Psychological Assessment
2014, Vol. 26, No. 2, 666–672

© 2013 American Psychological Association
1040-3590/14/$12.00 DOI: 10.1037/a0035246

BRIEF REPORT

Introducing the Geneva Emotion Recognition Test: An Example of Rasch-Based Test Development

Katja Schlegel, Didier Grandjean, and Klaus R. Scherer
Swiss Center for Affective Sciences, University of Geneva

Existing tests to measure the ability to recognize other people's emotional expressions (emotion recognition ability [ERA]) mostly focus on a single modality (usually the face) and include only a small number of emotions, restricting their ecological validity. Further, their reliability is often unsatisfactory. The goal of the present study was to develop a new ERA test (Geneva Emotion Recognition Test [GERT]) that (a) features dynamic and multimodal actor portrayals (short videos with sound), (b) contains a large number of emotions, and (c) is based on modern psychometric principles (item response theory). We asked 295 participants to watch 108 actor portrayals and to choose, for each portrayal, which of 14 emotions had been expressed by the actor. We then applied the Rasch model independently to each of the 14 emotion portrayal subsets to select 83 final items for the GERT. Results showed that the model fits the emotion subtests and the overall GERT and that measurement precision is satisfactory. Consistent with previous findings, we found a decline in ERA with increasing age and an ERA advantage for women. To conclude, the GERT is a promising instrument to measure ERA in a more ecologically valid and comprehensive fashion than previous tests.

Keywords: emotion recognition, test development, item response theory, emotional intelligence, age differences

Supplemental materials: http://dx.doi.org/10.1037/a0035246.supp

The recognition of emotions from another person's nonverbal expressions in the face, voice, or body is a key ability in social interactions, as it is the precondition for understanding, anticipating, and reacting appropriately to this person's behavior. Research on the recognition of emotional states from nonverbal cues therefore has a long tradition in psychology. A large part of the research has focused on the question of whether emotion recognition accuracy differs between cultures (e.g., Ekman & Friesen, 1971), men and women (e.g., Hall, 1978), age groups (e.g., Ruffman, Henry, Livingstone, & Phillips, 2008), and clinical vs. healthy populations (e.g., for a meta-analysis on schizophrenic patients, see Kohler, Walker, Martin, Healey, & Moberg, 2010). More recently, an increasing body of research has also examined how individual differences in emotion recognition ability (ERA) can explain success as well as deficits in social adjustment. In their meta-analysis, Hall, Andrzejewski, and Yopchick (2009) found that interpersonal sensitivity (which comprises the ability to accurately judge others' emotions) is positively related to social skills, cultural adjustment, workplace effectiveness, and relationship quality. From a clinical perspective, Marsh and Blair (2008) found deficits in ERA to be associated with antisocial behavior. With the advent of the emotional intelligence (EI) construct, ERA has sparked even more interest, being proposed as one of the basic components of the ability EI model (Mayer, Salovey, Caruso, & Sitarenios, 2003).

Given that the only widely used EI test, the Mayer-Salovey-Caruso Emotional Intelligence Test (Mayer et al., 2003), has been criticized in terms of its scoring procedure, inconsistent factor structure, and unclear concurrent validity, several scholars have recently suggested that the development of new EI measures should take into account long-existing ERA tests (e.g., Cherniss, 2010). In such tests, participants are typically presented a range of emotional expressions (such as pictures of faces or voice recordings) and are asked to choose, from a list of emotions, which emotion has been expressed in each portrayal. ERA is usually calculated as the proportion of correctly identified portrayals. Among the most widely used tests are the DANVA (Diagnostic Analysis of Nonverbal Accuracy; Nowicki & Duke, 1994), the JACBART (Japanese and Caucasian Brief Affect Recognition Test; Matsumoto et al., 2000), the ERI (Emotion Recognition Index; Scherer & Scherer, 2011), and the MERT (Multimodal Emotion Recognition Test; Bänziger, Grandjean, & Scherer, 2009). The DANVA consists of 24 pictures of facial expressions

This article was published Online First December 2, 2013.

Katja Schlegel, Didier Grandjean, and Klaus R. Scherer, Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland.

This research was supported by an Advanced Grant of the European Research Council, awarded to Klaus R. Scherer, and the Swiss Center for Affective Sciences.

Correspondence concerning this article should be addressed to Katja Schlegel, Swiss Center for Affective Sciences, University of Geneva, rue des Battoirs 7, CH-1205 Geneva, Switzerland. E-mail: Katja.Schlegel@unige.ch.

and 24 audio recordings displaying one of four emotions (anger, fear, happiness, sadness). The ERI features 30 pictures of faces and 30 audio recordings of five emotions (anger, fear, happiness, sadness, disgust). In the JACBART, 56 pictures of faces displaying one of seven emotions (anger, contempt, disgust, fear, happiness, sadness, surprise) are presented briefly, embedded in a neutral expression of the same person. The MERT consists of 30 portrayals of 10 emotions (irritation, anger, anxiety, fear, happiness, elated joy, disgust, contempt, sadness, despair) presented in four modalities (still picture, video, audio, audio-video).

As can be seen from these descriptions, all ERA tests use pictures of facial expressions, and in four cases, additionally vocal expressions as stimuli. Only the MERT includes multimodal (audio-video) emotion portrayals. Although unimodal stimuli are useful for comparing recognition rates between modalities, they are not the most common way to interact with others in real life. For measuring ERA as a predictor of social effectiveness, multimodal stimuli combining visual and auditory information are considered more ecologically valid (Hall, 1978). Furthermore, most ERA tests contain only few emotions and response alternatives. Thus, the measurement of ERA might be influenced by the use of exclusion and probability rules. For example, happiness, being the only positive emotion in most tests, can easily be distinguished from negative emotions in facial expressions, leading to high "recognition" rates for this emotion. Consequently, existing ERA tests are generally relatively easy, which might restrict their power to discriminate between individuals. Another limitation of most ERA tests is low internal consistency, which some researchers have explained by the huge variety of possible emotion expressions even within one emotion category. The content domain of ERA might thus be too diverse to meet the unidimensionality assumption for internal consistency (Scherer & Scherer, 2011).

In the present article, we describe the development of the new Geneva Emotion Recognition Test (GERT), which was designed to deal with some of these limitations. The GERT consists of 14 emotions including six positive ones and covers a larger spectrum of emotional expressions than previous tests. The stimuli are dynamic and multimodal (short audio-video clips) and were portrayed by 10 actors (five men and five women) of different ages, adding to the ecological validity of the test. Furthermore, we developed the GERT using the modern psychometric framework of item response theory (IRT). One major advantage of IRT is that subjects' responses are modeled at the item level, allowing for much more flexible item selection than classical test theory. Here, we specifically applied the Rasch model, a simple IRT model requiring only modest sample sizes. This approach is often used to select suitable items during test development when (a) the underlying construct to be measured is expected to be unidimensional, (b) responses are binary (e.g., correct/incorrect), and (c) when guessing is not expected to have a substantial influence on subjects' responses (Embretson & Reise, 2000). Given that (a) recent research suggests that ERA is essentially unidimensional (Schlegel, Grandjean, & Scherer, 2012), (b) guessing should have little influence in a test with 14 response options, and (c) responses on the GERT are coded as either correct or incorrect; the Rasch model is an appropriate choice for the development of the GERT.

In addition, in this article we provide first evidence for the construct validity of the GERT by examining gender and age differences. It has been consistently found that women score better in ERA tests than men (e.g., Hall, 1978). The existence of differential item functioning (DIF) in previous tests might be one reason for this finding. DIF occurs when members of different groups (e.g., men/women) with the same overall ability level do not have the same probability of solving a certain item of a test. For example, it might be that emotional expressions produced by women are recognized more easily by female than male subjects. With respect to age, meta-analysis has shown a decline in ERA from still pictures of faces and vocal stimuli for older subjects, particularly for anger, sadness, and fear (Ruffman et al., 2008). However, Ruffman (2011) and Phillips and Slessor (2011) recently pointed out that it remains largely unstudied whether, first, age differences persist also when multimodal stimuli are used and, second, age differences in positive emotions (other than happiness) are smaller than in negative emotions. We address these questions using the GERT in our study and test for age DIF in order to ensure equivalent measurement properties of the items for younger and older subjects.

Method

Subjects, Stimuli, and Procedure

Subjects were recruited through different German websites, predominantly through a panel for online studies founded by the psychology department of the Humboldt University of Berlin (www.psytests.de). Of the 454 subejcts (127 men, age 17–75 years; $M = 35.8$, $SD = 14.0$) who had started the study, 255 (56%) completed all 108 items. Here, we included the 295 subjects (65%, 82 men, age 17–74 years; $M = 37.1$, $SD = 13.9$) who completed at least 100 items to ensure sufficient task compliance. This sample size can be considered as sufficient for the planned Rasch analyses (Embretson & Reise, 2000). The retained sample did not differ from the sample who dropped out with respect to gender, educational background, and occupational status, but was significantly older (dropout sample age $M = 32.2$, $SD = 13.7$), $t(444) = 3.54$, $p < .000$. This might be due to the fact that many older subjects were recruited in an online forum for elderly people via personalized messages and were thus more motivated to complete the study. All subjects were either native German speakers or reported very good knowledge of German (which does not rule out possible differences in ethnic origin). More detailed characteristics of the full and of the retained sample are provided in supplementary Tables S1a and S1b. Subjects received feedback on their performance at the end of the study.

The stimuli were taken from the Geneva Multimodal Emotion Portrayals (GEMEP) corpus (Bänziger, Mortillaro, & Scherer, 2012), which contains 1,260 audio-video clips (duration = 1–4 s) of 18 emotions portrayed by 10 actors (five female) with different intensities and verbal contents. All actors were Caucasian, and their age ranged from 25 to 57 years ($M = 37$). As a starting point for the selection of the first GERT item pool, we only considered the 520 portrayals (a) in which emotions were expressed in normal intensity with one of two pseudolinguistic sentences as verbal content and (b) that represented the 14 emotions: joy, amusement, pride, pleasure, relief, interest, anger, fear, despair, irritation, anxiety, sadness, disgust, and surprise. The first 12 emotions were chosen because they can be evenly distributed on the four quadrants in the emotional valence-arousal space (Bänziger et al., 2012) and thus provide a balanced and varied set of emotions. Disgust and surprise were added because they are frequently used in other emotion recognition tests. From these 520 portrayals, we selected items on the basis of recognition percentages and believability ratings that were obtained by Bänziger et al. (2012). Our goal was to include portrayals that (a) were rated as sufficiently authentic displays of the respective emotion and (b) covered a wide range of item difficulties. We thus selected portrayals for which the target emotion (i.e., the emotion that the actor was asked to express) was the most frequently chosen response category and both the recognition accuracy and believability were above the 30th percentile of all portrayals in a given emotion category. Between six and nine portrayals per emotion were chosen (see Table 1), resulting in a pool of 108 items. In the present study, subjects watched all 108 clips in a random order (duration ~ 30 min). After each clip, the 14 emotion labels were presented on the screen, and subjects were asked to choose which of the 14 emotions had been expressed by the actor in the clip (forced-choice format). For each clip, responses were recoded into binary variables (0 = incorrect, 1 = correct). All instructions were provided in German. An English demo version of the GERT is available at www.affective-sciences.org/gert.

Data Analysis

In the Rasch model, the probability with which a person solves an item is determined by the location of this person on a latent trait dimension θ (in our case, ERA) and the item difficulty. Item difficulty determines the location of the logistic function describing the item's solving probability (item characteristic curve [ICC]) on θ. All ICCs are assumed to be parallel. Two conditions need to be met in order to interpret the parameters in the Rasch model: The Rasch model must fit (i.e., the observed response proportions for the items must match the postulated parallel ICCs) and the data need to be unidimensional; that is, item responses should not be influenced by any systematic factor apart from θ. As our item pool was very large (108 items) in comparison to our sample size, initial factor analyses did not yield stable and conclusive results with respect to whether the unidimensionality assumption was met and which items should be removed. We therefore decided to split the item pool into 14 more manageable subsets by bundling items belonging to one target emotion because, theoretically, the ability involved in correctly recognizing portrayals of one specific emotion, such as anger, should be unidimensional (Schlegel et al., 2012). Our goal was thus to create unidimensional item subsets that comply with the Rasch model and to eliminate nonfitting items on the subset level first before jointly analyzing the "cleaned" sets of retained items at a later stage. We tested the unidimensionality assumption for each of the 14 item subsets by running a one-factor comparative factor analysis (CFA). If model fit was insufficient (i.e., comparative fit index < .95, root-mean-square error of approximation > .05, and root-mean-square residual > .08), we eliminated items with negative or low factor loadings. In the second step, we fitted the Rasch model to each of the 14 unidimensional item subsets using the eRm package in R (Mair & Hatzinger, 2007). Model fit was evaluated by inspecting the weighted-fit or "Infit" and unweighted-fit or "Outfit" index for each item. These statistics indicate how much the observed ICC differs from the ICC that is theoretically expected, with values between .80 and 1.20 indicating "useful fit" (Wright & Linacre, 1994) and 1.00 representing perfect fit. From all items with "useful fit," we attempted to select six items per emotion, including three portrayals produced by female actors and three portrayals produced by male actors. This was done by removing the easiest items if more than three female or male items were available for an emotion.

The Rasch model was finally fit on all retained items, and Infit and Outfit indices were inspected. We evaluated the difficulty of the overall test and its measurement precision by inspecting the

Table 1
Number of Items Per Emotion (Produced by Female/Male Actors), Descriptive Statistics of the Initial Item Pool (108 Items) and the Final GERT (83 Items), and Correlations With Gender and Age

	Initial item pool			Final GERT			Correlations	
Emotion	No. of items (f/m)	*M*	*SD*	No. of items (f/m)	*M*	*SD*	Gender	Age
amusement	8 (4/4)	0.84	0.17	6 (3/3)	0.81	0.20	.07	−.17**
anger	8 (4/4)	0.64	0.24	6 (2/4)	0.55	0.29	.05	−.34**
disgust	7 (4/3)	0.56	0.22	6 (3/3)	0.50	0.24	.04	−.26**
despair	7 (4/3)	0.72	0.19	5 (3/2)	0.72	0.23	.20**	−.02
pride	8 (4/4)	0.68	0.20	6 (3/3)	0.63	0.24	.13*	−.29**
anxiety	8 (4/4)	0.75	0.18	6 (3/3)	0.71	0.21	.03	−.28**
interest	7 (4/3)	0.60	0.20	6 (3/3)	0.67	0.20	−.03	−.10
irritation	7 (4/3)	0.72	0.25	6 (3/3)	0.70	0.27	.01	−.43**
joy	9 (4/5)	0.63	0.18	6 (3/3)	0.75	0.22	.04	−.22**
fear	8 (4/4)	0.52	0.24	6 (3/3)	0.47	0.26	.16**	−.23**
pleasure	8 (4/4)	0.75	0.18	6 (3/3)	0.72	0.22	.00	−.10
relief	9 (4/5)	0.83	0.16	6 (3/3)	0.86	0.18	.10	.02
surprise	6 (3/3)	0.42	0.24	6 (3/3)	0.42	0.24	−.04	−.12*
sadness	8 (5/3)	0.77	0.19	6 (3/3)	0.80	0.21	−.03	−.05
total	108 (56/52)	0.67	0.09	83 (41/42)	0.67	0.10	.13*	−.46**

Note. Positive correlations with gender indicate higher scores for females, and positive correlations with age indicate higher scores with increasing age. GERT = Geneva Emotion Recognition Test.
* $p < .05$. ** $p < .01$.

test information curve (TIC), which shows the range of the latent dimension θ in which the test discriminates best, and the *standard error of measurement* (*SEM*) function, which can be used to calculate the confidence interval of a person's ability. Further, we calculated the total test score reliability from the IRT parameters following the method proposed by Dimitrov (2003) for binary data. Gender and age DIF were examined using the revised Angoff's delta method, which performs particularly well with smaller sample sizes and is implemented in the R package deltaPlotR (Magis & Facon, 2012). Gender and age differences in ERA were analyzed by calculating correlations with the 14 emotion subscores and the total GERT score.

Results

Mean recognition rates for the 14 emotions based on the 108 initial items ranged from .42 for surprise to .84 for amusement (see Table 1). In the course of the 14 CFAs, eight items (one item each for amusement, interest, joy, and sadness; two items each for despair and relief) were removed. The Rasch model fit each of the 14 subsets well, with only four items (one each for disgust, irritation, pleasure, and surprise) displaying both an Infit and Outfit below .80. The fit indices of the 14 CFAs and results of the 14 Rasch models are reported in supplementary Table S2. When we tried to select three male and female items per emotion, as described above, for disgust, irritation, interest, and surprise, only two items were left for one of the two genders. To maintain balanced actor gender, we decided to keep three of the items that displayed an Infit and Outfit below .80.[1] For anger, the final scale included four male and two female portrayals, as one of the two other available female portrayals did not fit the CFA model, and the other one was extremely easy ($M = .98$). The final despair scale consisted of only five portrayals because the other items did not load on the unidimensional model. The final GERT included 83 items—six items for each of the 14 emotions with the exception of despair that contained only five items. Recognition rates for all items are provided in supplementary Table S3.

In the Rasch analysis of the overall GERT, all items met the criterion of "useful fit" (see Table S3). We then inspected the item difficulty parameters and compared them with the ability estimates of the sample to evaluate overall test difficulty (see Figure 1). As can be seen on the left side of Figure 1, the ability estimates ranged from −1.52 to 1.80 on θ, with the mean being fixed to zero ($SD = 0.5$). The right side of Figure 1 shows the distribution of the item difficulties on θ, ranging from −2.83 for the item amu5 (the easiest item) to 2.82 for the item sur74 (the most difficult item), with a mean of −.82 ($SD = .99$). About 20% of the items had a difficulty above zero and discriminated best among individuals in the higher ability range, but the majority of the items measured most precisely in the lower ability range. Consequently, the GERT can be considered a comparatively easy test for the studied population. This was also evident from the TIC and the *SEM* function (see Figure 1; for the exact values corresponding to each raw score and ability estimate, see Table S4), showing that the GERT provided the highest measurement information for individuals with an ability of about −1. In the range of 95% of the person parameters (i.e., ca. between −1 and 1), the *SEM* ranged from −.24 for a θ of −1 to .32 for a θ of 1. Consequently, the 95% confidence interval for an ability estimate of 1 [0.37, 1.63] was by only .32 (about [1/2] standard deviation) larger than the confidence interval for an ability estimate of −1 [−1.47, −.53]. Thus, measurement precision was comparatively good for the large majority of the sample. Furthermore, test score reliability was excellent ($\rho = .92$). The Angoff's delta DIF tests flagged one irritation item (irr44) as being relatively easier to solve for older than for younger subjects and one surprise item (sur74) as relatively easier to solve for women than for men (see Table S2 for DIF statistics). As an inspection of these items did not reveal any obvious reason for DIF and only two items were concerned, we decided to keep them in the GERT.

Finally, the dimensional structure of the overall test was examined by comparing several competing CFA models that were run on the 14 emotion subscores (reported in Schlegel et al., 2012). Results showed that a model with one general ERA factor and additional subfactors for pairs of similar emotions (such as irritation and anger) fit the data well (see supplementary Figure S1 for details). The GERT can thus be considered an essentially unidimensional test, although future studies with bigger sample sizes are needed to assess whether unidimensionality also holds on the item level (i.e., when the 83 items instead of the 14 subscores are analyzed).

Table 1 shows the correlations of the GERT total score and emotion subscales with gender and age. Women tended to have significantly higher scores in recognizing despair, pride, and fear, and a small, but significant advantage in the GERT total score. Correlations with age were generally higher, revealing a decline in ERA with increasing age for three out of six positive emotions (amusement, joy, pride), surprise, all negative emotions except for despair and sadness, and the total score (see Figures S2 and S3 for a graphical representation). The correlation between the mean score of the positive emotion scales and age ($r = -.31$) was lower than for the mean of the negative emotion scales ($r = -.44$, Steiger's $Z = 2.24$, $p < .05$). We also tested whether actors' age was related to the mean recognition rate of their portrayals. This correlation was not significant ($r = -.14$, $p = .703$), suggesting that portrayals by younger and older actors were equally well recognized in our sample.

Discussion

In this article, we have presented a new ERA test that constitutes an advancement in the field on several levels. First, the GERT contains a large number of emotions that are expressed in dynamic and multimodal stimuli by 10 different actors. Thus, the GERT presumably captures ERA more broadly than previous tests, which have mostly relied on fewer emotions, fewer actors, and less stimuli from only one modality. Second, the GERT is the first test in the ERA domain to which IRT was applied. We found that the Rasch model, in which participants' performance is explained only by their underlying ERA and the difficulty of the items, fit our data sufficiently well. In addition, we tested whether GERT items function in a similar way for both genders as well as for younger and older subjects, and we concluded that DIF seems negligible. Our research has thus provided a more complete understanding of

[1] This decision can be justified considering that an Infit or Outfit value below 1 indicates that the respective item discriminates better than the average item on the test, which is a desirable feature.

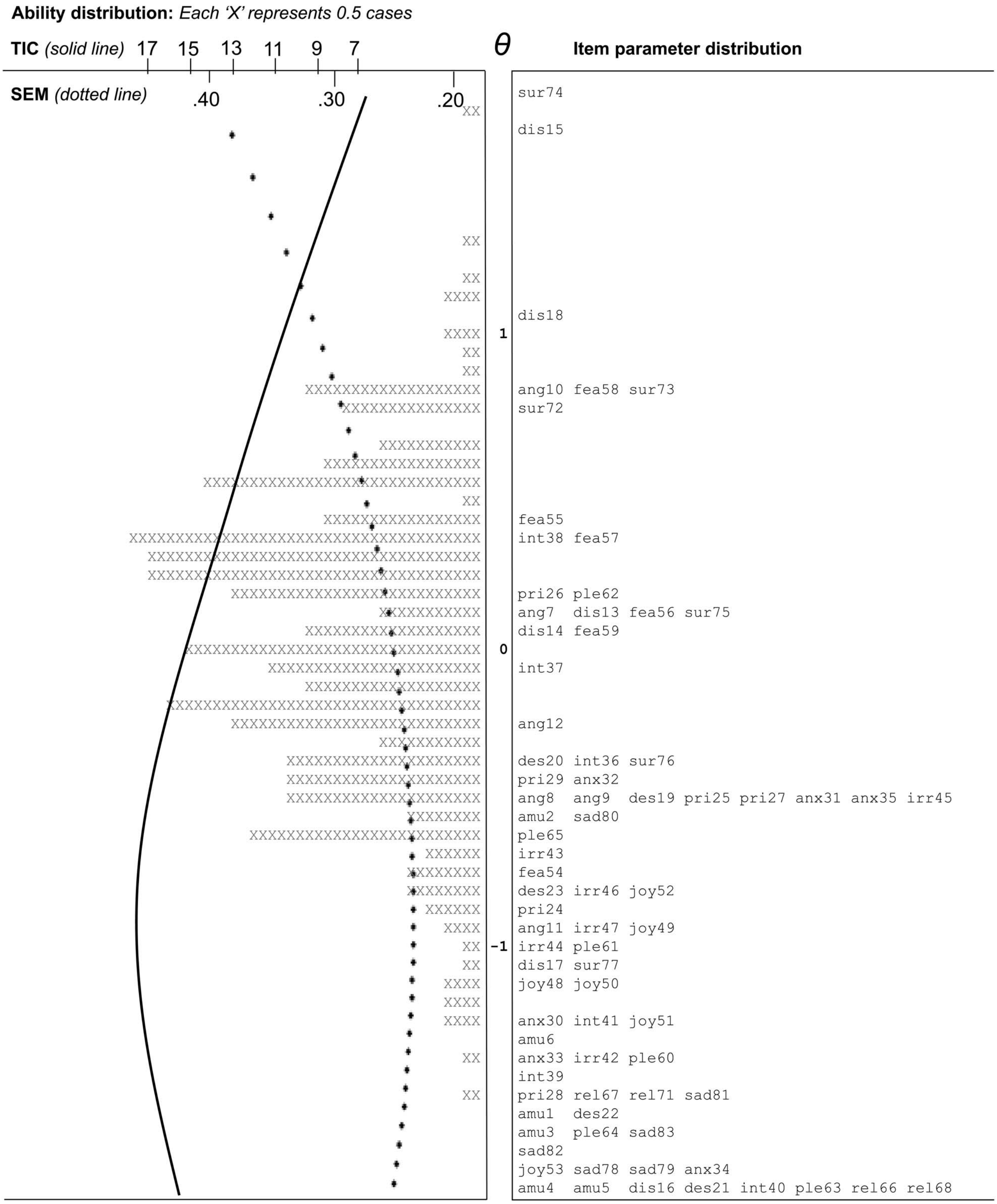

Figure 1. Wright Map displaying the distribution of the sample's ability estimates, the test information curve (TIC), and the standard error of measurement (SEM) on the left side and the item difficulties on the right side of the latent dimension θ. The labels on the right side represent the item numbers (1 to 83) and the respective emotion category. The mean of the ability distribution was fixed to zero. sur = surprise; dis = disgust; ang = anger; fea = fear; int = interest; pri = pride; ple = pleasure; des = despair; anx = anxiety; irr = irritation; amu = amusement; rel = relief; sad = sadness.

the measurement properties of the GERT than had been available for any previous ERA test.

In this study, we found age and gender differences in ERA that are generally consistent with previous findings and provide first evidence for the construct validity of the GERT. Specifically, we found a small, but significant advantage for women in overall ERA. This advantage seems to be somewhat smaller than what was found in the meta-analysis by Hall (1978) for multimodal stimuli. One reason might be that the stimuli used in the studies included in Hall's meta-analysis had not been tested for gender DIF and might have been relatively easier for women. With respect to age, we found a decline in ERA for older subjects for nine out of 14 emotions. Overall, these results speak for Ruffman's (2011) position that age differences persist even when rich information from multimodal stimuli is available. An exception might be the sadness/despair emotion family, for which we did not find age differences contrary to Ruffman et al.'s (2008) meta-analysis for facial and vocal stimuli. For these emotions, older subejcts seem to benefit from multimodal stimuli with complementary vocal, facial, and bodily cues, which attenuate the age effect observed in facial and vocal stimuli only. Our results also suggest that the age difference is smaller for positive than for negative emotions, which might be related to a positivity bias in older age that attenuates the ERA decline (Phillips & Slessor, 2011). These findings illustrate the usefulness of the GERT as an instrument to study the mechanisms of age and gender differences in ERA.

One shortcoming of the GERT is the relatively low difficulty as compared with the ability distribution in our sample. Several reasons may account for this. First, in our stimuli, visual and auditory cues complement each other in conveying an emotion expression that might be easier to recognize than in single-modality portrayals (Scherer et al., 2012). Second, the selection of items in ERA testing is guided by two seemingly opposed goals. On the one hand, the emotion expressions used as test items should not be recognized by the large majority of the sample, as they will not discriminate well. On the other hand, the expressions must contain enough cues to be recognized. This might often not be the case for difficult items, which might contain ambiguous or insufficient cues regarding the target emotion. As it is difficult to distinguish between "genuinely" high difficulty and high difficulty due to low stimulus quality (e.g., bad actor performance), emotion portrayals are usually preselected according to high rater agreement with respect to the target emotion, which in turn results in the selection of easy items. Here, we used believability ratings to disentangle difficulty and stimulus quality. However, even in large databases like the GEMEP (see the Method section), the number of believable and yet difficult portrayals is limited because they are "naturalistic" when compared with items used in other domains. Another reason for the rather low difficulty of the GERT might be that our sample did not cover the full ERA range in the population. The individuals who took part in our online study might have been particularly interested in this topic and presumably had a rather high ERA. Also, we excluded 35% of the subjects who completed fewer than 100 items, which might have eliminated a considerable number of lower ability subjects who found the task too difficult.

However, it should also be noted that the low difficulty level found in our study with "normal" adults might be optimal for measuring ERA in clinical populations for which recognition rates are, on average, 5%–20% lower (Marsh & Blair, 2008). Accordingly, the measurement precision of the GERT might be higher for patients than for healthy subjects. As most of the research in clinical populations with schizophrenia, depression, Parkinson's disease, and the like was conducted with still pictures of basic emotions, the field could benefit from using the GERT in terms of ecological validity. Further, the GERT allows measuring emotion-related response bias and confusion patterns in clinical groups more comprehensively than previous tests. Also, with an administration time of 15–20 min, the GERT is rather time efficient.

There are several limitations to our study. First, we exclusively recruited our subjects online. Although this allowed us to investigate a demographically diverse sample, we cannot ensure that subjects completed the study under the same circumstances (e.g., regarding the sound quality). In addition, it remains an open question whether especially our older subjects are comparable to people of the same age who do not use the Internet. Future studies should aim to replicate our findings with different samples and under more controlled conditions. Second, our sample size was rather small for IRT analyses and did not allow us to use more complex models than the Rasch model. Despite the fact that the Rasch model fit our data sufficiently well, more complex models like the two-parameter logistic model (2PL, allowing the ICC slopes to vary) or even the three-parameter logistic model (adding a guessing parameter to the 2PL) are likely to provide a more accurate account of the nature of ERA and to enhance our understanding of the underlying mechanisms. Future studies should therefore attempt to test larger samples, which would also allow examining the dimensional structure of the test on the item level and the psychometric properties of the total score. Furthermore, future studies should assess the test–retest reliability of the test scores and the equivalence of the measurement properties in clinical populations and other cultures and language groups. For example, given that the actors in the GERT were all Caucasian, it should be ensured that the measurement properties are equivalent for Caucasian and non-Caucasian test takers. Currently, data are being collected with English, French, and Dutch versions of the GERT, which will help resolve some of these questions in the near future.

The next validation steps also need to include the examination of the construct and predictive validity of the test scores. Regarding construct validity, it is particularly important to examine the overlap of the GERT with existing ERA and EI tests. If the GERT shares substantial variance with these measures, this would support our assumption that the GERT captures ERA more broadly than existing tests that correlate only to a low extent (Hall, 2001). With respect to predictive validity, future studies should focus on objective measures of social effectiveness, such as performance in face-to-face interaction tasks. As past research has mostly used indirect, subjective outcome measures such as supervisor ratings, little is known about the mechanisms that link ERA to success in private and professional life. To summarize, the GERT is a promising new measure of ERA, but further studies with larger samples involving different populations are needed to substantiate its psychometric quality. We hope that the present article will encourage other scholars in the field of ERA and EI to apply IRT to improve the measurement of these constructs.

References

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion, 9,* 691–704. doi:10.1037/a0017088

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion, 12,* 1161–1179. doi:10.1037/a0025827

Cherniss, C. (2010). Emotional intelligence: Toward clarification of a concept. *Industrial and Organizational Psychology, 3,* 110–126. doi:10.1111/j.1754-9434.2010.01231.x

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27,* 440–458. doi:10.1177/0146621603258786

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17,* 124–129. doi:10.1037/h0030377

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85,* 845–857. doi:10.1037/0033-2909.85.4.845

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior, 33,* 149–180. doi:10.1007/s10919-009-0070-5

Kohler, C. G., Walker, J. B., Martin, E. A., Healey, K. M., & Moberg, P. J. (2010). Facial emotion perception in schizophrenia: A meta-analytic review. *Schizophrenia Bulletin, 36,* 1009–1019. doi:10.1093/schbul/sbn192

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology, 65,* 302–321. doi:10.1111/j.2044-8317.2011.02025.x

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20,* 1–20.

Marsh, A. A., & Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: A meta-analysis. *Neuroscience and Biobehavioral Reviews, 32,* 454–465. doi:10.1016/j.neubiorev.2007.08.003

Matsumoto, D., LeRoux, J. A., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., . . . Goh, A. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior, 24,* 179–209. doi:10.1023/A:1006668120583

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2. 0. *Emotion, 3,* 97–105. doi:10.1037/1528-3542.3.1.97

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior, 18,* 9–35. doi:10.1007/BF02169077

Phillips, L., & Slessor, G. (2011). Moving beyond basic emotions in aging research. *Journal of Nonverbal Behavior, 35,* 279–286. doi:10.1007/s10919-011-0114-5

Ruffman, T. (2011). Ecological validity and age-related change in emotion recognition. *Journal of Nonverbal Behavior, 35,* 297–304. doi:10.1007/s10919-011-0116-3

Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews, 32,* 863–881. doi:10.1016/j.neubiorev.2008.01.001

Scherer, K. R., & Scherer, U. (2011). Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the Emotion Recognition Index (ERI). *Journal of Nonverbal Behavior, 35,* 305–326. doi:10.1007/s10919-011-0115-4

Schlegel, K., Grandjean, D., & Scherer, K. R. (2012). Emotion recognition: Unidimensional ability or a set of modality- and emotion-specific skills? *Personality and Individual Differences, 53,* 16–21. doi:10.1016/j.paid.2012.01.026

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370.

Received October 2, 2012
Revision received October 9, 2013
Accepted October 18, 2013 ■