Psychological Test and Assessment Modeling, Volume 58, 2016 (3), 497-530

Metric scales for emotion measurement

Martin Junge[1] *& Rainer Reisenzein*[2]

Abstract

The scale quality of indirect and direct scalings of the intensity of emotional experiences was investigated from the perspective of representational measurement theory. Study 1 focused on sensory pleasantness and disgust, Study 2 on surprise and amusement, and Study 3 on relief and disappointment. In each study, the emotion intensities elicited by a set of stimuli were estimated using Ordinal Difference Scaling, an indirect probabilistic scaling method based on graded pair comparisons. The obtained scale values were used to select test cases for the quadruple axiom, a central axiom of difference measurement. A parametric bootstrap test was used to decide whether the participants' difference judgments systematically violated the axiom. Most participants passed this test. The indirect scalings of these participants were then linearly correlated with their direct emotion intensity ratings to determine whether they agreed with them up to measurement error, and hence might be metric as well. The majority of the participants did not pass this test. The findings suggest that Ordinal Difference Scaling allows to measure emotion intensity on a metric scale level for most participants. As a consequence, quantitative emotion theories become amenable to empirical test on the individual level using indirect measurements of emotional experience.

Keywords: Emotion intensity, measurement of emotional experience, representational measurement theory, test of measurement axioms, quadruple axiom, scaling methods, Ordinal Difference Scaling, bootstrap test

[1] *Correspondence concerning this article should be addressed to:* Martin Junge, PhD, Institute of Psychology, University of Greifswald, Franz-Mehring-Straße 47, 17487 Greifswald, Germany; email: martin.junge@uni-greifswald.de

[2] University of Greifswald, Germany

Linguistic and phenomenological evidence indicates that emotional experiences differ from each other not only in quality (e.g., happiness versus fear), but also in intensity. More precisely, each emotion quality can be exemplified in different degrees or gradations, ranging from just noticeable to highly intense (see e.g., Frijda, Ortony, Sonnemans, & Clore, 1992; Reisenzein, 1994). This suggests that emotional experiences are quantitative magnitudes, or *quantities* (Michell, 1990); that is, continuous variables with a metric, or additive, structure (Hölder, 1901; Krantz, Luce, Suppes, & Tversky, 1971; Michell, 1990). Indirect support for this hypothesis is provided by the consideration (Reisenzein, 2012) that, in being a group of related phenomenal qualities graded in intensity, emotions are similar to sensations (e.g., of tone, touch, or temperature); sensations, however, are widely regarded as quantitative magnitudes (e.g., Kingdom & Prins, 2010; Schneider, 1982; Stevens, 1975).

If *emotions* are quantities, then *theories of emotion* should ideally be quantitative theories (Carnap, 1966), i.e. theories that connect emotions to their causes and consequences by quantitative laws. However, although the intensity aspect of emotions is acknowledged by most researchers and is considered in the most frequently used method of emotion measurement, ratings on quality-plus-intensity scales of emotional experience (e.g., Pekrun & Bühner, 2014; Reisenzein, 1994; Scherer, 2005), only few explicitly quantitative emotion theories have been proposed so far (e.g., Gratch, Marsella, Wang, & Stankovic, 2009; Mellers, 2000; Reisenzein, 2009). A main reason for this state of affairs is presumably the problem of measuring the intensity of emotions with sufficient precision to allow the testing of quantitative emotion theories. Suitable measurements must fulfill two requirements: First, they must be reasonably free of measurement error; second, they must have a metric, i.e. an interval or even (depending on the theory tested) a ratio scale level.

Do we have metric scales for emotion measurement? As concerns direct emotion intensity ratings (e.g., "How happy do you feel right now?" on a scale from 0 = "not happy at all" to 10 = "extremely happy"), emotion researchers typically do treat them as metric (interval) scales in their data analyses. Nevertheless, most of them would probably agree that the scale level of these ratings is more likely somewhere in between the ordinal and interval, and is possibly only ordinal (see e.g., Krantz et al., 1971). Although the critics of rating scales usually grant that sensations and similar subjective experiences, including emotional feelings, can be regarded as quantities, they believe that the information about their metric structure is partially or completely lost during the process of introspecting these quantities and mapping them into a rating scale (see e.g., O'Brien, 1985). However, as has long been argued (e.g., Krantz et al., 1971; Orth, 1982; Westermann, 1985), this claim – just as the converse claim that ratings are metric – should be tested rather than simply asserted. More importantly, even if category ratings do not yield metric measurements of emotion intensity, alternative methods may be able to do so.

Promising candidates for such alternative measurement methods have long been proposed in psychology in the form of *indirect scaling methods* (e.g., Borg & Staufenbiel, 2007; Maydeu-Olivares & Böckenholt, 2008; Titchener, 1905; Torgerson, 1958). The basic idea behind indirect scaling methods is to infer the intensity of subjective experiences from judgments 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 judgments used in indirect scaling procedures are ordinal pairwise comparisons of subjective intensities (e.g., in the case of emotion: "the intensity of relief elicited by event a is greater than that elicited by event b"; Junge & Reisenzein, 2013, 2015). From these data, the absolute intensities of the experiences caused by the stimuli are estimated with the help of appropriate scaling models, which are really miniature models of the judgment processes thought to underlie the pair comparison judgments. Hence, the intensities of the experiences elicited by the different stimuli are indirectly determined (i.e., inferred from the pair comparisons) rather than directly reported by the subject, as in category rating and other direct scaling methods (e.g., magnitude scaling; Stevens, 1975).

Supporting the advantages of indirect scaling methods in emotion research, Junge and Reisenzein (2013) found that indirect scalings of emotion intensity based on graded pair comparisons (GPCs; e.g., Bechtel & O'Connor, 1979) were more reliable than direct intensity ratings and had better fits to proposed quantitative emotion models. Part of the advantage of the indirect scalings was undoubtedly due to the reduction of random measurement error; but in part, it could also have resulted from the attainment of a metric (or at least close-to-metric) scale level. However, in our previous research we did not test whether the indirect and direct scalings of emotion intensity were metric. This is the aim of the studies reported in the present article. In three studies focusing on different kinds of emotional experience, we first tested whether emotion intensities estimated with an indirect scaling method tailored to GPCs (Ordinal Difference Scaling, ODS; see e.g., Boschman, 2001; Junge & Reisenzein, 2015), are metric. To determine whether this was the case, we took recourse to representational measurement theory (Krantz et al., 1971). Specifically, we tested whether the participants' judgments of intensity differences were in agreement with the quadruple axiom, a central axiom of difference measurement (e.g., Luce & Suppes, 1965). A modification of a parametric bootstrap test proposed by Maloney and Yang (2003) was used to decide whether observed axiom violations were systematic. For those participants who passed the test of the quadruple axiom and hence appeared able to provide metric indirect scalings, we then tested, using another bootstrap procedure, whether their direct scalings of emotion intensity correlated linearly with their indirect scalings up to measurement error, and hence might be metric as well.

A deductive approach to representational measurement

How can one decide whether or not a proposed measurement M of a latent variable is metric? The answer proposed by representational measurement theory (e.g., Krantz et al., 1971; Roberts, 1979) is that one needs to study the intrinsic structure of M, that is, the relations that hold between the different levels (including differences between and combinations of levels) of M, as exemplified by a set of objects. Furthermore, to avoid presupposing what needs to be shown, only nonquantitative (meaning, in the typical case, ordinal) relations must be allowed in this structural analysis. If a variable is metric, then its levels stand to each other in a determinate set of ordinal relations that together form an *additive structure* (Michell, 1990). Conversely, if the levels of a variable have an

additive structure, they can be represented by a metric scale. The meaning of "additive structure" was first spelled out in precise form by the mathematician Hölder (1901; see Michell, 1990; Michell & Ernst 1996, 1997). Subsequent representational measurement theorists have worked out the preconditions for the metric representation of many qualitative measurement structures (e.g., extensive structures, difference structures, bisymmetry structures, conjoint structures; see Krantz et al., 1971; Roberts, 1979; Luce & Suppes, 1965). In the studies reported in this article, we used a well-understood kind of measurement structure, called *difference structure* (Krantz et al., 1971; and earlier Alt, 1936; Block & Marschak, 1960; Debreu, 1958; Hölder, 1901; Luce & Suppes, 1965; Suppes & Winet, 1955). Difference structures describe the preconditions of difference measurement, which forms the basis of several indirect scaling methods including ODS (e.g., Junge & Reisenzein, 2015; Maloney & Yang, 2003).

It should be noted, however, that our application of representational measurement theory differs from its traditional use. In particular, following Westermann (1985, 1994) and Maloney and Yang (2003), we take what can be called a deductive rather than the traditional inductive approach to representational measurement, in the following sense: Classical applications of representational measurement theory begin with a set of qualitative (ordinal) empirical relations among objects (e.g., pairwise comparisons of intensity differences). These data are examined to determine whether they fulfill the axioms of a relevant measurement structure (e.g., a difference structure). If they do, the representation theorem of that measurement structure licenses the mapping of this structure into a homomorphic (structure-preserving) numerical representation. The actual measurement process – the assignment of numerical scale values to objects that represent the degrees of the measured attribute – is only performed in a subsequent step, often by applying a scaling algorithm (for illustrations of this approach using difference structures, see e.g., Orth, 1982; Schneider, 1982).

Our approach reverses this order of inquiry. We begin with a set of proposed numerical measurements of the latent variable (in our case, the intensities of an emotion elicited by a set of objects), estimated by a probabilistic scaling model (in our case, ODS of graded difference judgments). The obtained scale values are then used to select cases suitable for testing the axioms of an appropriate measurement structure. In our case, we test whether the participants' comparisons of intensity differences – which we assume to be based on the estimated emotion intensities – are in agreement with the quadruple axiom. If the proposed numerical measurement is indeed metric, then objects with scale values that fulfill the antecedent (if-) condition of the quadruple axiom, should also fulfill its consequent (then-) condition in the comparative judgments. In case of a positive outcome of the axiom test, we assume that the participants' indirect scalings are metric and proceed to test whether their direct emotion scalings of the same stimuli can be considered as linear transformations of their indirect scale values up to measurement error, and hence are metric as well.

The second but related difference of our approach to the classical representational measurement approach is that, different from the classical axiom tests but in agreement with more recent developments (e.g., Maloney & Yang, 2003; Karabatsos, 2005; Regenwetter, Dana, & Davis-Stober, 2011), our axiom test – a modified version of a test proposed

by Maloney and Yang (2003) – is based on an explicit error theory. This error theory is borrowed from the probabilistic scaling model used to scale the qualitative measurement structure (the GPCs), the ODS model (e.g., Boschman, 2001; Junge & Reisenzein, 2015).

The test of the quadruple axiom and the associated test of the metricity of the direct scalings are explained in detail in the Method. For the time being, we would like to point out an important implication of our deductive approach to representational measurement: In contrast to classical representational measurement theory, we do not interpret the "empirical" measurement structures for which the axioms are meant to hold, as directly observable entities. Rather, we interpret them as latent structures. Put differently, we reinterpret the axioms of representational measurement theory as describing, not the actual *performance* of "metric" participants, but their *competence* – their ability to respond correctly to test cases of the axioms. This ability, however, manifests itself only imperfectly in behavior due to random judgment errors.

Overview of the studies

To increase the generalizability of any potential findings, members of three different emotion families were studied: Feelings of sensory pleasantness and disgust evoked by pictures (Study 1), feelings of surprise and amusement induced by solutions to quiz items (Study 2), and feelings of relief and disappointment caused by lottery outcomes (Study 3). Relief and disappointment are widely held to be "cognitive" emotions because they presuppose beliefs and desires about the eliciting events (see Ortony, Clore, & Collins, 1988; Reisenzein, 2009). In contrast, it has been argued that disgust – at least the subform of disgust called "core disgust" by Rozin, Haidt, and McCauley (2008), which is elicited by objects such as spoiled food, body fluids, and maggots – is a "sensory" emotion because it is directly evoked by certain sensory features of the objects (Reisenzein, 2010; see also, Royzman & Sabini, 2001). Finally, surprise, as well as amusement, can be regarded as "fringe" cognitive emotions: Surprise – the emotional reaction to unexpected events – presupposes beliefs, but not desires (Reisenzein et al., 2012), whereas amusement seems to require the appraisal of the eliciting objects as both unexpected and "funny" (e.g., Suls, 1972).

Study 1: Measuring the intensity of sensory pleasantness and disgust

Method

Participants. Participants were 37 students (6 males and 31 females) from different faculties, with a mean age of 22.8 years (SD = 4.9), who responded to a posting on the student web forum of the University. The study was announced as an investigation of subjective reactions to pleasant and unpleasant pictures. Two additional participants had missing pair comparison data due to a technical glitch; these were excluded from the data analyses.

Materials. Twelve pleasant and 12 disgusting pictures intended to elicit different intensities of sensory pleasantness and disgust were used as stimuli. The pleasant pictures showed e.g. a laughing child, a sunflower field and a panda bear; the disgusting pictures showed e.g. a snake pit, a moldy piece of bread, and an overflowing ashtray. The pictures were 300 pixels wide and 360 pixels high and were presented on a 1280 * 1024 computer monitor.

Procedure. For both the pleasant and disgusting pictures, the participants completed three scaling tasks.

Direct scaling task I (Ratings). The first scaling task was the standard emotion rating. Half of the participants rated the pleasant pictures first and the other half the disgusting pictures. In each block, the pictures were separately presented in an individual random order on the computer monitor and the participants rated how pleasant (pleasant pictures) or disgusting (disgust pictures) they found the picture to be. Answers were given by moving an on-screen slider along a 100-point rating scale ranging from "0 = not at all pleasant [disgusting]" to "extremely pleasant [disgusting]". To encourage finely graded ratings, the currently selected scale value was displayed in numerical format immediately above the midpoint of the scale. The rating task was programmed using the experiment generator software WEXTOR (Reips & Neuhaus, 2002).

Indirect scaling task (Graded pair comparisons). Following the ratings, the participants completed a graded pair comparison task (e.g., Bechtel & O'Connor, 1969; Boschman, 2001; Junge & Reisenzein, 2015). GPCs are a variant of the classical pair comparison method (Thurstone, 1927; Torgerson, 1958). They differ from the classical pair comparison task in that participants judge not only which of the two stimuli in a pair dominates (is greater than) the other on the judgment dimension, but also how much the stimuli differ from each other, using an ordered category response scale. The participants judged all possible (12 * 11)/2 = 66 pairs of pleasant and all 66 pairs of disgusting pictures in two separate blocks, whose order was randomized. In each trial, the two compared pictures were presented side by side on the screen. The comparisons within each block were presented in a different random order to each participant; furthermore, 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. For each pair, the participants indicated which picture was more pleasant (disgusting), and how much more. Answers were given on a bipolar 12-category response scale ranging from "The left picture is extremely more pleasant [disgusting] than the right" to "The right picture is extremely more pleasant [disgusting] than the left". Intermediate scale points were labeled "very much more", "much more", "more", "a little more", and "just barely more". An "equally intense" answer was disallowed to encourage participants to discriminate even small intensity differences (see Böckenholt, 2001). We assumed that if the participants could not detect a difference, their responses would be determined by guessing. The response scale was positioned below the pictures in such a way that its left half extended below the left picture and its right half below the right picture. The GPC task was programmed using DMDX (Forster & Forster, 2003).

Direct scaling task II (Rank-rating). In the third part of the experiment, the participants performed another direct scaling task that combines elements of rating and ranking (e.g., Kim & O'Mahony, 1998). They received a set of small (4 cm * 4 cm) color prints of the pictures and were asked to place them on a table beside a 100 cm ruler according to the intensity of pleasantness (disgust) elicited by the pictures. The scaling task was again performed separately for the pleasant and disgusting pictures, with order randomized. Participants were encouraged to rearrange pictures until they were satisfied with the ordering.

It may be noted that there is little emotional adaptation to pleasant and disgusting pictures across repeated representations, at least in the short run (e.g., Codispoti, Ferrari, & Bradley, 2006; Junge & Reisenzein, 2013, 2015). Therefore, we may assume that genuine feelings of pleasantness and disgust were evoked in all parts of the experiment.

Scaling of the graded pair comparisons. To derive emotion intensities from the GPCs, as well as to estimate the judgment error (both parameters are needed for the subsequent axiom tests), we fitted the ODS model (Agresti, 1992; Boschman, 2001; Junge & Reisenzein, 2015) to the GPCs. ODS can be regarded as a descendant of the well-known Thurstonian scaling model (Thurstone, 1927; see Böckenholt, 2006) tailored to graded pair comparison judgments. In agreement with Thurstone (1927), the ODS model assumes that the graded responses are based on differences in latent scale values that are perturbed by random error. ODS is nonmetric because it requires only that the input data (the graded difference judgments) have an ordinal scale level. The statistical model underlying ODS can be described by the following two equations:

$$\Delta_{a,b} = \Psi_b - \Psi_a + \varepsilon, \text{ with } \varepsilon \sim N(0, \sigma^2) \tag{1}$$

$$R_{a,b} = j \text{ if } \theta_{j-1} < \Delta_{a,b} \leq \theta_j \tag{2}$$

$$\text{with } j = 1,\ldots,J \text{ and } -\infty = \theta_0 < \theta_1 < \ldots < \theta_{J-1} < \theta_J = +\infty$$

Ψ_a and $\Psi_b \in \{\Psi_1, \ldots, \Psi_n\}$ are the scale values of the two stimuli a and b compared in a trial of the GPC task (in Study 1, the intensities of pleasantness or disgust evoked by the pictures), and $\Delta_{a,b}$ is the internal decision variable on which the overt response $R_{a,b}$ is based. In addition, the ODS model contains $\theta_1, \ldots, \theta_{J-1}$ thresholds separating the response categories, which, like the scale values, must be estimated. Equation 1 assumes that the participant in a GPC task (implicitly) computes the difference between the scale values of the two presented stimuli, and that the judgment process – including the initial representation of the stimuli plus, in the case of emotional stimuli, the elicitation of feelings – is biased by independent random influences stemming from a normal distribution with constant variance σ^2. Equation 2 implies that, if the judgment error were zero, the decision variable $\Delta_{a,b}$ (which in our case represents the perceived difference between the emotion intensities elicited by stimuli a and b) would be mapped into category j of the response scale consisting of J ordered categories, whenever $\Delta_{a,b}$ lies between the thresholds θ_{j-1} and θ_j that mark the boundaries of j on the latent continuum. However, due to the presence of error, the wrong response category will occasionally be chosen, and this will happen more frequently, the closer the stimuli are on the judgment dimension. The

aim of ODS is to estimate, from the observable responses $R_{a,b}$ (the ordinal difference judgments), the latent scale values of the stimuli assumed to underlie these responses.

As just described, the ODS model is a special version of the ordered (or cumulative) probit model (e.g., McKelvey & Zavoina, 1975; Greene & Hensher, 2010), that can be obtained in a straightforward manner by applying the ordered probit model to GPCs (Agresti, 1992). More restrictive versions of the ODS model in which the threshold parameters are constrained to be symmetric around the middle of the scale have also been proposed (Boschman, 2001).

The ODS model can be estimated using widely available software for maximum likelihood estimation of cumulative link models (e.g., Christensen, 2013). The criterion variable consists of the GPC judgments, the predictors are the latent scale values of the stimuli. To set up the model, one specifies an $n \times p$ design matrix, where n is the number of graded pair comparisons (e.g., 66 in Study 1) and p is the number of the to-be-estimated scale values (e.g., 12 in Study 1). To make the model identifiable, the first column of the design matrix is dropped, implying that the scale value of the first stimulus is fixed at 0. The predictors are dummy-coded in a way that represents the occurrence of the stimuli in the different trials. For example, if stimuli 3 and 9 are compared in a trial, the respective row of the design matrix (with first column dropped) would be (0, 1, 0, 0, 0, 0, 0, -1, 0, 0, 0).

A technical difficulty that can arise when estimating cumulative link models, particularly with sparse data, is the occurrence of complete or quasi-complete separation. Separation is present, roughly speaking, if a predictor or combination of predictors allows the perfect or near-perfect prediction of the response (see e.g., Albert & Anderson, 1984; Allison, 2008; and specifically for the cumulative link model, Agresti, 2010; Kosmidis, 2014). In case of separation, unique maximum likelihood estimates of the coefficients of the responsible predictor variables do not exist. Fortunately, a solution to this problem is available in the form of bias-reducing maximum likelihood estimation (Firth, 1993; see also, Kosmidis & Firth, 2009). In this estimation procedure, a bias-reduced estimator based on adjusted score functions is used in place of the standard maximum likelihood estimator. Originally developed to reduce the bias inherent in maximum likelihood parameter estimates (Firth, 1993), the bias-reduced estimator also provides an effective solution to the separation problem (Heinze & Schemper, 2002). For cumulative link models, the method has been implemented in the R function *bpolr* (Kosmidis, 2014).[3]

Test of the quadruple axiom. Our test of measurement axioms focused on the quadruple axiom, a central axiom of difference measurement structures (e.g., Block & Marschak, 1960; Debreu, 1958; Luce & Suppes, 1965; Suppes & Winet, 1955; see also Orth, 1982; Petrusic, Baranski, & Kennedy, 1998).

Difference structures and the quadruple axiom. Difference structures are appropriate if the qualitative measurement operation used for probing the existence of metric structure consists of the ordinal comparison (symbolized $\succsim$) of differences between pairs of objects (ab; cd) from a set of stimuli A. Hence, $\succsim$ is defined on $A \times A$ and the (potential)

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

difference structure is <A × A, ≿>. Difference comparisons can be directly made by participants (e.g., "Is the difference in pleasantness elicited by stimuli a and b greater or less than the difference in pleasantness elicited by c and d?"; Junge & Reisenzein, 2015; see also Maloney & Yang, 2003; Schneider et al., 1974); but they can also be derived from GPCs (Roberts, 1979; Orth, 1982; see below for more detail). We chose the second option in our studies because we had decided to use GPCs in the indirect scaling task. The advantage of GPCs is that they are much more economical than direct difference comparisons (quadruple judgments), apparently without loss of information (Junge & Reisenzein, 2015).

The axioms of difference structures <A × A, ≿> impose constraints on the relation ≿ which, if met, entail the existence of a metric representation of the difference structure. That is, they entail the existence of a real-valued function Ψ defined on the set A that is unique up to a linear transformation, such that (Krantz et al., 1971):

$$ab \succsim cd \text{ if, and only if, } \Psi(a) - \Psi(b) \geq \Psi(c) - \Psi(d). \quad (3)$$

Several different axiomatizations of difference structures have been proposed (e.g., Block & Marschak, 1960; Debreu, 1958; Krantz et al., 1971; Suppes & Winet, 1955; Luce & Suppes, 1965). The standard axiomatization today is considered to be that proposed by Krantz et al. (1971); however, since the different axiomatizations are logically equivalent, the choice of a particular axiomatization is not crucial.

Two of the axioms of difference structures (Solvability, and the Archimedean condition) are not empirically testable, but are needed to achieve the desired representation and are plausible as idealizing assumptions (see Krantz et al., 1971; Michell, 1990). Of the remaining, testable axioms of difference structures, two are central in the standard axiomatization (Krantz et al., 1971): (1) the *weak ordering axiom*, which requires that ≿ is a weak order (i.e., transitive and connected); and (2) the *axiom of weak monotonicity* or the sextuple condition.[4] However, if the relation ≿ is derived from GPCs, as in our studies, the weak ordering axiom is necessarily fulfilled (Orth, 1982, p. 361).[5] It can be argued, however, that this axiom – which requires that participants are able to consistently order

[4] The axioms of difference structures proposed by Krantz et al. (1971) differ somewhat depending on whether A × A contains positive and negative intervals, only positive intervals, or absolute intervals (Krantz et al., 1971), although the central testable axioms are the same or comparable in the different subforms of difference structures. As in previous studies using difference structures (e.g., Orth, 1982; Petrusic et al., 1998; Schneider, 1980; Schneider, Parker, & Stein, 1974; Westermann, 1985), our difference data (the GPC judgments of intensity differences) are directional (it is meaningful to say that stimulus a elicits more or less of the emotion in question than b), but only one kind of differences were collected (i.e. the participants always judged which stimulus elicited the *more intense* emotion, and how much more intense it was). Therefore, our data can be regarded as a (potential) positive difference structure (e.g., Petrusic et al., 1998; Schneider et al., 1974), or as the positive half of an algebraic difference structure (e.g., Orth, 1982).

[5] The relation ≿ between GPC-derived differences is connected because all possible difference comparisons can be derived from n * (n-1)/2 GPCs. Transitivity of ≿ means that, if ab ≿ cd and cd ≿ ef, then ab ≿ ef. If the difference judgments are derived from GPCs, the judgments of ab, cd, and ef appearing in these three inequalities are identical; therefore, transitivity necessarily holds.

the differences between stimuli – is most likely fulfilled if they are able to consistently order (up to random error) the original stimuli, which in turn is plausible for sensations and emotional feelings, provided that the intensity differences between adjacent stimuli are not too small.[6] In fact, in discussions of the scale level of measurements of the intensity of sensations and emotions, the assumption that these measurements (even rating scales) have at least an ordinal scale level is usually taken for granted (e.g., O'Brien, 1985).

The weak monotonicity axiom is generally considered to be the central testable axiom of difference structures (Krantz et al., 1971). However, following Orth (1982) and Petrusic et al. (1998) we decided – partly to compensate for the non-testability of the weak ordering axiom with our data – to test a stronger axiom, the *quadruple axiom*, which replaces the weak monotonicity axiom in alternative axiomatizations of difference structures (e.g., Block & Marschak, 1960; Debreu, 1958; Luce & Suppes, 1965; Suppes & Winet, 1955). The quadruple axiom implies the weak monotonicity axiom, but not vice versa (see Block & Marschak, 1960; Debreu, 1958; Luce & Suppes, 1965).

As its name suggests, the quadruple axiom applies to sets of four stimuli (ab; cd), consisting of pairs ab and cd. The axiom claims that, for all quadruples of this kind for which $a \prec b$, $c \prec d$, and $b \prec d$[7]

$$\text{If } ab \succsim cd \text{ then } ac \succsim bd. \tag{4}$$

The quadruple axiom describes a necessary condition for the metric representation of difference structures, that is, it is implied by the assumption that a metric representation exists.[8]

Testing the quadruple axiom. To test the quadruple axiom, one selects pairs of quadruples (difference comparisons) (ab; cd) that fulfill the antecedent of the axiom ($ab \succsim cd$), and then checks, for each test case, whether the quadruple appearing in the axiom consequens (ac; bd) is correctly answered ($ac \succsim bd$). However, at this point a problem arises: Because the quadruple axiom (like all measurement axioms) is formulated deterministically, already a single violation of the axiom will disconfirm it. Rejecting the axiom if it is violated in at least a single case would however be acceptable only if the difference comparisons were error-free, because only then can an apparent axiom violation be taken

---

[6] In fact, if a participant is able to order stimuli in terms of increasing emotion intensity, s/he has implicitly already provided a transitive ordering for part of $\succsim$. For example, the ordering $a \succ b \succ c \succ d$ implies $ad \succ ac \succ ab$, $bd \succ bc$, $bd \succ cd$, $ad \succ bc$, and $bd \succ cd$.

[7] The quadruple axiom is formulated by some authors (e.g., Debreu, 1958; Petrusic et al., 1998) as a biconditional, i.e. as $ab \succsim cd$ if, and only if, $ac \succsim bd$. However, this formulation is equivalent to the simple conditional used by us and other authors (e.g., Block & Marschak, 1960; Luce & Suppes, 1965; Orth, 1982), because quantification ranges over all quadruples (x, y, z, u); it therefore also covers the case "if $ac \succsim bd$ then $ab \succsim cd$". The restriction $a \prec b$, $c \prec d$, and $b \prec d$ is not mentioned by some authors, whereas others require $a \prec b \prec c \prec d$ (Petrusic et al., 1998); however, the latter constraint is unnecessarily restrictive.

[8] This can be more formally shown as follows (e.g., Debreu, 1958): If the representation exists, then $ab \succsim cd$ implies $\Psi(a) - \Psi(b) \geq \Psi(c) - \Psi(d)$. Rearranging terms, we obtain $\Psi(a) - \Psi(c) \geq \Psi(b) - \Psi(d)$, which in turn implies $ac \succsim bd$.

at face value. In fact, however, human judgments are always contaminated by some amount of error. Therefore, a procedure is needed to decide whether an observed axiom violation reflects an underlying, systematic violation of the axiom, or is due to random judgment error. In classical applications of representational measurement theory, researchers used a "low error" decision rule; that is, they assumed that an axiom is fulfilled if the number of observed axiom violations remains below some low cutoff value (e. g., 10 % of the test cases; Orth, 1982). Although this method is acceptable for diagnosing axiom adherence (particularly if the number of axiom violations is very low), it is problematic for detecting axiom violators: Subjects who exceed the "low error" criterion may in fact obey the axiom, but make comparatively many (random) performance errors. In the case of the quadruple axiom, such errors are particularly likely to occur if the compared differences are small and therefore difficult to discriminate (Eqs. 1 & 2). To decide whether an observed axiom violation is systematic, one therefore needs to take the person's level of random error into account, by constructing an appropriate statistical test.

The problem of devising such a test has turned out to be difficult. However, during recent years, several solutions have been proposed (e.g., Karabatsos, 2005; Maloney & Yang, 2003; Regenwetter et al., 2011; Tsai & Böckenholt, 2006). In our studies, we used a modified version of an axiom testing procedure proposed by Maloney and Yang (2003; see also Knoblauch & Maloney, 2008, 2012). This is a parametric bootstrap test specifically developed to test the axioms of difference structures scaled by a probabilistic difference scaling model (Maloney and Yang [2003] used Maximum Likelihood Difference Scaling, a scaling model tailored to directly obtaine difference comparisons). Apart from applying the bootstrap test to the quadruple axiom, our adaptation of the Maloney-Yang test differs from the original in four respects: (a) The bootstrap test has been adjusted to take account of the fact that the difference comparisons are analytically derived from GPCs; (b) the scale values of the stimuli and the error variance are estimated from the GPCs using ODS; (c) percentage correct rather than the response likelihood (Maloney & Yang, 2003) is used as the index of axiom adherence; and (d) the cases (quadruples) used for testing the quadruple axiom are selected on the basis of the estimated scale values of the stimuli figuring in the quadruples, rather than the participant's responses to the quadruples.[9]

Of these differences, the first three are essentially technical. The first two result from the necessity to adapt the Maloney-Yang (2003) procedure to our kind of data (GPCs) and the associated scaling model (ODS), and the third is motivated by the desire to use the conventional index of axiom adherence.[10] In contrast, the last difference, which also

[9] That is, Maloney and Yang (2003) use the estimated scale values only to simulate responses to test cases of the axiom at issue, but not to select the test cases. Instead (if we correctly understand their procedure) the test cases are selected, just as in the classical approach, according to the participant's overt responses, i.e. as cases in which the participant affirms the axiom's antecedent. The bootstrapped distribution of the overall likelihood of the ideal observer's responses to the consequens of the axiom in the test cases is then compared to the likelihood of the response vector of the participant. However, as argued in the text, this selection procedure can lead to many wrongly included and wrongly excluded test cases.

[10] For comparison purposes, we also computed a GPC-adapted version of the Maloney-Yang (2003) likelihood statistic for our data, but again with the axiom test cases selected according to the estimated

constitutes (another) significant departure from the classical approach to testing measurement axioms (e.g., Orth, 1982; Schneider et al., 1974), is substantive. The reasoning behind this aspect of our axiom test is as follows: (1) The stimulus properties on which participants base their ordinal judgments (ab $\succsim$ cd) are, ultimately, the emotion intensities elicited by the stimuli (see Eqs. 1 & 2). (2) The best available estimates of these quantities are the scale values Ψ_a, Ψ_b, Ψ_c, Ψ_d estimated by the scaling program, as these are based on the complete set of GPC judgments. (3) Correspondingly, the best available prediction that the antecedent of the quadruple axiom ab $\succsim$ cd is fulfilled by a quadruple of stimuli presented in a trial, i.e. that the participant will perceive or believe that ab is greater than (or equal to) cd, is to assume that this is the case if ab is *in fact* greater than (or equal to) cd; that is, if $\Psi_b - \Psi_a \geq \Psi_d - \Psi_c$. In any case, this is a much better estimate of the participant's beliefs about the relation between the stimuli than his or her overt judgment that ab $\succsim$ cd, as this judgment is usually only made once in a difference judgment experiment and is therefore contaminated by (potentially large) error: The emotion intensities elicited by the stimuli in the trial when the judgment is made can deviate from their modal value, their intensities can be over- or underestimated, errors may occur when computing the differences between them etc.

In detail, our bootstrap test for the quadruple axiom comprises the following steps.

1. The scale values of the stimuli Ψ_1, …, Ψ_n for a given participant (the intensities of emotion elicited by the stimuli) are estimated from the GPCs using ODS.

2. The estimated scale values are used to identify the quadruples of stimuli that fulfill the antecedent of the quadruple axiom on the latent scale level. These are the test cases for the axiom.

Two specifications are made at this point. First, because scale values are estimated to seven decimal places by the ODS program, differences between them are in practice never identical, even if they are subjectively indiscriminable. However, there is much evidence that people are insensitive to small differences (e.g., Böckenholt, 2001; Falmagne, 1985; Luce, 1994). To take account of this fact, we introduced a (conservative) threshold of discriminability of 0.1 units of the ODS scale, corresponding to about 2 % of the typical subject's scale range. That is, pairs of stimuli ab and cd for which $|\Psi_b - \Psi_a| - |\Psi_d - \Psi_c| \leq 0.1$ were regarded as having subjectively equal distances.[11] This concerned between 8.5 % (pleasantness, Study 1) and 14.2 % (relief, Study 3) quadruples. Figure 1 shows the distribution of the absolute scale value differences between stimulus pairs obtained in Study 1 for disgust; the distributions for other emotions were quite similar.

scale values of the stimuli. In all three studies, the response likelihood yielded nearly the same results as the percent correct index (maximally 2 participants were classified differently).

[11] The choice of this threshold was based on a comparison of the effects of different thresholds on the percentage of correct responses and the number of cases for the quadruple axiom. The threshold of 0.1 turned out to be a good compromise. Still lower thresholds rapidly reduced the correct responses rates to the guessing level, indicating that the test cases included too many nondiscriminable stimulus pairs (see Figure 1). Higher thresholds would have been possible but would have reduced the number of available test cases, as well as the sensitivity of the test for small axiom violations.

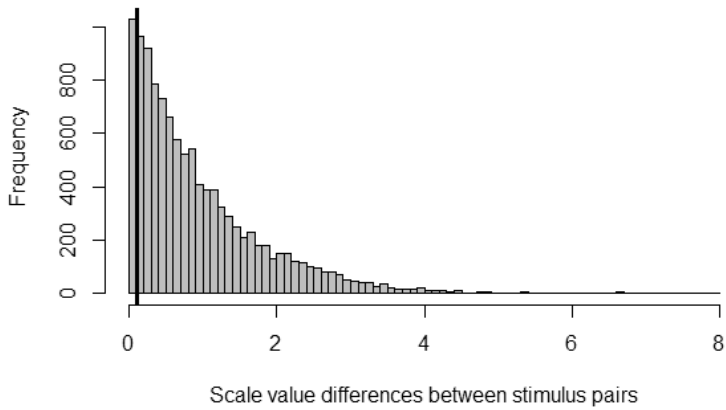

Figure 1:
Frequency distribution of the absolute scale value differences between stimulus pairs obtained in Study 1 for disgust. The vertical black line represents the chosen threshold of 0.1

Second, subjectively discriminable but small differences between intervals (that could be detected in direct difference comparisons) cannot reveal themselves in the GPC task because of the limited resolution of the GPC response scale. To account for this procedural limitation, we (following Orth, 1982) only included the "suprathreshold" quadruples in the axiom test. Hence, we tested a slightly weakened version of the quadruple axiom (if $ab \succ cd$ then $ac \succ bd$).

3. The estimated scale values $\Psi_1, \ldots, \Psi_n$ and the error variance σ^2 of the difference judgments, are plugged into the ODS model (Equations 1 and 2) and a set of GPC responses to the n * (n-1)/2 stimulus pairs is simulated. Next, the simulated GPCs are expanded into quadruple comparisons by assuming (see Roberts, 1979, p. 135; Orth, 1982) that, for all quadruples (ab; cd), $ab \succ cd$ (i.e., the intensity difference between a and b is judged as greater than the difference between c and d) if the rank of GPC(a, b) is greater than the rank of GPC(c, d). Responses are coded as 1 if $ab \succ cd$ and as 0 if $ab \prec cd$; if $ab \approx cd$ (i.e., if the two GPC judgments are equal, which can occur because of the limited resolution of the GPC response scale), they are randomly assigned to the "1" or "0" category.

4. Using the derived quadruple comparisons, simulated responses to the consequens of the quadruple axiom (ac; bd) are created for the test cases of the axiom selected in step 1. For each quadruple, this simulation corresponds to a Bernoulli experiment in which response 1 is generated with a probability π that depends on the scale values of the stimuli and σ^2. Whereas the error variance is assumed to be constant (Eq. 1), the scale values of the stimuli, and hence the response probability π, can differ for each quadruple. The simulated responses reflect the responses of an "ideal observer" (Maloney & Yang, 2003) to the test cases of the quadruple axiom, i.e. a hypothetical twin of the participant who responds to each pair in a quadruple (in the underlying GPC task) according to the ODS model with the participant's scale values and error variance.

5. The responses of the ideal observer to the test cases for the quadruple axiom are summarized in a performance index. We used percent correct (= 100 – percent of axiom violations), the classical index used in tests of measurement axioms.

6. Steps 3-5 are repeated numerous times (we used 10000 replications) and the performance index obtained in each simulation run is accumulated into a bootstrap distribution. This distribution reflects the variability of the responses of the "ideal observer" who responds repeatedly to the axiom test cases.

7. The percentage of correct responses attained by the participant is compared to the bootstrap distribution. If the participant's percentage of correct answers is improbable relative to this distribution ($p < .05$), we conclude that the participant systematically violates the quadruple axiom. Otherwise, we conclude that the null hypothesis – the participant responded in accordance with the quadruple axiom – can be retained.

Note that the described parametric bootstrap test takes account of the fact that the GPC-derived quadruple comparisons are not independent (because the same GPC judgment of a stimulus pair ab is used in every derived quadruple containing this pair). This dependency is taken into account by reusing (in step 4), the same derived response to a quadruple (ac, bd) in all occurrences of this quadruple in the axiom test cases (step 3). Although we did not systematically study the power of the bootstrap test, the fact that several hundred test cases for the quadruple axiom are obtained even with moderate numbers of stimuli (see studies 1 and 2) suggests that its power is high and hence, that small deviations from metricity can be detected.

Testing the metricity of the direct scalings. If one accepts that the ODS scale values of participants who pass the quadruple test are metric (interval) scales, it becomes possible to test whether their direct scalings (the rating, rank-ranking, and the combined scale formed by taking the mean of the two judgments) are metric as well. The logic of this test, which was inspired by a related (but nonstatistical) test proposed by Orth (1982), is as follows: If the emotion intensities estimated via ODS are interval-scaled, then any other error-free interval-scale measurement of the emotion intensities evoked by these stimuli is a linear transformation of the ODS values and hence its linear (Pearson) correlation to the ODS scalings is 1. The direct scalings, of course, are not error-free; therefore, their correlation to the ODS scalings could not be perfect even if they are reports of the same latent emotion intensities. Hence, the question is: Do the direct scalings correlate highly enough with the indirect scalings to be regarded as error-perturbed realizations of the ODS scale values?

To answer this question, we constructed another bootstrap test.

1. For each participant (as well as each emotion and each kind of direct scaling), 10000 sets of direct scalings of the emotional stimuli are generated from the ODS scale values, by perturbing them with error corresponding to that of the direct scaling. This simulates a hypothetical twin of the participant who operates with the ODS scale values when making direct scalings, but is subject to random performance errors corresponding to the participant's error level. To be able to run this simulation, we assumed, following Thurstone (1927), that the perceived emotion intensity in each trial of the simulated direct

Table 1:
Reliabilities of the Indirect and Direct Scalings[a]

	M[b]	*Min*	*Max*
Study 1 (*n* = 37)			
Indirect scalings			
Pleasantness	.96	.81	.98
Disgust	.97	.84	.99
Direct scalings (single scales)			
Pleasantness	.77	.32	.94
Disgust	.81	.01	.99
Direct scalings (combined scale)[c]			
Pleasantness	.82	.29	.96
Disgust	.87	.01	.99
Study 2 (*n* = 34)			
Indirect scalings			
Surprise	.97	.84	.99
Amusement	.98	.87	.99
Direct scalings (single scale)[d]			
Surprise	.85	.67	.94
Amusement	.87	.52	.98
Study 3 (*n* = 39)			
Indirect scalings			
Relief	.94	.41	.99
Disappointment	.94	.76	.98
Direct scalings (combined scale)[b]			
Relief	.77	.01	.99
Disappointment	.80	.01	.99

Notes.
a. Reliabilities were set to .01 when the correlations or Cronbach's α were negative.
b. Mean reliabilities computed using the Fisher Z-transformation
c. Cronbach's α
d. For Study 2, the reliability estimate used was the correlation of the ratings to the ODS scalings.

scaling task is drawn from an independent normal distribution with a mean corresponding to the ODS scale value of the judged stimulus, and constant variance σ^2_{er}. This assumption is compatible with the ODS model and has in fact been used to motivate that model (Boschman, 2001). The error variance σ^2_{er} was estimated from the reliabilities of the direct scalings as $\sigma^2_{er} = \sigma^2_{ods} (1- r_{xx})$ (see e.g., Guilford, 1954), where σ^2_{ods} is the variance of the ODS scale values, and r_{xx} is the reliability of the direct scalings reported later in Table 1 (i.e., the correlation between the ratings and rank-ratings for the single scales, and Cronbach's α for the combined scale).[12]

[12] These reliability estimates are appropriate if the ratings have an interval scale level, but can be biased if the ratings are only ordinal. While this bias can be reduced using polychoric correlations (e.g., O'Brien &

2. Each simulated set of direct scalings is linearly correlated with the ODS scale values (which are considered to be error-free in this simulation[13]) and the resulting correlations (the standardized slopes of the linear regression) are accumulated into a bootstrap distribution. This distribution reflects the expected variability of the correlation between the direct and ODS scalings for a participant who operates with the ODS scale values in the direct scaling task with an error corresponding to his direct scaling error.

3. The actually obtained correlation between the direct and ODS scalings is compared to the bootstrap distribution of this correlation. If the empirical correlation cuts off less than .05 of the bootstrap distribution, the null hypothesis that the direct scalings are linearly transformed, error-perturbed manifestations of the latent ODS scale values, and hence are also interval measurements, is rejected.

Note that the proposed test of the metricity of the direct scalings is (a) restricted to participants who passed the preceding test of the quadruple axiom, because only these can be taken to have metric ODS scale values and (b) in contrast to the quadruple axiom test, is based on only very few cases (the number of stimuli, 9-15 in our studies) and for this reason can be expected to have comparatively low power.

Results

Intensity range of pleasantness and disgust. The direct scalings of the pictures suggested that the emotions evoked by them spanned a reasonable range of intensity, although the range of disgust was greater than that of pleasantness: On the scale formed by averaging the ratings and rank-ratings, the means of the different pictures ranged from 29.8 to 79.9 (M = 55.9, SD = 28.7) for pleasantness and from 56.3 to 84.7 (M = 69.2, SD = 22.8) for disgust. Furthermore, there were large interindividual differences in how pleasant and disgusting the pictures were rated, supporting the proposal to use individual-level analysis in emotion research when possible (Junge & Reisenzein, 2013).

Reliabilities of the indirect and direct scalings. The reliabilities of the indirect and direct scalings were separately estimated for each participant. The reliabilities of the indirect scalings (ODS) were estimated using a parametric bootstrap procedure: The ODS model with the estimated scale values and error variance σ^2 was used to generate

---

Homer, 1987; Gadermann, Guhn, & Zumbo, 2012), this option was not available to us because of the low number of data points per subject (12). However, simulation studies (e.g., O'Brien, 1982; Bollen & Barb, 1981; see also Gadermann et al., 2012) suggest that the bias (typically an underestimation) of the correlation between latent metric variables estimated from ordinal measurements is small as long as five or more response categories are used (we used 101 in studies 1 and 3, and 11 in Study 2) and the distributions of the latent variables are not highly skewed. Based on these findings, we assumed that the estimated reliabilities would be close to the true ones even if the ratings are ordinal, and that the remaining bias would consist of a slight underestimation of the true reliabilities. In this case, the random error in the ratings is slightly overestimated and the metricity test is as a consequence slightly too liberal, i.e. it will diagnose too few participants as "nonmetric in the ratings".

[13] Although the ODS scalings are not perfectly reliable, they come close. More importantly, the quadruple axiom was tested for the estimated ODS scale values, not for a noise-perturbed version of them.

100 bootstrap samples of the 66 GPCs, these were subjected to ODS, and the median intercorrelation of the resulting scale values was used as the reliability estimate. As can be seen from Table 1, the reliabilities of the ODS scalings were high for most participants (M = .96 for pleasantness and .97 for disgust, computed using the Fisher Z-transformation).

As the estimate of the reliability of the two direct scalings (the rating and rank-rating), we used their correlation (see Footnote 10), based on the assumption that the two were sufficiently similar to be regarded as parallel measures. The obtained average reliabilities (M = .77 for pleasantness and M = .81 for disgust) were similar to the retest reliabilities of direct ratings of disgust obtained in a previous study (.76; Junge & Reisenzein, 2013, Study 2). We also created a combined scale by taking the mean of the rating and rank-rating. The reliability of this scale (Cronbach's alpha) was M = .82 for pleasantness and M = .87 for disgust. As expected, and in line with Junge and Reisenzein (2013, 2015), the reliabilities of the indirect scalings (ODS) were much higher than those of the direct scalings.

Test of the quadruple axiom for the indirect scalings. Using the threshold of 0.1 ODS scale units to select subjectively discriminable stimulus pairs (see Method), we obtained between 178 and 786 test cases (quadruples) for the quadruple axiom per participant for the pleasantness scalings (M = 427.6, SD = 143.3), and between 122 and 608 quadruples for the disgust scalings (M = 292.1, SD = 127.6). Using these axiom test cases, the bootstrap test described in the Method was carried out separately for each participant. The conventional significance level of α = .05 was used to decide if the participant failed the test.

Figure 2 shows the bootstrap distribution of the percentage of correct responses to the test cases of the quadruple axiom test for two participants, one who passed the test, and the other who failed the test. The results for the complete sample are summarized in Table 2. 36 of the 37 participants (97 %) passed the test of the quadruple axiom for pleasantness and 33 for disgust. However, three of the latter participants had relative frequencies of correct responses below 50 %. These participants were reclassified as not conforming to the axiom, leaving 30 of 37 (81 %) who passed the axiom test for disgust (Table 2). Furthermore, nearly all participants (29 of 30) who passed the quadruple axiom test for disgust also passed the test for pleasantness.

The average percentage of correct responses given to the test cases of the quadruple axiom was M = 68 % (SD = 10 %) for pleasantness and M = 62 % (SD = 12 %) for disgust. Using traditional cutoff values to decide on axiom adherence (e.g., 90 % correct; Orth, 1982), one would have to conclude that the majority of the participants did not conform to the quadruple axiom. However, the bootstrap test reveals that this conclusion is in most cases unwarranted, as the observed axiom violations can be explained by performance errors.

It is also instructive to look at the percentage of correct responses given to the antecedent of the quadruple axiom (rather than its consequens, as done so far). Because the same set of quadruples appears in the antecedent and consequens of the quadruple axiom (see Footnote 5), this percentage is identical to the percentage of correct responses to the

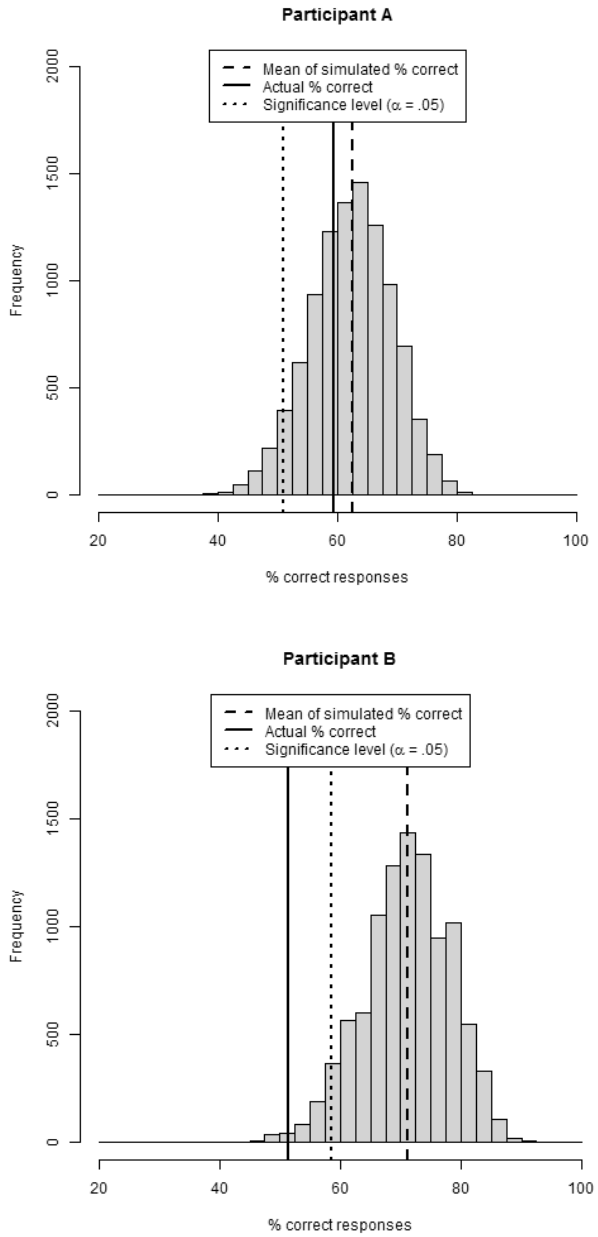

Figure 2:
Bootstrap distribution of the percentage of correct responses in the quadruple axiom test for two participants. Participant A passed the bootstrap test, participant B failed the test

Table 2:
Results of the Quadruple Test for the Indirect Scalings

	Number of participants who passed the test	Percent correct, actual participant[a]	Percent correct, ideal observer[b]
Study 1 (*n* = 37)			
Pleasantness	36 (97 %)	.68 (.10)	.68 (.07)
Disgust	30 (81 %)	.62 (.12)	.65 (.07)
Pleasantness & Disgust	29 (78 %)	.65 (.11)	.66 (.07)
Study 2 (*n* = 34)			
Surprise	30 (88 %)	.65 (.10)	.68 (.07)
Amusement	24 (71 %)	.61 (.11)	.65 (.07)
Surprise & Amusement	21 (62 %)	.63 (.11)	.67 (.07)
Study 3 (*n* = 39)			
Relief	29 (74 %)	.59 (.14)	.62 (.09)
Disappointment	38 (97 %)	.67 (.10)	.67 (.07)
Relief & Disappointment	29 (74 %)	.63 (.13)	.64 (.08)

Notes.
a. Mean percentage of correct responses for participants who passed the test. The standard deviation is given in parentheses.
b. Mean percentage of correct responses obtained in the ideal observer simulation (10000 simulations/participant). The standard deviation is given in parentheses.

consequens, i.e. M = 68 % for pleasantness and M = 62 % for disgust. What this suggests, however, is that the traditional way of testing the quadruple axiom – by selecting test cases according to the participant's responses to the axiom's antecedent – would have resulted in numerous wrongly selected items. In accord with this reasoning, the percentage of correct responses was markedly lower if the test cases were selected on the basis of the participant's responses to the antecedent, M = 59 % (SD = 14 %) for pleasantness, and M = 52 % (SD = 16 %) for disgust. In fact, in the case of disgust, performance dropped essentially to chance level.

Test of metricity of the direct scalings. As mentioned in the Method, the test of metricity of the direct scalings is restricted to participants who passed the test of the quadruple axiom, because only these can be taken to have metric ODS scale values. Hence, this test is unsuited to identify participants able to give metric ratings but unable to provide metric ODS scalings. However, this restriction is not very serious in our case because the maximum possible number of these participants was small (1 for pleasantness and 7 for disgust). Furthermore, given the higher difficulty of the direct ratings, one can argue that a person who fails the quadruple test is most likely unable to give metric ratings. The most optimistic assumption is probably that the proportion of participants able to make

metric ratings is the same among those who fail and those who pass the quadruple axiom test.

The results of the metricity test for the direct scalings are summarized in Table 3. As reported, 36 participants passed the test of the quadruple axiom for the ODS pleasantness scalings. The mean correlations between these participants' ODS scalings and their ratings, rank-ratings and the combined scale were .80 (*range* = .43 - .92), .91 (*range* = .46 - .99) and .92 (*range* = .54 - .99). 16 (44 %) of these participants also passed the metricity test for the direct pleasantness ratings, 27 (75 %) for the rank-ratings, and 25 (69 %) for the combined scale.

30 participants had passed the quadruple axiom test for the indirect disgust scalings. The mean correlations between these participants' ODS scalings and their ratings, rank-ratings and the combined scale were .81 (*range* = .18 - .96), .89 (*range* = .20 - .98) and .90 (*range* = .46 - .98). Seven (23 %) of these participants also passed the metricity test for the direct disgust ratings, 17 (57 %) for the rank-ratings, and 14 (47 %) for the combined scale.

Table 3:
Results of the Metricity Test for the Direct Scalings

	Number of participants who passed the quadruple test	Ratings (single scale)[a]	Rank-ratings (single scale)[a]	Combined scale[a]
Study 1 (*n* = 37)				
Pleasantness	36	16 (44 %)	27 (75 %)	25 (69 %)
Disgust	30	7 (23 %)	17 (57 %)	14 (47 %)
Pleasantness & Disgust	29	2 (07 %)	11 (38 %)	7 (24 %)
Study 2 (*n* = 34)				
Surprise	30	6 (20 %)	-	-
Amusement	24	11 (46 %)	-	-
Surprise & Amusement	21	2 (10 %)	-	-
Study 3 (*n* = 39)				
Relief	29	-	-	11 (38 %)
Disappointment	38	-	-	17 (45 %)
Relief & Disappointment	29	-	-	9 (31 %)

Notes.
a. Percentages are the proportion of participants with metric ratings of those who passed the quadruple test.

Discussion

Replicating previous findings (Junge & Reisenzein, 2013, 2015), the indirect scaling method yielded more reliable measurements of emotion intensity than the direct scalings. The results of the quadruple axiom test suggest that the indirect scalings of most participants (97 % for pleasantness and 81 % for disgust) can be regarded as metric (interval) measurements of emotion intensity. A subset of the participants with metric ODS scale values also passed the metricity test for the direct scalings. The number of these participants was markedly higher for the rank-ratings than for the simple ratings: On average, across pleasantness and disgust, 66 % of the participants passed the metricity test for the rank-ratings (59 % if one assumes that those who failed the quadruple test are unable to give metric ratings), but only 33 % (31 %) for the simple ratings typically used in emotion research. The better performance of the rank-rating method could have been due to two factors: (a) its combination of elements of rating and ranking facilitates metric intensity judgments; (b) it profited from having been applied last, after the direct ratings and GPCs. It is not possible to disentangle these factors in the present study.

Given the probable low power of the metricity test for the direct scalings, we refrain from concluding that the participants who passed this test are indeed able to give metric intensity ratings. However, the test allows us to conclude that about 70 % of our participants were unable to provide metric ratings, and about 35 % were unable to provide metric rank-ratings.

Study 2: Measuring the intensity of surprise and amusement

Method

Participants. Participants were 34 students (33 female and one male), with a mean age of 22.5 years (SD = 4.5). The study was announced as an investigation of the subjective experiences associated with answering quiz items.

Materials. The quiz items were taken from a pool of 120 items that had been previously compiled from quiz books, almanacs, the internet, and other sources with the aim of obtaining quiz items that elicit different intensities of surprise and amusement. For the present study, 15 surprise-eliciting and 15 amusing items that spanned the intensity range from low to high were selected. The items were presented using DMDX (Forster & Forster, 2003). The surprise items were formulated as questions together with the correct (according to our sources) answers, such as "How many trees have to be cut down for a Sunday New York Times? 63.000". The amusement items were formulated as statements, e.g. "Graham Bell, the inventor of the telephone, could never phone his wife or mother, because both were deaf".

Procedure. The procedure was similar to that of Study 1. Each participant completed four scaling tasks in this order: surprise ratings, surprise GPCs, amusement ratings, and amusement GPCs.

Direct scaling task. In the rating task, the 15 surprise and 15 amusement items were presented in random order to the participants, who rated the intensity of surprise (amusement) elicited by the items on 11-point rating scales ranging from "0 = not at all surprised [amused]" to "10 = extremely surprised [amused]". Responses were entered by pressing labeled keys (0-10) on the keyboard.

Indirect scaling task. The (15 * 14)/2 = 105 possible pairs of the surprise and amusement items were presented to each participant in an individual random order. In each trial, two text boxes displaying the items were shown side by side on the screen. The location of the items (left or right) was counterbalanced across trials. The participants were asked to indicate which of the two items was more surprising (amusing), and how much more surprising (amusing) it was. Answers were given on a 12-point bipolar category rating scale placed below the text boxes, ranging from "the left item is extremely more surprising [amusing]" to "the right item is extremely more surprising [amusing]". Intermediate scale points were labeled "very much more", "much more", "more", "a little more", and "just barely more". An "equally intense" answer was disallowed for reasons given in Study 1.

Results

Intensity range of surprise and amusement. The mean ratings of surprise intensity for the 15 surprise items ranged from a low of 0.12 to a high of 7.91 (M = 4.83; SD = 3.41). The mean ratings of amusement intensity for the 15 amusement items ranged from 0.38 to 7.03 (M = 3.82, SD = 3.14).

Reliabilities of the indirect and direct scalings. The reliabilities of the ODS scalings were estimated as in Study 1 and were found to be similarly high as in Study 1 (see Table 1). Different from Study 1, the reliabilities of the direct scalings could not be estimated via re-test correlation because the participants made the ratings only once. Still, an estimate of the reliability of the ratings is available in the form of their correlations to the ODS scalings, which were on average M = .85 for surprise and .87 for amusement (Table 1). These correlations are similar to the corresponding correlations obtained for disgust and pleasantness in Study 1 but somewhat higher than the single-scale re-test correlations obtained in that study, suggesting that they slightly overestimate the reliability of the ratings.

Test of the quadruple axiom for the indirect scalings. Using the threshold of 0.1 units on the ODS scale to identify subjectively discriminable intensity intervals, we obtained between 344 and 1868 cases suited for testing the quadruple axiom for the ODS surprise scalings (M = 1022, SD = 408.7), and between 394 and 1620 quadruples for the ODS amusement scalings (M = 857.4, SD = 324.5). As in Study 1, participants were classified as conforming to the quadruple axiom if they passed the bootstrap test and the frequency of correct responses was above 50 %. Again the significance level of α = .05 was adopted and 10000 bootstrap simulations were run. The results are shown in Table 2.

30 (88 %) of the 34 participants passed the test of the quadruple axiom for surprise, 24 (71 %) for amusement, and 21 (62 %) for both emotions. On average, these participants

responded correctly to M = 65 % (SD = 10 %) of the test quadruples for surprise and to M = 61 % (SD = 11 %) for amusement. As in Study 1, the percentage of correct responses was considerably lower, M = 56 % (SD = 13 %) for surprise and M = 52 % (SD = 14 %) for amusement, if the test cases were selected on the basis of the participant's responses to the antecedent of the quadruple axiom; in fact, in this case performance approached chance for both emotions.

Test of metricity of the direct scalings. The metricity of the direct scalings of surprise and amusement was again examined using the bootstrap test described in Study 1. Because individual reliability estimates for the ratings were not available in Study 2, we assumed that these reliabilities were (a) identical for all participants and (b) equal to the lower bound of the reliability suggested by the average correlation of the ratings to the ODS scalings (Table 1), i.e. r_{xx} = .85 for surprise and .87 for amusement. As in Study 1, the metricity test was performed only for participants who had passed the preceding test of the quadruple axiom (30 for surprise and 24 for amusement). The mean correlation of these participants' ODS scale values to their ratings (computed using the Fisher Z-transformation) was M = .86 (*range* = .67 - .94) for surprise and M = .88 (*range* = .52 - .98) for amusement. As shown in Table 3, 6 of these participants (20 %) passed the bootstrap test for the surprise ratings and 11 (46 %) for the amusement ratings. Two participants (10 %) passed the tests for both ratings.

Discussion

Pooled across emotions, fewer participants (80 %) passed the test of the quadruple axiom in Study 2 than in Study 1 (90 %). One might be tempted to attribute this difference to a greater power of the bootstrap test in Study 2, as the number of quadruples available for testing the axiom was on average 2.6 times larger than in Study 1. However, the lower average frequency of axiom-conforming participants was mainly due to amusement (71 %); the results for surprise (88 %) were even slightly better than those for disgust obtained in Study 1 (81 %).

From 20 % (surprise) to 46 % (amusement) of the participants who passed the test of the quadruple axiom also passed the metricity test for the ratings. These numbers drop to 18 % and 32 %, respectively, if one assumes that participants who failed the quadruple axiom test are unable to give metric ratings. Given the probable low power of the metricity test for the ratings, we again refrain from concluding that the participants who passed the test were able to give metric ratings. What we can conclude, however, is that from 54 % (amusement) to 80 % (surprise) of the participants were unable to provide metric ratings.

Study 3: Measuring the intensity of relief and disappointment

In Study 3, we studied the metricity of indirect and direct scalings of relief and disappointment caused by lottery outcomes. The data used in this analysis were collected by Junge and Reisenzein (2013, Study 1). Details of the method and design of the study are reported there. Here, we only summarize the main points.

Method

Participants. Participants were 39 students (6 males and 33 females) with a mean age of 22.3 years (*SD* = 4.8). The study was described as dealing with the subjective experience of gambling.

Design and materials. Relief and disappointment were induced using a lottery paradigm similar to that of Mellers, Schwartz, Ho, and Ritov (1997). Participants were presented with a set of wheels of fortune programmed with FLASH. In each trial, they could win or lose a small amount of money (-2, -.50, -.10, .10, .50, or 2€) indicated by a coin symbol at the center of the wheel, ostensibly with a probability (.05, .50 or .75) corresponding to the size of the gain sector (green) or the loss sector (red) of the wheel. Actually, the outcomes were determined by the experimental design. Interest focused on 9 potential gain lotteries (3 probabilities × 3 possible gains) and nine potential loss lotteries (3 probabilities × 3 possible losses) with zero outcomes, as these are occasions where relief (avoiding a possible loss) and disappointment (missing a possible gain) were primarily expected to occur. These lotteries were presented twice to increase the reliability of the direct ratings. To keep up the appearance of a real lottery, we also included 15 trials with nonzero outcomes.

Procedure. The procedure was similar to that used in studies 1 and 2.

Direct scaling task. The participants first played the 51 lotteries presented in a random order. In each trial, they were first asked to consider their chances of winning or losing and then to set the wheel in motion by pressing the "start" button. The wheel spun for about seven seconds before stopping in the gain, loss, or null (zero outcome) sector, as determined by the experimental design. Subsequently, the participants indicated how disappointed and relieved they felt about the outcome by moving sliders along 0-100 rating scales ranging from "0 = not at all disappointed [relieved]" to "100 = extremely disappointed [relieved]".

Indirect scaling task. Following the ratings, the participants were presented with all possible pairings of the zero-outcome lotteries from the first part of the experiment, separately for avoided losses (relief) and missed gains (disappointment). The 36 lottery pairs of each type where presented in an individual random order. Half of the participants judged the relief lotteries first and the other half, the disappointment lotteries. In each trial, the two money wheels were shown side by side on the screen, with location counterbalanced. Participants were asked to imagine that they again participated in the lottery

for real. They were asked to spin the left money wheel, wait until it stopped, and then do the same for the right wheel. Subsequently, they indicated which of the two outcomes would have caused stronger relief (disappointment) if they had played for real money, and how much more. Answers were given on a 12-point bipolar category rating scale ranging from "the left outcome is extremely more relieving [disappointing]" to "the right outcome is extremely more relieving [disappointing]". Intermediate scale points were labeled as in studies 1 and 2.

Results

Intensity range of relief and disappointment. To increase the reliability of the direct scaling, the two ratings of each lottery made in the first part of the study were averaged. On this combined scale, the mean intensity of relief evoked by the different zero-outcome lotteries ranged from 20.14 to 68.77 (M = 43.59, SD = 27.57) and the mean disappointment ratings ranged from 16.31 to 60.95 (M = 35.15, SD = 26.59).

Reliabilities of the indirect and direct scalings. As can be seen from Table 1, both the reliability of the ODS scalings and the reliability of the combined rating scale (Cronbach's α) were on average lower than in studies 1 and 2. This may have been due to the smaller number of stimuli scaled in this study (9 for disappointment and 9 for relief) and the ensuing greater instability of the reliability estimates. Again, the reliabilities of the direct ratings were much lower than those of the indirect scalings. Note that the reliabilities of the ratings are also lower than those reported in Junge and Reisenzein (2013) because we only included the ratings of nonzero outcomes into the present reliability computations.

Test of the quadruple axiom for the indirect scalings. Because of the complexity of the lottery comparison task, only 9 stimuli per emotion were used in Study 3. As a consequence, the number of quadruples available for testing the quadruple axiom was much smaller than in the first two studies: For relief, we obtained between 20 and 150 test cases per participant (M = 65.9, SD = 31.7) and for disappointment, between 28 and 142 cases (M = 77.3, SD = 31.1). These are only 20 % of the test cases available in Study 1 and 8 % of those available in Study 2. The same difference discrimination threshold as in Studies 1 and 2 (0.1 units of the ODS scale) and the same significance level (α = .05) was used. The results are shown in Table 2.

All 39 participants passed the test of the quadruple axiom for relief, but 10 had < 50 % correct responses and were therefore reclassified as having failed the test, leaving 29 (74 %) who passed the test. 38 of the 39 participants (97 %) passed the quadruple test for disappointment, and 29 (74 %) for both relief and disappointment. The relative frequency of correct responses of these participants was on average M = 59 % (SD = 14 %) for relief and M = 67 % (SD = 10 %) for disappointment. Similar to studies 1 and 2, percent correct dropped nearly to chance level if the test cases were selected according to the participant's responses to the antecedent of the quadruple axiom.

Test of metricity of the direct scalings. For relief, 11 (38 %) of the 29 participants with metric ODS scale values also passed the metricity test for the direct scaling. These par-

ticipants had a mean correlation of M = .78 (*range* = -.18 - .96) between their ODS scale values and ratings. For disappointment, 17 (45 %) of the 38 participants with metric ODS scale values passed the metricity test for the direct scaling. The mean correlation of their ratings to the ODS scaling was M = .80 (*range* = -.36 - .96). Seven participants passed the metricity tests for both the relief and disappointment rating.

Discussion

Like studies 1 and 2, Study 3 found that graded pair comparisons yielded metric ODS scale values for most participants. In contrast, from 55 % (disappointment) to 62 % (relief) failed the test of metricity of the direct ratings.

General discussion

Three studies investigated the scale level of indirect and direct scalings of the intensity of sensory pleasantness and disgust (Study 1), surprise and amusement (Study 2) and relief and disappointment (Study 3). In each study, we first examined the metricity of the indirect (ODS) scalings, by testing whether the participants' difference judgments, on which the scalings were based, fulfilled the quadruple axiom, a central axiom of difference measurement. For participants who passed this test, we proceeded to test the metricity of their direct scalings, by testing whether they correlated linearly with their indirect scalings up to measurement error.

The main results of the three studies were largely consistent and can be summarized in two points: (1) The indirect (ODS) scalings of emotion intensity yielded metric (interval) scales for most participants for all six emotions studied, ranging from 71 % (amusement, Study 2) to 97 % (pleasantness in Study 1 and disappointment in Study 3) (Table 2). (2) In contrast, again for all six emotions studied, the direct ratings of emotion intensity were found to be nonmetric for the majority of the participants (Table 3). On average (across studies and emotions) 64 % of the participants failed the metricity test (69 % if one assumes that those who failed the quadruple axiom test are unable to give metric ratings).

Before we discuss the implications of these findings, we need to address two issues: Possible biases caused by multiple tests, and the power of the bootstrap tests.

Control of the error rate

Because we conducted many significance tests (up to two for each participant) in each study, some of the apparently detected axiom violations could have been due to chance. To estimate the extent of this bias, we computed – separately for each study and emotion (e.g. pleasantness in Study 1) – p-values corrected for the number of tests, using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). If the corrected p-values are used to decide on axiom adherence, the average percentage of participants with (possibly) metric scale values increases from 85 % to 87 % for the indirect scalings, and from

36 % to 42 % for the direct scalings. Hence, controlling the error rate leaves the two main conclusions reached above essentially untouched. Note also that wrongly diagnosing nonmetric participants as metric is less problematic than the converse error (Westermann, 1985).

Power of the bootstrap tests

The power of the bootstrap test proposed by Maloney and Yang (2003; Knoblauch and Maloney, 2008), which we adapted to test the quadruple axiom for GPCs-based data, does not seem to have been systematically investigated. The main problem that arises in this context is to formulate an appropriate alternative to the null hypothesis, that the participant obeys the quadruple axiom. The most general alternative hypothesis is that the axiom is systematically violated; however, because the axiom can be violated in many different ways, corresponding to different possible systematic distortions of the metric latent variable (see e.g., O' Brien, 1982), this alternative is too unspecific to serve as the basis of a power calculation. What could be done, however, is to determine the power of the quadruple test for *specific kinds* of nonmetric distortions, specifically those that seem plausible or at least conceivable from a psychological perspective. To achieve this, one could create appropriately distorted metric scale values and then determine, via simulation, how sensitive the bootstrap test is to these specific deviations from metricity. Pending the results of these simulation studies, one can argue, as we did, that because of the large number of test cases used in the quadruple test in studies 1 (on average 350) and 2 (about 900), its power to detect at least moderate deviations from metricity in these studies was probably high. In fact, the finding that similar results were obtained in Study 3 with on average just 70 test cases suggests that the power of the quadruple test may be adequate even with a much smaller number of test cases.

Regarding the metricity test for the ratings, we argued on the same basis (number of test cases) that the power of this test is probably low, implying that too few participants are diagnosed as nonmetric by this test. Nevertheless, the finding that a subset of the participants came at least somewhat close to metric ratings is reassuring. Perhaps this finding can be taken to suggest an interindividual-differences specification of the hypothesis (e.g., Krantz et al., 1971) that ratings are in between the ordinal and interval scale level (at least for the case of emotion measurement): A subset of the participants seem to be able to give metric ratings, whereas the rest is not.

In sum, if one accepts the premises of our metricity tests, the lead question of this article – do we have metric scales for emotion measurement – can be answered affirmatively for most participants, provided that an indirect measurement method (ODS scalings of GPCs) is used. In contrast, for direct intensity ratings, the lead question has a negative answer for the majority of the participants, although the finding that a minority passed the metricity test remains noteworthy.

Implications for emotion research

A direct implication of the finding that indirect scalings of emotion intensity have not only a high reliability but also seem to attain a metric scale level is that it is possible to empirically test quantitative emotion theories using indirect scalings of emotion intensity. This conclusion is supported by the previous finding that, compared to direct ratings, indirect scalings of the intensity of relief and disappointment, as well as disgust, yielded substantially improved fits to quantitative models of these emotions (Junge & Reisenzein, 2013, Study 2). The present findings suggest that this improvement was partly due to the attainment of a metric (or close to metric) scale level, in addition to the reduction of random error. Because of their higher scale level and greater precision, indirect emotion intensity measurements invite the testing of emotion theories on the individual level (Junge & Reisenzein, 2013). Given that most psychological theories are formulated on the individual level, this is where they should ideally be tested.

The improved precision and metric scale level of indirect measurements of emotion intensity also recommend these measurement methods for investigating other questions of emotion psychology where increased measurement precision is crucial. For example, indirect scalings could provide an improved methodology for answering the contested question of the relation of emotional experiences to physiological (e.g., Mauss & Robinson, 2009) and expressive reactions (Reisenzein et al., 2013). A frequently proposed explanation for the, typically weak, correlations that have been obtained in this research attributes them to the lack of precision and other biases of measurements of emotional experience (e.g., Rosenberg & Ekman, 1994). This critique, however, targets the commonly used, direct ratings of emotional experience. Indirect scaling methods avoid at least part of these criticisms. For this reason, indirect scalings of emotion intensity can also be recommended for the investigation of correlations between subjective emotional experiences and brain states (e.g., Wager et al., 2013).

The test of the quadruple axiom used in our studies is easy to apply and requires no data apart from the GPCs needed for the indirect scalings. However, given our finding that 80-90 % of the participants pass the quadruple axiom test, it may not be necessary to apply the test to every new case of indirect emotion measurement, particularly because even participants who fail the quadruple test may approximate a metric scale level to a fair degree. This consideration suggests that, in addition to the statistical tests of metricity foregrounded in our article, a quantitative fit index that expresses a participant's "closeness to the metric scale level", or several indices expressing the amount of different possible distortions of metricity, would be useful.

Extensions

Junge and Reisenzein (2015) found that ODS scalings of GPCs of emotional stimuli yield similar scale values as the MLDS scaling of directly collected quadruple judgments (Maloney & Yang, 2003). They also found that the scaling of GPC-derived quadruple comparisons with MLDS yielded nearly identical scale values as the, theoretically more

appropriate, ODS scaling of the GPCs. In addition, Junge and Reisenzein (2013) found that scaling GPCs using an additive functional measurement model (AFM; Anderson, 1970) yielded similar scale values to those obtained with MLDS. Hence, different kinds of difference measurement methods yield similar results (scale values). Given these findings, it seems likely that the emotion intensities estimated with the other difference scaling methods mentioned will also pass the test of the quadruple axiom.

Both the indirect scaling and axiom testing method used in this article can be extended to other emotions and other emotion components, as well as from self-reports of emotion experiencers to emotion judgments of observers (Reisenzein et al., 2014). Beyond that, these methods can in principle be extended to the measurement of presumed psychological magnitudes beyond emotions, such as preferences, attitudes, and personality dimensions. Furthermore, the proposed test of metricity of direct scalings can be extended to other kinds of scalings, both direct and indirect. In this way, diverse currently used scaling methods (see e.g., Hein, Jaeger, Carr, & Delahunty, 2008 for an example) become amenable to an axiomatic test of their scale level.

Finally, the deductive, probabilistic axiom testing method used in this article can be extended to the testing of other measurement axioms of difference structures and beyond that, can be adapted to other measurement structures (see also, Knoblauch & Maloney, 2008, 2012).

References

Agresti, A. (1992). Analysis of ordinal paired comparison data. *Journal of the Royal Statistical Society. Series C* (*Applied Statistics*), *41*, 287-297.

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

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

Allison, P. (2008). Convergence failures in logistic regression. *SAS Global Forum.* Retrieved from http://www2.sas.com/proceedings/forum2008/360-2008.pdf

Alt, F. (1936). Über die Messbarkeit des Nutzens [On the measurability of utility]. *Zeitschrift für Nationalökonomie*, *7*, 161-169.

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review*, *77*, 153-170.

Bechtel, G. G., & O'Connor, P. J. (1979). Testing micropreference structures. *Journal of Marketing Research*, *16*, 247-257.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57*, 289–300.

Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theory of responses. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 97-132). Palo Alto, CA: Stanford University Press.

Böckenholt, U. (2001). Thresholds and intransitivities in pairwise judgments: A multilevel analysis. *Journal of Educational and Behavioural Statistics*, *26*, 269-282.

Böckenholt, U. (2006). Thurstonian-based analyses: Past, present, and future utilities. *Psychometrika*, *71*, 615-629.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232-239.

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

Boschman, M. C. (2001). DifScal: A tool for analyzing difference ratings on an ordinal category scale. *Behavior Research Methods, Instruments, and Computers*, *33*, 10-20.

Carnap, R. (1966). *Philosophical foundations of physics*. New York: Basic Books.

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

Codispoti, M., Ferrari, V., & Bradley, M. M. (2006). Repetitive picture processing: Autonomic and cortical correlates. *Brain Research*, *1068*, 213-220.

Debreu, G. (1958). Stochastic choice and cardinal utility. *Econometrica*, *26*, 440-444.

Falmagne, J. C. (1985). *Elements of psychophysical theory*. New York: Oxford University Press.

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

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, and Computers*, *35*, 116-124.

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* (pp. 60-89). Beverly Hills: Sage.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical and practical guide. *Practical Assessment, Research & Evaluation*, *17*, 1-13.

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)

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

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

Junge, M., & Reisenzein, R. (2013). Indirect scaling methods for testing quantitative emotion theories. *Cognition and Emotion*, *27*, 1247-1275.

Junge, M., & Reisenzein, R. (2015). Maximum likelihood difference scaling versus ordinal difference scaling of emotion intensity: A comparison. *Quality and Quantity, 49*, 2169-2185.

Hein, K. A., Jaeger, S. R., Carr, T., & Delahunty, C. M. (2008). Comparison of five common acceptance and preference methods. *Food Quality and Preference, 19,* 651–661.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine, 21,* 2409 - 2419.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass [the axioms of quantity and the theory of measurement]. *Berichte über die Verhandlungen der Königlichen Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, *53*, 1-64.

Kim, K. O., & O'Mahony, M. (1998). A new approach to category scales of intensity I: Traditional versus rank-rating. *Journal of Sensory Studies*, *13*, 241-249.

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

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

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

Kosmidis, I. (2014). Improved estimation in cumulative link models. *Journal of the Royal Statistical Society: Series B*, *76*, 169-196.

Kosmidis, I., & Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, *96*, 793-804.

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

Luce, R. D. (1994). Thurstone and sensory scaling: Then and now. *Psychological Review*, *101*, 217-277.

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. III (pp. 252-410). New York: Wiley.

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

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*, 573-585.

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

McKelvey, R., Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*, 103-120.

Mellers, B. A. (2000). Choice and the relative pleasure of consequences. *Psychological Bulletin*, *126*, 910-924.

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.

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

Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement. Translated from part I of Otto Hölder's German text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology*, *40*, 235-252.

Michell, J., & Ernst, C. (1997). The axioms of quantity and the theory of measurement. Translated from part II of Otto Hölder's German text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology*, *41*, 345-356.

O' Brien, R. M. (1982). Using rank-order measures to represent continuous variables. *Social Forces*, *61*, 144-155.

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

O'Brien, R. M., & Homer, P. (1987). Corrections for coarsely categorized measures: LISREL's polyserial and polychoric correlations. *Quality and Quantity*, *21*, 349-360.

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: University Press.

Pekrun, R., & Bühner, M. (2014). Self-report measures of academic emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 561-579). New York: Taylor & Francis.

Petrusic, W. M., Baranski, J. V., & Kennedy, R. (1998). Similarity comparisons with remembered and perceived magnitudes: Memory psychophysics and fundamental measurement. *Memory and Cognition*, *26*, 1041-1055.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42-56.

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.

Reisenzein, R. (2009). Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. *Cognitive Systems Research*, *10*, 6-20.

Reisenzein, R. (2010). Is disgust an emotion? [Abstract] *Review of Psychology*, *17*, 144-145.

Reisenzein, R. (2012). What is an emotion in the belief-desire theory of emotion? In F. Paglieri, L. Tummolini, R. Falcone, & M. Miceli (Eds.), *The goals of cognition: Essays in honor of Cristiano Castelfranchi* (pp. 181-211). London: College Publications.

Reisenzein, R., Junge, M., Studtmann, M., & Huber, O. (2014). Observational approaches to the measurement of emotions. In: R. Pekrun & L. Linnenbrink-Garcia (Eds.) *International Handbook of Emotions in Education* (pp. 580-606). Taylor & Francis / Routledge.

Reisenzein, R., Meyer, W.-U., & Niepel, M. (2012). Surprise. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior*, 2nd ed. (pp. 564-570). Waltham: Academic Press.

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

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

Rosenberg, E. L. & Ekman, P. (1994). Coherence between expressive and experiential systems of emotions. *Cognition and Emotion, 8,* 201-229.

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.

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: Guilford: Press.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*, 695-729.

Schneider, B. (1980). Individual loudness functions determined from direct comparisons of sensory intervals. *Perception and Psychophysics*, *28*, 493-503.

Schneider, B. (1982). The nonmetric analysis of difference judgments in social psychophysics: Scale validity and dimensionality. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 317-337). Hillsdale, NJ: Erlbaum.

Schneider, B., Parker, S., & Stein, D. (1974). The measurement of loudness using direct comparisons of sensory intervals. *Journal of Mathematical Psychology*, *11*, 259-273.

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

Suls, J. (1972). A two-stage model for the appreciation of jokes and cartoons: An information processing analysis. In J. Goldstein & P. McGhee (Eds.), *The psychology of humour* (pp. 81-99). New York: Academic Press.

Suppes, S., & Winet, M. (1955). An axiomatization of utility based on the notion of utility differences. *Management Science*, *1*, 259-270.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286.

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: Wiley.

Tsai, R.-C., & Böckenholt, U. (2006). Modeling intransitive preferences: A random-effects approach. *Journal of Mathematical Psychology*, *50*, 1-14.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C. W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *The New England Journal of Medicine, 368*, 1388-1397.

Westermann, R. (1985). Empirical tests of scale type for individual ratings. *Applied Psychological Measurement*, *9*, 265-274.

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.