DOI 10.3758/s13428-015-0646-4

Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation

Katja Schlegel[1] · Klaus R. Scherer[2]

Published online: 28 September 2015

Abstract The ability to accurately interpret others' emotional expressions in the face, voice, and body is a crucial component of successful social functioning and has been shown to predict better outcomes in private and professional life. To date, emotion recognition ability (ERA) has mostly been measured with tests that heavily rely on static pictures of the face and on few emotions, restricting their content validity. Recently, Schlegel, Grandjean, and Scherer (*Psychological Assessment*, *26*, 666–672, 2014) published a new test that measures ERA in a more comprehensive fashion, by (1) including a wide range of 14 positive and negative emotions and (2) using video clips with sound that simultaneously present facial, vocal, and bodily emotional cues. This article introduces the short version of the Geneva Emotion Recognition Test (the GERT-S), and presents two studies (total $N = 425$) that examine the internal consistency, factor structure, and convergent and discriminant validity of the test. The results show that the GERT-S is a unidimensional test with good internal consistency. Furthermore, the GERT-S was substantially positively correlated with other ERA tests, with tests of emotional understanding and emotion management, and with cognitive ability. Taken together, the present studies demonstrate the usefulness of the GERT-S as an instrument for the brief and reliable assessment of ERA. It is available, free of charge and in seven different languages, for academic research use. Given the brief test-taking time (approx. 10 min) and its possible administration via different online platforms, the GERT-S can easily be integrated by researchers into their own studies.

Keywords Emotion recognition ability · Emotional intelligence · Assessment · Testing · Factor analysis

The communication of emotions through nonverbal cues expressed in the face, voice, and body is a crucial element of everyday interpersonal interactions. In particular, the ability to accurately detect and interpret emotional expressions in another person helps anticipating his or her actions, adapting one's own actions accordingly, and consequently, promoting effective interpersonal behavior (McArthur & Baron, 1983). Individual differences in people's emotion recognition (ERA) have been studied for several decades in different domains of psychology, such as social, organizational, clinical, and developmental psychology. More recently, ERA has been proposed as an essential component of emotional intelligence (EI) or emotional competence (EC; for a review, see Roberts, MacCann, Matthews, & Zeidner, 2010; Scherer, 2007). For example, in the popular four-branch EI model by Mayer and Salovey (1997), ERA is considered the most basic branch underlying more complex skills like emotional understanding (knowledge about relationships between emotions and situations) or Emotion Management (the regulation of one's own and others' emotions). Previous research has demonstrated that higher ERA is linked to better social functioning in private and professional life. Numerous studies as well as meta-analyses showed that individuals with higher ERA are perceived as more likable, socially supportive, honest, and open by their peers; have more close relationships, and achieve higher workplace and academic performance (Elfenbein, Foo, White, Tan, & Aik, 2007; Elfenbein, Marsh, & Ambady,

✉ Katja Schlegel
k.schlegel@neu.edu

[1] Social Interaction Laboratory, Northeastern University, 125 Nightingale Hall, 360 Huntington Avenue, Boston, MA 02115, USA

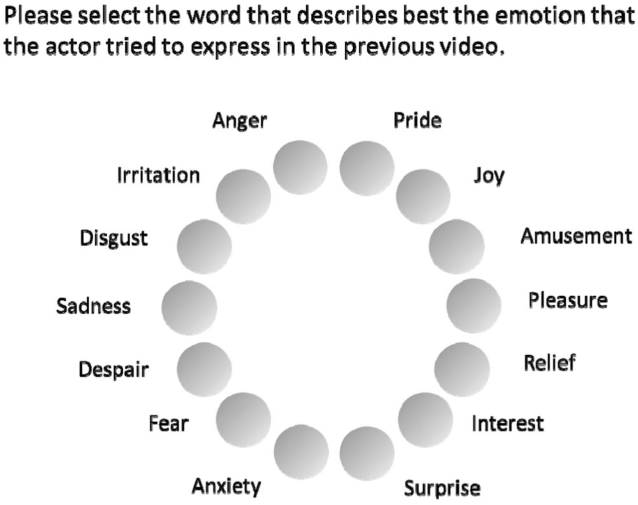
[2] Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

2002; Hall, Andrzejewski, & Yopchick, 2009). In contrast, low levels of ERA have been associated with higher levels of antisocial behavior and as well as mental disorders like schizophrenia and depression (Kohler, Walker, Martin, Healey, & Moberg, 2010; Marsh & Blair, 2008).

Despite widespread interest in the role of ERA in social functioning in the field of psychology, relatively little research has focused on the reliability and validity of the measurement instruments used. Typically, ERA is measured by presenting participants with a range of pictures or recordings of emotional expressions that are produced by actors. After seeing or hearing each portrayal, participants are asked to choose from a list of emotion words the one describing best the emotion expressed by the actor. Participants' responses are scored as correct or incorrect and are summed up into a total ERA score. Some of the most widely used tests using this format include the Diagnostic Analysis of Nonverbal Accuracy (DANVA; Nowicki & Duke, 1994), the Japanese and Caucasian Brief Affect Recognition Test (JACBART; Matsumoto et al., 2000), the Multimodal Emotion Recognition Test (MERT; Bänziger, Grandjean, & Scherer, 2009), and the Emotion Recognition Index (ERI; Scherer & Scherer, 2011). The DANVA consists of three subtests in which participants are presented with either pictures of the face, voice recordings, or body postures (with the face blurred), each depicting one of four emotions. In the JACBART, participants are required to pick one out of seven emotions for briefly displayed pictures of facial expressions embedded in neutral expressions of the same person. The MERT consists of portrayals of ten emotions that are presented either as still pictures, as videos without sound, as audio only, or as a video with sound. Finally, the ERI comprises a subtest with still pictures of facial expressions and a subtest with voice recordings, each measuring five emotions. In addition to the standard labelling paradigm used in these tests, some researchers have also used identification speed of emotional expressions as well as other paradigms such as matching (e.g., identifying which of three facial expressions displays a different emotion) or memorization to measure individual differences in ERA from faces (Herzmann, Danthiir, Schacht, Sommer, & Wilhelm, 2008; Palermo, O'Connor, Davis, Irons, & McKone, 2013; Wilhelm, Hildebrandt, Manske, Schacht, & Sommer, 2014).

As can be seen from this description, most available tests rely on still pictures of faces, and only few include vocal stimuli or postures and gestures. However, previous research suggests that accurate emotion recognition from all of these cue channels contributes to successful social functioning (Elfenbein & Ambady, 2002; Puccinelli & Tickle-Degnen, 2004; Tickle-Degnen, 1998). A second concern is that the different cue channels are assessed separately in most of the existing tests, whereas in real-life emotional expressions are usually communicated in a dynamic fashion and in multiple modalities at the same time (Phillips & Slessor, 2011; for a comparison of recognition accuracy in multimodal stimuli versus stimuli in single channels, see Scherer, Clark-Polner, & Mortillaro, 2011). Social effectiveness might therefore be better predicted by using multimodal expressions (Hall, 1978). Third, the available tests are predominantly restricted to a small number of basic emotion categories and only one positive emotion (happiness). As Frank and Stennett (2001) have noted, test-takers might thus be able to identify the correct response on the basis of category discrimination and exclusion (especially for happiness) rather than actual recognition. Relatedly, in real-life people are faced with a much larger range of positive and negative emotions than only the few basic ones. The content validity of existing ERA tests is therefore limited. In addition, the psychometric quality of these tests, in particular their reliability and factorial structure, has not been widely studied and remains to be solidly established.

The recently published Geneva Emotion Recognition Test (GERT; Schlegel, Grandjean, & Scherer, 2014) aimed to overcome some of the limitations of previous tests in terms of content validity and psychometric quality. In particular, this computer-administered test is based exclusively on multimodal emotion portrayals (i.e., portrayals in which facial, vocal, and bodily cues are presented simultaneously) and a large number of different emotions. It consists of 83 short video clips with sound in which ten actors enact 14 different emotions, six of which are positive. After each clip, participants choose which of the 14 emotions best describes the emotion the actor intended to express (see Fig. 1). The GERT takes about 20 min to complete. Several studies have provided first evidence for the good internal consistency of the test, as well as its construct validity (Schlegel, Fontaine, & Scherer, 2015; Schlegel, Grandjean, & Scherer, 2012, 2014). Furthermore, a recent study (Schlegel, Mehu, van Peer, & Scherer, 2015) found that higher GERT scores predicted higher monetary



Please select the word that describes best the emotion that the actor tried to express in the previous video.

Fig. 1 Response format of the GERT-S. Emotions are arranged in a circle to facilitate test-takers' orientation among the options (adapted from the Geneva Emotion Wheel instrument: Scherer, Shuman, Fontaine, & Soriano, 2013)

gains and higher ratings of cooperativeness and likability in a face-to-face negotiation over the terms of a fictional work contract. The GERT has been translated into different languages, including French, German, Dutch, Italian, Mandarin, and Hungarian, and it is currently being used by numerous research groups around the world via an online platform (for more information and a demo version, see www.affective-sciences.org/content/geneva-emotion-recognition-test-gert).

For some research settings in which participants' time is limited, administration of the full GERT might not be feasible or might be considered too onerous, so using a short version might be desirable. This is particularly true for settings including online studies, studies with a large battery of instruments, and studies with special populations, such as elderly people or patients. A short version of the GERT would also be particularly useful for studies aiming to comprehensively measure all EI components, which is rather time-consuming. This article introduces a short version of the GERT (GERT-S; with 42 items) that takes only about 10 min to complete. We present results from two studies investigating the psychometric quality of the GERT-S.

Study 1 was based on a large and diverse online sample and examined GERT-S's factor structure, internal consistency, and relationship with other emotional and cognitive abilities. In line with previous research on the full GERT, we expected that the GERT-S would show an essentially unidimensional factor structure and good reliability (Schlegel et al., 2012). Furthermore, we hypothesized that the GERT-S would be positively correlated with other components of EI, namely emotional understanding and emotion management. Given that ERA tests are performance-based measures and not self-report questionnaires, they are usually correlated with general mental ability (Murphy & Hall, 2011). We therefore expected the GERT-S to be positively correlated with cognitive intelligence. Finally, past research has often found women to perform somewhat better than men (Hall, 1978) and younger adults to perform better than older adults (Ruffman, Henry, Livingstone, & Phillips, 2008). Since gender and age differences had also been found for the full GERT (Schlegel et al., 2014), we expected to replicate this in Study 1.

Study 2 was conducted in a laboratory setting with undergraduate students and aimed to examine the internal consistency of the GERT-S and its relationship with gender, as well as its convergent validity with two widely used ERA tests.

Study 1

Method

Participants and procedure

Participants were recruited through the survey division of Qualtrics (Copyright © 2013 Qualtrics, Provo, UT, USA) from a panel of people living in the USA that regularly participate in online surveys. They had signed up on a panel and received the invitation by e-mail. The study was announced as investigating "emotional experiences and emotional competence," consisting of three sessions to be completed online on three consecutive days. Participants received gift vouchers for their participation in each session. Of the 443 people that started the survey, 360 completed the first session including the GERT-S, 211 completed the second session including measures of emotional understanding and emotion management, and 159 completed the third session including the measure of cognitive ability. For each session, data of some participants was excluded because of very low scores or extremely fast response times suggesting that they did not complete the tasks as instructed (see Table 1 below for the final *N*s, and the Measures section for details).

In the original sample (*N* = 443), ages ranged from 18 to 65 years, with a mean age of 45.4 (*SD* = 12.1). A total of 214 participants (48.3 %) were male and 229 (51.7 %) were female. The ethnic composition of the original sample was as follows: 65 % Caucasian, 8 % African-American, 7 % Asian/Asian-American, and 5 % Latino/Latin-American. Seven percent reported an ethnic background different from the above, and 6 % had mixed ethnic backgrounds. Three percent chose not to respond to this question. For the subpart of the sample that completed all three sessions, the demographic characteristics were very similar to those of the original sample.

All three sessions were administered in English using the survey tool of the Qualtrics Research Suite (Copyright © 2013 Qualtrics, Provo, UT, USA). Participants were instructed to complete each session in a quiet environment. In addition to the instruments reported in this article, participants also completed newly developed and unpublished questionnaires on emotional knowledge (Session 1), emotional experience and emotion regulation (Session 2), and a study on face perception (Session 3). These data are not part of the present analysis.

Measures

Geneva Emotion Recognition Test short version (GERT-S; Schlegel et al., 2014) The GERT-S consists of 42 short video clips with sound (duration 1–3 s), in which ten professional French-Swiss actors (five male, five female) express 14 different emotions. After each clip, participants are asked to choose which of the 14 emotions best describes the emotion the actor intended to express. Responses are scored into correct (1) and incorrect (0), yielding a total average GERT-S score that can range from 0 to 1. The clips for this test were taken from the Geneva Multimodal Emotion Portrayals database, which includes 1260 portrayals of 18 emotions in different intensities (GEMEP; Bänziger, Mortillaro, & Scherer, 2012). For the original GERT and the GERT-S, 12 emotions had been selected to evenly cover the four quadrants in the

Table 1 Descriptive statistics, reliability coefficients, and correlations with GERT-S in Study 1 and Study 2

	N	Mean (*SD*)	Cronbach's Alpha	Correlation With GERT-S
Study 1				
GERT-S	350	.45 (.15)	.80	
CFIT	128	33.4 (7.84)	.89	.44 ($p < .001$)
STEU (short form)	193	.64 (.20)	.83	.60 ($p < .001$)
STEM (short form)	191	.59 (.18)	.73	.45 ($p < .001$)
gender	350 (173 male, 177 female)			.02 ($p = .710$)
age	350	45.6 (12.1)		.08 ($p = .127$)
Study 2				
GERT-S	75	.67 (.16)	.83	
DANVA faces	73	.77 (.12)	.65	.48 ($p < .001$)
DANVA voices	73	.74 (.11)	.47	.50 ($p < .001$)
gender	75 (28 male, 47 female)			.34 ($p < .01$)

GERT-S = Geneva Emotion Recognition Test, short version; CFIT = Culture Fair Intelligence Test; STEU = Situational Test of Emotional Understanding; STEM = Situational Test of Emotion Management; DANVA = Diagnostic Analysis of Nonverbal Accuracy.

emotional valence–arousal space proposed by the GEMEP authors (Bänziger et al., 2012): joy, amusement, pride—high arousal/positive valence; pleasure, relief, interest—low arousal/positive valence; anger, fear, despair—high arousal/negative valence; and irritation, anxiety, sadness—low arousal/negative valence. Disgust and surprise were added because they are frequently used in other emotion recognition tests and studies, yielding a total of 14 emotions. In each video clip, the actors are visible from their upper torso upward (conveying facial and postural/gestural emotional cues) and pronounce a sentence made up of syllables without semantic meaning (conveying emotional cues through their voice). The recording procedure of those clips was embedded in a constant interaction between the actor and a professional director based on real-life scenarios for each emotion, to ensure a maximal emotion induction and high authenticity of the emotional expressions (for a discussion on the use of actor portrayals in research, see Scherer & Bänziger, 2010). The GEMEP database was validated by Bänziger et al. (2012) in a study that obtained believability ratings and data on how accurately the target emotion in each clip was recognized by external judges. Using those data, the clips for the full version of the GERT had been selected by Schlegel et al. (2014) on the basis of two criteria; first, the target emotion—that is, the emotion that the actor was asked to express—had to be the most frequently chosen response category, and second, both the recognition accuracy and believability had to be above the 30th percentile of all portrayals in a given emotion category. Furthermore, all clips in the GERT had been produced in the normal-intensity condition (the GERT also contains lower- and higher-intensity conditions).

The 42 items in the GERT-S were selected on the basis of data reported by Schlegel et al. (2014; Schlegel, Fontaine, & Scherer, 2015) for the full 83-item GERT from two samples, with a total *N* of 426. Item discrimination indices (i.e., the correlation of each item with the total GERT score) and mean recognition percentages were computed. For each emotion, first the most discriminative item that had not been recognized by more than 90 % of the sample was chosen. Second, for each emotion a second item portrayed by an actor of the opposite gender was chosen, again on the basis of higher item discrimination and preferably higher difficulty. Finally, a third item was selected on the basis of the same discrimination and difficulty criteria from all (male and female) actors, resulting in 42 items for the GERT-S. The goal of this strategy was to select an internally consistent set of items with a difficulty level that would allow optimal discrimination between participants with different ERA levels. Actor gender in the final GERT-S was balanced, with 21 items being portrayed by male and 21 by female actors. In the sample that was used to select these items from the full GERT ($N = 426$), the correlation between the 42 items that were retained for the GERT-S and the 83 items of the full GERT was $r = .89$ ($p < .001$). The correlation between the 42 GERT-S items and the 41 items of the full GERT that were not retained was $r = .67$ ($p < .001$).

In the present study, the data of ten participants with scores below .07 were excluded because they were below the score to be expected from random guessing (.071), as well as below the lowest score achieved in the previous two samples for the full GERT (total $N = 426$), which was .10 (Schlegel, Fontaine, & Scherer, 2015; Schlegel et al., 2014). Consequently, the sample size for the GERT-S in this study was $N = 350$.

Culture Fair Intelligence Test, Scale 2 (CFIT; Cattell & Cattell, 1957) The CFIT Scale 2 is a widely used measure of fluid intelligence for adults and consists of two parallel forms. Each form contains four timed tasks that require inferring complex relationships between elements of figures. These tasks include completing figure series (3 min), classifying figures (4 min), completing figure matrices (3 min), and inferring conditions (4 min). Here, we used Form A. Given that previous researchers noted that the difficulty of the CFT Scale 2 might be somewhat low (Weiss, 2006), we added the three last items of each subtest from Form B at the end of the respective subtest in Form A, yielding 57 items in total. Responses were scored into correct (1) and incorrect (0) and were summed to form a CFIT total score. Here, we excluded participants who stopped responding after less than 60 s or who responded to less than five items for any subtest, suggesting that they did not complete the task as instructed. The CFIT total score was calculated only for particiants who completed all four subtests (N = 128).

Situational Test of Emotional Understanding (STEU; MacCann & Roberts, 2008) The STEU is a performance-based test measuring knowledge of emotion antecendents and emotion features. Each item either asks the participant which of five emotion words corresponds best to a short written scenario, or which of five short scenarios best corresponds to a given emotion word. The correct answers are based on theoretical grounds and responses are scored as correct (1) or incorrect (0) accordingly, yielding an overall mean score ranging from 0 to 1. Here, we used the 25-item short version of the STEU (MacCann & Roberts, 2008). Eighteen participants who completed the whole session including STEU and STEM (see below) in less than 6 min were excluded because it is very unlikely that they attentively read all items in such a short time.

Situational Test of Emotion Management (STEM; MacCann & Roberts, 2008) The STEM is a performance-based test measuring the ability to regulate emotions in other people. Participants are asked to read short scenarios and to choose from four responses the most effective course of action. The STEM can be scored into correct and incorrect responses, or each response can be given a weight derived from expert ratings. Here, we used the 20-item short version of the STEM with dichotomous scoring of the responses (MacCann & Roberts, 2008). Eighteen participants were excluded for their extremely short response times (see above), and two participants were excluded because they clicked on the first response option for 18 out of 20 items.

Data analysis

To evaluate the psychometric quality of the GERT-S, we analyzed its (1) internal consistency, (2) factor structure, (3) test and item difficulty, and (4) convergent and discriminant validity.

Internal consistency was assessed by evaluating factor saturation using two coefficients omaga, ω_h and ω_t (Revelle & Zinbarg, 2009). Both coefficients are calculated on the basis of a factor analysis of the data with oblique factor rotation and a Schmid Leiman transformation. The ω_h value indicates the proportion of test variance due to a general factor, whereas ω_t is the proportion of test variance explained by all factors. Both coefficients were calculated with the psych package for R (Revelle & Zinbarg, 2009) using tetrachoric correlations between the binary test items. We also report Cronbach's alpha because it is typically used by researchers in the ERA field (but see Revelle & Zinbarg, 2009, for a critique). The *factor structure* of the GERT-S was assessed by testing a one-factor CFA model specifying the 42 binary items (1 = correct, 0 = incorrect) to load on one latent ERA factor. This model was tested because previous analyses with the full GERT had suggested that the test was essentially unidimensional (Schlegel et al., 2012). The CFA was performed with Mplus (Muthén & Muthén, 2011) using the weighted least squares means and variance-adjusted estimator, and model fit was evaluated by inspecting the comparative fit index (CFI), the Tucker–Lewis Index (TLI), and the root mean square error of approximation (RMSEA).

Test and item difficulty were assessed using item response theory (IRT), which is a psychometric framework based on the idea that the probability of solving a test item depends on a person's ability and on item parameters such as item difficulty (Embretson & Reise, 2000). Here, we used the Rasch model (the simplest IRT model) that had also been applied for the development of the full GERT. It assumes that the probability of solving an item depends only on a person's location on a latent ability dimension θ and on the difficulty of an item, which are displayed on the same latent dimension. The location of an item on θ corresponds to the ability level at which this item discriminates best between participants—that is, at which it has the lowest error of measurement. By comparing the distributions of person parameters (ability scores) and of item difficulties on θ, it is possible to evaluate the difficulty level of the overall test as well its measurement precision in relation to the particular sample that was studied. Specifically, the *test information curve* shows the range of the latent dimension θ over which the test discriminates best among test-takers and has the lowest measurement error. The fit of the Rasch model was evaluated by inspecting the weighted fit, or "Infit," and unweighted fit, or "Outfit," index for each item, with values between 0.80 and 1.20 indicating "useful fit" (Wright & Linacre, 1994), and 1.00 representing perfect fit. Analyses were carried out with the eRm package in R (Mair & Hatzinger, 2007). Finally, c*onvergent and discriminant validity* were examined by calculating correlations between the GERT-S and the CFIT, STEU, STEM, age, and gender.

Results and discussion

The calculation of the coefficients omega showed that the variance percentage in the GERT-S explained by one general factor (ω_h) was 68 % and that the total reliable variance in the test explained by all factors (ω_t) was 89 %. Cronbach's alpha was .80. These results suggest that although additional factors explain more variance, the contribution of one general factor is strong. This was confirmed by the one-factor CFA model that showed good fit ($\chi^2 = 920.662$, $df = 819$, $p = .008$, CFI = .952, TLI = .950, RMSEA = .019). It can therefore be concluded that the GERT-S is a unidimensional test, which confirmed previous results (Schlegel et al., 2012) and created the precondition for conducting the IRT analysis. The Infit indices for the 42 items ranged from 0.81 to 1.31, and the Outfit indices ranged from 0.86 to 1.17. Four items had Infit values that were slightly above the boundary of 1.20 indicating useful fit, but given that the respective Outfit values were below 1.20, the fit of the Rasch model to the GERT-S can be considered as satisfactory overall. Figure 2 shows the item difficulty parameters as well as the sample's ability estimates on the same metric of the latent dimension θ. The ability estimates for the individual participants (with the mean fixed to zero and a standard deviation of 1) ranged from –2.22 to 1.77 (left side of the figure). The item difficulties (right side of the figure) ranged from –1.47 for the item dis9 (the easiest item) to 2.03 for the item dis8 (the most difficult item) and were well distributed over the range of θ. The mean of the 42 item difficulties was 0.29, which is slightly above the mean ability level of the studied sample. Accordingly, the test information curve had its maximum at the same point, indicating that the GERT-S provides its most precise measurement at an ability level of less than a third of an *SD* above the sample mean. Taken together, these results suggest that the GERT-S has an appropriate level of difficulty for the studied sample, allowing for a precise measurement of ERA for most tested participants.

The correlations of the GERT-S with other instruments, as well as descriptive statistics and reliability coefficients for all measures, are presented in Table 1. As predicted, the GERT-S

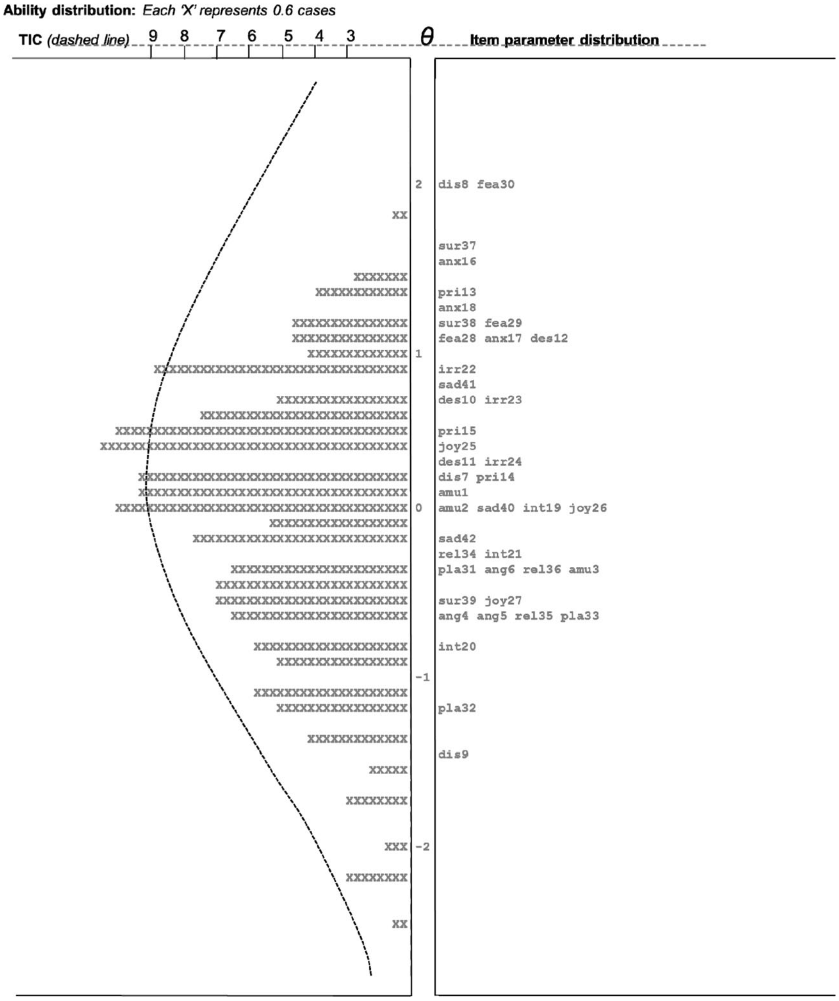


Fig. 2 Person–item map displaying the distribution of the sample's ability estimates and the test information curve (TIC), on the left side, and the item difficulties, on the right side, of the latent dimension θ. The labels on the right side represent the item numbers (1 to 42) and the respective emotion category: amu = amusement, ang = anger, dis = disgust, des = despair, pri = pride, anx = anxiety, int = interest, irr = irritation, fea = fear, ple = pleasure, rel = relief, sur = surprise, sad = sadness. The mean of the ability distribution was fixed to zero

was substantially positively correlated with other emotional competencies (emotion management as measured by the STEM, and emotional understanding as measured by the STEU) and with cognitive ability as measured by the CFIT. Given that the correlation of the GERT-S with the CFIT was of a magnitude similar to the test's correlations with the STEU and STEM, we calculated partial correlations with the STEU and STEM while controlling for CFIT scores. The partial correlation was .48 ($p < .001$) between GERT-S and STEU and .33 ($p < .001$) between GERT-S and STEM. These results suggest that the GERT-S is related to emotion management and emotional understanding skills beyond what can be explained by general cognitive ability. These findings support the theoretical notion that ERA might be the most basic component of EI and is crucial for good performance in the more complex components, such as understanding and managing emotions in other people (Mayer & Salovey, 1997).

Contrary to our predictions, age and gender were not correlated with GERT-S performance. To date, few studies have investigated the relationship between age and emotion recognition in multimodal, dynamic emotional expressions. Some researchers have suggested that the information in such stimuli is richer and more ecologically valid, and can thus compensate for the decline with age observed when using still pictures of emotional faces (Phillips & Slessor, 2011). In addition, an age-related decline in ERA has often been observed only in adults above the age of 65 (Ruffman et al., 2008). In the present study, ages ranged from 18 to 65, and all individuals in this age group might perform equally well. As for the gender effect in ERA, the effect size in favor of women is generally small and has not always been found (Hall, 1978).

Taken together, this study provides evidence for the excellent psychometric quality of the GERT-S in a large sample drawn from an online panel of respondents. The goal of Study 2 was to replicate the test's internal consistency in a different sample and setting, and to evaluate its convergent validity with other ERA tests. More specifically, in Study 2 the GERT-S was administered to undergraduate psychology students in a laboratory setting, along with two widely used ERA tests.

Study 2

Method

Participants and procedure

A total of 75 psychology undergraduate students at Northeastern University participated in this study for course credit. Of these, 28 were male (37 %) and 47 were female (63 %). Ages ranged from 18 to 25 ($M = 19.6$, $SD = 1.5$). In all, 57 % were Caucasian, 26 % were Asian, and 12 % were Latin-American or Latino. The remainder of the sample reported a different ethnic background or chose not to respond to this question. Participants completed the GERT and DANVA (faces and voices) tests in a random order as part of a larger study that investigated a new method to train emotion recognition. All measures were administered through Qualtrics in a small laboratory with two desktop computers. Maximally, two participants were tested at the same time.

Measures

GERT-S The GERT-S was presented in the same version as in Study 1.

Diagnostic Analysis of Nonverbal Accuracy (DANVA; Nowicki & Duke, 1994) Here, we administered the DANVA faces and voices subtests. The DANVA faces consist of 24 photographs of facial expressions of students that express happiness, sadness, anger, or fear. The DANVA voices consist of 24 audio recordings in which actors say the sentence "I am going out of the room now but I'll be back later" in a happy, fearful, sad, or angry tone. After each picture or recording, participants are asked to choose which of the four emotions had been expressed. Responses are scored as correct (1) or incorrect (0) and form a total score for each of the two subtests. Two participants did not complete the DANVA tests for technical reasons.

Data analysis

Due to the small sample size in this study, it was not feasible to run IRT analyses as well as factor analyses—hence, it was not feasible to report omega, which is based on factor analysis. We therefore only report Cronbach's alpha. The relationship between the GERT-S and the DANVA subtests was analyzed with Pearson correlations.

Results and discussion

The descriptive statistics and Cronbach's alphas of all instruments, as well as the correlations of DANVA and gender with the GERT-S, are presented in Table 1. The Cronbach's alpha of the GERT-S was .83, which is slightly higher than in Study 1. Together with the results of Study 1, the internal consistency of the GERT-S can be considered rather good relative to other tests in the ERA domain. In a recent meta-analysis, the average reported reliability (Cronbach's alpha) of ERA tests was .60 (Boone & Schlegel, 2015). In addition, the high Cronbach's alpha in Study 2 indicates that the GERT-S also differentiates well between participants of a rather homogeneous sample of similar age and educational background. As predicted, both DANVA tests were substantially positively correlated with the GERT-S. This suggests that the GERT-S as a multimodal ERA test taps both vocal and facial ERA, providing further evidence for its construct validity. In this study, we also found a small gender difference, with women performing somewhat better than men, which supports previous meta-analytic findings (Hall, 1978). Taken together, this study provided further support for the good psychometric quality of the GERT-S in a setting and sample that were very different from those of Study 1. Notably, the sample in Study 2 substantially outperformed the sample in Study 1 (see Table 1). Several reasons might account for this difference. First, the sample in Study 2 consisted of young undergraduate students who are selected by schools on the basis of their SAT or ACT scores and presumably have higher cognitive abilities than the average (older) population. As was shown in Study 1, higher cognitive ability is related to better performance on the GERT-S. Second, half of the sample in Study 2 completed a short training intended to improve ERA before taking the GERT-S and DANVA tests, which improved this subsample's performance on the GERT-S as compared to the untrained participants. The mean GERT-S score of the untrained participants was .60 (*SD* = .15), which is closer to the score of the sample in Study 1. Additional factors that might have contributed to the difference in GERT-S scores between the two studies include differences in motivation or interest (e.g., due to the presence of an experimenter, or due to being enrolled in a social psychology course), being in a laboratory versus in an uncontrolled setting, and being used to completing similar judgment tasks to obtain course credit.

General discussion

Recently, the GERT has been proposed as a new performance-based test to measure individual differences in people's ability to accurately detect and interpret emotional expressions in others (Schlegel et al., 2014). In contrast to previous ERA tests, the GERT is entirely based on multimodal video clips containing facial, vocal, and bodily cues, and on a large number of both positive and negative emotions. In consequence, the GERT arguably has better content validity than existing tests that largely focus on static facial expressions and basic emotions. Furthermore, the GERT has been shown to have a good psychometric quality with a unidimensional factor structure and good internal consistency.

The present article introduced and validated a short version of the GERT, the GERT-S. The GERT-S consists of 42 items that were selected from the full GERT and takes about 10 min to complete. In two studies, we showed that the GERT-S has good internal consistency and a unidimensional factor structure. Moreover, we found evidence for the convergent validity of the test as indicated by substantial correlations with other ERA tests, tests of other emotional competencies (emotional understanding and emotion management), and cognitive ability. These results were found in two diverse settings and samples, namely in an online study with online survey panel members of a wide age range as well as in a laboratory setting with undergraduate students. In both studies, the GERT-S demonstrated good psychometric quality as indicated by high internal consistency and theoretically meaningful correlations with other tests.

Taken together, our results suggest that the GERT-S can be considered a measure of general ERA within the framework of EI or EC. Because it contains only three items per emotion, the GERT-S is less suitable as a measure of emotion-specific emotion recognition skills and we generally recommend using only the total score. This is supported by the unidimensional structure of the test, showing that a person's performance in recognizing any of the included emotions is influenced by a common underlying ability.

The GERT-S is administered online and is available in different language versions free of charge to researchers for academic purposes (details are available at www.affective-sciences.org/content/geneva-emotion-recognition-test-gert). In addition, the GERT-S can be integrated by researchers in their own surveys through different online survey tools. With these options, the short test-taking time, and its good psychometric properties, the GERT-S is a useful addition to the range of available ERA tests. It can be recommended when researchers want to measure participants' general ERA across emotions and modalities. However, if scores for the different emotions are needed, we recommend using the full GERT because of its larger number of items per emotion.

There are many potential applications for this test in different domains of psychology. For example, the GERT-S can be used in aging research to study a potential decline in ERA with increasing age. Relevant work in this field to date heavily relied on emotion recognition of static facial expressions. This is problematic as it has been suggested that age differences might be attenuated when more ecologically valid tests are used (Phillips & Slessor, 2011). Our own findings concerning this issue—Schlegel et al. (2014), where we found a decline in ERA for older adults, versus Study 1 of the present article, where we did not find any age difference up to 65 years—suggest that, if anything, the relationship is complex, possibly involving curvilinear functions or moderating factors. Another field that is likely to benefit from using the GERT-S is assessment in organizational and work psychology. Due to the growing popularity of the EI construct, there has been increasing interest in investigating ERA as a predictor of workplace performance (Elfenbein et al., 2007). At the same time, there is a shortage of performance-based EI measures that meet the criteria for good psychometric quality (Roberts et al., 2010). In the absence of more comprehensive assessment data, the GERT-S as a brief measure of one of the most basic EI components might serve as a proxy measure for EC. Other fields in which this test can potentially be applied include clinical psychology and rehabilitation, social psychology, and affective neuroscience.

Future research on the GERT and the GERT-S should focus on further investigating their construct and predictive validity. In particular, it should be systematically examined to what extent these tests complement other existing ERA tests and correlate with tests for other components of EI/EC, as well as with measures of general intelligence and personality. Studying this nomological network would allow carving out more precisely what facets of EC the GERT measures and that social outcomes it is therefore likely to predict. Although first evidence shows that the GERT predicts better performance in interpersonal negotiation (Schlegel, Mehu, et al., 2015), other studies will be needed to assess the extent to which it is related to outcomes in other social contexts, such as close relationships or group interaction.

Because several language versions of the GERT exist, future research should compare these with respect to their psychometric quality and potential intercultural differences in test difficulty. Although a recent review of studies on the universality and cultural specificity in the expression and perception of emotion in different modalities (Scherer et al., 2011) showed that ERA is quite stable across different cultures, evidence is growing of "dialect" effects on emotion expression and recognition (e.g., Elfenbein, Lévesque, Beaupré, & Hess, 2007). In conclusion, we hope that the GERT-S, as a brief ERA assessment instrument with good psychometric properties, can contribute to the further development of the sorely needed repository of tests for EI/EC.

Author note This research was supported by an ERC Advanced Grant in the European Community's 7th Framework Program, under Grant Agreement 230331-PROPEREMO (Production and perception of emotion: an affective sciences approach) to K.R.S.

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

Boone, R. T., & Schlegel, K. (2015). Is there a general skill in perceiving others accurately? In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately*. Cambridge, UK: Cambridge University Press.

Cattell, R. B., & Cattell, A. K. S. (1957). *Test of "g": Culture Fair (Scale 2, Form A)*. Champaign, IL: Institute for Personality and Ability Testing.

Elfenbein, H. A., & Ambady, N. (2002). Predicting workplace outcomes from the ability to eavesdrop on feelings. *Journal of Applied Psychology, 87,* 963–971.

Elfenbein, H. A., Foo, M. D., White, J., Tan, H. H., & Aik, V. C. (2007a). Reading your counterpart: The benefit of emotion recognition accuracy for effectiveness in negotiation. *Journal of Nonverbal Behavior, 31,* 205–223.

Elfenbein, H., Lévesque, M., Beaupré, M., & Hess, U. (2007b). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion, 7,* 131–146.

Elfenbein, H. A., Marsh, A. A., & Ambady, N. (2002). Emotional intelligence and the recognition of emotion from facial expressions. In L. F. Barrett & P. Salovey (Eds.), *The wisdom in feeling: Psychological processes in emotional intelligence* (pp. 37–59). New York, NY: Guilford Press.

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

Frank, M. G., & Stennett, J. (2001). The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology, 80,* 75–85.

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.

Herzmann, G., Danthiir, V., Schacht, A., Sommer, W., & Wilhelm, O. (2008). Toward a comprehensive test battery for face cognition: Assessment of the tasks. *Behavior Research Methods, 40,* 840–857. doi:10.3758/BRM.40.3.840

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.

MacCann, C., & Roberts, R. D. (2008). *The brief assessment of emotional intelligence: Short forms of the situational test of emotional understanding (STEU) and situational test of emotion management (STEM)*. Princeton, NJ: Educational Testing Service.

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 & Biobehavioral Reviews, 32,* 454–465. doi:10.1016/j.neubiorev.2007.08.003

Matsumoto, D., LeRoux, J., 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.

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 3–31). New York, NY: Basic Books.

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review, 90,* 215–238.

Murphy, N. A., & Hall, J. A. (2011). Intelligence and interpersonal sensitivity: A meta-analysis. *Intelligence, 39,* 54–63.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

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.

Palermo, R., O'Connor, K. B., Davis, J. M., Irons, J., & McKone, E. (2013). New tests to measure individual differences in matching and labelling facial expressions of emotion, and their association with ability to recognise vocal emotions and facial identity. *PLoS ONE, 8,* e68126. doi:10.1371/journal.pone.0068126

Phillips, L., & Slessor, G. (2011). Moving beyond basic emotions in aging research. *Journal of Nonverbal Behavior, 35,* 279–286.

Puccinelli, N. M., & Tickle-Degnen, L. (2004). Knowing too much about others: Moderators of the relationship between eavesdropping and rapport in social interaction. *Journal of Nonverbal Behavior, 28,* 223–243.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74,* 145–154.

Roberts, R. D., MacCann, C., Matthews, G., & Zeidner, M. (2010). Emotional intelligence: Toward a consensus of models and measures. *Social and Personality Psychology Compass, 4,* 821–840.

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.

Scherer, K. R. (2007). Component models of emotion can inform the quest for emotional competence. In G. Matthews, M. Zeidner, & R. D. Roberts (Eds.), *The science of emotional intelligence: Knowns and unknowns* (pp. 101–126). New York, NY: Oxford University Press.

Scherer, K. R., & Bänziger, T. (2010). On the use of actor portrayals in research on emotional expression. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 166–178). Oxford, UK: Oxford University Press.

Scherer, K. R., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology, 46,* 401–435.

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. *Journal of Nonverbal Behavior, 35,* 305–326.

Scherer, K. R., Shuman, V., Fontaine, J. R. J., & Soriano, C. (2013). The GRID meets the wheel: Assessing emotional feeling via self-report. In J. R. J. Fontaine, K. R. Scherer, & C. Soriano (Eds.), *Components of emotional meaning: A sourcebook* (pp. 281–298). Oxford, UK: Oxford University Press.

Schlegel, K., Fontaine, J. R., & Scherer, K. R. (2015). Psychometric quality and construct validity of the French and Dutch version of the Geneva Emotion Recognition Test. Manuscript in preparation.

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.

Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the Geneva Emotion Recognition Test: An example of Rasch-based test development. *Psychological Assessment, 26,* 666–672.

Schlegel, K., Mehu, M., van Peer, J. M., & Scherer, K. R. (2015). Sense or sensibility: Which is more important for successful negotiation? Manuscript submitted for publication.

Tickle-Degnen, L. (1998). Working well with others: The prediction of students' clinical performance. *American Journal of Occupational Therapy, 52,* 133–142.

Weiss, R. H. (2006). Grundintelligenztest Skala 2 (CFT 20–R) mit Wortschatztest (WS) und Zahlenfolgentest (ZF). *Intellektuelle Hochbegabung, 80.*

Wilhelm, O., Hildebrandt, A., Manske, K., Schacht, A., & Sommer, W. (2014). Test battery for measuring the perception and recognition of facial expressions of emotion. *Frontiers in Psychology, 5,* 404. doi:10.3389/fpsyg.2014.00404

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