or superadditive effect. Labelling the four cells of the 2 (natural vs. unnatural colour) × 2 (small vs. large picture) design from top left to bottom right *a*, *b*, *c*, *d*, this assumption implies

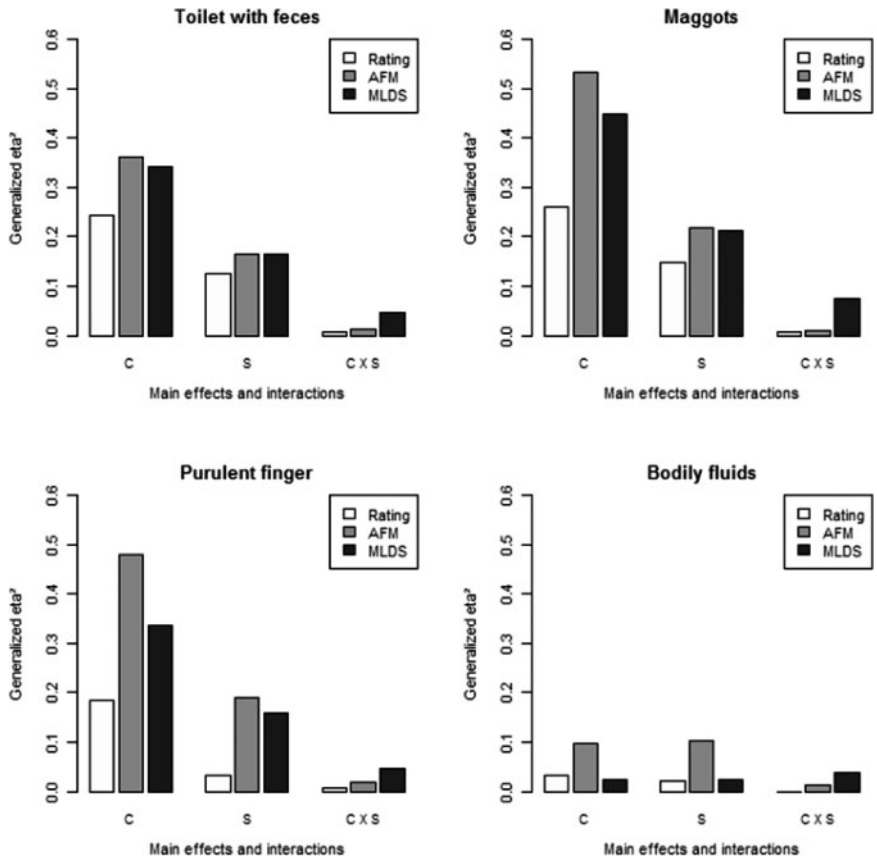

Figure 5. *Generalised η^2 for the main effects of colour (C) and size (S) and the Colour × Size interaction (C × S), shown separately for the four disgust objects.*

$a < b$, $c < d$ and $a > c$, $b > d$. That is, for both levels of colour, large pictures should be more disgusting than small pictures; likewise, for both levels of size, naturally coloured pictures should be more disgusting than unnaturally coloured pictures.

As shown in Table 3, for all four disgust pictures, the number of participants whose patterns of ratings or scale values conformed to the ordinal predictions of the disgust model increased substantially from the ratings to the scalings. On average (across the four pictures), the percentage of participants consistent with the disgust model was 30% for the ratings, 51% for the MLDS scale values, and 85% for the AFM scalings. In two cases (toilet and purulent finger), more than 90% became model-consistent if the AFM scale values were used. McNemar's test revealed that the increases in the number of model-consistent participants were significant (ps $< .05$) for all disgust pictures for the AFM scalings, and for one picture (purulent finger) for the MLDS scalings (Table 3).

Discussion

Study 2 conceptually replicated the central findings of Study 1 for the emotion of disgust. Analogous to Study 1, disgust intensity measurements obtained using indirect scaling procedures (AFM and MLDS) yielded greatly improved fits to the proposed model of sensory disgust. The advantage of the indirect scaling methods for the measurement of disgust was evident at both the group and individual levels of analysis. At the group level, it was reflected in a greater sensitivity for interaction effects (4 of 4 for the indirect scalings vs. 2 of 4 for the ratings) and in substantially increased effect sizes (with the exception of the main effects obtained with the MLDS scale values for plate with bodily fluids). At the level of individuals, the superiority of the indirect scaling methods was evident in the greatly increased number of participants whose data were consistent with the disgust model.

Finally, also replicating the findings of Study 1, we obtained better model fits using AFM than MLDS.

Table 3. *Number of participants conforming to the disgust model for the direct and indirect scalings*

	Ratings		*AFM*		*MLDS*		*Ratings–AFM*		*Ratings–MLDS*		*AFM–MLDS*	
Object	*N*	%	*N*	%	*N*	%	$\chi^2(1, N = 24)$		$\chi^2(1, N = 24)$		$\chi^2(1, N = 24)$	
Toilet	9	38	23	96	14	58	12.07	***	1.23	*ns*	7.11	**
Maggots	13	54	21	88	17	71	4.90	*	0.90	*ns*	2.25	*ns*
Purulent finger	6	25	22	92	14	58	14.06	***	4.90	*	4.08	*
Bodily fluids	1	4	15	63	4	17	10.56	**	0.80	*ns*	9.09	**

Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

GENERAL DISCUSSION

Two studies investigated the utility of indirect scaling methods, based on graded pair comparisons, for the testing of quantitative emotion theories: A quantitative belief–desire model of relief and disappointment (Study 1) and a semi-quantitative theory of disgust as a sensory emotion (Study 2). The findings of the studies suggest that both indirect scaling methods (AFM and MLDS) yield more precise measurements of emotion intensity than the direct ratings typically used in emotion research. The superiority of the indirect intensity measurements was evident in higher reliabilities as well as in better fits of the emotion models. The beneficial effects of the indirect scaling methods were particularly pronounced at the level of individuals. In Study 1, the median fit index increased from $R^2 = .76$ (ratings) to .97/.95 (AFM/MLDS) for relief and from .86 to .94/.95 for disappointment; furthermore, whereas only 15% (relief) and 18% (disappointment) of the participants showed the predicted significant Belief × Desire interaction using the direct ratings, these percentages rose to 44/41% for the AFM scale values and to 54/36% for the MLDS scale values (plus another 25% showed a data pattern close to the predicted fan shape). Similarly, in Study 2, the number of participants consistent with the proposed disgust model increased on average (across the four disgust pictures) from 30% when disgust was measured using direct ratings to 51% for the MLDS and 85% for the AFM scalings.

At the group level, the superiority of the indirect scalings was evident in increased effect sizes (see Study 2 in particular) and in a higher sensitivity for predicted interactions: Whereas all predicted interactions were detected with the indirect scaling methods, the direct ratings missed one of two interactions in Study 1, and two of four in Study 2.

Differences between AFM scaling and MLDS

Judged by the fit of the emotion models, AFM scaling was generally superior to MLDS, although the difference was not large.[10] Because both scaling methods were ultimately based on the same raw data (graded pair comparisons), the reason for this difference can be traced to the scaling models. One possible explanation is that AFM scaling exploits additional reliable information contained in the graded pair comparisons that is disregarded in MLDS. This additional information concerns the size of the intervals (differences) between scale values. For example, assume that the interval (a,b) is rated 6 (e.g., in Study 1, outcome b is judged as "extremely much more relieving" than a), (c,d) is rated 4 and (e,f) is rated 2. Then the difference between the intervals ab and cd is 2 and that between ab and ef is 4. The AFM model explicitly takes these differences in interval size into account when estimating the scale values, whereas the MLDS model uses only the ranks of the differences, to decide which of two intervals is larger. Accordingly, the response to both the quadruple (ab,cd) and the quadruple (ab,ef) in the above example is coded 1 in MLDS. Thus, some of the information about interval size differences is ignored in MLDS (although not all of this information is ignored, as larger intervals will more often dominate smaller intervals). The better performance of AFM scaling could therefore indicate that graded pair comparisons contain more than just ordinal information about intensity differences—they also contain some amount of metric information.

How does indirect scaling achieve its effects?

O'Brien (1985) distinguished three kinds of error that can contaminate direct intensity ratings of a latent quantity: Random measurement errors, transformation errors resulting from the mapping of the latent metric variable into a manifest

[10] Interaction effects were larger for MLDS than for AFM in Study 2 (see Figure 5) and partly also in Study 1 (for relief). However, without additional investigation (e.g., simulation studies), we find it difficult to say whether this finding reflects greater sensitivity of MLDS for interaction effects or a tendency of MLDS to overestimate the size of these effects.

ordinal scale, and errors of categorisation (or coarsening) resulting from a reduction of the level of resolution of the manifest variable (the rating). Although it was not the goal of the present studies to determine to which degree indirect scaling differentially reduces these different kinds of errors, the findings provide some information relevant to this question. In particular, the higher reliabilities and larger effect sizes obtained for the indirect scaling methods indicate that they reduced *random measurement error*. This reduction was in part undoubtedly due to the larger number of judgements from which the indirect scale values were derived; but in part it could also have been due to the fact that these judgements were comparative rather than absolute (cf. the introduction). In addition, it seems likely that the graded pair comparisons reduced *categorisation errors*. The reason is that graded pair comparisons of objects enable and encourage participants to make finer distinctions than can be made (or at least are typically made) with rating scales (see Reisenzein & Schimmack, 1999, Study 3).

Finally, the hypothesis that the indirect scaling methods reduced *transformation errors* receives some support from the better performance of AFM compared to MLDS: As argued above, this finding could mean that the graded pair comparison judgements contain metric information. However, this support is rather indirect. Inspection of the statistical fit of the scaling models is similarly inconclusive. Theory dictates that, if the judgement models underlying the AFM and MLDS scaling methods are correct, then the estimated scale values form an interval scale (Bechtel, 1967; Knoblauch & Maloney, 2008). However, although the obtained statistical fits of the scaling models to the data (not reported in this article) were good, they were by no means perfect; and even a high global fit of a scaling model is compatible with systematic violations of metric structure (Michell, 1990; see also, Birnbaum, 2011).

A more sensitive method for probing the metric character of a scale is the testing of measurement axioms. For the case of MLDS, a set of necessary and sufficient axioms for the existence of a metric (interval) representation of the input data (quadruple judgements) is available in the form of the axioms for *additive difference structures* (Krantz et al., 1971). These axioms can be tested with the same data that are used to estimate the MLDS scale values; hence an additional benefit of graded pair comparisons is that they allow us to test whether the resulting scale is metric. Although the measurement axioms are formulated deterministically (i.e., do not consider random error), methods for axiom testing that take random error into account have recently been developed (e.g., Karabatsos, 2005; Maloney & Yang, 2003). Specifically, along with introducing MLDS, Maloney and Yang (2003) proposed a parametric bootstrap test for the central axiom of additive difference structures, the *weak monotonicity axiom*. This test allows us to decide whether the responses of a person to quadruples of objects (ab,cd) conform to the requirement of weak monotonicity up to the person's level of random error, estimated as a by-product of MLDS. This bootstrap test can be extended to other axioms of additive difference structures (Knoblauch & Maloney, 2008). Furthermore, it seems possible to extend this axiom testing method to AFM scalings, and even to rating scales (Junge & Reisenzein, 2013b; see also Orth, 1982; Westermann, 1994). Tests of the measurement axioms for additive difference structures for several kinds of emotion measurements will be reported elsewhere (Junge & Reisenzein, 2013b).

Implications for emotion research

Although indirect scaling methods have been around for a long time, they have been rarely used in emotion research. As mentioned in the introduction, one reason for this may have been the absence of studies that demonstrate the benefits of indirect scaling techniques in emotion research. Filling this research lacuna, our studies show that indirect scaling methods based on graded pair comparisons can substantially improve the testing of emotion theories relative to direct ratings of emotion intensity. This finding, together with the possibility afforded by pair comparisons

to probe the metric character of the obtained scales, recommend the indirect scaling techniques as alternative methods of emotion measurement (see Böckenholt, 2004; Maydeu-Olivares & Böckenholt, 2008, for additional advantages). However, the indirect scaling methods are more expensive than ratings, and they have limitations of their own. Therefore, the question arises under which circumstances indirect scaling methods should be used. This question is discussed below. In addition, we give some practical advice about how to conduct indirect scaling studies using graded pair comparisons.

When should indirect scaling methods be used?

The central advantage of indirect scaling methods is their increased precision. Accordingly, one central consideration in deciding whether (or when) to use or not to use indirect scaling methods for the measurement of emotion intensity is how much measurement precision is needed. The answer depends on the precision of the hypotheses one wishes to test (quantitative, ordinal, or purely qualitative). As mentioned in the introduction, testing quantitative emotion theories requires reasonably error-free, metric scales. The results of our studies suggest that indirect scaling methods are more likely to provide such scales than direct ratings, particularly at the level of the individual, on which emotion theories should preferably be tested; although direct ratings are still useful to estimate the scale origin (see Study 1; other methods for identifying the scale origin are discussed in Böckenholt, 2004). However, tests of ordinal and qualitative hypotheses (where the dependent variable is quantitative) will also profit from more precise measurement because the power of statistical tests is increased (cf. Study 2).

The main disadvantage of indirect scaling methods is that they are more expensive, in terms of both administration time and computational effort, than direct ratings. Although the method of graded pair comparisons used in our studies is more economical than are some other indirect scaling methods, in particular the method of quadruple comparisons (Maloney & Yang, 2003), it still requires many more judgements than direct scaling methods do. However, as explained below, this difference can be further reduced by scaling only a subset of the stimuli. Furthermore, although indirect scaling methods require a larger investment in measurement, they lead to savings in terms of participants because precise measurements allow the detection of effects with smaller samples (see also, Nagengast et al., 2011).

Apart from their greater costs, indirect scaling methods have certain limitations of applicability that one should be aware of. One limitation concerns the number of stimuli that can be scaled. Because the number of possible pair comparisons increases with the square of the number of stimuli-for n stimuli, there are $(n^2 - n)/2$ pairs-collecting the complete set of graded pair comparisons in a single session becomes impractical beyond, say, 16 stimuli (120 pair comparisons, cf. Study 2). However, more stimuli can be scaled by dividing the experiment into several sessions, or by presenting only a subset of the pair comparisons (subsampling). Subsampling is possible because the probabilistic scaling methods used in our studies allow the estimation of scale values from incomplete input data. Subsampling can either be done randomly (Maloney & Yang, 2003), or systematically, using so-called incomplete cyclic designs (Burton, 2003). Subsampling studies conducted by these authors suggest that the number of pair comparisons needed to derive reliable scale values can be reduced to at least a third, meaning that the number of stimuli judged by a participant even in a single session can be increased to at least 25. If care is taken that the stimuli are optimally spaced across the intensity continuum, this number should be sufficient to allow estimation of even fairly complicated quantitative functions (e.g., the inverted S-shaped probability weighting functions of prospect theory; Stott, 2006). Alternatively, the reduction in the number of pair comparisons afforded by subsampling could be used to scale more than one emotion, either in the same or in different trials. Still, it should be acknowledged that even when these options are exploited to the full, indirect

scaling methods are not well suited for measuring multiple emotions at a time.

A second possible limitation of the usability of pair comparisons in emotion research stems from the assumption of the scaling models that the intensity of an emotion elicited by a stimulus on repeated presentations remains constant up to a randomly fluctuating component (treated as part of the measurement error). This assumption implies that participants do not quickly adapt to the stimuli, which may be unrealistic for some kinds of emotion elicitors. One way to overcome this problem is to ask participants to recall or imagine emotion-eliciting events (that may have been presented before) and compare these recalled or imagined events (cf. Study 1). This method presupposes that the emotional effects of recalled and imagined events are comparable to those of real events, which may not always be justified. Note, however, that the problem of adaptation to stimuli exists for all kinds of repeated emotion measurements including ratings, although it is less salient in the latter case because the number of repeated presentations of the same stimuli is usually very small.

Despite their limitations and greater costs, we believe that indirect scaling methods based on graded pair comparisons could be the method of choice for measuring emotion intensity in many research situations. Apart from their advantages over direct ratings, graded pair comparisons also have advantages over other indirect scaling methods. Compared to binary pair comparisons, they are similarly economical but generate more information and can also be used for the scaling of clear suprathreshold intensity differences. Compared to quadruple comparisons, they are much more economical but seem to contain largely the same information (see also Footnote 3). Therefore, graded pair comparisons can also be scaled using MLDS, which requires only that the difference judgements are on an ordinal scale level. Other scaling models for graded pair comparisons that offer the same advantage are also available (the cumulative probit model and nonmetric multidimensional scaling; though see Footnotes 2 and 7 for limitations). Although our studies suggest that AFM scaling can perform even better than MLDS, this could be different with other emotions or in different settings. Furthermore, in our view the use of nonmetric scaling methods should always be considered as an alternative to AFM scaling, or as a check on its results, to avoid the strong assumption of a metric response function.

Finally, note that the indirect scaling methods advocated in this article are not restricted to self-reports of specific feelings but can be used to measure all aspects of emotion that are consciously accessible, including appraisals, action tendencies and bodily symptoms.

Practical considerations in conducting indirect scaling studies

Graded pair comparisons can in principle be made using paper and pencil and a printed questionnaire, and AFM-type scale values (for complete data) can in principle be computed by hand (Oishi et al., 1998). Nevertheless, to speed up the collection and analysis of the data, to allow the estimation of scale values from incomplete data, and to be able to use more complex scaling models such as MLDS, we recommend the computer-aided implementation of graded pair comparison measurement. Basically, any experiment generator software, including the freeware programs DMDX (Forster & Forster, 2003) and WEXTOR (Reips & Neuhaus, 2002) used in our studies, can be used to program indirect scaling experiments for a wide variety of emotional stimuli (e.g., pictures, sounds, text-based scenarios, video clips). Likewise, basically any statistical analysis system that includes a module for generalised linear models can be used to estimate the scale values of the AFM and MLDS models. We recommend R (R Development Core Team, 2011) for several reasons: It is free; it already contains several ready-to-use add-on packages for specific scaling models including MLDS (Knoblauch & Maloney, 2008); it makes it comparatively easy to program one's own implementations and extensions of scaling models or to modify existing programs; and it allows the creation of a seamless, customised analysis workflow ranging

from reading the graded pair comparison judgements into R to combining the obtained individual scale values with other data from the same participants.

Manuscript received 21 September 2012
Revised manuscript received 27 February 2013
Manuscript accepted 1 March 2013
First published online 22 April 2013

REFERENCES

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review*, *77*, 153–170. doi:10.1037/h0029064

Anderson, N. H. (1982). *Methods of information integration theory*. New York, NY: Academic Press.

Anderson, N. H. (1989). Information integration approach to emotions and their measurement. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research and experience* (Vol. 4, pp. 133–186). New York, NY: Academic Press.

Bagozzi, M. (1980). *Causal models in marketing*. New York, NY: Wiley.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379–384. doi:10.3758/BF03192707

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York, NY: Wiley.

Bechtel, G. G. (1967). The analysis of variance and pairwise scaling. *Psychometrika*, *32*, 47–65. doi:10.1007/BF02289404

Bechtel, G. G., & O'Connor, P. J. (1979). Testing micropreference structures. *Journal of Marketing Research*, *16*, 247–257. doi:10.2307/3150688

Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations Research*, *33*, 1–27. doi:10.1287/opre.33.1.1

Birnbaum, M. H. (2011). Testing theories of risky decision making via critical tests. *Frontiers in Psychology*, *2*, 315. doi:10.3389/fpsyg.2011.00315

Böckenholt, U. (2001). Thresholds and intransitivities in pairwise judgments: A multilevel analysis. *Journal of Educational and Behavioral Statistics*, *26*, 269–282. doi:10.3102/10769986026003269

Böckenholt, U. (2003). Thurstonian-based analyses: Past, present, and future utilities. *Psychometrika*, *71*, 615–629. doi:10.1007/s11336-006-1598-5

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, *4*, 453–465. doi:10.1037/1082-989X.9.4.453

Bollen, K. A., & Noble, M. D. (2011). Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 15639–15646. doi:10.1073/pnas.1010661108

Borg, I., & Staufenbiel, T. (2007). *Theorien und Methoden der Skalierung [Theories and methods of scaling]*. Bern: Huber.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik [Nonparametric methods in biostatistics]* (3rd ed.). Heidelberg: Springer.

Boschman, M. C. (2001). DifScal: A tool for analyzing difference ratings on an ordinal category scale. *Behavior Research Methods, Instruments, & Computers*, *33*, 10–20. doi:10.3758/BF03195343

Burton, M. L. (2003). Too many questions? The uses of incomplete cyclic designs for paired comparisons. *Field Methods*, *15*, 115–130. doi:10.1177/1525822X03015002001

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin and Review*, *15*, 692–712. doi:10.3758/PBR.15.4.692

Critchlow, D. E., & Fligner, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, *3*, 517–533. doi:10.1007/BF02294488

Curtis, V., Aunger, R., & Rabie, T. (2004). Evidence that disgust evolved to protect from risk of disease. *Proceedings of the Royal Society London B*, *271*(Suppl.), 131–133. doi:10.1098/rsbl.2003.0144

Curtis, V., & Biran, A. (2001). Dirt, disgust and disease: Is hygiene in our genes? *Perspectives in Biology and Medicine*, *44*, 17–31. doi:10.1353/pbm.2001.0001

Davis, W. (1981). A theory of happiness. *Philosophical Studies*, *39*, 305–317. doi:10.1007/BF00354361

De Beuckelaer, A., Kampen, J. K., & Van Trijp, H. C. M. (2013). An empirical assessment of the cross-national measurement validity of graded paired comparisons. *Quality and Quantity*, *47*, 1063–1076. doi:10.1007/s11135-011-9583-1

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy.

Behavior Research Methods, Instruments, & Computers, *35*, 116–124. doi:10.3758/BF03195503

Fox, C. R., & Poldrack, R. A. (2009). Prospect theory and the brain. In P. W. Glimcher, C. F. Camerer, E. Fehr & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 145–174). London: Elsevier/Academic Press.

Frijda, N. H., Ortony, A., Sonnemans, J., & Clore, G. L. (1992). The complexity of intensity: Issues concerning the structure of emotion intensity. In M. S. Clark (Ed.), *Review of personality and social psychology* (Vol. 13, pp. 60–89). Beverley Hills, CA: Sage.

Galanter, E. (1990). Utility scales of monetary and nonmonetary events. *American Journal of Psychology*, *103*, 449–470. doi:10.2307/1423318

Gonzalez, R., & Wu, G. (1999). On the shape of probability weighting function. *Cognitive Psychology*, *38*, 129–166. doi:10.1006/cogp.1998.0710

Gratch, J., Marsella, S., Wang, N., & Stankovic, B. (2009). *Assessing the validity of appraisal-based models of emotion*. International Conference on Affective Computing and Intelligent Interaction. Amsterdam, IEEE, 2009. (Retrieved from: http://www.ict.usc.edu/ ~marsella/publications/ACII09-appraisal.pdf)

Green, O. H. (1992). *The emotions: A philosophical theory*. Dordrecht: Kluwer.

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

Gridgeman, N. T. (1959). Pair comparison, with and without ties. *Biometrics*, *15*, 382–388. doi:10.2307/2527742

Guilford, J. P. (1954). *Psychometric methods*. New York, NY: McGraw-Hill.

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions*. College Station, TX: Stata Press.

Junge, M., & Reisenzein, R. (2010). *Kleine Wahrscheinlichkeiten—große Emotionen! [Small probabilities—Big emotions!]*. Talk presented at the 47th Congress of the German Psychology Association in Bremen.

Junge, M., & Reisenzein R. (2013a). *Graded pair comparisons are an equivalent substitute for quadruple comparisons in emotion scaling experiments*. Talk presented at the 55th Conference of Experimental Psychologists in Vienna.

Junge, M., & Reisenzein, R. (2013b). *Metric scales for emotion measurement*. Manuscript in preparation, University of Greifswald.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. doi:10.2307/1914185

Karabatsos, G. (2005). The exchangeable multinomial model as an approach to testing deterministic axioms of choice and measurement. *Journal of Mathematical Psychology*, *49*, 51–69. doi:10.1016/j.jmp.2004.11.001

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice Hall.

Knoblauch, K., & Maloney, L. T. (2008). MLDS: Maximum likelihood difference scaling in R. *Journal of Statistical Software*, *25*, 1–28.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. Vol. 1: Additive and polynomial representations*. New York: Academic Press.

Lazarus, R. S. (1991). *Emotion and adaptation*. New York, NY: Oxford University Press.

Loomes, G., & Sudgen, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *Review of Economic Studies*, *53*, 271–282. doi:10.2307/2297651

MacKay, D. B., & Zinnes, J. L. (1986). A probabilistic model for the multidimensional scaling of proximity and preference data. *Marketing Science*, *5*, 325–344. doi:10.1287/mksc.5.4.325

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*, 573–585. doi:10.1167/3.8.5

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, *23*, 209–237. doi:10.1080/02699930802204677

Maydeu-Olivares, A., & Böckenholt, U. (2008). Modeling subjective health outcomes: Top 10 reasons to use Thurstone's method. *Medical Care*, *46*, 346–348. doi:10.1097/MLR.0b013e31816dd8d9

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*, 423–429. doi:10.1111/j.1467-9280.1997.tb00455.x

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.

Michell, J. (2006). Psychophysics, intensive magnitudes, and the psychometrician's fallacy. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *17*, 414–432. doi:10.1016/j.shpsc.2006.06.011

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the

"X" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science*, *22*, 1058–1066. doi:10.1177/0956797611415540

Nesse, G., & Williams, C. (1995). *Evolution and healing*. London: Weidenfeld & Nicolson.

O'Brien, R. (1985). The relationship between ordinal measures and their underlying values: Why all the disagreement? *Quality and Quantity*, *19*, 265–277.

Oishi, S., Schimmack, U., Diener, E., & Suh, E. M. (1998). The measurement of values and individualism–collectivism. *Personality and Social Psychology Bulletin*, *24*, 1177–1198. doi:10.1177/01461672982411005

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*, 434–444. doi:10.1037/1082-989X.8.4.434

Orth, B. (1982). A theoretical and empirical study of scale properties of magnitude-estimation and category-rating scales. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 351–377). Hillsdale, NJ: Erlbaum.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Prelec, H. (1998). The probability weighting function. *Econometrica*, *66*, 497–528. doi:10.2307/2998573

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. (ISBN 3-900051-07-0, URL http://www.R-project.org/)

Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, and Computers*, *34*, 234–240.

Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and Social Psychology*, *67*, 525–539. doi:10.1037/0022-3514.67.3.525

Reisenzein, R. (2000). Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition and Emotion*, *14*, 1–38. doi:10.1080/026999300378978

Reisenzein, R. (2009a). Emotions as metarepresentational states of mind: Naturalizing the belief–desire theory of emotion. *Cognitive Systems Research*, *10*, 6–20. doi:10.1016/j.cogsys.2008.03.001

Reisenzein, R. (2009b). Emotional experience in the computational belief desire theory of emotion. *Emotion Review*, *1*, 214–222. doi:10.1177/1754073909103589

Reisenzein, R. (2010). Is disgust an emotion? [Abstract]. *Review of Psychology*, *17*, 144–145. (Abstract of a talk presented at the 9th Alps Adria Psychology Conference, Klagenfurt, Austria)

Reisenzein, R., & Junge, M. (2006). *Überraschung, Enttäuschung und Erleichterung: Emotionsintensität als Funktion von subjektiver Wahrscheinlichkeit und Erwünschtheit [Surprise, disappointment and relief: Emotion intensity as function of subjective probability and desirability]*. Talk presented at the 45th Congress of the German Psychological Association in Nuremburg.

Reisenzein, R., & Junge, M. (2012). Language and emotion from the perspective of the computational belief–desire theory of emotion. In P. A. Wilson (Ed.), *Dynamicity in emotion concepts* (Lodz Studies in Language, Vol. 27, pp. 37–59). Frankfurt am Main: Peter Lang.

Reisenzein, R., & Junge, M. (2013a). *Testing the quantitative belief–desire theory of emotion using mixed nonlinear models*. Manuscript in preparation, University of Greifswald.

Reisenzein, R., & Junge, M. (2013b). *Disgust as a sensory emotion*. Manuscript in preparation, University of Greifswald.

Reisenzein, R., & Schimmack, U. (1999). Similarity judgments and covariations of affects: Findings and implications for affect structure research. *Personality and Social Psychology Bulletin*, *25*, 539–555. doi:10.1177/0146167299025005001

Reisenzein, R., Studtmann, M., & Horstmann, G. (2013). Coherence between emotion and facial expression: Evidence from laboratory experiments. *Emotion Review*, *5*, 16–23. doi:10.1177/1754073912457228

Ritz, C., & Streibig, J. C. (2008). *Nonlinear regression with R*. Berlin: Springer.

Roberts, F. S. (1979). *Measurement theory: With applications to decision making, utility, and the social sciences*. Reading, MA: Addison-Wesley.

Royzman, E. B., & Sabini, J. (2001). Something it takes to be an emotion: The interesting case of disgust. *Journal for the Theory of Social Behaviour*, *31*, 29–59. doi:10.1111/1468-5914.00145

Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 757–776). New York, NY: Guilford Press.

Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92–120). New York, NY: Oxford University Press.

Stevens, S. S. (1975). *Psychophysics*. New York, NY: Wiley.

Stott, H. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*, 101–130. doi:10.1007/s11166-006-8289-6

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286. doi:10.1037/h0070288

Titchener, E. B. (1905). *Experimental psychology: A manual of laboratory practice. Vol. II. Quantitative experiments. Part II. Instructor's manual*. London: Macmillan.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley.

Tukey, J. (1949). One degree of freedom for non-additivity. *Biometrics*, *5*, 232–242. doi:10.2307/3001938

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352. doi:10.1037/0033-295X.84.4.327

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. doi:10.1007/BF00122574

Westermann, R. (1994). Measurement-theoretical idealizations and empirical research practice. In M. Kuokkanen (Ed.), *Idealization VII: Structuralism, idealization and approximation* (Poznan studies in the philosophy of the sciences and the humanities, Vol. 42, pp. 271–284). Amsterdam: Rodopi.

Zheng, B., & Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, *19*, 1771–1781. doi:10.1002/1097-0258(20000715)19:13%3C1771::AID-SIM485%3E3.0.CO;2-P