Effectiveness of a short audiovisual emotion recognition training program in adults

Katja Schlegel[1]

Ishabel M. Vicaria[2]

Derek M. Isaacowitz[2]

Judith A. Hall[2]

[1]Institute of Psychology, University of Bern, Switzerland

[2] Northeastern University, Boston, Massachusetts, USA

This research was supported by an Early Postdoc.Mobility fellowship awarded to K. Schlegel by the Swiss National Science Foundation, grant number P2GEP1_155698.

Correspondence regarding this article should be addressed to Katja Schlegel, Institute of Psychology, University of Bern, Fabrikstrasse 8, CH-3012 Bern, Switzerland, email: katja.schlegel@psy.unibe.ch.

Parts of the data in this manuscript have been presented at the 17th Annual Convention of the Society of Personality and Social Psychology in February 2016, at the 27th Annual Convention of the Association for Psychological Science in May 2016, and at the 2015 Meeting of the International Society for Research on Emotion.

Abstract

The ability to recognize emotions from others' nonverbal behavior (emotion recognition ability, ERA) is crucial to successful social functioning. However, currently no self-administered ERA training for non-clinical adults covering multiple sensory channels exists. We conducted four studies in a lifespan sample of participants in the laboratory and online (total $N = 531$) to examine the effectiveness of a short computer-based training for 14 different emotions using audiovisual clips of emotional expressions. Results showed that overall, young and middle-aged participants that had received the training scored significantly higher on facial, vocal, and audiovisual emotion recognition than the control groups. The training effect for audiovisual ERA persisted over four weeks. In older adults (59 to 90 years), however, the training had no effect. The new, brief training could be useful in applied settings such as professional training, at least for younger and middle-aged adults. In older adults, improving ERA might require a longer and more interactive intervention.

Keywords: Emotion recognition, training, aging, emotional competence, emotion perception

Effectiveness of a short audiovisual emotion recognition training program in adults

The perception of others' nonverbal cues and the attributions people make based on these perceptions are fundamental and adaptive mechanisms in human interaction (McArthur & Baron, 1983). For example, accurately recognizing the type and intensity of another person's emotional state from facial, vocal, and bodily cues allows to anticipate his or her actions, to adapt one's own actions accordingly, and consequently, to make the interaction more predictable and manageable (Van Kleef, 2010; Hall, Mast, & West, 2016). Although non-clinical individuals are often relatively accurate at making judgments about others' emotions, important individual differences in this ability, hereafter referred to as emotion recognition ability (ERA), have been observed. Researchers in various domains of psychology, such as clinical, developmental, organizational, and social psychology have studied the correlates and effects of individual differences in ERA. ERA has also been proposed as a central component in different models of emotional and social intelligence (e.g., Mayer & Salovey, 1997; Scherer, 2007). Previous studies showed that higher ERA is associated with a wide range of psychosocial benefits, such as better relationship quality, mental health, social adjustment, social skills, mental health, and academic and workplace performance (Hall, Andrzejewski, & Yopchick, 2009; Elfenbein, Foo, White, Tan, & Aik, 2007; Schlegel, Grandjean, & Scherer, 2013). Conversely, lower ERA is related to maladaptive traits such as trait anger, anxiety, and alexithymia (Schlegel, Fontaine, & Scherer, 2017), and is characteristic of various mental disorders including schizophrenia (Kohler, Walker, Martin, Healey, & Moberg, 2010), bipolar disorder (Derntl, Seidel, Kryspin-Exner, Hasmann, & Dobmeier, 2009), and borderline personality disorder (Domes, Grabe, Czieschnek, Heinrichs, & Herpertz, 2011).

Given the multitude of positive correlates associated with higher ERA, an important question is whether ERA can be improved through training. For example, training ERA could benefit various professions in which ERA has been found to predict job outcomes, such as

healthcare providers, security officers, customer service representatives, and managers (e.g., (Blanch-Hartigan, 2012; Hall, 2011; Hurley, Anker, Frank, Matsumoto, & Hwang, 2014; Rubin, Munz, & Bommer, 2005; Matsumoto & Hwang, 2011). Another large group of the general population that might specifically benefit from such training is older adults. Many studies have found a substantial decline in emotion recognition accuracy on standard tests of face, voice, and body ERA (for a meta-analysis, see Ruffman, Henry, Livingstone, & Phillips, 2008). However, older adults do not report poorer interpersonal functioning or reduced life quality (Lansford, Sherman, & Antonucci, 1998). Because studies that use dynamic (i.e., audiovisual clips) as opposed to static (i.e., photographs) stimuli do not show such drastic declines in performance, researchers are now exploring more ecologically valid assessments of older adults' social perception skills (Krendl & Ambady, 2010).

A rich history of training programs in other domains of interpersonal perception such as deception detection suggests that ERA might be a trainable skill (for an overview, see Blanch-Hartigan, Andrzejewski, & Hill, 2012; Blanch-Hartigan, Andrzejewski, & Hill, 2016). However, to date surprisingly few interventions have been developed for improving ERA, especially for non-clinical, healthy populations. Currently, ERA training programs exist for certain clinical populations, such as children with autism spectrum disorder (Golan et al., 2010), adults suffering from schizophrenia (Silver, Goodman, Knoll, & Isakov, 2004), or patients with body dysmorphic disorder (Buhlmann, Gleiß, Rupf, Zschenderlein, & Kathmann, 2011). These programs are targeted at improving emotion perception on a very basic level, using only few emotion categories and focusing on static facial expressions. Furthermore, specific programs have been developed to train healthcare providers to recognize affective cues in patients (Blanch-Hartigan, 2012; Ruben, Hall, Curtin, Blanch-Hartigan, & Ship, 2015). In addition, Matsumoto and colleagues (Matsumoto & Hwang, 2011; Hurley, 2012; Hurley et al., 2014) have created a training to teach the detection of facial micro-expressions of basic emotions that has successfully been used with security officers.

Finally, emotion perception is also part of some broad and comprehensive socio-emotional training programs for school children (Rivers, Brackett, Reyes, Elbertson, & Salovey, 2013) and for adults (Herpertz, Schütz, & Nezlek, 2016).

However, to date no training that specifically and exclusively focuses on ERA in a non-clinical lifespan sample exists. In fact, in their meta-analysis on person perception accuracy interventions in young adults, Blanch-Hartigan et al. (2012) did not identify a single study on emotion recognition. The main goal of the present research was therefore to validate a new training intended to improve ERA in multiple sensory modalities (face, voice, and body) for a wide range of emotions that can be used with the general adult population for many different purposes. This training was designed as a self-administered and short computer program in order to facilitate its future applicability for both research and applied settings.

The design of this new training was informed by Blanch-Hartigan et al.'s (2012) meta-analytic findings regarding the effectiveness of different training elements. Specifically, their analysis revealed that trainings combining *practice* with *feedback* regarding the correctness of one's responses are particularly effective, and that *instructions* about the cues signaling the correct responses can potentially further enhance trainings effectiveness, especially in participants with lower baseline. Further, they found that length of the training did not influence effectiveness, and that even interventions of one hour or less can yield a substantial increase in performance. Accordingly, we developed a short, computer-based, and self-administered training program incorporating the elements *instruction* (descriptions of facial, vocal, and bodily cues signaling each of 14 emotions and example video clips for each emotion), *practice* (guessing which emotion was being expressed for a set of short video clips produced by actors), and *feedback* (learning whether a chosen emotion was correct or not and seeing the correct answer after two incorrect guesses). We chose a self-led format because it allows for a very flexible usage of the training at a low cost which can be an advantage in

various applied settings. As it does not require the presence of an instructor, it can be completed online and can be easily implemented in curricula of professional trainings, e.g., for nurses or teachers. Previous studies demonstrated that self-led trainings can improve person perception (e.g., Hurley, 2012).

Here, we validate this new ERA training in four studies while also addressing the following limitations of previous research on training person perception: First, previous studies typically used the same or similar stimuli both to train the respective skill and to assess the training outcome. For example, Blanch-Hartigan (2012) trained participants to recognize affective cues in patients using one half of a standard test to measure affect recognition in patients (the Patient Emotion Cue Test, PECT; Blanch-Hartigan, 2011), and measured training outcomes using the other half of the PECT, i.e., with similar stimuli. In their meta-analysis, Blanch-Hartigan et al. (2012) found that training effects were generally significant only when outcomes were measured with the same or similar stimuli as those used in the training; interpersonal perception training had no significant effects on tests using different stimulus material. Previous research therefore implies that interpersonal perception trainings might not have transfer effects beyond the specific training material, but results are inconclusive because many studies did not assess such transfer effects.

As a second limitation, to our knowledge no study to date has trained interpersonal perception skills such as ERA in older adults. As older adults tend to score lower on ERA than younger adults and might focus their attention differently when perceiving emotions compared to younger adults (Murphy & Isaacowitz, 2010), the same training might not be equally efficient in both groups. However, to date it remains largely unknown whether lower baseline accuracy in person perception will be associated with lower or higher training effects. In their meta-analysis, Blanch-Hartigan et al. (2012) speculated that individuals with lower baseline accuracy might benefit more from training as there is more room for

improvement. This would mean that older adults should benefit more from ERA training than younger adults if their average ERA is lower.

The present research

Four studies were conducted to assess the effectiveness of the new ERA training (Training Emotion Recognition Ability, TERA), in each of which participants were randomly assigned to either the TERA group or one of various control conditions and completed ERA tests after the intervention. In order to address the question of transfer effects to different measures of ERA, each study used three different ERA tests as outcome measures; one test using multimodal emotional expressions similar to the training context, one test using static facial expressions of basic emotions, and one test using vocal expressions of basic emotions. Given that the new intervention trains ERA in multiple modalities and for a wide range of (more than just basic) emotions, we assume that trained participants will also perform better on the facial and vocal ERA tests with fewer emotions.

In Study 1, ERA performance of the TERA group was compared to an untreated control group to establish first evidence for the effectiveness of the intervention. This study was conducted in the laboratory with undergraduate students. In Study 2, two additional control conditions were tested to rule out that first, TERA was effective simply because it increased participants' familiarity with, and awareness of, emotional expressions, and second, the training was effective because it enhanced participants' ability to discriminate between different categories independent of the emotional content. This study was conducted online on Amazon's Mechanical Turk platform. In Study 3, a slightly modified version of the training was tested in a large sample of undergraduate students, some of who completed a follow-up session four weeks after the intervention in order to examine long-term training effects. Finally, in Study 4, the same training version as in Study 3 was tested in a sample of community-dwelling older adults in the laboratory.

Study 1

The goal of this study was to test the first version of the TERA and to examine whether trained individuals show higher ERA performance in comparison to individuals in an untreated control group.

Method

Participants. One hundred three undergraduate students at Northeastern University (61% female) with a mean age of 19.56 (*SD* = 1.72) participated in the study for partial course credit in an introductory psychology course. Students represented a wide variety of different academic majors. Ethnic composition was as follows: Fifty-one percent White, 27% Asian, seven percent Hispanic, three percent Black, three percent Arabic, and eight percent reported mixed ethnicity or chose not to report their ethnicity.

Procedure. Upon arrival in the laboratory, participants were randomly assigned to either the TERA group (*N* = 52) or the untreated control group (*N* = 51). Participants in the training group first underwent the TERA (duration about 35 minutes) and then completed three ERA tests that were presented in a random order: the Geneva Emotion Recognition Test short form (GERT-S; Schlegel & Scherer, 2016), the Diagnostic Analysis of Nonverbal Accuracy adult faces test (DANVA2-AF, hereafter referred to as "DANVA Face"; Nowicki, 2006), and the Diagnostic Analysis of Nonverbal Accuracy adult voices test (DANVA2-AV, hereafter referred to as "DANVA Voice"; Nowicki, 2006). Participants in the untreated control group filled in several personality questionnaires that were unrelated to the present study and roughly required the same amount of time and concentration as the training. They them completed the same three ERA tests as the training group in random order.

Materials.

Training for Emotion Recognition Ability (TERA). The TERA consists of two parts, an instruction part and a practice-with-feedback part (total duration about 35 minutes). It uses short video clips with sound (duration 1-3 s) from the validated Geneva Multimodal Emotion Portrayals (GEMEP) database (Bänziger, Mortillaro, & Scherer, 2012) in which 10 younger,

middle-aged, and older actors (5 female, 5 male) express 18 different emotions while saying a sentence without meaning in a pseudo-language. Each clip shows the actor's upper body and head, and thus conveys facial, postural, gestural, and vocal information. The clips were recorded in an interactive setting with a director in which the actors acted out each emotion based on scenarios that they had received prior to the session. This acting technique was used to ensure that the emotional expressions were as authentic as possible. For the training, 14 of the 18 emotions from the GEMEP were used (6 positive: pride, amusement, joy, pleasure, relief, interest; 7 negative: anger, irritation, fear, anxiety, disgust, despair, sadness; and surprise as an emotion of neutral valence). Twelve of them were selected to cover evenly the four quadrants in the emotional valence-arousal space (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; irritation, anxiety, sadness – low arousal/ negative valence. Disgust and surprise were added because they are frequently studied in the emotion field, yielding a total of 14 emotions.

In the instruction part of the training, for each of the 14 emotions participants see 1) a written description of the meaning of the word (e.g., for sadness "Sadness typically occurs after the loss of a person, place, or thing and describes a state of unhappiness and misery with low physical arousal"), 2) a description of the nonverbal cues that can signal this emotion (e.g., for sadness "Slow and low voice, lip corners pulled down, inner eyebrows lifted, frowning, slouched posture, arms hanging, little body movement", and 3) two example videos from the GEMEP. The two example videos were chosen among the best recognized portrayals for each emotion based on data from a pilot study (Bänziger et al., 2012) in order to provide relatively unambiguous displays to illustrate the respective nonverbal cues. The two selected videos were portrayed by different actors. Participants were instructed to pay attention to the nonverbal cues that had been described earlier and were also told that the two videos for one emotion can be somewhat different because people can express the same emotion in different

ways. Participants could watch each example video up to three times and were able to reread the description of the nonverbal cues while watching. The nonverbal cue descriptions were created based on a literature search about the facial, vocal, and bodily cues associated with these emotions. For some less commonly studied emotions such as interest and relief the descriptions were additionally based on observations of the example videos.

The instruction part of the training is followed by a practice-with-feedback part in which participants watch 42 GEMEP clips and are asked, after each clip, to guess which of the 14 emotions had been expressed by the actor or actress. Participants make their choice by clicking on one of the 14 emotion words that are arranged in a circle roughly corresponding to the affective valence-arousal circumplex to facilitate participants' orientation among the options. After making a choice, participants receive feedback about whether their choice was correct or not. If their response is correct, they are told to continue with the next clip; if their response is incorrect, they are told to watch the same clip again and make a new choice. If their response is still incorrect, participants see the correct emotion on the screen and then proceed to the next clip. The 42 clips in this part include three clips for each of the 14 emotions. The three clips per emotion were selected to cover different levels of difficulty to make the training useful for individuals with lower and higher baseline ERA. Thirty-six of the 42 clips were selected from the Geneva Emotion Recognition Test (GERT; Schlegel, Grandjean, & Scherer, 2014; see below for a description), which is a standard ERA test consisting of GEMEP portrayals. The 36 selected clips did not overlap with the items of the short form of the GERT (GERT-S; Schlegel & Scherer, 2016; see below for a description). For six emotions, one clip each was additionally selected from the GEMEP for the training because the respective GERT clips were either too easy or too hard; yielding a total of 42 clips in the practice with feedback part of the training. The difficulty level of each item (i.e., recognition rate) based on which the 42 clips were selected was taken from the studies conducted by Schlegel et al. (2014) and Bänziger et al. (2012). These clips were not the same

ones appearing in the example videos in the instruction part of the training. The TERA is available for academic research purposes upon request.

Geneva Emotion Recognition Test short form (GERT-S; Schlegel & Scherer, 2016). The GERT-S is a standard test to measure ERA and consists of 42 short video clips with sound in which 10 actors express 14 different emotions (pride, amusement, joy, pleasure, relief, interest, anger, irritation, fear, anxiety, disgust, despair, sadness, surprise). These clips were taken from the GERT (Schlegel et al., 2014) that uses clips from the GEMEP corpus (Bänziger et al., 2012) and the 14 emotions are the same as in the TERA. However, no clip from the training appears in the GERT-S. All 42 clips are multimodal, i.e., they show the upper body with arms and hands as well as the head and face, and they include the voice of the actor. 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 as correct (1) or incorrect (0), yielding a total average GERT-S score that can range from zero to one and represents the proportion of items judged correctly. Correct answers correspond to the emotion that the actor had been instructed to portray in the GEMEP database. Eight studies including the GERT and the GERT-S conducted in German, French, English, and Dutch provided substantial evidence for the high reliability (Cronbach's alpha around .80 in all studies) and construct validity of these tests (Schlegel et al., 2014; Schlegel & Scherer, 2016; Schlegel et al., 2017). Specifically, they were positively correlated with a range of other ERA tests, performance measures of emotional intelligence, and with adaptive self-reported personality traits (e.g., empathy), whereas they were negatively correlated with maladaptive affective traits such as alexithymia, trait anxiety, and trait anger. In the present four studies (combined in one dataset), Cronbach's alpha was .83. Over all four studies, a small but significant gender difference favoring women was found in line with previous studies ($r = .09$, $p < .05$).

Diagnostic Analysis of Nonverbal Accuracy (DANVA) Adult Face and Adult Voice tests (Nowicki, 2006). The DANVA Face test consists of 24 photographs of facial expressions of students that express happiness, sadness, anger, or fear. The DANVA Voice test consists 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. The DANVA tests have been very widely used since the development of the first version in the late 1980ies (see Nowicki & Duke, 1994) and showed predictive validity for a variety of positive psychosocial outcomes (e.g., see meta-analysis by Hall et al., 2009). Cronbach's alpha for the DANVA Face in Studies 1 and 2 (it was not used in studies 3 and 4) was .78, and Cronbach's alpha for the DANVA Voice test across all four studies was .65. Over Studies 1 and 2, women performed better on the DANVA Face ($r = .18$, $p < .01$). No gender differences were found on the DANVA Voice across the four studies ($r = 0.05$, $p = .299$). In Study 1, the DANVA Face and Voice test were correlated at $r = .42$ ($p < .001$). The GERT-S correlated with the DANVA Face at $r = .45$ ($p < .001$) and with the DANVA Voice at $r = .49$ ($p < .001$).

Results and Discussion

Descriptive statistics for the ERA training and control groups on the three ERA tests are shown in Table 1. Independent-samples *t*-tests showed that the training group performed significantly better on the GERT-S, $t(101) = 5.94$, $p < .001$, and the DANVA Voice, $t(99) = 2.16$, $p = .03$, than the untreated control group. The training group also performed better on the DANVA Face test, but the difference between the two groups was not significant, $t(99) = 1.47$, $p = .15$. Given that overall women tend to perform better on ERA tests than men, possible gender effects in training effectiveness were explored in a two-way ANOVA with group and gender as factors. Results did not reveal a significant group by gender interaction

for any ERA test, suggesting that the training is similarly effective in men and women. These results provided first evidence for the effectiveness of the TERA in improving multimodal, vocal, and to some extent facial ERA.

Study 2

The goal of Study 2 was to compare the effectiveness of the TERA against two additional control conditions that allowed ruling out potential alternative explanations for increased ERA performance in the training group. One alternative explanation could be that the TERA improves performance simply because participants get sensitized to emotional expressions or get familiar with the actors who appear in the GERT-S, and not because the training provides instructions, practice, and feedback. Another alternative explanation could be that the TERA improves participants' ability to discriminate between different categories (i.e., category learning) and does not specifically enhance emotion-specific knowledge and skills. We hoped that accuracy would improve more in the training condition than in either of these control conditions.

In the "familiarity" control condition, participants viewed all videos that appeared in the TERA in a random order and answered a multiple-choice question about the appearance of the actor after each clip. This condition was intended to familiarize participants with the style of the videos used in the training and emotional expressions, without conveying any knowledge on emotions or nonverbal cues. The "category learning" control condition was a "cloud training" condition in which participants were trained to recognize 14 different cloud types from pictures. This training had the same structure as the TERA, consisting of instruction followed by practice with feedback, and used 14 cloud types instead of 14 emotion categories.

Method

Participants. One hundred fifty-nine participants (61% female) were recruited through Amazon's Mechanical Turk for $2.00. Mean age was 33.39 (SD=10.19) and ethnic

composition was Seventy-four percent White, 7% Black, 5% Hispanic, 5% Asian, and 9% mixed or other.

Procedure. Participants were randomly assigned to an untreated control condition (N = 41), the familiarity condition (N = 44), the cloud training (N = 40), or the TERA (N = 34). The Ns were not equal in the four conditions due to technical problems with playing the videos online, as a result of which some participants were unable to continue the study. As in Study 1, participants in the untreated control condition completed several unrelated personality questionnaires. All participants completed the GERT-S, DANVA Face, and DANVA Voice in a random order at the end. In addition, just before completing the ERA tests, all participants filled in a short measure of momentary affect (see below). Participants in all conditions except the untreated control were also asked how interesting they found the previous task on a five-point scale from "not at all interesting" to "very interesting." These measures of affect and interest were collected to examine whether the different interventions were similarly engaging.

Materials.

***TERA*.** The same version as in Study 1 was used.

Familiarity control condition. Participants watched the 42 videos from the TERA in a random order and, after each video, were asked to answer a yes/no question about the appearance of the actor, e.g., "Did the actor wear a ring?" or "Did the actress have gray hair?" The questions were different for each video. Participants did not receive feedback on the correctness of their answers.

***Cloud training control condition*.** This condition consisted of an instruction part in which participants read about the features that characterize each of 14 different cloud types such as Cirrus, Cumulus, or Nimbostratus, and saw two example pictures for each type. The descriptions included information about the altitude, opaqueness, and shape of the cloud type. This information as well as all example and training pictures were obtained through an

internet search. Next, participants completed a practice-with-feedback part in which they saw 42 other pictures of clouds and, after each, were asked to guess which cloud type had been shown. As in the TERA, the 14 response options were arranged in a wheel shape, and each cloud name was displayed together with a small example picture to facilitate orientation. For each picture, participants received feedback about whether their response was correct. If their response was incorrect, the same picture was presented a second time and they were asked to guess again. If the second response was still incorrect, the correct answer was displayed.

GERT-S, DANVA Face and Voice. See Study 1 for a description of these tests. In Study 2, the DANVA Face and Voice were correlated at $r = .48$, and the correlations with the GERT-S were $r = .53$ (DANVA Face) and $r = .55$ (DANVA Voice), respectively (all p-values $< .001$).

Measure of momentary affect. Participants rated how they felt right now on a five-point scale. Five adjectives (distressed, afraid, interested, proud, determined) were taken from the Positive and Negative Affect Schedule (Watson, Clark, & Tellegen, 1988), and two adjectives (sad and bored) were added because they seemed potentially relevant to the interventions. One average affect score was computed after reversing the items distressed, sad, afraid, and bored, with higher scores indicating more positive affect.

Results and Discussion

Descriptive statistics for the TERA and control groups on the three ERA tests and analysis of variance ANOVA results are shown in Table 1. A one-way ANOVA with condition (untreated control, familiarity control, cloud training, TERA) as the independent variable and GERT-S as the dependent variable showed a significant effect of condition, $F(3, 155) = 3.78$, $p = .01$. Pairwise post-hoc comparisons with the Scheffé test revealed that the familiarity control group ($p = .02$) and marginally the cloud training group ($p = .07$) achieved lower scores on the GERT-S than the TERA group. However, the difference between the TERA group and the untreated control group was smaller in this study (0.08 as compared to

0.15 in Study 1) and not significant. This might be due to the online setting of this study in which we had less control over how much attention participants were paying to the training. In addition, it could have been that participants in the untreated control condition were more alert and motivated to perform well as they completed the ERA tests earlier in the experiment than the other groups that underwent an intervention. Condition also had a significant effect on DANVA Face scores, $F(3, 155) = 2.62$, $p = .05$, but not on DANVA Voice scores, $F(3, 155) = 0.17$, $p = .92$. Scheffé tests did not reveal any significant group differences for the DANVA tests. These findings suggest that the TERA did not generalize to these other tests, but given that this study was conducted in an online setting and the sample sizes in each group were relatively small, further studies are needed to reach a definitive conclusion about transfer effects.

The finding that the familiarity control and cloud training groups performed worse on the GERT-S than the TERA group suggests that the TERA improves ERA not through increased familiarity with emotional stimuli or the actors, or through non-specific category learning and discrimination. These data support the idea that ERA performance is increased through teaching emotion-specific knowledge and through practice with feedback. Participants in the four conditions did not differ in positive affect and the three interventions (TERA, cloud training, and familiarity condition) were rated as similarly interesting (see Table 1). These results suggest that the interventions were similarly engaging to participants and the TERA was not effective simply because it was more interesting. Like in Study 1, additional ANOVAs including gender as a factor did not show any significant group by gender interactions, suggesting that the interventions did not have different effects depending on gender.

Study 3

The goal of Study 3 was to examine the short- and long term effects of the TERA on a larger sample of undergraduate students. This study adds to the first two studies in several

ways. First, about half of the participants completed a follow-up session four weeks after the training session in order to examine whether higher GERT-S performance in the trained group persisted over time and whether trained participants perceived themselves as more emotionally sensitive, more emotionally intelligent, and reported higher well-being than untrained participants. We also examined whether the difference in individuals' GERT-S scores between the training session and the follow-up was related to cognitive intelligence, well-being, self-rated emotional sensitivity, and emotional intelligence. Second, we used the Emotion Recognition Index (ERI) Face subtest (Scherer & Scherer, 2011) instead of the DANVA Face. Third, the TERA was modified in that the instructions were presented as short videos in addition to text to make it more engaging. We believed that this format would be more enjoyable and better capture the attention of participants.

Method

Participants. One hundred sixty-eight undergraduate students (50% female) were recruited as in Study 1. Mean age was 18.65 (*SD* = 1.11) and ethnic composition was 43% White, 27% Asian, seven percent Black, four percent Hispanic, and 18 % reported a mixed or other ethnicity or chose not to report it.

Procedure. The study consisted of two sessions. In session 1, participants were randomly assigned to either the TERA (*N* = 83) or the cloud training (*N* = 85). We chose the cloud training as the control condition for this study because it had similar structure and duration as the TERA. After the respective training, participants indicated on a five-point scale how interesting they found the training and filled in the Positive and Negative Affect Schedule (PANAS) short form (Thompson, 2007). Participants then completed the GERT-S, the DANVA voices, and the ERI Face in a random order. After these tests, participants were paired with another participant to complete a negotiation exercise. The data on the negotiation outcomes and behaviors is presented elsewhere (Schlegel, manuscript in preparation).

Ninety out of the 168 participants (those who participated early in the semester) were invited for session 2, of whom 87 accepted. The session took place approximately four weeks after session 1 in the laboratory (46 from the original cloud condition and 41 from the original TERA condition). Participants completed the GERT-S, a short cognitive intelligence test (Cattell Culture Fair Intelligence Test subtest 1; CFIT; Cattell & Cattell, 1957), the Trait Emotional Intelligence Questionnaire (TEIQue; Petrides, 2009), the emotional sensitivity subscale of the Social Skills Inventory (SSI; Riggio, 1986), and the World Health Organization brief general well-being questionnaire (WHO-5; Topp, Østergaard, Søndergaard, & Bech, 2015). The CFIT was administered to assess whether long-term benefits of the TERA in terms of GERT-S performance are higher for individuals with higher cognitive intelligence. The other questionnaires were administered to evaluate whether participants' self-evaluated emotional skills and well-being were affected by the TERA. The 87 participants that completed session 2 did not differ significantly from the 81 participants that only participated in session 1 with respect to their mean scores on the ERA tests (GERT-S, DANVA Voice, ERI Face), interest ratings, or positive and negative affect measured in session 1.

Materials.

TERA and cloud condition. Modified versions of these interventions were used in which all instructions were presented in short video clips before the written text for each emotion. In these clips, a female experimenter is shown saying the same things that were presented as text in Studies 1 and 2. This was done to make the interventions more enjoyable and because we assumed this format might be more engaging for older adults in the other planned study (Study 4).

GERT-S and DANVA Voice. See Study 1 for descriptions of these tests.

Emotion Recognition Index Face (ERI Face; Scherer & Scherer, 2011). The ERI Face consists of 30 pictures of posed expressions from the Pictures of Facial Affect set

(Ekman & Friesen, 1976) that are presented for 3 s each. After each portrayal, participants are asked to choose which out of five emotional states (sad, fearful, angry, happy, and neutral) had been expressed. A total score is calculated from the number of items in which the participant's response matched the target emotion. The ERI has been validated in several studies that found correlations with other ERA tests (e.g., Bänziger, Grandjean, & Scherer, 2009) as well as expected differences on demographic variables such as gender and profession (Scherer & Scherer, 2011). Cronbach's alpha in studies 3 and 4 was .44. Although this value is below common recommendations for good reliability, it matches the average reliability of ERA tests of .48 found in a recent meta-analysis (see Schlegel, Boone, & Hall, 2017, for a detailed discussion of the issue of reliability in ERA tests). Across Studies 3 and 4, women performed better on the ERI Faces than men ($r = .16$; $p < .01$).

Positive and Negative Affect Schedule short form (PANAS; Thompson, 2007). The PANAS short form contains 10 adjectives (determined, alert, attentive, inspired, active, afraid, nervous, upset, shamed, and hostile), for each of which participants indicate on a five-point scale how much they feel that way right now. One score for positive affect (mean of the first five adjectives) and one score for negative affect (mean of the last five adjectives) was calculated.

Culture Fair Intelligence Test (CFIT; Cattell & Cattell, 1957). The CFIT is a widely used measure of fluid intelligence for adults and consists of two parallel forms with four subtests each. Here, we used subtest 1 from form A which requires inferring complex relationships between elements of figures and is timed (3 min). For each item, participants see a series of three abstract figures and have to choose, from five other figures, the one that completes the series. There are 12 items in this subtest. Given that previous researchers noted that there might be a ceiling effect to the CFIT (Weiss, 2006), we added the three last items of subtest 1 of form B, yielding a total of 15 items that have to be solved in the same period of

time. Responses were scored as correct (1) and incorrect (0) and were summed to form a total score.

Trait Emotional Intelligence Questionnaire (TEIQue; Petrides, 2009). The TEIQue is a self-report questionnaire measuring four broad factors of trait emotional intelligence (well-being, self-control, emotionality, and sociability) and 15 more specific facets. Participants are asked to state their agreement with various statements on a seven-point Likert scale ranging from "disagree completely" to "agree completely," Here, we used the TEIQue short form that consists of 30 items.

Social Skills Inventory (SSI; Riggio, 1986). The SSI is a self-report questionnaire to measure six dimensions of socio-emotional competencies. Here, we used the Emotional Sensitivity subscale of the SSI that consists of 15 statements for each of which participants state how characteristic they are of them on a five-point scale. This scale reflects self-perceived attunement to others' feelings.

World Health Organization well-being questionnaire (WHO-5; Topp et al., 2015). The WHO-5 is a five-item self-report questionnaire in which participants indicate how often in the past three to four weeks they have been feeling cheerful and in good spirits, calm and relaxed, etc. on a seven-point scale.

Results and Discussion

Descriptive statistics for all measures in the two conditions, as well as the results of *t*-tests, are provided in Table 1. The mean scores for each emotion, unbiased hit rates (Wagner, 1993), and confusion matrices for the GERT-S, ERI Face, and DANVA Voice for both conditions are provided in the Supplementary Material (Tables S1 to S7). The zero-order correlations between all variables are shown in supplementary Table S8.

In session 1 the TERA group performed significantly better on the GERT-S, $t(165) = 9.46$, $p < .001$, and the ERI Face, $t(166) = 4.06$, $p < .001$, than the cloud training group immediately after the training. The DANVA Voice did not differ significantly between the

groups. Participants in the two groups did not differ in positive and negative affect after their respective intervention, but unlike in Study 2, the cloud training was rated as significantly less interesting (M = 2.93, SD = 1.08) than the TERA (M = 3.53, SD = 0.94); $t(166) = -3.84$, $p <$ .001. Additional ANOVAs including gender as a factor revealed one significant group by gender interaction when predicting ERI Face scores. In this analysis, there was a main effect of group, $F(1,164)=18.06$, $p < .001$, but no significant main effect of gender, $F(1,164) = 2.61$, $p = .108$. The significant interaction of group and gender ($F(1,164)=7.33$, p =.008) showed that the difference in ERI Face scores between the two groups was larger in women (mean difference .07) than in men (mean difference .02), suggesting that ERA training was more effective in improving facial ERA in women.

In session 2, after four weeks, the TERA group still performed significantly better on the GERT-S than the cloud training group, $t(85) = 4.11$, $p < .001$. Also, the mean score of the TERA group had not changed since session 1 (M = .79 at both time points). At the same time, the cloud training group significantly improved in GERT-S performance compared to the first session (mean difference .05, SD = .08), $t(45) = 4.25$, $p < .001$, but this group's scores were still substantially lower than those of the TERA group. The improved performance in the cloud training group is in line with previous studies showing that the repeated administration of the same ERA test without any feedback can increase performance (e.g., Bänziger et al., 2009), and similarly, that ERA practice enhances performance to some extent even without feedback (Blanch-Hartigan, 2012). These results suggest that training effects persist over four weeks, although it must be noted that only the GERT-S and no other ERA tests were administered at T2.

The two groups did not differ significantly in the WHO-5 wellbeing scale and trait emotional intelligence as measured by the TEIQue at session 2. However, the cloud training group perceived itself as more sensitive to others' emotions than the TERA group as measured with the SSI Emotional Sensitivity scale, $t(85) = 2.57$, $p = .01$. The same finding

was reported by Blanch-Hartigan (2012) in her training. She argued that receiving feedback on errors in emotion recognition ability might lead participants to question their own ability and to have less confidence in it, although they objectively perform better.

Analyzing both groups simultaneously, the difference in people's GERT-S scores between session 1 and 2 was not correlated with cognitive intelligence as measured by the CFIT, well-being, and trait emotional intelligence. However, the GERT-S difference score was correlated with emotional sensitivity such that the more participants' performance improved over time, the higher they rated their emotional sensitivity ($r = .38$, $p < .001$). This result might suggest that individuals have some insight into how their performance changed over the two time points.

Meta-Analysis of Studies 1-3

Taken together, Studies 1 to 3 provide evidence for the effectiveness of the TERA in improving performance on the GERT-S. In addition, Study 1 suggested improved performance in vocal ERA (DANVA Voice) and Study 3 suggested improved performance in facial ERA (ERI Face). Given that *p*-values are affected by the differing sample sizes, we conducted a fixed effects mini-meta-analysis (Goh, Hall, & Rosenthal, 2016) to summarize the results across these three studies that examined younger adults. For this analysis, we treated all control conditions (untreated, familiarity, and cloud) across the studies as one condition and compared it against the TERA condition on the three tests that had been used in more than one study (i.e., GERT-S, DANVA Face, and DANVA Voice). The combined *p*-values were calculated using the Stouffer method (Mosteller & Bush, 1954). Results showed that the TERA group performed significantly better on the GERT-S (3 studies, $Z = 9.21$, $p < .001$) and the DANVA Voice (3 studies, $Z = 2.18$, $p = .03$), but not the DANVA Face (2 studies, $Z = 1.11$, $p = .13$). The magnitude of the average effect size (Pearson correlation) across these studies can be considered large for the GERT-S ($r = .45$), and small for the DANVA Voice ($r = .11$) and the DANVA Face ($r = .08$; Cohen, 1988). However, the

DANVA Face test might have failed to detect ERA improvements due to ceiling effects, and the other facial ERA test that was used in Study 3 showed significantly higher scores in the TERA group with a medium effect size of $r = .30$ ($p < .001$). Taken together, these results suggest that the training improved multimodal ERA as well as facial and vocal ERA which were measured with tests that differ in their stimulus material (modality, actors, emotions included) from the training material. Notably, the effect sizes were smaller in Study 2 than in Studies 1 and 3, which might be related to the fact that Study 2 was conducted online and not in the laboratory, or to the difference in participants' mean age.

Study 4

While the first three studies mostly examined young adults, Study 4 aimed to investigate whether the TERA also improves ERA in older adults. Similarly to Study 3, this study consisted of one training session and one voluntary follow-up questionnaire to measure self-perceived emotional competencies and well-being.

Method

Participants. Ninety-eight participants (64% female; age range 59 to 90 years; M_{age} = 69.53, SD_{Age} = 6.53) were recruited from the greater Boston area for a payment of $20. Fifty-seven percent were between 59 and 69 years old, 35% were between 70 and 79 years old, and 8% were between 80 and 90 years old. Upon phone recruitment, all participants completed the 26-item telephone version of the Mini-Mental State Examination (T-CogS; Newkirk et al., 2004). All participants scored above a threshold of 21 points which corresponds to the typically used cutoff of 23 points on the in-person Mini-Mental State Examination (Kraemer, Taylor, Tinklenberg, & Yesavage, 1998). Ethnic composition was as follows: Sixty-four percent White, 13% Black, four percent Native American, and 18% other, mixed or no reported ethnicity. In terms of educational level, 46% had obtained postgraduate degrees (Masters or higher), 27% had a college degree, and 26% had high school degree or general education diploma. All participants reported having at least some familiarity in using a

computer. The data from thirteen participants were excluded for the analysis for various reasons (see results section below), leading to a final sample size of $N = 85$.

Procedure. The study consisted of one session in the laboratory and one optional paper-pencil or online follow-up questionnaire. In the laboratory, participants' visual acuity was tested with the Snellen and Rosenbaum eye charts (Snellen, 1862) and contrast sensitivity was assessed with the Pelli-Robson eye chart (Pelli, Robson, & Wilkins, 1988). Participants also completed the digit span subtests (both forward and backward) of the Wechsler Adult Intelligence Scale–Revised (WAIS–R; Wechsler, 1981) and the Shipley Vocabulary Test (Zachary, 1986). In the digit span tests, participants repeat a list of digits read by experimenter in forward or backward order. In the vocabulary test, participants identify, from a list of six words, the word that has the same meaning as a given target word.

After completing these tests, participants were randomly assigned to the TERA condition (N=49) or the cloud control condition (N=49) and completed the respective intervention. After the intervention, participants rated how interesting they found the training on a five-point scale and filled in a short measure of momentary affect (PANAS short form). Participants then completed the GERT-S, the DANVA Voice, and the ERI Face in a random order. At the end of the laboratory session, participants were asked to rate how friendly and supportive their experimenter had been on a five-point scale. Finally, participants were invited to take part in a short follow-up four weeks later. The follow-up questionnaire consisted of the TEIQUE and the WHO well-being questionnaire, and was sent to participants by email as a link to an online survey or by mail as a paper-pencil version. Fifty-five participants (64.71%) returned the completed questionnaire or filled in the online survey.

Materials.

TERA and cloud training. The same versions as in Study 3 were used.

GERT-S, DANVA Voice, and ERI Face. See Studies 1 and 3 for a description.

PANAS short form, TEIQue, and WHO-5. See Study 3 for a description.

Results and Discussion

Data from 13 of the 98 participants were excluded for the following reasons, leaving a total sample size of $N = 85$: Three participants did not finish their respective intervention or expressed strong displeasure about the intervention; eight participants had a visual acuity of 20/100 or less as measured by the Snellen and Rosenbaum eye charts; one person reported suffering from dyslexia; and one person had a very low digit span score of 11 points (forward plus backward).

Descriptive statistics and results of *t*-tests are presented in Table 1. The mean scores for each emotion on the GERT-S, DANVA Voice, and ERI Face, unbiased hit rates (Wagner, 1993), and confusion matrices for each condition are provided in the Supplementary Material (Tables S9 to S15). The zero-order correlations of all variables are provided in Table S16. Unlike in Studies 1 to 3, participants in the two conditions did not significantly differ in their overall performance on any of the three ERA tests. Additional ANOVAs including gender did not yield any significant group by gender interactions. Participants rated the TERA as marginally more interesting than the cloud training, $t(82) = -1.93$, $p = .06$. The two groups did not differ in positive affect and evaluated their experimenter as similarly friendly and supportive, but the cloud training group reported significantly more negative affect after the training than the TERA group, $t(75) = 2.10$, $p = .04$. Among the participants who participated in the follow-up survey, there were no differences between the TERA group and the cloud training group in the TEIQue or WHO-5 four weeks after the training.

Thus, overall the TERA was not effective in improving ERA in older adults. In order to better understand the differences between older and younger adults, we further examined the results of the GERT-S by inspecting the number of correct answers for each of the 14 emotions, the unbiased hit rates, and the confusion matrices for both groups (Tables S9 to S15). *T*-tests showed that the TERA group reached significantly higher scores than the cloud

training group in recognizing fear, $t(83) = 2.14$, $p = .04$, sadness, $t(83) = 2.13$, $p = .04$, and pride, $t(83) = 2.14$, $p = .04$. At the same time, the TERA group achieved lower scores for anxiety, $t(83) = -2.28$, $p = .03$, and a lower unbiased hit rate for amusement, $t(83) = -2.11$, $p = .04$, than the cloud training group. The differences in the confusion matrices of the two groups (Table S13) revealed that participants who received the TERA less often confused fear with anxiety, sadness with despair, and pride with pleasure and interest. This pattern is similar to the one found in younger adults who received the training in Study 3 (Table S5). However, older adults in the TERA group (but not in the cloud training group) often mistook anxiety for relief, and also more often confused amusement with joy and pleasure, driving the lower overall scores for amusement and anxiety. In contrast, younger adults in Study 3 *less often* confused amusement with joy, and almost never confused anxiety with relief.

These post-hoc emotion findings indicate that the training was somewhat effective for older adults for some emotions, whereas for other emotions the training had no effect or even a reverse effect. It may be the case that differences between emotions of the same family (e.g., anxiety and fear) might have been too subtle and too difficult to remember for the entire training, though future research will need to examine this question empirically.

Additional analyses also revealed that the extent to which older participants found their respective intervention interesting positively predicted their performance on the GERT-S ($r = .26$, $p < .01$) and the ERI Face ($r = .33$, $p < .01$), while this was not the case in younger adults (see Table S8). Furthermore, cognitive abilities (verbal ability and working memory) were strongly correlated with ERA in older adults (Table S16). These findings coupled with previous research demonstrating that older adults perform better in social perception contexts that they deem relevant (Richter & Kunzmann, 2011), suggest that individuals with lower levels of cognitive ability and engagement with the task may have performed more poorly. We expect that these preliminary results will spur future research to continue to investigate

the mechanisms behind age differences in ERA in order to develop more effective training interventions.

General Discussion

Although ERA has been widely studied as a predictor of social, professional, and health outcomes in various domains of psychology, to date no ERA training for the general, non-clinical adult population existed. To our knowledge, the present research is the first to validate a short, self-administered training program intended to improve ERA in the face, voice, and body across a wide range of positive and negative emotions. Three studies showed that the new TERA increased ERA in younger and middle-aged adults when completed in the laboratory and online, and that the training effects persist over at least four weeks. Furthermore, results showed that the effects cannot be explained simply by familiarity with the training material or by category learning, but that the elements of instruction, practice, and feedback are crucial for improving participants' ERA. Overall, the studies also suggested that the ERA training is similarly effective in both men and women.

Importantly, a meta-analytic summary of the first three studies demonstrated that the training improved multimodal ERA using a test that was similar to the material used in the training (i.e., same actors, same emotions, same pseudolanguage, but not the same clips), as well as on facial and vocal ERA tested using completely different instruments. Such transfer effects on different tests had often not been studied or reported, and when they were reported, they were overall not significant as shown in Blanch-Hartigan et al.'s (2012) meta-analysis. The presence of transfer effects in this ERA training increases the possibility that being trained might also affect participants' behavior in social interactions as well as psychosocial outcomes. These results also imply that interventions focusing on many emotions and on multiple sensory modalities simultaneously might be more ecologically valid than more specialized interventions and might thus be particularly useful for increasing general ERA in

the non-clinical adult population. However, future studies are needed to examine whether this training indeed affects real-life behaviors and outcomes.

Although the TERA proved effective in the first three studies, Study 4 revealed older age as a boundary condition of training effectiveness that might potentially apply to person perception trainings in general. In its current form, the training did not improve ERA in older adults. This finding sheds some light on the previously open question whether individuals with lower initial person perception skills benefit more or less from training than individuals with a higher baseline (Blanch-Hartigan et al., 2012). Although individual baseline ERA was not measured, older adults overall scored lower on the GERT-S and the DANVA Voice than younger adults in Studies 1 and 3 (see Table 1), suggesting that a certain minimal baseline ERA level could be required for the present training to be effective. However, future research including assessments of baseline ERA is needed to better understand for which individuals training is (in)effective. For individuals with a relatively low baseline ERA, interactive training elements such as discussions with other participants and the assistance of a trainer might be necessary to improve their skills and self-administered programs might be less effective. Previous research suggests that such interactive elements enhance training effectiveness (Hurley, 2012; Ruben et al., 2015). Another limitation of self-directed trainings such as the TERA is that individuals with lower initial ERA skills might not complete the training as they get more negative feedback in the practice section and are generally less likely to participate in respective interventions (Sheldon, Ames, & Dunning; 2014). On the other hand, the TERA is easy and cheap to implement on a large scale in applied settings such as professional training.

Despite a rapidly expanding field of research investigating age differences in person perception through various paradigms, the underlying mechanisms for the age decline have not been definitively identified (Freund & Isaacowitz, 2014). In order to optimize ERA in particular, interventions must train the component that actually declines. It could be that the

underlying mechanism through which older adults recognize emotions is qualitatively different from the mechanism in younger adults. For example, eye-tracking studies have demonstrated that older adults look more at the mouth and less at the eyes of emotionally expressive faces compared with young adults (e.g., Murphy & Isaacowitz, 2010), an effect which is stronger in older men than women and is correlated with ERA performance (Sullivan, Campbell, Hutton, & Ruffman, 2017). The inquiry of whether this finding extends to video stimuli, and whether manipulating eye gaze increases performance, are exciting avenues for future research. This also relates to the more general question of which components of the TERA (instruction or practice with feedback) drives the improvement in emotion perception. Previous research suggests that instruction might be less important to training effectiveness than practice with feedback (Blanch-Hartigan et al., 2012). However, given that the TERA includes a large number of emotions, instruction might have been crucial as it specifically focused on subtle differences between the emotions and demonstrated these in example videos. Future research might tease apart the effects of the instruction and the practice with feedback parts in the TERA. If the instruction part individually contributes to its effectiveness, other elements such as instructions regarding eye gaze may be promising additions.

Because the training in its current form improved ERA in young and middle aged adults but not in older adults, future research may look to the qualities of successful trainings in older adults in other cognitive domains. Recent reviews on cognitive training interventions recommend that interventions for older adults be conducted in group settings, and consist of repetitive and adaptive training, including long-term follow ups and booster sessions (Kelly et al., 2014). With regard to training ERA in particular, we encourage future research to empirically investigate potential mechanisms, such as cognitive ability, in order to design more effective trainings for improving ERA in older adults.

While the present studies yielded first evidence for the effectiveness of the new TERA, there are several limitations that need to be addressed in future research. First, the effects of the training on real-life behavior and psychosocial outcomes need to be examined in detail to complement the current validity evidence. Although we did not find any effects on self-rated wellbeing in Study 3, the training might nevertheless have affected other outcomes. For example, the training might have increased participants' awareness of and attunement to emotional expressions of others in their everyday life, which could have resulted in more successful and rewarding social interactions.

Second, although objective training outcomes are desirable, the effects of training on self-perceived emotional sensitivity or emotional intelligence also need to be considered and addressed in future studies. In their recent meta-analysis, Joseph, Jin, Newman, and Boyle (2015) found that self-perceived emotional intelligence predicted higher job performance independently of people's actual performance-based emotional intelligence, which was partly explained through higher self-efficacy. Given that in Study 3 participants rated their emotional sensitivity four weeks after the training as *lower* than the control group although their objective ERA had improved, future interventions should make this improvement more explicit to participants in order to enhance their self-perceived competence.

Third, the present studies only examined training outcomes four weeks after the intervention, although some studies that assessed more comprehensive and multi-session socio-emotional training programs investigated training effects for up to six months (e.g., Herpertz et al., 2016). In addition to examining more long-term effects of the TERA, future studies could also directly compare it to other existing person perception interventions. To our knowledge, no study has directly compared different existing person perception trainings, especially single-session and multi-session programs, on the same outcome measures in order to identify the minimal duration and elements necessary to achieve long-lasting improvements.

Taken together, the present studies have yielded promising evidence for the effectiveness of a novel training in improving multimodal ERA for younger and middle-aged adults. Importantly, the short duration of less than one hour as well as the easy accessibility and administration as an online program that does not require the presence of a trainer makes the training useful as a tool in many research and applied settings such as professional training and development for healthcare providers, teachers, managers, customer service representatives, etc. Given the broad orientation of the new training in that it targets ERA across modalities and many emotions without being limited to a specific population, this intervention might have the potential to positively affect a variety of psychosocial outcomes across these different settings.

Compliance with Ethical Standards

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent was obtained from all individual participants included in the study.

Funding

This research was funded by an Early Postdoc.Mobility fellowship awarded to K. Schlegel by the Swiss National Science Foundation, grant number P2GEP1_155698.

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*(5), 691–704.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*, *12*(5), 1161–1179.

Blanch-Hartigan, D. (2011). Measuring providers' verbal and nonverbal emotion recognition ability: Reliability and validity of the Patient Emotion Cue Test (PECT). *Patient Education and Counseling*, *82*(3), 370–376.

Blanch-Hartigan, D. (2012). An effective training to increase accurate recognition of patient emotion cues. *Patient Education and Counseling*, *89*(2), 274–280.

Blanch-Hartigan, D., Andrzejewski, S. A., & Hill, K. M. (2012). The Effectiveness of Training to Improve Person Perception Accuracy: A Meta-Analysis. *Basic and Applied Social Psychology*, *34*(6), 483–498.

Blanch-Hartigan, D., Andrzejewski, S. A., & Hill, K. M. (2016). Training people to be interpersonally accurate. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The Social Psychology of Perceiving Others Accurately* (pp. 253–269). Cambridge University Press.

Buhlmann, U., Gleiß, M. J. L., Rupf, L., Zschenderlein, K., & Kathmann, N. (2011). Modifying emotion recognition deficits in body dysmorphic disorder: an experimental investigation. *Depression and Anxiety*, *28*(10), 924–931.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Derntl, B., Seidel, E.-M., Kryspin-Exner, I., Hasmann, A., & Dobmeier, M. (2009). Facial emotion recognition in patients with bipolar I and bipolar II disorder. *British Journal of Clinical Psychology*, *48*(4), 363–375.

Domes, G., Grabe, H. J., Czieschnek, D., Heinrichs, M., & Herpertz, S. C. (2011). Alexithymic Traits and Facial Emotion Recognition in Borderline Personality Disorder. *Psychotherapy and Psychosomatics*, *80*(6), 383–385.

Elfenbein, H. A., Foo, M. D., White, J., Tan, H. H., & Aik, V. C. (2007). Reading your counterpart: The benefit of emotion recognition accuracy for effectiveness in negotiation. *Journal of Nonverbal Behavior*, *31*(4), 205–223.

Freund, A. M., & Isaacowitz, D.M. (2014) Aging and social perception: So far, more similarities than differences. *Psychology and aging* 29 (3), 451-453.

Goh, J. X., Hall, J. A.., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10,* 535-549.

Golan, O., Ashwin, E., Granader, Y., McClintock, S., Day, K., Leggett, V., & Baron-Cohen, S. (2010). Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces. *Journal of autism and developmental disorders*, *40*(3), 269–279.

Hall, J. A. (2011). Clinicians' accuracy in perceiving patients: Its relevance for clinical practice and a narrative review of methods and correlates. *Patient Education and Counseling*, *84*(3), 319–324.

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior*, *33*(3), 149–180.

Hall, J. A., Mast, M. S., & West, T. V. (2016). *The Social Psychology of Perceiving Others Accurately*. Cambridge University Press.

Herpertz, S., Schütz, A., & Nezlek, J. (2016). Enhancing emotion perception, a fundamental component of emotional intelligence: Using multiple-group SEM to evaluate a training program. *Personality and Individual Differences*, *95*, 11–19.

Hurley, C. M. (2012). Do you see what I see? Learning to detect micro expressions of emotion. *Motivation and Emotion*, *36*(3), 371–381.

Hurley, C. M., Anker, A. E., Frank, M. G., Matsumoto, D., & Hwang, H. C. (2014). Background factors predicting accuracy and improvement in micro expression recognition. *Motivation and Emotion*, *38*(5), 700–714.

Joseph, D. L., Jin, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology*, *100*(2), 298–342.

Kelly, M. E., Loughrey, D., Lawlor, B. A., Robertson, I. H., Walsh, C., & Brennan, S. (2014). The impact of cognitive training and mental stimulation on cognitive and everyday functioning of healthy older adults: a systematic review and meta-analysis. *Ageing Research Reviews*, *15*, 28-43.

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*(5), 1009–1019.

Kraemer, H. C., Taylor, J. L., Tinklenberg, J. R., & Yesavage, J. A. (1998). The stages of Alzheimer's disease: a reappraisal. *Dementia and Geriatric Cognitive Disorders, 9*(6), 299–308.

Krendl, A. C., & Ambady, N. (2010). Older adults' decoding of emotions: Role of dynamic versus static cues and age-related cognitive decline. *Psychology and Aging, 25*(4), 788–793.

Lansford, J. E., Sherman, A. M., & Antonucci, T. C. (1998). Satisfaction with social networks: an examination of socioemotional selectivity theory across cohorts. *Psychology and Aging, 13*(4), 544-552.

Matsumoto, D., & Hwang, H. S. (2011). Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*, *35*(2), 181–191.

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: Basic Books.

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review*, *90*(3), 215–238.

Mosteller, F. M., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Volume I. Theory and method*. Cambridge, MA: Addison-Wesley.

Murphy, N. A., & Isaacowitz, D. M. (2010). Age effects and gaze patterns in recognising emotional expressions: An in-depth look at gaze measures and covariates. *Cognition and Emotion, 24*, 436-452.

Newkirk, L. A., Kim, J. M., Thompson, J. M., Tinklenberg, J. R., Yesavage, J. A., & Taylor, J. L. (2004). Validation of a 26-point telephone version of the Mini-Mental State Examination. *Journal of Geriatric Psychiatry and Neurology*, *17*(2), 81–87.

Nowicki, S. Jr., & Duke, M. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior, 18*, 9-35.

Nowicki, S. (2006). *A manual and reference list for the Diagnostic Analysis of Nonverbal Accuracy (DANVA2)*. Department of Psychology, Emory University, Atlanta: GA.

Pelli, D. G., Robson, J. G., & Wilkins, A. J. (1988). The design of a new letter chart for measuring contrast sensitivity. *Clinical Vision Science*, 2, 187–199.

Petrides, K. V. (2009). Psychometric properties of the Trait Emotional Intelligence Questionnaire. In C. Stough, D. H. Saklofske, and J. D. Parker, *Advances in the Assessment of Emotional Intelligence* (pp. 85-101). New York: Springer.

Richter, D., & Kunzmann, U. (2011). Age differences in three facets of empathy: performance-based evidence. *Psychology and Aging, 26*(1), 60-70.

Riggio, R. E. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology*, *51*, 649–660.

Rivers, S. E., Brackett, M. A., Reyes, M. R., Elbertson, N. A., & Salovey, P. (2013). Improving the social and emotional climate of classrooms: A clustered randomized controlled trial testing The RULER Approach. *Prevention Science*, *14*(1), 77–87.

Ruben, M. A., Hall, J. A., Curtin, E. M., Blanch-Hartigan, D., & Ship, A. N. (2015). Discussion increases efficacy when training accurate perception of patients' affect. *Journal of Applied Social Psychology*, *45*(6), 355–362.

Rubin, R. S., Munz, D. C., & Bommer, W. H. (2005). Leading from within: The effects of emotion recognition and personality on transformational leadership behavior. *Academy of Management Journal*, *48*(5), 845–858.

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*(4), 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: Oxford University Press.

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*(4), 305–326.

Schlegel, K. (manuscript in preparation). *How emotion recognition training affects negotiation behavior and outcomes.*

Schlegel, K., Boone, R. T., & Hall, J. A. (2017). Individual differences in interpersonal accuracy: A multi-level meta-analysis to assess whether judging other people is one skill or many. *Journal of Nonverbal Behavior*, *41*(2), 103–137.

Schlegel, K., Fontaine, J. R., & Scherer, K. R. (2017). The nomological network of emotion recognition ability: Evidence from the Geneva Emotion Recognition Test. *European Journal of Psychological Assessment,* 1-12.

Schlegel, K., Grandjean, D., & Scherer, K. R. (2013). Constructs of social and emotional effectiveness: Different labels, same content? *Journal of Research in Personality*, *47*(4), 249–253.

Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the Geneva Emotion Recognition Test: An example of Rasch-based test development. *Psychological Assessment*, *26*(2), 666–672.

Schlegel, K., & Scherer, K. R. (2016). Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation. *Behavior Research Methods*, *48*, 1383–1392.

Sheldon, O. J., Dunning, D., & Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: Reactions to feedback about deficits in emotional intelligence. *Journal of Applied Psychology*, *99*(1), 125–137.

Silver, H., Goodman, C., Knoll, G., & Isakov, V. (2004). Brief emotion training improves recognition of facial emotions in chronic schizophrenia. A pilot study. *Psychiatry Research*, *128*(2), 147–154.

Snellen, H. (1862). Probebuchstaben zur Bestimmung der Sehschärfe, Utrecht, The Netherlands.

Sullivan, S., Campbell, A., Hutton, S. B., & Ruffman, T. (2015). What's good for the goose is not good for the gander: Age and gender differences in scanning emotion faces. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 72 (3), 441-443.

Thompson, E. R. (2007). Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology*, *38*(2), 227–242.

Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO-5 Well-Being Index: a systematic review of the literature. *Psychotherapy and Psychosomatics*, *84*(3), 167–176.

Van Kleef, G. A. (2010). The emerging view of emotion as social information. *Social and Personality Psychology Compass*, *4*(5), 331–343.

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*(1), 3–28.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale–Revised*. New York: Psychological Corporation.

Zachary, R. (1986). *Shipley Institute of Living Scale, Revised Manual.* Los Angeles: Western Psychological Services.

Table 1

Descriptive Statistics (Means and Standard Deviations) and Results of Group Comparisons in the Four Studies

Measures	*N*	untreated control	familiarity control	cloud control	ERA training	*t*/ *F*	Cohen's d/ η2
Study 1							
GERT-S	103	.60 (.15)	-	-	.75 (.11)	$t(101)=5.94^{***}$	$d=1.14$
DANVA Face	101	.76 (.13)	-	-	.79 (.08)	$t(99)=1.47$	$d=0.28$
DANVA Voice	101	.72 (.11)	-	-	.77 (.11)	$t(99)=2.16^{*}$	$d=0.64$
Study 2							
GERT-S	159	.57 (.15)	.53 (.17)	.55 (.14)	.65 (.19)	$F(3,155)=3.78^{*}$	$\eta 2=.07$
DANVA Face	159	.78 (.10)	.71 (.18)	.80 (.13)	.76 (.16)	$F(3,155)=2.62^{*}$	$\eta 2=.05$
DANVA Voice	159	.71 (.14)	.71 (.16)	.70 (.14)	.71 (.17)	$F(3,155)=0.17$	$\eta 2=.003$
interest rating of intervention	120	-	3.18 (1.02)	3.40 (1.33)	3.39 (1.18)	$F(2,117)=0.46$	$\eta 2=.008$
positive affect	161	3.97 (0.65)	3.99 (0.64)	3.83 (0.89)	3.71 (0.61)	$F(3,157)=1.32$	$\eta 2=.03$
Study 3 – session 1							
GERT-S	167	-	-	.65 (.10)	.79 (.09)	$t(165)=9.46^{***}$	$d=1.47$
ERI Face	168	-	-	.68 (.09)	.72 (.07)	$t(166)=4.06^{***}$	$d=0.50$
DANVA Voice	168	-	-	.76 (.07)	.78 (.08)	$t(166)=1.25$	$d=0.27$
interest rating of intervention	168	-	-	2.93 (1.08)	3.53 (0.94)	$t(166)=3.84^{***}$	$d=0.59$
PANAS positive affect	166	-	-	2.73 (0.90)	2.89 (0.76)	$t(164)=1.16$	$d=0.19$
PANAS negative affect	166	-	-	1.21 (0.35)	1.16 (0.29)	$t(164)=1.07$	$d=0.16$
Study 3 – session 2 (after 4 weeks)							
GERT-S	87	-	-	.71 (.10)	.79 (.08)	$t(85)=4.11^{***}$	$d=0.84$
CFIT	87	-	-	.85 (.14)	.87 (.11)	$t(85)=0.78$	$d=0.16$
TEIQue	87	-	-	4.88 (0.75)	4.92 (0.62)	$t(85)=0.23$	$d=0.06$
SSI emotional sensitivity	87	-	-	0.58 (.09)	0.55 (.09)	$t(85)=2.57^{*}$	$d=0.33$
WHO-5 wellbeing questionnaire	85	-	-	3.50 (0.97)	3.55 (1.13)	$t(83)=0.23$	$d=0.05$
Study 4 – session 1							
GERT-S	85	-	-	.50 (.12)	.51 (.14)	$t(83)=0.54$	$d=0.08$
ERI Face	85	-	-	.70 (.10)	.71 (.10)	$t(83)=0.40$	$d=0.10$
DANVA Voice	85	-	-	.69 (.14)	.67 (.11)	$t(83)=0.90$	$d=0.16$
interest rating of intervention	84	-	-	3.86 (1.13)	4.27 (0.78)	$t(82)=1.93$	$d=0.42$
friendliness rating of experimenter	65	-	-	4.33 (1.00)	4.61 (0.61)	$t(63)=1.34$	$d=0.34$
PANAS positive affect	80			3.67 (0.86)	3.83 (0.57)	$t(78)=0.84$	$d=0.22$
PANAS negative affect	77			1.35 (0.58)	1.12 (0.35)	$t(75)=2.10^{*}$	$d=0.48$
Study 4 – session 2							
TEIQue	48			4.02 (0.44)	3.92 (0.31)	$t(46)=0.85$	$d=0.26$
WHO-5 wellbeing questionnaire	47			2.55 (1.02)	2.69 (0.94)	$t(45)=0.48$	$d=0.14$

Note. GERT-S = Geneva Emotion Recognition Test, DANVA = Diagnostic Analysis of Nonverbal Accuracy, PANAS = Positive and Negative Affect Schedule, ERI = Emotion Recognition Index, CFIT = Culture Fair Intelligence Test, TEIQue = Trait Emotional Intelligence Questionnaire, SSI = Social Skills Inventory, WHO-5 = World Health Organization wellbeing questionnaire, ERA = Emotion Recognition Ability, $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.