Emotion
2009, Vol. 9, No. 5, 691–704

1528-3542/09/$12.00 DOI: 10.1037/a0017088

Emotion Recognition From Expressions in Face, Voice, and Body: The Multimodal Emotion Recognition Test (MERT)

Tanja Bänziger, Didier Grandjean, and Klaus R. Scherer
University of Geneva

Emotion recognition ability has been identified as a central component of emotional competence. We describe the development of an instrument that objectively measures this ability on the basis of actor portrayals of dynamic expressions of 10 emotions (2 variants each for 5 emotion families), operationalized as recognition accuracy in 4 presentation modes combining the visual and auditory sense modalities (audio/video, audio only, video only, still picture). Data from a large validation study, including construct validation using related tests (Profile of Nonverbal Sensitivity; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979; Japanese and Caucasian Facial Expressions of Emotion; Biehl et al., 1997; Diagnostic Analysis of Nonverbal Accuracy; Nowicki & Duke, 1994; Emotion Recognition Index; Scherer & Scherer, 2008), are reported. The results show the utility of a test designed to measure both coarse and fine-grained emotion differentiation and modality-specific skills. Factor analysis of the data suggests 2 separate abilities, visual and auditory recognition, which seem to be largely independent of personality dispositions.

Keywords: emotion recognition, emotional intelligence, test, multimodal

Emotionally competent individuals are characterized by optimal functioning of the emotion mechanism in two major domains—emotion *production* and emotion *perception* (Scherer, 2007). Whereas emotion production competence refers to the appropriateness of the total pattern of bodily and behavioral changes as an adaptive response to a relevant event, allowing the organism to successfully cope with its consequences, emotion perception competence refers to the ability to accurately perceive and interpret the emotional state of others in social intercourse. The latter competence is generally acknowledged as a central factor of emotional intelligence (Matthews, Zeidner, & Roberts, 2007; Mayer & Salovey, 1993). However, even in emotional intelligence tests that claim to study actual abilities or competences rather than self-report of adjustment (see Scherer, 2007), this ability is not assessed. Thus, the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT; Mayer, Salovey, Caruso, & Sitarenios, 2003), considered as the reference for this approach, contains only a handful of expression items. In addition, the expressions used in this test are highly ambiguous as to the underlying emotions, making it difficult to evaluate how modal and representative the mean of a reference or "expert" group is and what it means.

Although emotion psychologists, in studying emotional expression, have extensively studied the capacity of individuals to recognize emotions from facial and vocal expressions (Ekman, 1972; Ekman & Rosenberg, 2005; Scherer, Johnstone, & Klasmeyer, 2003), often trying to establish the universality of affect expression programs (Ekman, 1992), there has been a surprising lack of concern with the development of psychometrically sound and construct validated test instruments capable of diagnosing individual differences in this important ability. The same has been true in the area of nonverbal communication research, although some attempts have been made to measure *nonverbal sensitivity*, defined as the ability to recognize emotions or interpersonal attitudes from nonverbal cues in face, body, and/or voice (Hall & Bernieri, 2001). Given the long tradition and the pervasiveness of interest in emotion recognition ability from different nonverbal modalities, it is surprising that one finds only a limited number of established instruments in the literature that are designed to measure this capacity, several of which have not been thoroughly assessed for reliability and validity. In particular, most of the tests in this domain focus on a single modality, in general the face or the voice. Some instruments draw exclusively on still photographs of facial expressions[1] (e.g., Japanese and Caucasian Facial Expressions of Emotion; JACFEE; Biehl et al., 1997), whereas other instruments also include vocal expressions (e.g., Diagnostic Analysis of Nonverbal Accuracy; DANVA; Nowicki & Duke, 1994; Emotion

Tanja Bänziger, Didier Grandjean, and Klaus R. Scherer, Swiss Centre for Affective Sciences, University of Geneva, Switzerland.

The research reported here has been supported by the German and the Swiss National Research Foundations and the Swiss Center for Affective Sciences. We acknowledge the important contributions by Rainer Banse, Heiner Ellgring, and Harald Wallbott in recording and analyzing the original Munich corpus. We thank two anonymous reviewers for helpful suggestions.

Correspondence concerning this article should be addressed to Klaus R. Scherer, Swiss Centre for Affective Sciences, University of Geneva, rue des Battoirs 7, CH–1205 Geneva, Switzerland. E-mail: klaus.scherer@unige.ch

[1] It is not our aim here to review the literature in this area and discuss all existing instruments. In addition to the tests mentioned (which are used in this validation study), other tests have been proposed, such as Buck's (1974) Communication of Affect Receiving Ability Test (CARAT) and the Aprosodia Battery (Ross, Thompson, & Yenkosky, 1997).

Recognition Index ERI; Scherer & Scherer, 2008). There also has been a noticeable absence of studies on the ability to infer emotions from gestures and body posture (but see Wallbott, 1998).

With the exception of the Profile of Nonverbal Sensitivity (PONS; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979), which systematically presents face, voice, body, and multimodal versions of the portrayals of a single encoder, no instrument makes use of simultaneously recorded dynamic expressions, producing multimodal stimuli (face, voice, and body) produced by the same encoders. Yet, examining the ability of individuals to recognize emotions with differential degrees of accuracy from different communication channels or modalities seems to be of great interest. An answer to the question of whether there is one general factor (such as *r* in intelligence research) underlying emotion sensitivity and recognition ability or whether there are separate modality-specific abilities has both theoretical importance (for the possible origins of the individual differences) and practical implications (for the interpretation of results and the design of future tests).

An additional drawback of the currently available instruments is that they are generally based on a limited number of emotional categories and, consequently, request the testee to choose between a restricted number of alternative answers (e.g., four in DANVA, two in PONS) rather than to recognize the emotions expressed. It is doubtful whether studies using a small number of response alternatives in an emotion recognition test actually study recognition or whether, more likely, the psychological process involved is discrimination or categorization among a small number of alternatives. Clearly, the latter allows judges to arrive at the correct answer by using exclusion and probability rules. This possibility is particularly problematical if, as is often the case, only one positive emotion (joy) and three negative emotions diametrically opposed on an easily recognized arousal dimension (fear or anger vs. sadness) are used. The need to identify a particular emotion from a larger set of target emotions (e.g., >10) reduces the likelihood of judges relying on such simple combinatorial discrimination techniques. Obviously, real life generally requires true emotion recognition rather than emotion discrimination. Even though specific contexts may reduce the probability of encountering certain emotions, we are still able to detect them if they occur. In consequence, the ecological validity of recognition rates can be expected to increase greatly with the number of alternatives (see Banse & Scherer, 1996).

In this article, we introduce a new instrument for the assessment of the perception of dynamic facial, vocal, and bodily expressions that tries to remedy some of these shortcomings—the Multimodal Emotion Recognition Test (MERT). From the outset, we specify that our ambitions are limited and that the purpose of the new test is highly focused. As described earlier, the recognition of emotion in others is a central component of emotional competence or intelligence as it is the fundamental basis for empathy and the ability to interpret a person's reactions and predict the probability of ensuing behaviors. This ability is of fundamental importance for smooth social interaction, the establishment and maintenance of relationships, and specific encounters such as negotiations. As one might expect, the nature and origins of this competence are extremely complex and the ability to efficiently use this competence in real life will depend on many context factors. Thus, the accuracy of emotion recognition may depend greatly on whether expressions of familiar persons or strangers are to be interpreted, to what extent attention is directed at this task, the nature of the communication situation, the amount of contextual information available, as well as cues from other persons. In addition, the recognition process may unfold in a dynamic fashion as a function of the interaction and the sender's reactions to nonverbal cues of the receiver in the process.

It would seem impossible to measure a person's ability to perceive and interpret emotions in realistic face-to-face interaction situation by means of a standardized test, as many of the factors outlined above are dependent on the nature of the interaction partners, the communication situation, contextual factors, and multiple sources for individual differences. In consequence, our intention is to focus on a core skill that we believe to be a basic constituent of the complex skill outlined above, namely the ability to recognize and interpret fuzzy sets of prototypical expression patterns for certain modal emotions (Scherer, 1994) that are widely shared on a cultural, and to some extent, even universal level (Ekman, 1992). We suggest that these prototypical sets of expression configurations are used when individuals want to display a certain type of emotion for strategic reasons of impression management or when actors produce expressions of certain emotions as part of the authentic performance of a role in a given plot. It seems reasonable to assume that the ability to recognize the meaning of such prototypical expressions for modal emotion as part of a cultural, or even universal code, is a necessary competence to interpret the emotion expressions shown by familiar persons as well as complete strangers (including media performers) in viewing situations in which no interaction takes place and little context information is available. We suggest that this is a core skill that also plays a role in the more complex exercise of emotion inference in face-to-face communication situations, as it seems likely that the observed expression patterns are compared and evaluated with respect to prototypical patterns, especially as regards their authenticity. Given the restriction of our intention to measuring this core skill, it is appropriate to use decontextualized presentations of expression prototypes in different modalities that reproduce the communication channels in real life (seeing and hearing someone in an interaction or in a movie, hearing someone on the telephone, seeing someone's face from afar, or seeing a still photograph of an emotion expression). In consequence, the observers are not asked to judge what emotion the poser "experiences" (although the use of Stanislavski techniques may well induce the appropriate emotions in actors) but rather which emotion the expression "represents."

The MERT instrument was developed on the basis of these considerations. It includes 10 actor-portrayed emotions (anxiety, panic fear, happiness, elation, cold anger, hot anger, sadness, despair, disgust, and contempt), which represent two variants each for five major emotion families (differing on the arousal/intensity dimension). Each emotion is instantiated by three film clips and presented in four modes: video only (facial cues), audio only (vocal cues), audio/video (integrating facial and vocal cues), and still photographs (extracted from the film clips). In what follows, we present the development of the test and results of a study in which we examined item difficulty and test reliability as well as construct validity, using established tests of nonverbal sensitivity and emotion recognition. In addition to allowing us to report the results of this first validation study, the data also allow us to discuss, for the first time, results on modality and emotion effects

for a set of standard emotion portrayals, as well as a more detailed analysis of individual differences.

Method

To create the MERT, we selected 30 emotion portrayals, videotaped in the context of a long-term research project on the recognition of emotion. The construction of the audiovisual corpus established in this project, from which MERT portrayals were extracted (the Munich corpus), has been described extensively in other publications (Banse & Scherer, 1996; Scherer & Ellgring, 2007). For the convenience of the reader, we reproduce some elements of this description in the following paragraphs. The selection criteria for including portrayals as items for the MERT are integrated into this condensed description of the corpus.

Actors Included in the Test

The emotion portrayals in the corpus were produced by 12 professional stage actors (six men and six women). All actors were native speakers of German. They had all graduated from professional acting schools and were regularly employed in radio, TV, and stage work. They were paid for their participation. Of the original 12 actors, 10 are still represented in the selection of portrayals used for MERT. The actors contributed with differential frequency to the test items: one portrayal (two actors), two portrayals (three), three portrayals (two), five portrayals (two), and six portrayals (one). No actor was used twice for the same emotion category. We chose this unequal representation of actors for the emotions to decrease the possibility of associating a specific actor with a given set of emotion categories and to thereby use exclusion rules. Exactly half of the portrayals (15) in MERT are produced by women or men.

Emotions Included in the Test

In the original corpus, 14 different emotions were used. Eight emotions represented four basic emotion families (see Ekman, 1994), having similar emotion qualities but differing in intensity and arousal levels: the anger family (hot anger, cold anger), the fear family (panic fear, anxiety), the sadness family (despair, sadness), and the happiness family (elation, happiness). In addition, interest, boredom, shame, pride, disgust, and contempt were included. Although reviews of emotion recognition studies show that the basic families are generally well discriminated on the basis of nonverbal cues, only few studies have examined the capacity to discriminate emotions within families. One of these studies, using the Munich corpus (Banse & Scherer, 1996), showed a relatively high degree of within-family confusion, but also documented the existence of recognizable differences in nonverbal displays within emotion families. Starting from this basis, we postulate that genuinely high ability to recognize emotional expressions should go beyond the capacity to discriminate between the basic emotion families that are used in most emotion recognition tasks, an ability we suggest labeling as *coarse recognition*. Given our aim to construct a test that is sensitive to the whole gamut of emotional competence levels, we decided to select the four pairs of emotions from the corpus that consist of two variants for each of the four "basic" families that are most frequently included in emotion research: anger, fear, sadness, and happiness (hot anger/cold anger, panic fear/anxiety, despair/sadness, elation/happiness). In addition, we included portrayals of disgust and contempt, two emotions that are often considered to be related, although more distantly (disgust being less intense and generally provoked by sensory stimuli; contempt as more intense and triggered by moral evaluation of others, with an underlying common dimension of moral repulsion). We postulate that the capacity to recognize emotions within each family constitutes evidence for fine-grained recognition.

Definitions of Emotions for the Actors

The actors were provided with brief scenarios featuring the core theme of each emotion category. Details of the recording procedure can be found in previous publications on the corpus (e.g., Banse & Scherer, 1996). All scenarios represent typical antecedent situations for the elicitation of the respective emotion. They were selected from a database of situations collected in several large intercultural studies on emotional experience (Scherer & Wallbott, 1994; Scherer, Wallbott, & Summerfield, 1986). The respective situation descriptions were rewritten in such a way as to render the scenarios stylistically similar across emotions.

Verbal Content of Portrayals in the Test

The use of standardized language material for the construction of the corpus avoided effects of differences in phonemic structure on the acoustic variables. The following two standard sentences from an earlier study (Scherer, Banse, Wallbott, & Goldbeck, 1991) were used: (a) "*Hät sandig pron you venzy*," and (b) "*Fee gött laich jonkill gosterr.*" These meaningless utterances resemble normal speech. Listeners generally have the impression of listening to an unknown foreign language. Fourteen portrayals included in MERT feature the second sentence, and 16 portrayals feature the first sentence, the selection criteria precluding the occurrence of systematic associations between emotions and sentences.

Selection of Test Stimuli

The original corpus from which MERT portrayals are extracted comprises 224 audio/video recordings of the actors' faces and voices. The 224 portrayals of the corpus were selected in several steps. Expert ratings were performed first by 12 acting students; the most "authentic" and "believable" portrayals were selected. A rating study with 12 naïve participants was then used to make the final selection. Hence, the 224 recordings in this corpus were selected for being most readily recognized as belonging to the categories intended by the actors and the researchers. For MERT, we randomly selected 30 recordings of this corpus by using the following criteria: three portrayals for each of the 10 emotion categories to be included in the test; no actor producing the same emotion twice; and equal gender representation and approximate equal verbal content throughout the test. Although the original stimuli had been recorded on high-quality U-Matic tapes, we had to digitize the recordings from an edited VHS tape that contained the edited portrayals selected for further processing after extensive rating studies (e.g., Banse & Scherer, 1996). In consequence, the quality of the audio/video recordings on this tape was not always optimal. We chose to exclude, a priori, recordings with technical

defects (such as a flickering picture or a noisy sound track). This exclusion restricted the choices to be made on the basis of the additional criteria for selection outlined earlier.

We extracted still pictures from the videos to create a "static" subset of stimuli for the test. This was done in an iterative process, with one research collaborator extracting a few expressive stills from each video and other collaborators (all researchers in emotion psychology) selecting the most expressive pictures. In some cases, no picture was perceived as being emotionally expressive; in such cases, further stills were retrieved. In the end, one picture (perceived as the most expressive but still representing a given emotion) for each video was selected. We took care to avoid still pictures in which the eyes were closed or in which the articulation was predominant. We did include stills however, with both open and closed mouths.

Test Design

A computer interface was created to present the digitized test stimuli to the participants on a computer screen. The 30 test stimuli were presented in four different modes: audio/video, video only, audio only, and picture only (a single still extracted from the videos). The resulting 120 stimuli were presented in one of two fixed random orders. The orders were constrained to never include the same stimulus repeated in immediate sequence (in different modes). We further restricted the random sequence by excluding that more than two successive portrayals were produced by the same actor. We also excluded more than two successive portrayals featuring the same emotion, as well as more than two successive portrayals presented in the same mode.

The two different sequences were created partly with the assumption that they would reduce potential order effects on the level of group analysis but primarily for the test–retest reliability assessment (each participant taking the test twice with items in a different order at an interval of 6 weeks).

The answers were recorded by a computer interface, which also presented the 10 alternative targets (emotion categories) in the form of buttons to click on screen. Each participant was always required to give an answer (forced choice) to proceed to the next stimulus. Instructions to the participants were also presented on screen. They were short and mainly required the participants to focus on the stimuli and select the best available emotion descriptor (label) after each presentation. The participants were informed that MERT is a test of nonverbal sensitivity (i.e., an assessment of skills in a specific domain).

Tests Used for Construct Validation

The criteria used for choosing tests that are appropriate for construct validation is linked to our focus on identifying the ability to identify the meaning of prototypical expression configurations for modal emotions. In consequence, we chose instruments that use actor portrayals of such modal emotions and that have been validated in published empirical research.

DANVA

We used a digital version of DANVA (DANVA2–AF and DANVA2–AP; Nowicki & Duke, 1994) with instructions and response alternatives translated into French. DANVA includes 24 audio recordings of vocal expressions and 24 photographs of facial expressions. Facial expressions are portrayals of four emotion categories (anger, fear, joy/happiness, sadness) with two intensities (weak or strong). The test includes three portrayals for each combination of emotion and intensity. Portrayals recognized by more than 80% of the respondents in a pilot study are used in this test. The senders are young Americans (not professional actors); each sender appears maximally two times in the selected portrayals. Vocal portrayals are produced by two professional actors (one man, one woman) and represent the same four emotions with two intensity levels. The emotions are always portrayed in the following sentence: "I'm going out of the room now, and I'll be back later." Portrayals recognized by more than 70% of the respondents in a pilot study are used in the test.

Test participants are asked to select one of four categories (either anger, fear, joy, or sadness) for each portrayal. The 24 facial portrayals are displayed first in fixed order; each photograph being shown for 2 s. The 24 vocal portrayals are then presented in fixed order as well. Accuracy scores are computed separately for facial and for vocal scores as a proportion of correct answers (i.e., answers matching the target category).

PONS

The PONS test (Rosenthal et al., 1979) is composed of 20 audio/video recordings. One sender (a young female research collaborator) portrays 20 attitudes (e.g., "trying to seduce someone," "saying a prayer," "admiring nature," "expressing jealous anger"). They are classified as either dominant or submissive attitudes and either positive or negative attitudes. Each recording is shown in 11 different modes (channels): 1 face alone; 2 body (from neck to knees); 3 face and torso (head/face and body down to the waistline, shows hand gestures); 4 low-pass filtered speech alone (no picture); 5 randomized-spliced speech (Scherer, 1971) alone (no picture); 6 to 11 combinations of the three visual recordings with the two manipulated voice/speech recordings. The 220 portrayals are presented in fixed order and the test participants are required to select one of two potential attitudes for each portrayal. One alternative corresponds to the intention of the sender, whereas the other alternative was chosen randomly by the authors of the test among the remaining 19 attitudes.

JACFEE

For this test, we followed the descriptions provided by Matsumoto, LeRoux, Wilson-Cohn, and Raroque et al. (2000) for the Japanese and Caucasian Brief Affect Recognition Test (JACBART). The test includes 56 photos of posed facial expressions from the JACFEE picture set. On each picture, a different individual (14 Caucasian males, 14 Caucasian females, 14 Asian males, and 14 Asian females) portrays one of seven basic emotions with configurations of facial features following the descriptions of Ekman and Friesen (1975). Consequently, eight different individuals (with balanced gender and ethnicity) posed the same expression (identical facial muscle contractions). The seven emotions are labeled surprise, sadness, anger, happiness, fear, disgust, and contempt. In the test, each expressive picture is displayed for 200 ms between two 500-ms exposures of the same face without expression. This

procedure (display of a neutral face, followed by a relatively brief display of the expressive face, which is "covered" by the neutral face again) does not create an illusion of motion; it is more comparable to a masking procedure. We chose an exposure of 0.2 s, which is long enough to consciously see the expressive face. This exposure is the longest used by the authors of this test (see Matsumoto et al., 2000). The 56 sequences, neutral-expressive-neutral, are presented in fixed order. The participant is required to select one of seven alternative answers.

ERI

The ERI test (Scherer & Scherer, 2008) uses pictures from the Ekman and Friesen Pictures of Facial Affect (PFA; Ekman and Friesen, 1976) series for a 30-item/five-category (anger, fear, joy, sadness, disgust) facial recognition test (FACIAL–INDEX) and vocal emotion portrayals selected from a large corpus of portrayals produced by four German professional radio actors (see Scherer et al., 1991) for a 30-item/five-category (sadness, fear, anger, joy, neutral) vocal recognition test (VOCAL–INDEX), both without time limitation, presented automatically on a computer screen. The test has been validated with $N = 1{,}380$ employees in international companies at various levels of management (see Scherer & Scherer, 2008).

Personality Inventories

To investigate the influence of personality traits as well as other individual differences related to emotional sensitivity, we asked participants to complete the following personality inventories: (a) the NEO Five-Factor Inventory (NEO-FFI; Neuroticism [N], Extraversion [E], Openness [O]; Rolland, Parker, & Stumpf, 1998), a brief version of the NEO Personality Inventory–Revised, investigating the five personality factor model proposed by Costa and McCrae (1988); (b) the scale developed by Carver and White (1994), assessing the interindividual sensitivity of Behavioral Activation/Inhibition systems (BIS-BAS) based on Gray's model; (c) the State–Trait Anxiety Inventory (STAI), measuring trait anxiety (Spielberger & Sydeman, 1994); and (d) the Management Potential (Mp) special purpose scale of the California Personality Inventory (CPI; Gough, 1984), consisting of an empirical selection of items that successfully separated groups of people with different degrees of supervisory responsibility.

Participants

Seventy-two participants, 63 women and 9 men,[2] with a mean age of 22 years (*SD* 4 years) took part in this study. They were all undergraduate psychology students and participated in exchange for course credit. They also were promised and given personalized feedback on their results for all nonverbal sensitivity tests (MERT, DANVA, ERI, JACFEE, and PONS).

Procedure

The tests and questionnaires in this study were all administered on individual computers. The study took place in a computer room habitually reserved for the students to work on their exam papers or to do research on the Internet. The computers were equipped with standard headphones. The sound level was equal on all computers and adjusted for the amplification of the test stimuli. The participants were scheduled to participate in successive groups of 10 (minimum) to 20 (maximum) participants. Every participant was expected to return for three separate sessions. In the first session, MERT was administered and was followed by a request to complete a series of questionnaires, which included self-report measures of personality (NEO–FFI), sensitivity of behavioral activation/inhibition systems (BIS/BAS scales), state anxiety (STAI), and the CPI Mp subscale.

In the second session (the following week), the same participants were requested to complete the other instruments of nonverbal sensitivity, including PONS ($N = 68$ participants from the original pool), ERI Vocal ($N = 68$) and Facial ($N = 72$), DANVA ($N = 70$), and JACFEE ($N = 67$). Finally, a retest session of MERT was organized about 6 weeks after the first session (with 67 participants from the original pool). In the first session, the participants were allocated randomly to one sequence of stimuli presentation for MERT; in the retest session, they completed MERT with the other sequence of items.

The tests and questionnaires were run form a distant server and results were uploaded via file transfer protocol (FTP). There were no technical problems reported during the sessions, but a few result files were not correctly uploaded on the distant server. Some participants also failed to show up at one or sometimes two of three sessions, which resulted in some missing data points. For instance, the second administration of MERT was completed by 67 of 72 participants who completed the test 6 weeks earlier.

Computation of Accuracy Scores in the Four Tests

Accuracy was computed as the proportion of correct answers given by a participant. Each participant obtained a global score in each test (based on all items in the test). For MERT, the overall accuracy score is computed on 120 items (30 emotion portrayals, times four presentation modes). For DANVA, the score is computed on 48 items (24 facial portrayals and 24 vocal portrayals). The two ERI subscales have 30 items each. For PONS, the global accuracy score is computed on 220 items and for JACFEE on 56 items.

Whenever a test allows, we also computed subscores for separate expressive channels (e.g., facial portrayals vs. vocal portrayals). To simplify the presentation and to make the subscores more comparable across tests, we grouped the two voice modes in PONS (40 items; 20 portrayals, times two vocal manipulations) to compute one accuracy score based on vocal portrayals only. An accuracy score for PONS facial portrayals (20 items) was computed, as well as a score based on portrayals combining face and voice (40 items).[3] For DANVA, we computed an accuracy score for facial portrayals (24 items) and an accuracy score for vocal portrayals (24 items).

For MERT, four mode-specific accuracy subscores were computed (30 items each) for the following modes of presentation of

[2] The imbalance in the gender ratio directly reflects the proportion of the two sexes in psychology courses in most parts of the world. As we required participants to come to the lab repeatedly and to respond to a large number of rather demanding performance tests, we felt that it was essential to have a homogeneous sample and use the same incentives and conditions for all participants that is, participating in the study as part of their course work.

[3] Other presentation modes are included in PONS (11 combinations in total), but we are not reporting those scores in this article because they do not compare with the scores obtained in the other tests.

the same set of 30 portrayals: (a) dynamic visual input (mostly face and upper torso) but without sound (video only); (b) stills extracted from the videos (still picture only); (c) vocal presentation without picture (audio only); and (d) a combination of video and audio (audio/video).

In addition to the global score for MERT, which we will refer to as the individual emotion recognition (IR) score, we computed a family recognition (FR) score, which involves a different scoring of accuracy. Here, a response is considered as correct if the target emotion or the other member of the respective emotion family is reported. In other words, this score represents coarse recognition, as described in the introduction, that is, the ability to identify the major emotion categories in discriminating expressions.

One source of incompatibility between tests stems from differing numbers of response options on the answer sheet. To render the accuracy percentages comparable across tests despite differential answer formats, we computed the one-sample effect size estimator called the proportion index, or *pi* (Hall, Andrzejewski, Murphy, Schmid Mast, & Feinstein, 2008; Rosenthal & Rubin, 1989). *Pi* converts any mean accuracy that originates as a proportion, no matter how many response options each item had, to its equivalent proportion were it to have been based on two options.

Results

Item Difficulty

Detailed information on item difficulty and item discrimination indices for MERT and the three tests used for construct validation can be found in the supplementary material for this article (downloadable from www.affective-sciences.org/supplement_material). MERT has 28% of items with very high recognition rates (which we defined as those with a correct response rate above 80%) as compared to 52 to 62% in the four other tests. The percentage of items recognized below *chance level,* defined as a function of the number of alternative answers in a given test was 4% (compared to percentages ranging from 0% to 16% for the other tests). Thus, overall, the level of difficulty of the MERT items is somewhat higher for MERT in comparison to the other tests, mostly due to the smaller number of "easy" items. We feel that this is advantageous as a percentage of more than 50% of easy items may limit the capacity of a test to discriminate a wide range of individuals.

Reliability (Interrater Agreement)

To assess interrater agreement for the MERT, we computed the average interparticipant correlation. The values obtained were: audio/video = .35, audio = .38, video = .40, and still picture = .39, yielding an average of .38 across all modes. The amount of agreement among participants is large, showing that the pattern of responses is consistent across participants despite individual differences (see Zentner, Grandjean, & Scherer, 2008).

Test–Retest Reliability

Participants were required to complete MERT on two successive occasions with 6 weeks between test sessions and a different order of items on both occasions; 67 participants completed both tests. Correlations of scores between the first and the second test administration were as follows: audio/video $r = .55$, audio $r = .56$, video $r = .46$, and still picture $r = .56$. The correlation for the total scores reached $r = .78$. The higher correlation for the total scores in test–retest comparison probably reflects the general observation that scores computed on a larger number of items are more reliable overall than are those for a smaller *N*. The average accuracy percentages in the retest condition were about 3 to 5% higher than for the first administration. Given that participants differentially improved their scores, as is normally observed in test–retest situations involving learnable competences, we consider the stability of the MERT over time as quite satisfactory. Additional indexes and a more detailed explanation are provided in the supplementary materials.

Accuracy of Recognition in Different Modalities

Table 1 shows the mean accuracy scores over presentation modes and emotion. As one would expect, the FR score, requiring only coarse discrimination, is much higher than the IR score. It must be noted, of course, that the chance level also increases somewhat. We find it interesting that the range of the FR score (minimum = .63 to maximum = .85) is about a third lower than that of the IR score (minimum = .43 to maximum = .76), suggesting that individuals differ less on coarse than on fine-grained discrimination ability.

We statistically tested the differences in accuracy for different modes and different emotions. A repeated-measures analysis of variance (ANOVA) of these data with mode (audio/video, audio, video, and still picture), intensity/arousal level (high, low), and emotion (five emotions), as factors revealed the following effects:

- mode, $F(3, 213) = 119.16$, $p < .001$, dynamic facial portrayals (video with or without voice) were better recognized than still pictures or audio only;

Table 1
Descriptive Summaries for MERT Accuracy Subscores

Variable	*M*	*SD*	*pi*
Total scores			
Individual recognition	.61	.07	.93
Family recognition	.75	.05	.96
Mode-specific scores (IR)			
Audio/video	.70	.09	.95
Audio	.49	.10	.90
Video	.69	.08	.95
Picture	.56	.10	.92
Emotion-specific scores (IR)			
Low-intensity arousal			
Anxiety	.42	.14	.87
Happiness	.73	.19	.96
Cold anger	.55	.27	.92
Sadness	.59	.11	.93
Disgust	.61	.19	.93
Total	.59	.18	.92
High-intensity arousal			
Panic fear	.56	.27	.92
Elation	.72	.18	.96
Hot anger	.87	.09	.98
Despair	.41	.22	.86
Contempt	.66	.11	.95
Total	.64	.17	.94

Note. $N = 72$. MERT = Multimodal Emotion Recognition Test; *pi* = proportion index; IR = individual emotion recognition.

• intensity/arousal level, $F(1, 71) = 30.88$, $p < .001$, the more intense emotions were better recognized than the less intense ones; and

• emotion, $F(4, 284) = 78.68$, $p < .001$, as in earlier research in this area, there are sizable differences in recognition accuracy for different emotions.

All interactions were significant Mode × Emotion, $F(12, 852) = 41.48$, $p < .001$, Mode × Intensity, $F(3, 213) = 10.61$, $p < .001$, Emotion × Intensity, $F(4, 284) = 65.75$, $p < .001$, and Mode × Emotion × Intensity, $F(12, 852) = 25.67$, $p < .001$. Figure 1 shows the average percentages of correct answers for the emotions across all four modes. In the interest of easier comparison, we grouped the presentation of the more intense/aroused family members (panic fear, elation, hot anger, despair, contempt) with the less intense/aroused (anxiety, happiness/contentment, cold anger, sadness, despair, disgust). The data show clear interactions between presentation mode and emotions. Most noticeably, for disgust and elation, portrayals presented in audio mode (vocal portrayals) are less well recognized than those including video (facial) cues. Both hot and cold anger are less well recognized from still pictures than from dynamic video clips.

We also computed confusion matrices separately for each of the different modes. Given the number of large tables necessary to report these data, both tables and text can be found in the supplementary materials. The data show, as one would expect, that most confusions occur between family members (which explains why FR scores are much higher than IR scores). This pattern is particularly pronounced for symmetric confusions between anxiety and panic fear on the one hand and asymmetric confusions, despair portrayals being taken for sadness and contempt being taken for disgust but not the reverse, on the other.

Individual Differences in Accuracy

A cluster analysis on the participants' performance as measured by the correct responses across the different emotions for each participant for the different modes (see supplementary materials for a figure of the dendrogram) revealed three groups of participants based on the linkage distance (superior to 2). A repeated-measures ANOVA on the performance data with mode (audio/visual, audio, visual, and still picture) and group (Group 1, $N = 30$; Group 2, $N = 23$; and Group 3, $N = 19$) revealed main effects of mode, $F(3, 207) = 157.19$, $p < .001$, and groups, $F(2, 69) = 66.55$, $p < .001$, and more crucially an interaction effect Mode × Groups, $F(6, 207) = 6.09$, $p < .001$. The mean performance scores of the three groups across modes are shown in Figure 2. This analysis indicates that Group 2 showed the best performance, with audio and still picture being inferior in accuracy. Group 1 had a comparably good performance for the audio/video and video modes, but differs significantly from Group 2 in the audio mode, $F(1, 69) = 7.99$, $p < .01$, and the still picture mode, $F(1, 69) = 45.51$, $p < .001$. The performance of Group 3 is globally inferior compared with the two other groups for all modes, but shows the same profile over modes as that of Group 2. One might speculate that members of Group 1 rely particularly on dynamic cues in the visual domain and do as well as Group 2 in those cases, but are at a loss when only static configurational cues of facial expression are available, as in still pictures.

Correlations Between Subscores and Factor Structure

Table 2 presents the correlations between the four mode-specific subscores. All of these are significant, indicating sizable overlap in the performance for the four presentation modes. This is hardly surprising, as there is a strong dependence between these scores, which are based on judgments of the same portrayals presented in different modes. At the same time, the amount of shared variance does not exceed 25%, suggesting that there is ample room for individual differences in the extent to which their recognition skills are specialized for certain modes or are relatively uniform across all modes.

A central issue in test construction, especially in the area of emotional intelligence testing, is the issue of whether only one general competence factor (small g) underlies the performance or whether there are domain-specific factors (Mayer et al., 2003; Roberts, Zeidner, & Matthews, 2001). It is thus of interest to determine the factor structure of MERT. A principal component analysis on the summary scores (over emotions) for the four separate modes with 72 participants produces one underlying factor (i.e., with an eigenvalue > 1), which does not explain much of the variance (55.5%). If one adds one more factor, the solution accounts for 74.6% of variance. The load-

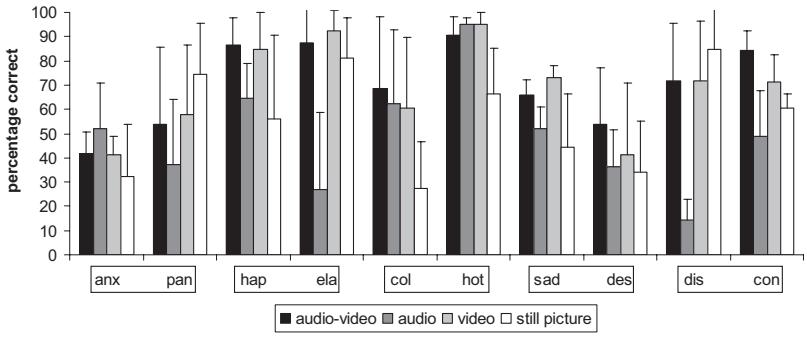

Figure 1. Mean accuracy scores for the basic emotion family pairs in four modes. anx = anxiety; pan = panic fear; hap = happiness; ela = elation; col = cold anger; hot = hot anger; sad = sadness; des = despair; dis = disgust; con = contempt.

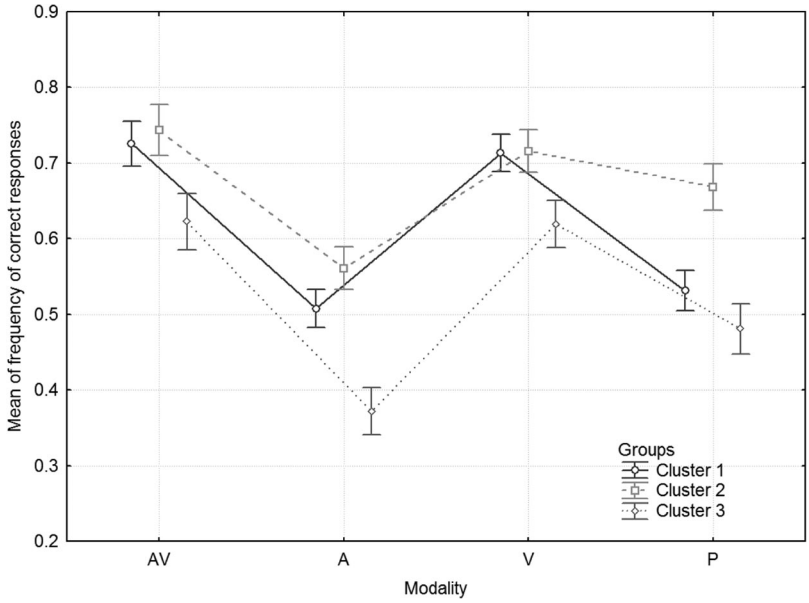

Figure 2. Mean frequency of correct responses by mode for the three clusters of participants. AV = audio/video; A = audio; V = video; P = still picture.

ings (after varimax rotation) are shown in Table 3. These results suggest that there may be two separate factors in emotion recognition competence, one representing visual and the other auditory competence.

It is interesting to note that the audio/video score loads more highly on the auditory factor than on the visual one. A stepwise regression analysis of the individual mode scores on the audio/video scores shows that audio is entered in the first step (r^2 = .196) joined by video in the second step (bringing the r^2 to .302, audio β = .348, p = .001; video β = .339, $p < .01$); still picture does not enter the equation. Thus, despite its overall lower accuracy performance, the auditory impression seems to weigh heavily as a component in combined auditory-visual perception/inference.

Personality Correlates of Recognition Accuracy

We examined to what extent individual differences in total test score and mode specific subscores can be predicted with the help of personality tests. We did not find significant correlations between self-rated personality traits on the NEO–FFI or Trait Anxiety (STAI). For the BIS/BAS scale, we found significant positive correlations between the BAS Drive subscale and MERT audio (.263), video (.314), and audio/video (.260) scores suggest that MERT scores might be partially dependent on "achievement motivation" of the testees, especially given the high level of difficulty of the test.

We also examined (using ANOVAs) whether the three testee groups, as revealed by the cluster analysis of the confusion patterns, show systematic differences on the personality measures. A main effect for Bas–D, $F(2, 69) = 9.00$, $p < .001$, $\eta^2 = .21$, confirms and specifies the significant correlations, showing that Group 3 (low performers, M 2.9) is significantly lower on motivational drive than are Groups 1 (high performers except for static input, 4.1) and 2 (overall high performers, 3.9). The only difference between Groups 1 and 2, and a possible explanation for the difficulties Group 1 has with static stimuli, is a main effect for the CPI Mp scale, $F(2, 69) = 3.26$, $p < .05$, $\eta^2 = .08$, showing that Groups 1 (M = 49.1) and 3 (47.3) are significantly lower than Group 2 (56.0) on the CPI scale. As described in Scherer (2007), this CPI subscale shows highly significant correlations with emotional stability, emotion regulation, stress resilience, functional

Table 2
Correlations Between Total Accuracy Subscores for Different Modalities in the Multimodal Emotion Recognition Test (MERT)

MERT	Audio-video	Video	Audio
Video	.436**		
Audio	.443**	.279*	
Still picture	.336**	.535**	.411**

Note. N = 72.
* $p < .05$ (two-tailed). ** $p < .001$ (two-tailed).

Table 3
Factor Structure of the Multimodal Emotion Recognition Test, Principal Component Analysis of Total Accuracy Scores per Mode

Mode	Component 1	Component 2
Video	.887	
Still picture	.799	
Audio		.889
Audio/video	.326	.736

Note. Varimax rotation with Kaiser normalization, factor loadings of four presentation modes (loadings smaller than .3 are not displayed).

coping, and objective measures of fluid intelligence, suggesting that high scores indicate general good adjustment, emotional competence, and "reasonableness." It seems appropriate that this group of individuals also performs well in emotion recognition tasks such as the MERT.[4]

Assessment of Convergent Validity Across Instruments (Construct Validity)

Table 4 shows a descriptive summary of the global accuracy scores in the three tests used for construct validation (the respective values for MERT are included for the sake of comparison). The average raw accuracy score for MERT is lower than the average accuracy scores in the other tests, reflecting the lower chance level and the higher difficulty of the test items (see Table 1). However, once the differential chance levels due to different answer alternatives are taken into account by using the *pi*, MERT fares just as well or even better than other tests. Only the PONS, with only two answer alternatives, has a somewhat lower *pi* than the other tests.

Table 5 presents descriptive summaries of the accuracy subscores for different modes computed in our sample for DANVA, PONS, and ERI.

Tables 6, 7, and 8 show the convergence of scores and mode-specific subscores obtained in different tests. One-tailed tests for significance have been used, as there is an a priori assumption that the ability tests correlate with each other.

Table 6, presenting the correlations of the global scores across tests, shows that the global scores in MERT, JACFEE, and PONS are substantially correlated ($r = .45$ to .59), indicating that they are probably all reflecting an underlying competence of nonverbal emotion recognition. The ERI total score correlates reasonably highly with the MERT FR score (probably because of a comparable type of emotions and response alternatives) and with JACFEE (which is also based on still pictures). The scores for DANVA converge relatively less with those obtained in the other tests; only one correlation gets close to 10% shared variance ($r = .30$ with JACFEE).

Table 7, presenting the correlations of subscores based on dynamic video portrayals (facial expression with or without voice), shows that only the scores obtained for dynamic video portrayals (facial expression with or without voice) are significantly correlated in PONS and in MERT. The scores obtained in the two tests on the basis of the same dynamic facial portrayals, but without vocal portrayals, are not correlated. This observation seems at first difficult to explain. However, it may well be related to the fact that, within the PONS test, the scores obtained for facial portrayals with sound are not correlated with the scores obtained for facial portrayals without sound (the correlation is not significant), whereas the scores for these two modes are correlated in MERT (see Table 4).

Table 8 presents correlations of scores based on static faces and on audio only. As shown in Table 8, there are reasonably high and significant correlations between the scores based on still pictures obtained in MERT, ERI, and JACFEE. Again, DANVA scores based on stills are not as consistently correlated with other comparable subtests. For scores based on recognition in the audio mode (Table 8), scores obtained for MERT audio correlate significantly with scores obtained for DANVA and ERI audio stimuli. The pertinent scores obtained for DANVA, ERI, and PONS do not correlate highly with each other.

Overall, the correlations reported in Tables 7 and 8 are weaker than the correlations reported in Table 6, despite the fact that the modes are shared. This is probably due to the smaller number of items included in the calculation of scores for the different modes within tests, as scores computed on a large number of items for different tests are more reliable than scores obtained for different modes within tests.

On the whole then, the data reported here suggest that construct validity can be firmly asserted for the MERT, especially given that it measures the ability to recognize emotions in the face and the voice in a much more comprehensive fashion, using dynamic, multimodal presentation, and a list of emotions that is not only larger, but also much more subtle and differentiated than that in other tests.

Table 4
Descriptive Summaries of Accuracy Scores for the Four Tests

Test	*M*	*SD*	Options	*pi*
DANVA	.77	.054	4	.91
ERI	.71	.059	5	.91
JACFEE	.76	.097	7	.95
MERT	.61	.071	10	.93
PONS	.77	.046	2	.77

Note. $N = 72$ (except JACFEE, $N = 69$). *pi* = proportion index; DANVA = Diagnostic Analysis of Nonverbal Accuracy; ERI = Emotion Recognition Index; JACFEE = Japanese and Caucasian Facial Expressions of Emotion; MERT = Multimodal Emotion Recognition Test; PONS = Profile of Nonverbal Sensitivity.

Table 5
Descriptive Summaries of Accuracy Subscores in the Other Tests

Test	Mean	*SD*	*pi*
DANVA[a]			
Face	.83	.071	.94
Voice	.71	.085	.88
PONS[b]			
Face	.81	.074	.81
Voice	.62	.084	.62
Face and voice	.84	.061	.84
ERI[c]			
Facial	.75	.062	.92
Vocal	.67	.087	.89

Note. For DANVA, the two subscores were obtained on independent portrayals (not produced by the same senders), whereas for the Multimodal Emotion Recognition Test and PONS the same portrayals were presented repeatedly in different modes. *pi* = proportion index; DANVA = Diagnostic Analysis of Nonverbal Accuracy; PONS = Profile of Nonverbal Sensitivity; ERI = Emotion Recognition Index.
[a] $N = 72$. [b] $N = 70$. [c] Facial $n = 72$, vocal $n = 68$.

[4] Scherer and Scherer (2008) reported a correlation of $r = .084$ ($N = 1{,}302$, $p < .01$) between the ERI and the CPI scale in a study with an international group of employees at the management level, suggesting that the relationship is reliable although the effect size is small.

Table 6
Correlations Between Total Scores in Different Tests

Test	MERT IR	MERT FR	DANVA	ERI	JACFEE
MERT FR	.854*				
DANVA	.217*	.224*			
ERI	.277*	.383**	.159		
JACFEE	.455**	.433**	.302**	.303**	
PONS	.505**	.420**	.144	.087	.589**

Note. *N*s vary between 65 and 72 (see the Method section). MERT = Multimodal Emotion Recognition Test; IR = individual recognition; FR = family recognition; DANVA = Diagnostic Analysis of Nonverbal Accuracy; ERI = Emotion Recognition Index; JACFEE = Japanese and Caucasian Facial Expressions of Emotion; PONS = Profile of Nonverbal Sensitivity.
* $p < .05$ (one-tailed). ** $p < .001$ (one-tailed).

Discussion

In this section, we first discuss potential shortcomings of the MERT test and the procedures used for its validation. We conclude that, given the specific purpose of the test, the complexity of the issues, and the satisfactory results of post hoc analyses, none of these points pose a threat for the nature of the test or the validation undertaken here. We then summarize the results on item difficulty, reliability, and construct validation and conclude that general psychometric standards are fulfilled and construct validity amply established. We then discuss the factorial structure of the test and adduce additional evidence for discriminably separate skills for using video and audio cues for emotion recognition. Finally, we describe the utility of different subscores of the MERT (such as the IR and the FR) as well as two relative scores that index the potential of a person to achieve satisfactory recognition on the basis of partial information.

Shortcomings

Stimulus Quality

As the actor portrayal corpus from which the MERT test stimuli were drawn was recorded several years ago, using black and white video, as the selected items were later copied to different formats of videotape in the extensive research process conducted with the corpus, and that these clips were then digitized, the audio and visual quality of the stimuli is not uniformly excellent. However, none of the testees in our sample complained and most found it a challenging and valuable experience. In addition, the stimulus quality does not seem to have affected the results in any way. We are thus confident that this test is a viable instrument, especially because, to our knowledge, no comparable test currently exists.

Table 7
Correlations Between Scores Based on Facial Dynamic Portrayals (With and Without Voice)

Test	PONS dynamic facial	PONS dynamic face and voice
MERT dynamic video	.181	.087
MERT dynamic audio/video	.235*	.335**

Note. $N = 68$. PONS = Profile of Nonverbal Sensitivity; MERT = Multimodal Emotion Recognition Test.
* $p < .05$ (one-tailed). ** $p < .001$ (one-tailed).

Table 8
Correlations of Scores Based on Still Pictures of Facial Portrayals or Vocal Portrayals

	MERT		DANVA		ERI	
Test	Photo	Audio	Facial	Vocal	Facial	Vocal
DANVA facial[a]	.224*					
ERI facial	.296**[a]		.146[b]			
JACFEE	.325**[c]		.278*[c]		.300**[c]	
DANVA vocal		.323**				
ERI vocal		.274*[d]		.013[e]		
PONS vocal		.252*[e]		.235[a]		.032[e]

Note. MERT = Multimodal Emotion Recognition Test; DANVA = Diagnostic Analysis of Nonverbal Accuracy; ERI = Emotion Recognition Index; JACFEE = Japanese and Caucasian Facial Expressions of Emotion; PONS = Profile of Nonverbal Sensitivity.
[a] $N = 70$. [b] $N = 72$. [c] $N = 67$. [d] $N = 66$. [e] $N = 68$.
* $p < .05$ (one-tailed). ** $p < .001$ (one-tailed).

Length

Another shortcoming might be the length of the test, requiring approximately 45 min to complete, depending on the time taken by individual participants to respond. However, the participants in our validation study did not voice complaints in this respect. The variability of the items and their interest in their own results probably provides sufficient motivation to complete the task, although this may depend on testee motivation. It is difficult to reduce the size of the test if one aims to reasonably sample both modes and emotions and to compute subscores for different modes. As it is, we have only three items per emotion and thus advise against computing subscores for emotions (e.g., to examine whether some individuals are more sensitive to static facial configurations or dynamic facial or vocal cues for specific emotions). We believe that it is problematic to attempt to measure an important competence, such as ability to recognize subtle emotion expressions correctly, through very few items.

Gender Ratio

One may regret that the gender ratio of the participants is imbalanced as some gender differences in nonverbal sensitivity have been reported in the literature (Hall & Bernieri, 2001). However, the effect sizes for these differences are generally so small (see also Scherer & Scherer, 2008) that a very large number of participants would be required to yield significance. It would have been virtually impossible to run this very time-consuming validation study with a very large number of participants. Furthermore, it was not the purpose of this validation study to produce norm values and examine systematic individual differences, including age and gender differences.

As we wanted to make sure that there are no systematic differences between male and female participants in this validation sample, we used resampling or bootstrapping methods to check for

significant differences (given the small N for men). Applying bootstrapping to the mean differences between men and women (10,000 samples, $p < .001$) for the different factors that we tested in the ANOVA described above, we found no gender differences for the different levels of emotion (IC anger = −.44 to .42; IC happiness = −.50 to .36; IC fear = −.24 to .61; IC sadness = −.52 to .25; IC disgust–contempt = −.39 to .66), no differences for the intensity (IC low intensity = −.20 to .37 and IC high intensity = −.32 to .29), and no differences for the modalities (IC audio = −.49 to .27; IC audio/visual = −.35 to .39; IC picture = −.37 to .36, and IC visual = −.25 to .64). Given that all confidence intervals included the 0 value, we conclude that the gender factor does not detract from the successful validation of the test.

Accuracy Criterion

One might raise the issue of the criterion used to determine that a response is correct. In this test, the communicative intentions of the actors define the emotion expressed. Moreover, the portrayals used have been extracted from a set of recordings preselected so as to ensure reasonable recognition of portrayals produced in audiovisual mode (see Banse & Scherer, 1996). The preselection of items, a standard procedure in research using actor portrayals, is sometimes criticized as biasing the results in a specific direction. However, item preselection by experts is a standard procedure in most test development, as the appropriateness of the item for the underlying construct or category needs to be established. As long as the normal psychometric indicators testify to appropriate item difficulty and discriminability (which, as shown earlier, is the case for MERT), the fundamental criterion for the diagnosis of individual differences is met. In addition, our results show that preselection on the basis of audiovisual portrayals does not necessarily facilitate the recognition of the same portrayals when they are presented in other modes. As outlined earlier, it may well be that some modes are better suited for the communication of certain emotions. Furthermore, if one wants to maintain the design feature of using the same sender and the same portrayal in different modalities, it would be impossible to preselect test stimuli in such a way that recognition rate distribution is standardized across presentations in all modes.[5]

Acting

Actor portrayals used in emotion expression research are often seen as lacking realism, authenticity, and believability. One can contest this widespread prejudice with a number of counterarguments, for example, by highlighting the fact that individuals constantly act out emotions in real life (applying display rules or manipulating expressions for Machiavellian purposes; Goffman, 1959; Griffiths, 2003) and by demonstrating that actor portrayals that are produced with appropriate induction techniques (e.g., using Stanislavski or method acting procedure, which involve memory retrieval and empathy) may well achieve a reasonable degree of authenticity and believability (see also Banse & Scherer, 1996; Bänziger & Scherer, 2007, 2008). Apart from these considerations, given the absence of even minimally controlled natural emotion expression records for the same senders and a standard set of emotions, test construction has no choice but to rely on actor portrayals. This is particularly true given our focus on measuring the core skill of being able to recognize shared representations of prototypical expression patterns.

Motivation Bias

One of the reviewers of the first version of this article felt that the relatively small but significant positive correlation of the MERT score with the BAS–D scale, measuring achievement orientation, might be a shortcoming. We do not think so. We believe that all performance tests are subject to differential effects of individual achievement motivation. Unfortunately, this effect is almost never quantitatively assessed. We can show that the BAS–D score actually predicts group differences in recognition accuracy. In consequence, we believe that this or similar measures should always be used in conjunction with performance test to allow an estimation to what extent test scores are based on ability or motivation.

Item Characteristics and Reliability

Given the explicit aim to construct a test that is able to discriminate the whole range of the underlying competence continuum, with a special emphasis on high skill levels necessary to distinguish subtle difference between members of the same emotion family that differ mostly in arousal/intensity, our evaluation of item difficulty and discriminability has yielded satisfactory results, comparing well with the same indexes for the other emotion recognition tests we examined. In consequence, the MERT should allow the determination of individual difference with a high level of discrimination.

The reliability of the MERT seems established. The results for test–retest reliability are satisfactory, especially given that learning effects cannot be avoided, as is the case in any test–retest design of learnable skills. In this case, this is even more of a problem, as each portrayal is shown four times, albeit in different presentation modes. As three of the four modes contain similar facial configuration cues, one can expect some carryover effects and a higher potential for learning. Because it is not possible to compute standard item reliability in the psychometric sense, we used mean interrater agreement correlations to obtain an index of replicability. Here, too, the results were satisfactory, showing high agreement on the profile of accuracy differences across emotion, despite sizable individual differences in level of accuracy.

Convergence and Construct Validity

We have concluded that the correlations of total scores and subscores between MERT and the other tests used in this study demonstrate ample construct validity for the MERT.

Subscore Correlations and Factor Structure

The correlation matrix and the factor structure of the MERT suggest two separate factors, one for facial and one for vocal recognition. This concept supports earlier data in the literature that show a similar picture. Thus, Scherer (2007) reported for a large international group of employees at a managerial level, tested with

[5] Apart from the suitability of certain modalities for certain emotions, actors also often seem to be "specialized" to express a certain emotion more convincingly in a particular modality.

the ERI instrument (among many other assessment instruments), a significant but relatively low correlation between the two subtests for facial and vocal recognition, $r = .24$ ($N = 1{,}264$; $p < .001$). Thus, although both tests seem to tap into a common facet of emotion recognition competence, the joint variance is only about 6% and there are clearly mode-specific competence differences. A composite score for fluid intelligence, based on four appropriate tests, correlates with facial recognition, $r = .12$ ($N = 1{,}231$; $p < .001$), and with vocal recognition, $r = .18$ ($N = 1{,}311$; $p < .001$; Scherer, 2007). If one controls for fluid intelligence, the correlation between facial and vocal recognition ability drops to $r = .14$ ($p < .001$). Thus, it seems likely that emotion recognition tests should separately assess these two competences, as it seems unlikely that there is a single, strong underlying factor.

The MERT results reported in this article, based on the same senders for both modalities (which eliminates the possibility that low correlations in other tests such as ERI and DANVA are due to different encoders in both channels), suggest that the relative ability to correctly decode both dynamic cues and static configurations in the face (as visible in a photo) may vary across individuals. Thus, in addition to identifying one group of testees that were generally less skillful than the best group, there is a third group that seems to have generalized problems to decode static visual cues. The MERT approach of using the same senders and portrayals shown in different modes, and the computation of mode subscores, will allow researchers to unravel these issues in future, allowing the development of a much more fine-grained diagnostic scheme for emotional and interpersonal sensitivity.

Test Scores

The MERT, like the other tests used in this validation study, yields accuracy scores reflecting the degree of competence of the testee in identifying target emotions as portrayed by actors and as validated by expert preselection and empirical rating studies (see Banse & Scherer, 1996). However, in the case of the MERT, the IR scores reflect real recognition ability rather than just the ability to discriminate among a few categories, as is often the case when only a few basic emotions are used. The latter is approximated by the FR score, which evaluates the ability of an individual to make a coarse discrimination of the overall quality of an expressed emotion. As the restricted range of the FR suggests, coarse discrimination ability, probably helped by chance, is a rudimentary ability that is present, at least to a certain degree, in most individuals. This is not surprising, as coarse discrimination is probably an indispensable skill for any kind of social interaction. However, whether this is a good basis to diagnose emotional intelligence in the sense of perception competence, as described at the outset of this article, is questionable. It seems that a more sensitive instrument, able to measure the ability for more fine-grained analysis is necessary. Based on the results reported earlier, we suggest computing an index that expresses IR as a proportion of FR (IR/FR), which expresses the potential for such subtle differentiation (subtle differentiation index, SDI). In other words, the higher this index, the higher the individual's ability to go beyond coarse differentiation in the direction of fine-grained recognition. For example, for a testee with an FR of .85 and an IR of .79, the SDI reaches .89 (suggesting that the person is capable of finely differentiating most of the time), whereas a testee with an FR of .78 and an IR of .54 has an SDI of .69, indicating a larger gap between coarse and fine recognition. These detailed mode-specific scores are provided as part of the MERT results to allow detailed feedback about strength and weaknesses of each candidate that may inform on the need for skills development.

A major advantage of MERT is that in addition to the total scores (IR and FR), subscores for four modes allow differentiation of the emotion recognition ability for the mode of stimulus presentation. Figure 3 shows these subscores for all participants in the study, sorted by the audio/video accuracy score in descending order. As indicated

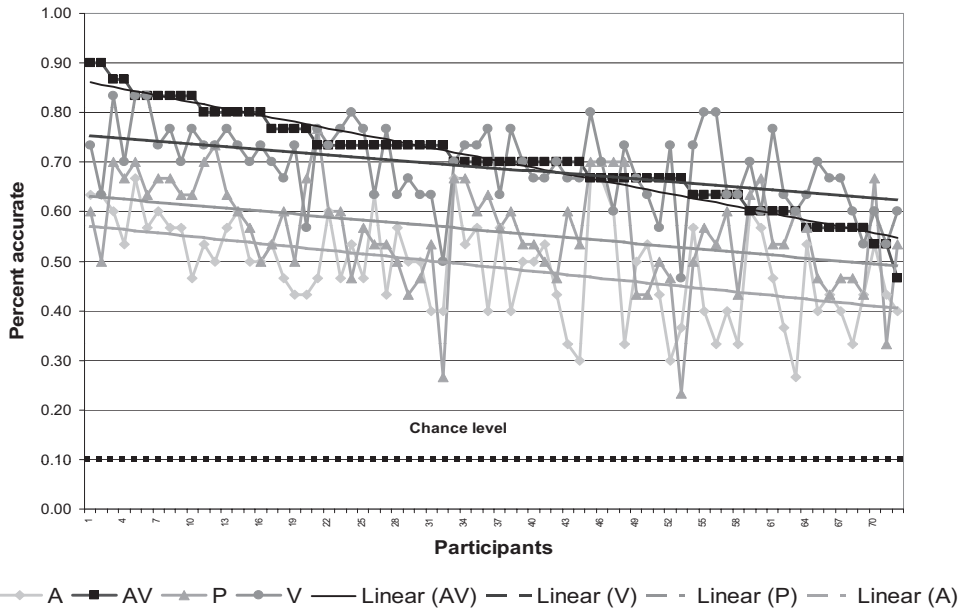

Figure 3. Plot of the mode-specific accuracy subscores for all participants. AV = audio/video; A = audio; V = video; P = still picture.

by the correlations shown in Table 5, the subscores share common variance, as illustrated by the descending trend lines. Yet, there is a remarkable degree of variability in the respective scores across individuals, illustrating some of the findings reported earlier. Although the scores for the audio mode are generally the lowest, some individuals have remarkably high scores, coming close to the audio/video scores. As expected by the groups that are revealed by a cluster analysis of the confusion patterns, some individuals seem to have a special problem with inferring the correct emotion from still pictures. Finally, there are some individuals whose scores in the video conditions actually exceed their scores in the audio/video condition, which generally yields the highest accuracy, given that it provides a maximum of cues. One possible explanation for this finding is that their intuitive sense for the correct visual cues (e.g., facial actions) is confused by erroneous inferences from the auditory channel. This powerful effect of the auditory cues in combination with visual cues is also suggested by the fact that the audio/video scores load higher on the auditory ability factor and that audio scores predict audio/video scores more strongly than do visual cues. As this is the first attempt to unravel the differential emotion recognition competence regarding preferences or specific abilities for certain communication channels, these results can only serve as illustration, requiring more in-depth empirical study.

Given the importance of the differential diagnosis of such modality-specific aptitudes, we have added mode-specific potential indexes to the MERT feedback information, following the model of the SDI described earlier, expressing the audio-only, video-only, and still picture-only scores as a proportion of the audio/video score. Thus, the Auditory Potential Index (API) and the Visual Potential Index indicate the proportion of a person's overall recognition capacity, given exclusive access to the respective channel. For example, with a common audio/video score of .67, one testee has an A score of .33, producing an API of .50, whereas another testee who has an A of .60 reaches an API of .90, suggesting a greater capacity to use only audio cues to reach a result that is similar to the case of having access to all auditory and visual cues. The Still Picture Potential Index indicates the person's recognition potential based on cases in which dynamic cues are lacking and only static, configurational cues in the visual modality are accessible (as is the case in most expression tests in which still photographs are used). Group 1 of our testees has consistently lower values for this index. We expect these indexes to be of considerable utility in diagnosing an individual's specific strengths and weaknesses and in determining remedial training needs.

Conclusions

This first evaluation of MERT has shown very promising results, emphasizing the importance of including several communication modalities and a larger number of emotion categories in tests of emotional sensitivity for assessing true and fine-grained recognition of emotional expressions and gaining better understanding of the underlying skill(s). We are currently involved in an attempt to test MERT in different cultures to examine the justification of generalizing the results obtained with German actor portrayals and French-speaking Swiss testees to other language and culture groups. As facial expressions are considered to be largely universally recognized and we have used language-free phonetic material for the vocal utterance, we are fairly confident that the test will be relatively culture free. However, lower levels of accuracy are possible, especially in the audio-only mode, with more differences between the standard prosodic and voice quality features of another language (Scherer, Banse, & Wallbott, 2001).

Most important, the predictive validity of the MERT will need to be established in examining the relationships of the recognition accuracy tested with other competencies, adaptation, and adjustment (e.g., emotional stability, life satisfaction); task-oriented behavior (e.g., negotiation skills) and interpersonal interaction (e.g., conflict management); and long-term relationships (e.g., successful marriages). Predictive validity investigations conducted with other tests in this domain have generally reported positive results (Hall, Andrzejewski, & Yopchick, 2009; Hall & Bernieri, 2001). Scherer (2007) provided an interesting example for the ERI Vocal–Index: If one compares over 1,300 testees in higher management positions with those in lower echelons, one finds more than a three-percentage point advantage (t significant at $p < .01$) for employees in positions below the executive level for the recognition of vocal anger expression, suggesting that people in power may be less aware of vocal signs of anger in their interlocutors (which may well be due to the fact that they encounter a smaller number of appropriate instances from their interaction partners). Clearly, valid tests of recognition ability are essential for the diagnosis of different aspects of emotional competence and as a basis for appropriate remedial programs.

Given the earlier results reported in the literature, we are quite confident that MERT will indeed be shown to have appreciable predictive power. It is in the interest of further studies on the predictive validity in a wide variety of applications that the MERT be administered by researchers wanting to use it in a noncommercial fashion to examine the role of emotion recognition ability and its various facets in human affect and behavior. Therefore, the MERT is freely accessible on the Internet (www.affective-sciences.org/MERT), allowing individuals to take the test and to obtain their overall accuracy scores. The use of the test in research with groups of participants can be arranged with the authors.

The data reported in this article suggest that the aim of objectively measuring an important aspect of emotional perception competence, namely, emotion recognition ability, is not an impossible dream. However, such assessment does have a cost if it is to yield valid and reliable data, requiring solidly constructed tests with a sufficient number of appropriate items and reflecting a large and subtly differentiated gamut of emotions, as well as a ground truth as a criterion. Adding a few judgments of photographs selected on the basis of uncertain criteria and comparing testees' responses to reference group means is unlikely to yield the required diagnostic information that can be useful for selection and training. Further development of this approach includes the construction of a second generation of the test described here. Our group is currently working on a set of stimuli with uniformly excellent digital quality for a second edition of the MERT, using portrayals from the Geneva Multimodal Emotion Portrayal corpus (GEMEP, Bänziger & Scherer, 2007, 2008), which has been designed on the basis of the research results and experiences with MERT. It is expected that with this new instrument, and others of its kind, further empirical work will allow a determination of which parts of a testee's performance are a constitutional ability or a learnable and trainable skill (see Scherer, 2007). Most important, future work will need to establish the predictive validity of the MERT in terms of behavioral impact and interactional outcomes.

References

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70,* 614–636.

Bänziger, T., & Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In A. Paiva, R. Prada, & R. Picard (Eds.), *Affective computing and intelligent interaction 2007: Lecture notes in computer science: Vol. 4738* (pp. 476–487). Berlin, Germany: Springer-Verlag.

Bänziger, T., & Scherer, K. R. (2008). *Studying emotional expression through actor portrayals: The Geneva Multimodal Emotional Portrayal (GEMEP) corpus.* Manuscript submitted for publication.

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., et al. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21,* 3–21.

Buck, R. (1974). *Communication of Affect Receiving Ability Test (CARAT).* Unpublished manuscript, University of Connecticut.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology, 67,* 319–333.

Costa, P. T., & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology, 54,* 853–863.

Ekman, P. (1972). Universals and cultural differences in facial expression of emotion. In J. R. Cole (Ed.), *Nebraska symposium on motivation: Vol. 19*(pp. 207–283). Lincoln: University of Nebraska Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6,* 169–200.

Ekman, P. (1994). Moods, emotions, and traits. In P. Ekman & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp. 56–58). New York: Oxford University Press.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face.* Englewood Cliffs, NJ: Prentice-Hall.

Ekman, P., & Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS*) (2nd ed.). New York, NY: Oxford University Press.

Goffman, E. (1959). *The presentation of self in everyday life.* New York: Doubleday Anchor.

Gough, H. G. (1984). A managerial potential scale for the California Psychological Inventory. *Journal of Applied Psychology, 69,* 233–240.

Griffiths, P. E. (2003). Basic emotions, complex emotions, Machiavellian emotions. In A. Hatzimoysis (Ed.), *Philosophy and the emotions* (pp. 39–67). Cambridge, England: Cambridge University Press.

Hall, J. A., Andrzejewski, S. A., Murphy, N. A., Schmid Mast, M., & Feinstein, B. A. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality, 42,* 1476–1489.

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

Hall, J. A., & Bernieri, F. J. (2001). *Interpersonal sensitivity: Theory and measurement.* Mahwah, NJ: Erlbaum.

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

Matthews, G., Zeidner, M., & Roberts, R. (2007). The science of emotional intelligence: Knowns and unknowns. New York: Oxford University Press.

Mayer, J. D., & Salovey, P. (1993). The intelligence of emotional intelligence. *Intelligence, 17,* 433–442.

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

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy. *Journal of Nonverbal Behavior, 18,* 9–35.

Roberts, R. D., Zeidner, M., & Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? Some new data and conclusions. *Emotion, 1,* 196–231.

Rolland, J. P., Parker, W. D., & Stumpf, H. (1998). A psychometric examination of the French translations of the NEO–PI–R and NEO–FFI. *Journal of Personality Assessment, 71,* 269–291.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test.* Baltimore: John Hopkins University Press.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106,* 332–337.

Ross, E. D., Thompson, R. D., & Yenkosky, J. (1997). Lateralization of affective prosody in brain and the callosal integration of hemispheric language functions. *Brain and Language, 56,* 27–54.

Scherer, K. R. (1971). Randomized splicing: A simple technique for masking speech content. *Journal of Experimental Research in Personality, 5,* 155–159.

Scherer, K. R. (1994). Toward a concept of "modal emotions." In P. Ekman & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp. 25–31). New York: Oxford University Press.

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., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32,* 76–92.

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion, 15,* 123–148.

Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion, 7,* 113–130.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, & H. Goldsmith (Eds.), *Handbook of the affective sciences* (pp. 433–456). New York: Oxford University Press.

Scherer, K. R., & Scherer, U. (2008). *Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the Emotion Recognition Index (ERI).* Manuscript submitted for publication.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 66,* 310–328.

Scherer, K. R., Wallbott, H. G., & Summerfield, A. B. (Eds.). (1986). *Experiencing emotion: A crosscultural study.* Cambridge, England: Cambridge University Press.

Spielberger, C. D., & Sydeman, S. J. (1994). State–Trait Anxiety Inventory and State–Trait Anger Expression Inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 292–321). Hillsdale, NJ: Erlbaum.

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology, 28,* 879–896.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Exploring and differentiating music-induced emotions. *Emotion, 8,* 494–521.

Received November 24, 2008
Revision received May 20, 2009
Accepted June 24, 2009 ■