Journal of Nonverbal Behavior (2021) 45:67–81
https://doi.org/10.1007/s10919-020-00340-4

ORIGINAL PAPER

Exploring Emotion Recognition and the Understanding of Others' Unspoken Thoughts and Feelings when Narrating Self-Experienced Emotional Events

Anders Flykt[1] · Tina Hörlin[1] · Frida Linder[1] · Anna-Karin Wennstig[1] · Gabriella Sayeler[1] · Ursula Hess[2] · Tanja Bänziger[1]

Published online: 24 January 2021

Abstract
Emotion decoding competence can be addressed in different ways. In this study, clinical psychology, nursing, or social work students narrated a 2.5–3 min story about a self-experienced emotional event and also listened to another student's story. Participants were video recorded during the session. Participants then annotated their own recordings regarding their own thoughts and feelings, and they rated recordings by other participants regarding their thoughts and feelings [empathic accuracy, EA, task]. Participants further completed two emotion recognition accuracy (ERA) tests that differed in complexity. The results showed that even though significant correlations were found between the emotion recognition tests, the tests did not positively predict empathic accuracy scores. These results raise questions regarding the extent to which ERA tests tap the competencies that underlie EA. Different possibilities to investigate the consequences of method choices are discussed.

Keywords Emotion recognition · Empathic accuracy · Narratives · Self-experienced emotional events

The ability to recognize emotion expressions (i.e., emotion decoding) is at the core of the Emotional Intelligence (EI) concept. Salovey and Mayer (1990) describe EI as a hierarchically composed competence based on the ability to recognize emotional expressions (see also Mayer et al. 2008; Salovey et al. 2008, p. 536). Emotion recognition is considered to be the most reliable aspect of EI (Elfenbein and Ambady 2002) and the results from a

Tanja Bänziger passed away on June 3, 2017. This study is the realization of one of the studies that she considered necessary to promote knowledge about how emotional understanding of others can be improved in professional helpers. This study was also part of the second to the fifth authors' exam work to become clinical psychologists. The order of these four authors is purely alphabetical.

✉ Anders Flykt
Anders.Flykt@miun.se

[1] Department of Psychology and Social Work, Mid Sweden University, Kunskapensväg 8, House P, 831 91 Östersund, Sweden

[2] Department of Psychology, Humbodlt-University of Berlin, Berlin, Germany

meta-analysis (Schlegel et al. 2017) suggest that emotion recognition ability (ERA) test scores appear to have the highest correlations (despite being rather low in absolute terms) with other domains of interpersonal accuracy, including empathic accuracy (EA, Ickes 1993, 2016). EA describes a person's ability to accurately infer others' unspoken feelings and thoughts. The notion that this ability is partially based in nonverbal emotion recognition accuracy is supported by studies that tested EA using either muted or filtered speech (Gesn and Ickes 1999; Hall and Schmid Mast 2007; Zaki et al. 2008), which found at least some accuracy in the inferences about others unspoken feelings and thoughts (i.e., EA).

Yet, research on emotion recognition accuracy (ERA) has encountered a number of methodological problems (Bänziger et al. 2009). One of these problems is that ERA is often assessed by asking the participant to choose from a list of emotion words the one label that best fits an emotion expression. It can be argued that this task assesses emotion discrimination rather than emotion recognition, especially when the proposed list of labels is short (Bänziger et al. 2009; Nelson and Russell 2016). Specifically, with few alternatives to choose from, participants may be able to deduce the correct answer by excluding obviously incorrect alternatives rather than by positively identifying the correct alternative. To reduce this potential problem, a larger list of choices could be presented, or open answers could be used.

When trying to generalize from an ERA test to real life recognition, a further problem is encountered: in real life situations emotion expressions are often regulated (see e.g., Gross 1998). Regulation may serve to increase some emotion expressions and to reduce or even mask others. For example, in therapy sessions clients might smile (or even laugh, see Marci et al. 2004) when talking about negative self-experienced emotional events as a means to protect themselves from strong negative emotional experiences. Such reasoning finds support by the results by Ansfield (2007), who showed that strongly disgusting pictures elicited more smiling in participants than less disgusting pictures. Similarly, Hess and Bourgeois (2010) found that both speakers and listeners smiled consistently during an anger narrative. During speech, facial actions may also be used to emphasize different aspects of a narrative (see e.g., Chovil 2005, for a report on how many different ways an eyebrow raise can be used in emphasizing verbal content), and this may interfere with concurrent emotion expressions. Thus, in situations where people talk about self-relevant emotional events, their nonverbal emotion expressions might reflect conversational strategies and emotion regulation more than actual emotional state. Conversely, in such situations, the accurate decoding of emotion expression is much more complex than the situation in standard ERA tests.

As such, the question arises of whether nonverbal displays alone suffice to correctly understand the thoughts and feelings of others in complex real-life situations. Gesn and Ickes (1999) showed 3-min videos of women who talked about their real-life problems in a simulated therapy session. Participants then had to make inferences about the women's thoughts and feelings (i.e., testing EA). When the speech was filtered—and verbal content could not be understood—participants were only about 1/3 as accurate as when the speech was understandable. Hall and Schmid Mast (2007) did a refined replication of this study and found that inferences about thoughts may be more dependent on verbal cues, whereas inferences about feelings may be more dependent on nonverbal cues. In a further replication, Zaki et al. (2008) found similar results. This suggests that, at least for the stimulus materials used in these three studies (i.e., video material with or without filtered speech), some level of understanding was achieved.

Also, results from Schlegel et al. (2017) suggest that there is some covariation between different tests assessing ERA, suggesting some underlying shared variance. Together, these

findings suggest that even though the task of assessing emotional feeling states from conversations (EA) and classic ERA tasks have very different surface characteristics, they may nonetheless share a common underlying emotion recognition skill.

By contrast, Buck et al. (2017) understand ERA and EA as two functionally different processes. This does not, however, imply that high competence in emotion recognition does not facilitate the ability to make correct inferences concerning thoughts and feelings of others. Further, failures to show covariation between ERA and EA test scores could be due to the fact that ERA tests may assess discrimination rather than decoding accuracy, as suggested by Bänziger et al. (2009). That is, they do not provide an accurate test of the ability to decode emotion expressions.

Moreover, to our knowledge, no study has tested Ickes' (1993, 2016) conceptualization of EA in live interactions (as opposed to filmed material, Gesn and Ickes 1999) and related EA to an ERA test where the participants have several answer alternatives, both for positive and negative emotions, and with different intensity emotions. Using a live interaction rather than filmed material should provide a richer source of information for the EA task, whereas an ERA test that reduces reliance on discrimination rather than decoding should provide a more adequate ERA measures. As such, this approach should maximize chances to detect shared variance due to an underlying emotion perception skill.

The Present Research

The present study had the aim to investigate the covariation between ERA as assessed by two different standardized computer tests and EA. Due to the special position that emotion decoding ability should have as a base for understanding others, it ought to be of importance for professions aiming to help others in problematic situations or/and in different degrees of distress. We, therefore, chose an EA task that simulated an excerpt of a therapy session and participants who were students in helping professions. We predicted that EA scores co-vary with ERA scores from standardized tests. To address the possibility of tapping other processes, in particular, emotion discrimination instead of emotion recognition as suggested by Bänziger et al. (2009), we compared a well-established test with four response options, the Diagnostic Analysis of Nonverbal Accuracy (DANVA, Nowicki and Duke 1994), with a newer test that offers 12 different emotion alternatives, the Emotion Recognition Assessment in Multiple modalities (ERAM, Hovey et al. 2018; Holding et al. 2017). An additional aim was to explore whether completing an emotion recognition test prior to the interaction improves EA by making the notion of accuracy salient and potentially increasing accuracy motivation. For the EA task, participants freely reported what thoughts and emotions they thought the other person was experiencing at specified time points.

Method

Participants

Sixty students (36 women) enrolled in clinical psychology ($n = 22$), nursing ($n = 9$) and social work ($n = 18$) programs participated in the study. Their mean age was 25 years with a range of 19–40 years, and they were in their 1st to 6th semester. Exclusion criterion was

prior knowledge of the computer tests for emotion recognition used in the study. A sensitivity analysis using G*Power yielded a power of .80 to detect a meaningful correlation of .35 or higher.

Material

The Emotion Recognition Assessment in Multiple modalities (ERAM) is a computerized test developed by Bänziger and Laukka (and previously used by Hovey et al. (2018), Holding et al. (2017), based on a selection of emotional expressions from the Geneva Multimodal Emotion Portrayals database (GEMEP, Bänziger and Scherer 2010; Bänziger et al. 2012). The recording of the stimulus materials for the GEMEP was done using the Stanislavski approach (Moore 1960) to induce emotions instead of posing expressions. The selection of items from the GEMEP-database for the ERAM had the following constraints: (1) Only the two pseudo-language sentences of the database were used (the "aaa"-expressions were not used) and were presented an equal number of times; (2) Twelve different emotional expressions (pride, interest, relief, joy, pleasure, irritation, anger, disgust, anxiety, panic fear, sadness, and despair) were used; (3) The three recording modalities were used (video without sound for the upper part of the body, voice recording only, and both voice and video for the upper part of the body); (4) Each of the emotions were presented with an easy and a difficult item for each of the emotions, in each modality; (5) There were an equal number of female and male encoders (5 + 5); (6) Each emotion included 3 male and 3 female speakers and the two sentences 3 times each; and (7) Sound levels were normalized within each actor. Thus, ERAM consists of 24 video recordings of facial emotional expressions without sound (i.e., no voice), 24 emotional expressions with voice (the actors used two different pseudo-speech sentences), and 24 video recordings with combined facial and vocal emotional expressions in a fixed order (i.e., in total 72 items). Stimulus length varied between approximately 1 and 2 s. The 10 actors posed 12 different emotional expressions; pride, interest, relief, joy, pleasure, irritation, hot anger, disgust, anxiety, panic fear, sadness, and despair. The participants' task was to choose the most appropriate of these 12 labels.

The ERAM has good psychometric properties. Data from collected from 260 participants across eight unpublished data collections shows a close to perfect normal distributions for all three modalities (See Table 1). The correlation between the three different modalities (video without sound, as voice recording only, and both video and voice in the same recordings) showed low but significant correlations suggesting that the three different modalities addressed three different, but related, aspects of emotion decoding.

From the Diagnostic Analysis of Nonverbal Accuracy (DANVA, Nowicki and Duke 1994), which is also available as a computerized test, 24 color photos of facial emotional expressions and 24 voice recordings of emotional expressions of happiness, anger, fear, and sadness posed by male and female actors were chosen. In the voice recordings the actors utter two phrases" I'm going out of the room now," and "I'll be back later." The photos were shown for 2 s. The participants' task was to choose one of the four emotion labels that best describes the emotion presented. Both DANVA and ERAM were presented on PC- laptops with over-ear headphones.

Table 1 The statistical properties of ERAM based on an accumulation of 260 participants from eight unpublished data collections

N = 260 ERAM	Visual modality	Auditory modality	Both modalities
Mean	12.93	10.19	15.08
Median	13.00	10.00	15.00
Mode	14.00	11.00	16.00
Minimum	4.00	3.00	5.00
Maximum	21.00	16.00	21.00
SD	2.84	2.76	3.11
Percentiles			
25	11.00	8.00	13.00
50	13.00	10.00	15.00
75	15.00	12.00	17.00
Skewness	− 0.19	0.07	− 0.46
Curtosis	− 0.17	− 0.39	0.17
Correlations (Pearsons *r*)			
Visual modality			
Auditory Modality	.41*		
Both Modalities	.47*	.33*	

*$p < .01$

Procedure

The study was announced during lectures that were part of clinical psychology, nursing, and social work programs. Students interested in participation provided contact information and were later contacted by phone. Interested participants received an email informing them about the study and asking for informed consent. Importantly, they were instructed to prepare a story of a self-experienced emotional event to narrate (2.5–3 min) during the experiment. For each session, two participants were recruited. Some participants were aware of each other from class, whereas others were strangers.

To measure EA, we modified Ickes (2016) unstructured dyadic interaction paradigm. Specifically, we used a simulated therapy session as the context of the interaction. Participants were instructed to remember a self-experienced emotional event and to prepare a narration of this event. They were instructed to talk about an emotionally charged, but not traumatic, positive or negative event for 2.5 to 3 min. (Retrospectively it turned out that many of the negatively valenced narratives were based on self-experienced events that could be considered as quite upsetting; severe accidents, physical abuse, and serious illness.)

During the session, participants first provided informed consent and then rated how well they knew their interaction partner (acquaintance was rated on a scale from total strangers to close friends), as prior knowledge of a person can impact EA (Stinson and Ickes 1992). Half the participants completed the computerized emotion recognition tests first, the other half completed these following the interaction.

Interaction Task

Two digital cameras (NIKON D3300) with a resolution of 1920 × 1080 pixels were used to video record the "therapy session." Next to each camera there was a stand with a studio light with a bouncer to provide good light conditions without sharp shadows. Each camera had an external microphone (Røde VideoMic GO).

Participants were seated opposite each other at a table. One participant assumed the role of the client and narrated the emotional event they had prepared (encoder/"client") while the other participant listened (decoder/"professional helper"). Then the roles were reversed.

Following the interaction, each participant saw a video of the interaction and annotated both parts (telling their emotional story and listening to the other participant). The participants were not informed about this part of the procedure until after the stories were presented. The instructions suggested a minimum of five and a maximum of ten annotations of thoughts and feelings. The participants were provided with a list of emotion words in case they needed help to find an appropriate emotion label as informal pretests of the procedure had revealed that some participants were unsure of which word to use. The emotion words on the list were taken from both the ERAM-test and the emotion words form the Swedish version of The Circumplex Structure of Affect (Knez and Hygge 2001, see also Larsen and Diener 1992; Russell 1980).

Following this, participants watched the two videos showing the other participant, first when the other participant acted as professional helper and second when the other participant told his/her own story (the client role). The experimenter stopped the video at the time points where both thoughts and feelings had been indexed by the participant shown on the video. The other participant was asked to indicate both a feeling and a thought for each time point.

At the end of the experiment, participants were debriefed and received a gift card that could be used at a variety of Swedish shops (corresponding to approx. 10 Euros). Participants were also given a second informed consent form to complete, which asked for their agreement for the use of the videos for future research. The experimenter reminded participants that it was possible to get feedback on their results within the next few months and that they would be contacted by email with the specific times and dates. Finally, they were asked not to talk about the study with others. Participants were given the option of support in case the study conditions had been perceived as distressing. No participant requested this support. Each session took approximately 1.5 h in total.

Data Treatment and Analysis

The ERAM test was scored such that the correct answer received 2 points, the lower or higher activation of the same feeling received 1 point (if applicable e.g., sadness and despair), whereas the same valence received .5. The sum for each positive emotion was then divided by 5 (as the test included five positive emotions) and the sum for each negative emotion was divided by 7 (as the test included seven negative emotions). This was done separately for the three different modalities of the test (video only, audio only, video with audio). The DANVA includes only one positive emotion and three negative emotions, we therefore followed the standard procedure for this test by calculating an average score based on the correct answers only (separate for face and voice).

To exclude that the emotion recognition in the dyadic interactions (i.e., the simulated therapy session) was based on emotion discrimination instead of emotion recognition we used open answers. To code the open answers, we used a procedure adapted from Ickes (2001, 2016) which gives accuracy points for the mind-reading of thoughts.

For thoughts "accuracy points" were given depending on how similar the words to describe the narrator's thoughts used by the encoder and decoder were (see Ickes 2001). The similarity was established by consensus expert-ratings by authors 2 to 5. Zero points were given if the content was "essentially different" (i.e., when the answer given by the decoder did not correspond to the meaning of the reported thought), 1 point was given if the content was "similar, but not the same" (i.e., when of answer given by the decoder did correspond to the gist of the meaning of the reported thought) and 2 points if the content was "essentially the same" (i.e., when answer given by the decoder did closely match the meaning of the reported thought).

The empathic accuracy of the emotions reported by the participants were scored based on the emotion circumplex by Russell (1980) using a Swedish version (Knez and Hygge 2001, see Fig. 1). When the participants who narrated the emotional event and the decoder both used emotion words to describe the feelings of the narrator that fell into the same

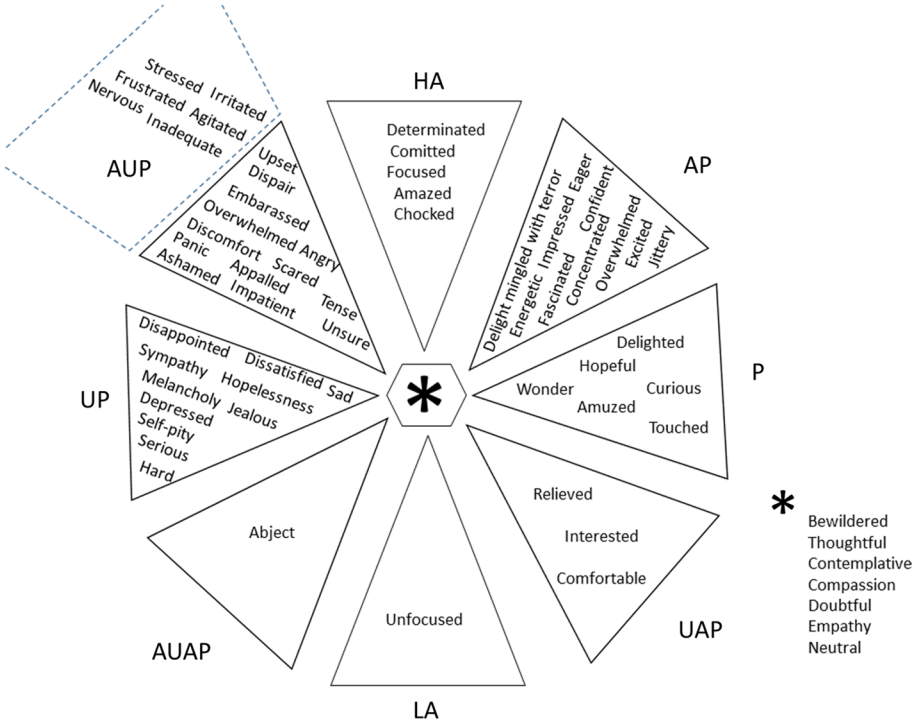

Fig. 1 Listing of the emotion words reported during the mind-reading/empathic accuracy organized together with emotion words in the Emotion Circumplex model originating from Russell (1980, see also Larsen and Diener 1992: Knez and Hygge 2001). Words reported as feeling, but which couldn't be categorized either as high or low on arousal, or with any valence, are located in the middle of the model and denoted with an *. Abbreviations: HA = High Activation, AP = Activation Pleasant, P = Pleasant, UAP = UnActivated Pleasant, LA = Low Activation, UAUP = UnActivated UnPleasant, UP = UnPleasant, AUP = Activated UnPleasant

group on the circumplex (see Fig. 1), the decoder received 2 points. If the decoders used words that were in groups to either side of the group where the word reported by the narrator was place, the decoder scored 1 point, if the decoders used words that were up two steps from the correct group of words, they received .5 points. Words falling into any other group of emotion words scored zero. Words that were not included in the circumplex were placed into it based on consensus expert ratings by author 2-5 (see Fig. 1). The coders scored the answers together and resolved disagreements by discussion.

Results

Preliminary Analyses

Only the EA scores from the "professional helper" trying to understand the "client" were included in the analyses. We first verified whether results differed as a function of the order in which the emotion recognition and the interaction task were completed. No significant differences were found (see Table 2).

We correlated the scores on the EA-task for each dyad with each other to find out if someone who easily understood other's unspoken feelings and thoughts was also easy to understand. We found no substantial explained variance ($r^2 < .01$) for feelings, but a reasonable sized explained variance for thoughts ($r^2 = .11$). The reliabilities for the two ERA tests showed Cronbach's $\alpha = .79$ for ERAM and Cronbach's $\alpha = .56$ for DANVA over all items. Bivariate correlations between the EA scores and the different scores of ERA tests were computed to check for multicollinearity (see Table 3).

Main Analyses

As we found no indication of multicollinearity, we analyzed the data using regression models. A first linear multilevel regression analysis was conducted with the *feeling* EA score as outcome variable and the five different emotion recognition measurements (the different modalities of the two tests) as level one predictor variables, as well as the rated level of acquaintance with the other person in the dyad and semester of study as level two predictor variables, all nested within dyad. The results showed that the five emotion recognition measures tended to predict EA of feelings, $F(5, 54) = 2.09$, $p = .08$, and explained 16

Table 2 Average of empathic accuracy scores for the "Clients" feelings and thoughts divided on whether the participants had the emotion decoding tests before or after the Mind reading, and the t-value of difference between the two different groups

	Emotion perception tests	Average score	*SD*	*t*	df	*p*	Cohen's *d*
EA score Feeling				−0.93	58	.36	−0.24
	Before	1.12	0.22				
	After	1.19	0.32				
EA score Thought				−1.53	58	.13	−0.40
	Before	0.81	0.38				
	After	0.97	0.47				

Table 3 Bivariate correlations, and indications of *p* values, between the scores for the measured variables

	1	2	3	4	5	6	7	8	9
1 Empathic Accuracy Feeling	–								
2 Empathic Accuracy Thought	.46***	–							
3 Video only ERAM	− .26*	− .03	–						
4 Audio only ERAM	.04	.04	.04	–					
5 Video & Audio ERAM	.01	.14	.55***	.24	–				
6 Face DANVA	− .08	− .01	.28*	.35**	.43***	–			
7 Voice DANVA	.18	.21	.23	.03	.31*	.27*	–		
8 Semester	.21	.24	− .36**	.07	− .21	− .04	− .09	–	
9 Acquaintance	.18	.16	− .06	− .02	.16	.08	− .06	.08	–

The numbers in the top row correspond to the variable indexation in the first column

*$p < .05$, **$p < .01$, ***$p < .001$

percent of the variance (8.5 percent for R^2_{adjusted}). Semester of study and the rated degree of acquaintanceship with the other person in the dyad at level 2 did not add any significant explained variance, $F(2, 51) = 1.27$, *n.s.*. The emotion recognition measure that significantly contributed to the model was the ERAM video channel score (see Table 2, left panel). Yet, notably, the ERAM video channel scores were negatively related to the EA score. Moreover, the DANVA scores on emotion recognition showed a tendency to a positive correlation (see Table 4, left panel). When adding the level two predictors, this variable reached significance.

A second linear multilevel regression analysis was conducted, with the *thought* EA score as outcome variable and the five different emotion perception accuracy measurements as level one predictor variables, as well as semester of study, and the rated level of acquaintance with the other person in the dyad as level two predictor variables.

The results showed that the five emotion recognition scores did not predict thought EA scores, $F(5, 54) = 1.02$, *n.s.*. Semester of study and the rated degree of acquaintanceship with the other person in the dyad at level 2 showed a non-significant tendency to increase the amount of explained variance, $F(2, 52) = 2.49$, $p = .09$. It was the semester of study and DANVA scores that tended to contribute to the model (see Table 5, right panel). However, the entire model did not approach significance, $F(7, 52) = 1.48$, $p = .20$.

Discussion

Participants who completed the computerized ERA tests before the simulated therapy session did not differ in their EA (neither for feelings nor for thoughts) from those who completed the ERA tests afterwards. Thus, completing ERA tests before the EA task did not seem to prime the ability or motivation to perform the EA task.

The results also indicated—at least for the EA of thoughts—that the length of study has a positive impact. This speaks in favor of the notion that the professional training of mental health professionals does indeed foster EA skills for thoughts. This finding is in line with findings from Hall et al. (2015) and Ruben et al. (2015). Yet, as in these two studies, the explained variance for length of professional training (number of semesters) was low in the present study as well (about 4%, in studies by Hall et al. and Ruben et al. the effect of

Table 4 The standardized coefficients (β), t-value (*t*), probability (*p*), and VIF for the predictor variables in multilevel regression analysis with EA feeling scores as outcome variable in the left panel and EA thought scores as outcome variable in the right panel

Empathic accuracy feelings thoughts						Empathic accuracy					
Model	Variables	β	*t*	*p*	VIF	Model	Variables	β	*t*	*p*	VIF
1	Video only ERAM	−.39	−2.57	.01	1.48	1	Video only ERAM	−.17	−1.05	.30	1.48
	Audio only ERAM	.05	0.35	.73	1.19		Audio only ERAM	.03	0.24	.81	1.19
	Video and Audio ERAM	.19	1.17	.25	1.73		Video & Audio ERAM	.21	1.25	.22	1.73
	Face DANVA	−.14	−0.94	.35	1.39		Face DANVA	−.13	1.58	.12	1.39
	Voice DANVA	.25	1.87	.07	1.15		Voice DANVA	.22		.12	1.15
2	Video only ERAM	−.31	−1.94	.06	1.66	2	Video only ERAM	−.05	−0.28	.79	1.66
	Audio only ERAM	.06	0.41	.69	1.21		Audio only ERAM	.03	0.20	.84	1.21
	Video & Audio ERAM	.15	0.89	.38	1.84		Video & Audio ERAM	.19	1.09	.28	1.84
	Face DANVA	−.16	−1.08	.29	1.40		Face DANVA	−.15	−1.03	.31	1.40
	Voice DANVA	.27	2.02	.05	1.17		Voice DANVA	.24	1.75	.09	1.17
	Acquaintance	.16	1.22	.23	1.09		Acquaintance	.13	1.01	.31	1.09
	Semester	.13	0.94	.35	1.17		Semester	.26	1.91	.06	1.17

In Model 2 Semester of study and rated Acquaintance are added predictor variables

training was about 5–6%). Moreover, some caution must be taken, as length of study could be influenced by self-selection. That is, those who find it easier to understand other people's unspoken thoughts and feelings might be more likely to continue their studies.

Furthermore, there was a negative correlation between semester of study and the video-only modality of ERAM. This may indicate that with training, help professionals may realize that when discussing issues that potentially can cause emotional turmoil, smiles and other facial behaviors may serve emotion regulatory functions. Therefore, an implicit strategy that discounts facial expressions may develop in later semesters of studies.

Importantly, the results gave no indication that high scores on ERA tests predict high EA scores. To the contrary, higher scores on facial emotion recognition in emotion recognition tests were negatively associated with EA scores for feelings. Facial emotion recognition scores showed negative associations with the EA score for both ERAM and DANVA, even if the beta value only was significant for ERAM. Yet, given that ERAM uses videos and DANVA stills, this consistency across face measures is suggestive of a reliable finding. That facial emotion recognition accuracy negatively predicts empathic accuracy of feelings in a real dyadic interaction may suggest that strategies that are useful for the decoding of highly prototypical expressions (i.e., pattern matching) may even be counter-productive for EA of feelings.

The results from Hall and Schmid Mast (2007) may shed some light on this finding. These authors found that nonverbal cues were considered to express feelings more than thoughts, and that verbal information was more important for inferences about thoughts than about feelings. If this belief holds true for people in general, it might not be surprising that participants with high scores on ERA for facial expressions have lower scores on EA. That is, a belief that nonverbal cues are useful for detecting feelings would lead participants to use this strategy for the feeling EA task. However, this strategy becomes problematic when facial expressions are used to regulate emotions while talking about emotional issues (see Marci et al. 2004). Thus, those participants who are better at labeling facial expressions (high ERA) are more likely to be misled by expressions that served a regulatory function rather than an emotion expression function, such as smiling when talking about a harrowing event.

Voice ERA scores did not correlate with EA, suggesting that, at the least these scores do not reflect counter-productive strategies. Such reasoning is in line with the finding by Kraus (2017) that only using the voice for making emotion inferences is better than face and voice together. In sum, these findings suggest that classic ERA tests—which show more or less intense prototypical expressions bereft of context—may not be an ideal means of assessing the emotion decoding abilities that are relevant to real-life interactions. First, the expressions shown in a conversation may reflect conversational rules and emotion regulation more than emotional states. Second, given the above, even if the emotions expressed reflect an underlying state, they are unlikely to be prototypical.

Specifically, the EA tasks used here tap processes linked to perspective-taking more than simple pattern matching, i.e., the careful observation of facial or auditory patterns (see Hess 2015, for a discussion of these two forms of emotion decoding). Participants had access to the full narrative and could leverage their understanding of the situational context and their impressions of the other's personality when "mind-reading." As such, they were able to base their efforts to understand others' thoughts and feelings about a rich social context.

By contrast, ERA tests are reasonably best performed by using a different strategy. As noted before, Bänziger et al. (2009) pointed out that when participants are given very few labels to choose from, they can use emotion discrimination strategies. Specifically, they

only need to pick up on one facial feature that does not match with a label to exclude the label and then keep the one that remains. That is, the format invites simple pattern matching and reasoning. However, this explanation cannot be used for the ERAM test that arguably taps a process whereby participants try to understand the fine-grained aspects of the emotion expressed (a process also involving, for example, differentiating between expressions with the same core relational themes, but with different intensities).

Despite this difference between the two ERA tests, the covariation between the different modalities in these emotion perception tests was higher (except for the audio-only in ERAM) than the correlation between the ERA tests and the scores of the EA task, suggesting that the latter taps a different skill. For example, high scores on the video-only part of ERAM and low scores on EA for feelings are compatible with the idea that the participant might have misinterpreted laughter (Marci et al. 2004) or smiling (Ansfield 2007) when talking about negative events, which helps regulating the negative emotions aroused by the narration, as an instance of positive emotion expression. When attending to the verbal content of an interaction—as in the EA task—such a misattribution is considerably less likely. Moreover, in real-life situations, facial actions like an eyebrow raise may also be used as a marker of emphasis (see e.g., Chovil 2005), and this nonverbal signal may then interfere with concurrent emotion expressions. Yet, when attending to speech content such confusions are less likely. As such, the present results also provide support for Buck et al. (2017) and Ickes' (2016) suggestion that ERA and EA are supported by two functionally different processing systems.

There are a number of inherent problems with the present design. The design's clear advantage is the ecologically valid setting. Yet, this also means that the emotional stories told by the participants had different contents and were different both in intensity and valence. Furthermore, the "clients" were different in their emotional expressivity, thus, some were harder to understand with respect to their unspoken thoughts and feelings than others (see Ickes 2016; Zaki et al. 2008). All of these factors add noise, which could also explain why the standard ERA tests showed so little predictive value for EA-scores.

To control these sources of variance, an alternative design would be to use pre-recorded videos and thereby provide the same information for all decoders, similar to Ickes standard paradigm (Ickes 2001, 2016). However, this approach necessarily reduces the ecological validity of the situation for a professional helper.

Yet another potential problem is the difference in methodology. The ERA tests used answer alternatives, whereas the EA task had open answers. This might be one reason for the very low or negative correlation between the ERA tests and the EA scores. Using the same answer format would have allowed for stronger support of Buck et al.'s (2017) notion that these test do test two different abilities. Because few answer alternatives are used in the EA, finding strong correlations between ERA and EA may assess other abilities, like response discrimination. Thus, the most convincing results would be come from a design where both ERA and EA are assessed with open answers.

Moreover, even though the emotion expressions in the ERAM-test were induced via emotion induction using the Stanislavski approach (Moore 1960), the test only includes stimuli with clearly visible unregulated expressions. By contrast, in the EA task, emotion expressions were regulated and masked. A use of an ERA test with masked emotions could be an alternative way of addressing the relation between ERA and EA. A further alternative would be to use samples from videos like the ones recorded in this study and filter the speech to create an ERA-test. However, if it were necessary to use essentially the same stimulus material for the ERA and EA tasks to find some a reasonably strong positive correlation between them, the question arises of whether something like a general emotion

decoding ability can be measured in a meaningful way at all. The circumstances under which people are required to understand the emotions of others in everyday life vary necessarily from situation to situation, and if emotion recognition ability changes with every contextual shift, it would be hard to argue for the general adaptive value of this skill.

One additional problem that some participants reported was that they had to differentiate between what they felt when they talked about the event during the session and what they felt during the emotionally charged situation at the time it happened. As they (but not the helper) knew what they originally felt, and this knowledge may have influenced their ratings, the annotation may have been biased. However, this problem also occurs in actual therapy sessions. That is, clients may, at times, experience a mix between the relived emotion of the event they are taking about and other emotions relating to the current situation. For example, for a strong negative emotional event, negative emotions might be elicited when reliving the event, as well as positive emotions from the relief that the situation is in the past (and handled). Such interplay among emotions and thoughts demands much more from the professional helper than does the decoding of an emotion expression in the face or the voice in a standard test.

The results from this explorative study do not provide clear cut answers. However, the results suggest that some forms of ERA may not only not relate to EA but may even be detrimental. Most people would have both skills of pattern-matching and perspective-taking, but some may rely much more on pattern-matching than on taking into account the rich social context information provided in an interaction. The present research suggests that emotion decoding is not a simple skill. It has different facets (see e.g., Schlegel et al. 2017), which may be of more or less use in a given situation. Especially when complex emotional reactions are to be recognized, a test based on simple, context-free stimuli and providing a limited choice of answer alternative may not be a good predictor. As such, this research falls in line with research on emotion recognition that uses more complex stimuli and departs from the use of forced-choice labels to better predict actual social interaction competence (cf. Hess et al. 2016). The competence to understand the unspoken thoughts and feelings of others (i.e., EA) is considered valuable in helping professions, as it provides additional information about the needs of the clients. The present results suggest that standardized ERA test may not in fact assess this skill.

Funding Open access funding provided by Mid Sweden University.

References

Ansfield, M. E. (2007). Smiling when distressed: When a smile is a frown turned upside down. *Personality and Social Psychology Bulletin, 33*(6), 763–775. https://doi.org/10.1177/0146167206297398.

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. https://doi.org/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*(5), 1161–1179. https://doi.org/10.1037/a0025827.

Bänziger, T., & Scherer, K. R. (2010). Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 271–294). Oxford, England: Oxford University Press.

Buck, R., Powers, S. R., & Hull, K. S. (2017). Measuring emotional and cognitive empathy using dynamic, naturalistic, and spontaneous emotion displays. *Emotion, 17*(7), 1120–1136. https://doi.org/10.1037/emo0000285.

Chovil, N. (2005). Measuring conversational facial displays. In V. Manusov (Ed.), *The source book of nonverbal measures: Going beyond words* (pp. 173–188). Mahwah, NJ: LEA.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*(2), 203–235. https://doi.org/10.1037/0033-2909.128.2.203.

Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology, 64,* 83–93. https://doi.org/10.1037/0022-3514.77.4.746.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2,* 271–299. https://doi.org/10.1037/1089-2680.2.3.271.

Hall, J. A., & Schmid Mast, M. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion, 7*(2), 438–446. https://doi.org/10.1037/1528-3542.7.2.438.

Hall, J. A., Ship, A. N., Ruben, M. A., Curtin, E. M., Roter, D. L., Clever, S. L., et al. (2015). The test of accurate perception of patients' affect (TAPPA): An ecologically valid tool for assessing interpersonal perception accuracy in clinicians. *Patient Education and Counseling, 94,* 218–223. https://doi.org/10.1016/j.pec.2013.10.004.

Hess, U. (2015). Nonverbal communication. In H. Friedman (Ed.), *Encyclopedia of mental health* (vol. 2nd Ed.).

Hess, U., & Bourgeois, P. (2010). You smile—I smile: Emotion expression in social interaction. *Biological Psychology, 84,* 514–520. https://doi.org/10.1016/j.biopsycho.2009.11.001.

Hess, U., Kafetsios, K., Mauersberger, H., Blaison, C., & Kessler, C.-L. (2016). Signal and noise in the perception of facial emotion expressions: From labs to life. *Personality and Social Psychological Bulletin, 42,* 1092–1110. https://doi.org/10.1177/0146167216651851.

Holding, B. C., Laukka, P., Fischer, H., Bänziger, T., Axelsson, J., & Sundelin, T. (2017). Multimodal emotion recognition is resilient to insufficient sleep: Results from cross-sectional and experimental studies. *Sleep, 40,* (11). https://doi.org/10.1093/sleep/zsx145.

Hovey, D., Henningsson, S., Cortes, D. S., Bänziger, T., Zettergren, A., Melke, J., et al. (2018). Emotion recognition associated with polymorphism in oxytocinergic pathway gene ARNT2. *Social Cognitive and Affective Neuroscience, 13,* 173–181. https://doi.org/10.1093/scan/nsx141.

Ickes, W. (1993). Empathic accuracy. *Journal of Personality, 61,* 587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x.

Ickes, W. (2001). Measuring emphatic accuracy. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 219–241). Mahwah, NJ: Erlbaum.

Ickes, W. (2016). Empathic accuracy: Judging thoughts and feelings. In J. A. Hall, M. S. Mast & T. V. West (Eds.), The social psychology of perceiving others accurately; the social psychology of perceiving others accurately (pp. 52–70, Chapter xviii, 430 Pages) New York, NY: Cambridge University Press, New York, NY. https://doi.org/10.1017/cbo9781316181959.003

Knez, I., & Hygge, S. (2001). The circumplex structure of affect: A Swedish version. *Scandinavian Journal of Psychology, 42*(5), 389–398. https://doi.org/10.1111/1467-9450.00251.

Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *American Psychologist, 72*(7), 644–654. https://doi.org/10.1037/amp0000147.

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Emotion* (pp. 25–59, Chapter ix, 326 Pages). Thousand Oaks, CA.: Sage Publications, Inc.

Marci, C. D., Moran, E. K., & Orr, S. P. (2004). Physiologic evidence for the interpersonal role of laughter during psychotherapy. *Journal of Nervous & Mental Disease, 192*(10), 689–695. https://doi.org/10.1097/01.nmd.0000142032.04196.6.

Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology, 9,* 507–536. https://doi.org/10.1146/annurev.psych.59.103006.093646.

Moore, S. (1960/1984). The first and simplified guide to Stanislavski's teachings, the Stanislavski system, the professional training of an actor. Harrisonburg, VA: Penguin Books

Nelson, N. L., & Russell, J. A. (2016). A facial expression of pax: Assessing children's "recognition" of emotion from faces. *Journal of Experimental Child Psychology, 141,* 49–64. https://doi.org/10.1016/j.jecp.2015.07.016.

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. https://doi.org/10.1007/BF02169077.

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,* 355–362. https://doi.org/10.1111/jasp.12301.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39,* 1161–1178. https://doi.org/10.1037/h0077714.

Salovey, P., Detweiler-Bedell, B. T., Detweiler-Bedell, J. B., & Mayer, J. D. (2008). Emotional intelligence. In M. Lewis, J. M. Haviland-Jones, & L. Feldman Barrett (Eds.) *Handbook of emotion*, 3rd ed. New York: Guilford Press

Salovey, P., & Mayer, J. D. (1990). *Emotional intelligence*. New York: Baywood Publishing Company Inc.

Schlegel, K., Boone, R. T., & Hall, J. A. (2017). Individual differences in interpersonal accuracy: Using meta-analysis to assess whether judging other people is one skill or many. *Journal of Nonverbal Behavior, 41,* 103–137. https://doi.org/10.1007/s10919-017-0249-0.

Stinson, L. & Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of Personality and Social Psychology, 62*(5), 787–797. https://doi.org/10.1037//0022-3514.62.5.787.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19,* 399–404. https://doi.org/10.1111/j.1467-9280.2008.02099.x.