Learning and Individual Differences 52 (2016) 129–136

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Measuring two aspects of emotion recognition ability: Accuracy vs. sensitivity☆

Dmitry Lyusin [a,b,*], Victoria Ovsyannikova [a]

[a] National Research University Higher School of Economics, Myasnitskaya 20, Moscow 101000, Russia
[b] Russian Academy of Sciences, Institute of Psychology, Yaroslavskaya 13, Moscow 129366, Russia

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form 15 January 2015
Accepted 29 April 2015

Keywords:
Emotion recognition
Emotion intelligence
Videotest of Emotion Recognition

ABSTRACT

The present paper aims at showing the necessity to distinguish two aspects of emotion recognition ability, accuracy of the recognition of emotion types that constitute the emotional state of the perceived person and sensitivity to the intensity of the perceived person's emotions. A new technique that measures these two aspects of emotion recognition, the Videotest of Emotion Recognition, is proposed. The accuracy and sensitivity indices provided by the Videotest of Emotion Recognition have high reliability and yield different correlation patterns with other cognitive and personality variables.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Emotion recognition has been widely studied for decades in psychology. In modern psychology, emotion recognition is often conceptualized and measured in the frame of emotional intelligence research. Broadly, emotional intelligence refers to the set of abilities that allows understanding and managing of emotions. Emotion recognition is widely considered to be one of the basic emotional intelligence components. The well-known emotion intelligence model proposed by Mayer, Salovey, Caruso, and Sitarenios (2001) identifies four branches of emotional intelligence; two of them, Emotion Perception and Emotion Understanding, are related to emotion recognition. Emotion Perception includes skills concerned with accurate detection and identification of emotions in oneself and others. Emotion Understanding concerns the ability to understand relationships between emotions, emotion language and signals conveyed by emotions. According to this model, four branches are ordered hierarchically, the basic branch being Emotion Perception (Salovey & Grewal, 2005). It seems that distinguishing Emotion Understanding from Emotion Perception is artificial and has an intuitive rather than theoretical background.

Another approach to the conceptualization of emotion skills, proposed by Scherer and Scherer (2011), understands emotion perception as one of the three major domains of emotional competence along with emotion production and emotion regulation. Emotion perception is considered a central socio-emotional competence essential for many different types of occupation.

One of the important directions in emotion recognition research is developing methods for measuring emotion recognition ability. Most of these methods focus on accuracy of emotion recognition. The present paper aims at showing the necessity to distinguish between the two aspects of the ability to recognize emotion, namely accuracy and sensitivity; a technique for measuring accuracy and sensitivity is also proposed.

1.1. Tests for measuring emotion recognition ability: diversity and problems

The number of studies on measuring emotion recognition ability has been growing in the recent decades. Most of the new measurement instruments have been developed in the context of emotion intelligence assessment. Two types of assessment methods are traditionally distinguished in the research on emotional intelligence, objective tests and self-report questionnaires. They correspond to the two types of emotional intelligence models that are usually called ability and mixed models (Mayer, Salovey, & Caruso, 2000). Ability models understand emotional intelligence as a set of cognitive abilities and competencies analogous to other types of intelligence such as verbal or spatial. Mixed models, also called trait models, define emotional intelligence more broadly, as an array of cognitive, personality, and motivational traits that provide better emotion understanding and management, and finally result in higher levels of adaptation and well-being of an individual. For measuring emotional intelligence, proponents of ability models use objective tests similar to traditional intelligence tests with answers that can be assessed

☆ This study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics in 2015.

* Corresponding author at: Leningradskoe shosse 48-1-29, Moscow 125212, Russia.
E-mail addresses: ooch@mail.ru (D. Lyusin), v.ovsyannikova@gmail.com (V. Ovsyannikova).

http://dx.doi.org/10.1016/j.lindif.2015.04.010
1041-6080/© 2015 Elsevier Inc. All rights reserved.

as right or wrong. Mixed models proponents prefer self-report questionnaires similar to personality inventories. Some exceptions from this correspondence between the two types of models and approaches to measurement are possible. For example, the EmIn Questionnaire developed by Lyusin (2006a, 2006b) that will be described below is based on the ability model. The author claims that it measures perceived emotional intelligence, understood as a cognitive ability, rather than personality traits.

The limitations of self-report assessment are broadly known; hence this paper will focus on objective tests that evaluate emotion recognition ability independent of an individual's self-concept and beliefs about his or her behavior. There is a large diversity of such tests in modern psychology. They differ in stimuli, item formats, indices, and scoring procedures. For instance, stimuli can be photographs of facial expressions, videos with various types of behavior, voice recordings, vignettes describing emotional situations, and even thoroughly non-human stimuli such as geometric figures.

The problem of scoring is one of the hardest in performance-based assessment of socio-emotional abilities. Unlike traditional intelligence tests, there are no obvious logical foundations for establishing correct answers in most emotion recognition tests. Three major approaches to scoring have been suggested, namely expert, consensus, and target scoring. Expert scoring is based on expert opinions about the correct or best choice among suggested answers. The main difficulty is to decide who has expertise in this domain. In most cases, emotion researchers are suggested for this role, but it is often questioned if they or any other professionals such as psychotherapists, counselors, and actors qualify as emotion experts. Some authors even claim that the emotion domain is one of those ill-defined knowledge domains where no objective standards for verification exist and, therefore, no qualified experts can be suggested (Legree, Psotka, Tremble, & Bourne, 2005). Consensus scoring is based on the opinion of the majority of the participants about correct answers. It is often supposed that consensus scoring reflects cultural biases in beliefs about emotions. Moreover, it is regarded as logically unacceptable to establish correct answers to the intelligence test items, especially to the difficult ones, on the basis of the consensus opinion. In target scoring, the correct response is set by a target person who creates the stimuli. These target persons can be actors portraying emotions for photographs or voice recordings, authors of the vignettes who define a priori which emotion should be experienced by a certain character, etc. Target scoring can be applied only to a limited range of stimuli, and it can always be questioned if the target emotion was adequately portrayed or expressed in the stimuli. All three approaches have their own limitations, but they are used in psychological research and assessment for the lack of better solutions.

An important feature of emotion recognition items, as well as of any emotional and social abilities items, is the difficulty in establishing one correct response. Several responses to the same item can often be regarded as correct with different levels of confidence. This situation is quite normal for the psychological content being measured since emotional states are often ambiguous and constitute a mixture of various emotion types. The stimuli cannot represent all individual and situational features that result in a certain emotional state. Two important consequences result from this. First, it makes sense to use rate-the-extent format of responses similar to the Likert-type scales, rather than just to classify responses as correct and incorrect. Secondly, the unidimensional format of responses when a participant estimates the presence of only one emotion in the stimulus is less appropriate as compared to the multidimensional format that allows estimating the presence of an array of emotions in the stimulus.

Different approaches to scoring and different response formats (unidimensional or multidimensional) are used in modern emotion abilities tests. The following brief review of emotion recognition tests summarizes the main tendencies in this field.

One of the most prominent early techniques for emotion recognition is the Profile of Nonverbal Sensitivity (PONS; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979). It consists of twenty audio/video recordings in which one female person represents twenty attitudes (such as expressing jealousy, asking for a favor). The participant must assess the attitude expressed by the character. Attitudes are set initially by the test developer and are classified as dominant versus submissive and positive versus negative. Each recording is represented by eleven channels of expression (face, speech, etc.). The 220 portrayals are presented to the participant in a fixed order. For each portrayal, the participant is required to select one of two alternative answers. The accuracy index is calculated as the percentage of correct answers of the total number of test stimuli.

The Diagnostic Analysis of Nonverbal Accuracy was designed to assess the sensitivity to nonverbal expressions of emotions (DANVA; Nowicki & Duke, 1994). Twenty-four photographs of facial expressions and 24 voice recordings of four emotions (anger, fear, joy, sadness) are used as stimuli. Each emotion category is presented in two intensities, low and high. The emotions are portrayed by professional actors. The participant has to choose one of the four emotion categories for each stimulus. The accuracy scores are calculated as the percentage of correct responses separately for both types of stimuli and for the whole test.

A notable feature of the Japanese and Caucasian Brief Affect Recognition Test is the use of the images of people of different races as stimuli (JACBART; Matsumoto et al., 2000). The photographs of European and Japanese facial expressions of seven basic emotions (anger, joy, sadness, contempt, disgust, fear, and surprise) are presented to participants who have to assess the presence of each of the seven emotions in the portrayals by means of nine-point scales. The average values for each emotion category obtained in the American sample are considered to be standard. Accuracy scores are calculated as correlations between the participants' responses and the standard estimates. An interesting feature of the technique is the possibility to calculate different accuracy scores separately for each emotion category, for different races and sexes.

The most famous measure of emotion recognition is the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT; Mayer, Salovey, Caruso, & Sitarenios, 2003). The test is based on Salovey and Mayer's model of emotional intelligence that regards it as a set of hierarchically organized cognitive abilities. The MSCEIT consists of four subtests. The first and third subtests, Emotion Perception and Emotion Understanding, measure abilities related to emotion recognition. The Emotion Perception subtest includes two types of tasks with photographs of facial expression and pictures of landscapes and abstract designs as stimuli. The participant must assess the degree of presence of several emotions in each stimulus using Likert five-point scales. The Emotion Understanding subtest consists of the Blends task and the Changes task. In the Blends tasks, the participant must identify which emotions will result from the blend of several other emotions and select one of the response options. In the Changes tasks, the participant must select the emotion from the list of emotions that may result from the situation described. The weights based on expert and consensus ratings are attributed to each response option. An accuracy index is calculated by averaging the weights of the responses selected by the participant.

Recently, the Emotional Intelligence Measure (AEIM; Warwick, Nettelbeck, & Ward, 2010) was developed, which is, actually, a revised version of the MSCEIT. The two scales, Emotion Perception and Emotion Management, have been changed. The principles of stimuli selection and scoring methods are similar to the MSCEIT.

The Situational Test of Emotional Understanding (STEU; MacCann & Roberts, 2008) consists of the descriptions of situations related to different emotions. The STEU items were developed according to Roseman's appraisal theory of emotions (Roseman, 2001). The test authors set the correct responses on the basis of this theory. The accuracy index is calculated as the percentage of correct responses.

The Multimodal Emotion Recognition Test (MERT; Bänziger, Grandjean, & Scherer, 2009) consists of the presentations of expressions of five emotion families in four formats, video with sound, video without sound, audio without image, and photo taken from video. Emotional expressions are portrayed by professional actors. In total, there are 120 stimuli. The participant has to select one of ten emotion categories; accuracy is calculated as the percentage of correct responses. It is possible to obtain the accuracy scores separately for different types of emotions and different formats of stimuli.

The Emotion Recognition Index (ERI; Scherer & Scherer, 2011) consists of two subtests, the Index of Facial Emotion Recognition (FACIAL-I) and the Index of Vocal Emotion Recognition (VOCAL-I). Each subtest includes 30 items that present facial and vocal expressions of five basic emotions. Emotion recognition accuracy is calculated as the percentage of the participant's correct responses.

Emotion recognition tasks may form a subtest in the tests that measure other cognitive abilities. For example, the battery of face cognition measures includes three tests on emotion recognition out of the 18 tests that make up the battery (Herzmann, Danthiir, Schacht, Sommer, & Wilhelm, 2008). In the facially expressed emotion decision task the participants have to decide whether the stimulus face expresses happiness or anger. The reaction time of the testee is measured. In the emotional odd-man-out task a set of three faces is presented to the subject. Two of the faces show the same emotional expression, while the third one portrays a different expression. The participant has to choose the face with a different emotional expression. The reaction time of the participant is measured. The facially expressed emotion labeling task measures speed and accuracy of recognition of six basic emotions under rapid stimulus presentation. The accuracy index is calculated as the percentage of correct responses.

Table 1 summarizes the characteristic features of these tests with an emphasis on the methods of obtaining accuracy indices. The majority of techniques use a unidimensional response format, and the test scores are calculated as the percentage of correct responses. As mentioned above, we claim that a unidimensional response format is not quite adequate for emotion recognition tasks because of the ambiguous nature of most emotional states; ignoring this fact reduces the ecological validity of the measurement.

1.2. Emotion recognition accuracy indices for multidimensional response format of test items

If the multidimensional response format has been chosen by a test developer, a test score should reflect the degree of similarity between the participant responses and the correct responses to a test item. The similarity index can be obtained in different ways. For further discussion, we will, as an example, take a typical test item that requires assessment of the stimulus (e.g., the emotional state of a video character) with several scales representing different emotion categories. In this case there is a set of a participant responses and a set of the standard estimates that are regarded as correct responses.

Fig. 1 demonstrates hypothetical responses of a participant to an item that consists of the fifteen Likert six-point scales. The solid line represents the profile of correct responses; the dashed line represents the profile of responses of Participant 1. What is the best way to assess the degree of similarity between these two profiles? A simple and often used measure of similarity for non-metric data is a so-called ‘city-block metric’ (Reis & Judd, 2000) that is calculated as the sum of the absolute values of deviations of the participant responses from the correct responses on each scale. It can be defined as

$$D = \sum |Q_i - R_i|$$

where Q_i is a participant's response on Scale$_i$, and R_i is a correct response on the same scale.

The greater the D value, the less accurate the participant's evaluations of the character's emotional state. For the data presented in Fig. 1, $D = 27$.

This measure of similarity is often used, but it seems that two essentially different aspects of emotion recognition are mixed in it. It can be illustrated by the hypothetical responses of Participant 2 presented in Fig. 2. D value is equal to 27 as it was the case with Participant 1. However, Participant 2 identifies emotions very accurately in a certain sense. He or she gives higher estimates on Scales 6, 8, 10, 14, and 15 and lower estimates on Scales 3, 7, and 8. Thus, the shape of the participant's response profile perfectly corresponds to the correct response profile. The only difference concerns the average level of these two profiles, the participant's profile being noticeably higher.

Therefore, it seems important to introduce two different indices. The first indicates the accuracy of recognition of various emotion types that constitute the emotional state of the observed person. The second reflects the perceiver's sensitivity to the intensity of the perceived person's emotions. Notably, in this work we understand intensity as a feature of emotion that is different from arousal. Practically all dimensional models of affect describe arousal as one of the most important dimensions that distinguish between emotions (Fontaine, Scherer, Roesch, & Ellsworth, 2007; Russell, 1980). Arousal is sometimes taken into account in the models of natural and automatic emotion recognition (e.g., Wöllmer, Kaiser, Eyben, Schuller, & Rigoll, 2013). Some authors use the terms arousal and intensity as synonyms (e.g., Gunes & Pantic, 2010). However, a number of studies show that intensity of

Table 1
Measures of emotion recognition ability and methods of obtaining indices of emotion recognition accuracy.

Measure	Stimuli	Method of scoring	Response format	Calculation of the accuracy index
PONS (Rosenthal et al., 1979)	Video recording of emotion expression and its components (only faces, only speech, etc.)	Target	Unidimensional	Proportion of correct responses
DANVA (Nowicki & Duke, 1994)	Photographs of faces and voice recordings	Target	Unidimensional	Proportion of correct responses
JACBART (Matsumoto et al., 2000)	Photographs of faces	Consensus	Multidimensional	Correlation between the standard estimates and a participant's responses
MSCEIT (Mayer et al., 2003); AEIM (Warwick et al., 2010)	Photographs of faces and other images (Emotion Perception subtest); descriptions of situations and other verbal tasks (Emotion Understanding subtest)	Expert and consensus	Multidimensional (Emotion Perception subtest); unidimensional (Emotion Understanding subtest)	The averaged weights of a participant's responses
STEU (MacCann & Roberts, 2008)	Descriptions of situations	Target	Unidimensional	Proportion of correct responses
MERT (Bänziger et al., 2009)	Video recordings with or without sound, audio recordings, photographs	Target	Unidimensional	Proportion of correct responses
ERI (Scherer & Scherer, 2011)	Photographs of faces and voice recordings	Target	Unidimensional	Proportion of correct responses
Emotion recognition tests of the battery of face cognition measures (Herzmann et al., 2008)	Photographs of faces	Target	Unidimensional	Reaction time and proportion of correct responses

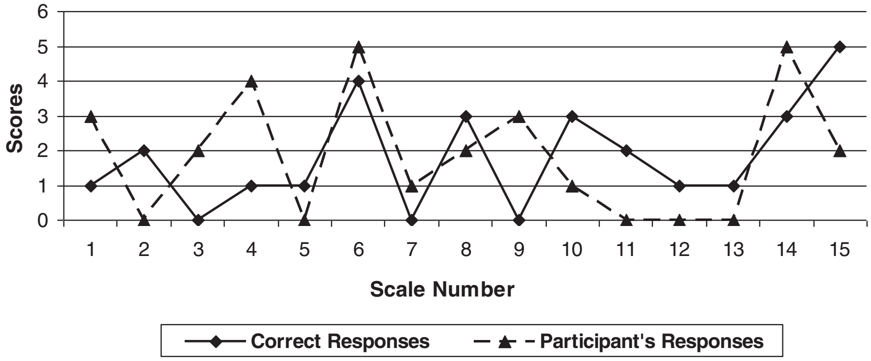

Fig. 1. Hypothetical responses of Participant 1. $D = 27$.

emotion is independent from arousal (Feldman Barrett & Russell, 1999; Kuppens, Tuerlinckx, Russell, & Feldman Barrett, 2013; Reisenzein, 1994). For instance, one can feel intensely bored or tired.

We propose to use a standard deviation of the differences between the participant's responses and correct responses on each scale as the accuracy index; it can be designated by *A*. The sensitivity index *S* can be calculated as the sum of deviations of the participant's responses from the correct responses on each scale; unlike in the formula for the *D* index, signs of the deviations should be taken into account. Dividing this sum by the number of scales would result in putting its range into limits defined by the number of points of the chosen Likert scales. The *S* index can be defined as

$$S = \sum (Q_i - R_i)/m$$

where Q_i is a participant's response on Scale$_\text{i}$, R_i is a correct response on the same scale, and *m* is the number of scales, i.e., of emotion categories used for assessment.

It is important to note that the *A* index is inverse, i.e., the larger are its values, the less accurate is the participant. A zero value would mean that the participant is perfectly accurate in recognition. The theoretical range of the *S* index would lie within the limits defined by the chosen Likert scales. The *S* values would be positive if the participant overestimates the intensity of the observed person emotions, and negative if the participant underestimates the intensity of emotions. The indices *A* and *S* are mathematically independent which allows assessing accuracy and sensitivity as two independent aspects of emotion recognition. It can be illustrated by the hypothetical data presented in Fig. 2 where $A = 0.54$ which indicates high accuracy in emotion recognition; however, $S = 1.8$ which means the obvious tendency to overestimate emotion intensity, raising its estimates on almost two scores out of possible five. The hypothetical responses presented in Fig. 3 show the opposite case. Participant 3 is quite inaccurate in emotion recognition ($A = 2.14$), however, this participant does not overestimate or underestimate the emotion intensity ($S = -0.07$). It is remarkable that the *D* values are identical for Participants 2 and 3. Their emotion recognition ability could be erroneously regarded as similar without the use of the suggested index of accuracy (*A*) and index of sensitivity (*S*). Nevertheless, the data presented in Figs. 2 and 3 clearly show that there are two different ability structures beyond the identical *D* indices.

2. The Videotest of Emotion Recognition

The review of modern emotion recognition measures revealed their typical limitations. One of the aims of the present study is to develop a new emotion recognition test that would overcome some of them. First, the stimuli used in emotion recognition assessment often lack ecological validity. In real life, people rarely identify emotions on the basis of separate aspects of human behavior, such as only mimics or voice sound. Most typically, human behavior is perceived holistically and within a certain situation. The knowledge of this situational context allows the understanding of factors influencing the person, rules restricting or prescribing possible emotion expressions, etc. Thus, a rich diversity of information sources is usually involved in emotion recognition. This is why we decided to use video recordings showing various aspects of the character's behavior including facial expressions, movements, speech, and situational context that should be understandable at least roughly. To further improve the ecological validity of the stimuli, excerpts from the natural behavior should be presented in the video recordings. Emotional behavior portrayed by the actors should not be used, since their emotional expressions are often either exaggerated or too much tuned to the cultural standards which makes them not natural enough.

Secondly, as demonstrated above, many limitations of the existing measures can be overcome by using the multidimensional response

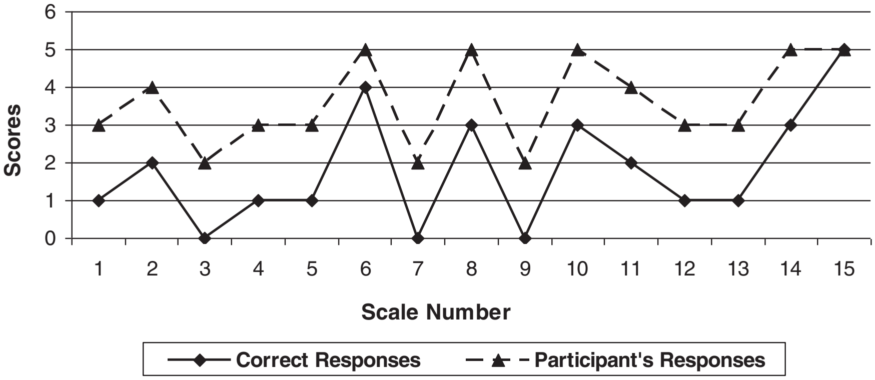

Fig. 2. Hypothetical responses of Participant 2. $D = 27$, $A = 0.54$, $S = 1.80$.

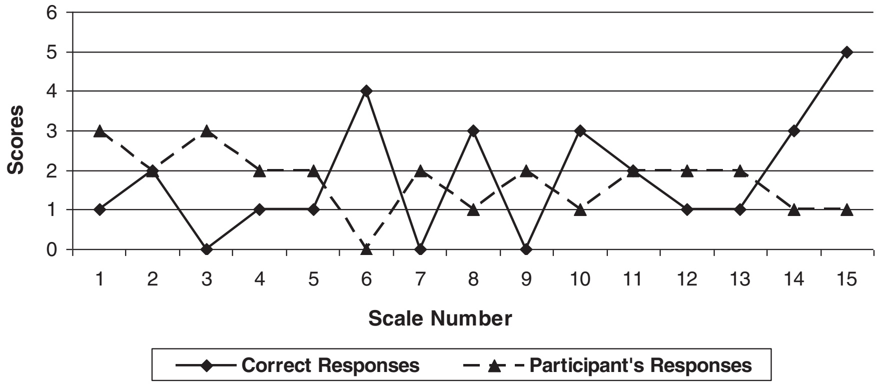

Fig. 3. Hypothetical responses of Participant 3. $D = 27$, $A = 2.14$, $S = -0.07$.

format that gives the participants the opportunity to estimate the intensity of different emotions in the stimulus.

Thirdly, we wanted to develop a technique that would allow one to measurer separately accuracy and sensitivity of emotion recognition with the use of two suggested indices, *A* and *S*.

In line with these ideas, the Videotest of Emotion Recognition has been developed. Video recordings of natural behavior in various real-life situations were taken as stimuli. The selection of these recording was based on the following criteria.

1. Each video recording must represent human behavior in natural situations, rather than in a laboratory setting.
2. The target character of the video must be in some emotional state. However, this emotional state should not be too intense in order not to make it obvious which emotion is experienced. We also avoided test item with intense emotions because they could be so simple that the variability of the responses would be too low.
3. Diverse types of information must be available from the videos, including facial expressions, movements, speech, and reactions of other characters. The situational context of the behavior should also be comprehensible at least in a general way.

The video recordings were between 10 and 60 s long; the target characters were both males and females.

Participants had to assess the characters' emotional states with a set of 15 scales representing different emotion categories. The categories were selected in the pilot study so that they corresponded to characters' emotions. The selection procedure was described in detail in Ovsyannikova (2007). Each scale is a unipolar Likert-type six-point scale with points from 0 to 5, where '0' means that this emotion category does not correspond to the character's state at all, '1' means that this emotion category corresponds to the character's state minimally, and '5' means that this emotion category describes the character's state perfectly. The list of the aforementioned scales is presented in Fig. 4.

The Videotest consists of seven video recordings selected from a large number of recordings on the basis of judges' estimates. Judges were seven counseling psychologists with more than ten years of professional experience. The judges assessed the target characters' emotional states using the set of fifteen scales described above. The internal consistencies of their estimates of each recording were assessed with Cronbach's alphas. The recordings selected for the final version had alphas in the range from .82 to .95. For each scale in each recording the medians of judges' estimates were calculated. It yielded standard estimates that were considered correct responses.

The testing procedure consists in the demonstration of the video recordings in a fixed order. Before each recording, the testee is informed who the target is. After each recording, the testee assesses the character's emotional state by using the set of fifteen scales. Two indices of emotion recognition ability are calculated, accuracy of the recognition of various emotion types in the stimulus (*A* index) and sensitivity to the intensity of emotions in the stimulus (*S* index).

The Videotest of Emotion Recognition as well as some other ability and personality measures were administered to a rather large sample (N = 684). We expected the two suggested indices of accuracy and sensitivity (1) to be independent or, at least, not highly intercorrelated and (2) to yield different correlation patterns with other psychological measures. Such a result would confirm our understanding of accuracy and sensitivity as two different aspects of emotion recognition ability.

3. Method

3.1. Participants

A total of 684 young adults (463 female), with an average age of 21.5 (*SD* = 5.82), participated in the study. They were undergraduate students, high school students, and adults of different professional occupations. Participation was voluntary and anonymous; the participants were informed about the purpose of the study and were debriefed after the session.

Anger	0 1 2 3 4 5
Relaxation	0 1 2 3 4 5
Surprise	0 1 2 3 4 5
Contempt	0 1 2 3 4 5
Shame	0 1 2 3 4 5
Anxiety	0 1 2 3 4 5
Disgust	0 1 2 3 4 5
Interest	0 1 2 3 4 5
Displeasure	0 1 2 3 4 5
Arousal	0 1 2 3 4 5
Suffering	0 1 2 3 4 5
Happiness	0 1 2 3 4 5
Fear	0 1 2 3 4 5
Calmness	0 1 2 3 4 5
Guilt	0 1 2 3 4 5

Fig. 4. List of the Videotest scales.

3.2. Measures and procedure

All participants were administered the Videotest of Emotion Recognition. In addition, subsamples of different sizes completed two emotional intelligence measures, an intelligence test, and two personality questionnaires.

3.2.1. Emotional intelligence measures

The first measure was the Russian adaptation of the Emotion Perception branch of the MSCEIT that consists of the Faces and Pictures subtests (Mayer, Salovey, & Caruso, 2002; Sergienko & Vetrova, 2010). It was chosen because it measures practically the same construct as the Videotest does. Emotionally laden stimuli, such as faces, landscapes, and geometric designs, were administered to 45 participants who had to assess which emotions were present in these stimuli. The second measure was the EmIn Questionnaire, a Russian self-report measure of emotional intelligence that allows for the assessment of people's beliefs about their emotional abilities (Lyusin, 2006a, 2006b). It consists of 46 items with 4-point Likert scale response format, from “completely disagree” to “completely agree”. These items form four scales: Interpersonal EI (e.g., “I understand other people's inner states without words”), Intrapersonal EI (e.g., “I know what to do to improve my mood”), Emotion Comprehension (e.g., “Often, I don't find the words to describe my feelings to my friends”), and Emotion Management (e.g., “If I hurt somebody's feelings, I don't know how to restore the good relationship with him”). The EmIn Questionnaire was completed by 274 participants.

3.2.2. Intelligence measure

Two-hundred and thirty participants completed the Raven's Advanced Progressive Matrices (with a 40 minute time limit) as a measure of general intelligence (Raven, Raven, & Court, 1998).

3.2.3. Personality questionnaires

The Russian adapted version of the NEO Five-Factor Inventory (Costa & McCrae, 1989) was used as a measure of personality traits (66 participants). Also, the Russian adapted version of the Mehrabian and Epstein's Questionnaire Measure of Emotional Empathy (Mehrabian & Epstein, 1972; Tutushkina, 1996) was completed by 55 participants.

The Videotest of Emotion Recognition, the Emotion Perception subtests of the MSCEIT, and Raven's Advanced Progressive Matrices were administered individually. The questionnaires were administered either individually, or in small groups.

4. Results

The data allowed us to use both the expert and consensus approach to establish correct responses. We calculated the test scores both ways, taking the medians of expert estimates and the medians of participants' answers as correct responses. *S* indices, based on expert and consensus approaches, are linear transformations of each other. *A* indices calculated through both approaches were highly correlated ($r = .92$, $p < .01$). This means that it makes no difference which approach to use. In the subsequent analyses, the medians of expert estimates were used as correct responses.

Descriptive statistics for the accuracy and sensitivity indices are presented in Table 2. Since the distributions of both indices did not match the normal distribution (Kolmogorov–Smirnov's test), nonparametric statistical tests were used in further data analysis.

Table 2
Descriptive statistics for the Videotest indices.

	Accuracy index (*A*)	Sensitivity index (*S*)
Mean	1.17	0.20
Standard deviation	0.24	0.42
Minimum	0.67	−0.73
Maximum	2.00	1.56

Table 3
Reliability of the Videotest indices.

Reliability measures	n	Accuracy index (*A*)	Sensitivity index (*S*)
Internal consistency (Cronbach's α)	684	.74	.93
Test–retest reliability (Spearman's correlation)	48	.79	.86

To assess the reliability of the accuracy and sensitivity indices, the internal consistency and test–retest reliability coefficients were calculated (Table 3). Reliability of the *A* index is somewhat lower than of the *S* index; in general, however, reliability of both scores is satisfactory.

The reliability coefficients are comparable to those of other emotional intelligence tests. For example, Cronbach's alphas reported for the Emotion Perception branch of the MSCEIT, the most similar by its content to the Videotest, were .68 for the Faces subtest and .80 for the Pictures subtest (Roberts et al., 2006). The authors of the MSCEIT (Mayer et al., 2003) obtained higher internal consistency indices for these subtests, .82 and .87 respectively. Cronbach's alpha for the STEU is .71 (MacCann & Roberts, 2008); it varies from .86 to .92 for the JACBART (Matsumoto et al., 2000).

Sex differences for the Videotest scores were analyzed by using the Mann–Whitney U test (see Table 4); no significant differences were found.

As noted above, the *A* index is inverse, therefore, we multiplied its values by −1 to make the interpretation of correlation coefficients easier. After this transformation, positive correlation coefficients meant direct relationships between variables and negative coefficients meant inverse relationships.

The Spearman's correlation between the accuracy and sensitivity scores was −.39 ($p < .01$). This moderate negative correlation means that the accuracy and sensitivity indices are not statistically independent; however, they are definitely not identical and can be regarded as reflecting different aspects of emotion recognition.

To assess the validity of the Videotest, Spearman's correlations of its indices with emotional intelligence, general intelligence and personality traits were calculated. The results of this analysis are presented in Table 5.

The accuracy and sensitivity indices correlated with the Emotion Perception branch of the MSCEIT in different ways; accuracy was positively related to Emotion Perception ($r = .38$, $p < .01$), whereas sensitivity had a negative correlation with Emotion Perception ($r = -.32$, $p < .05$). General intelligence did not have any statistically significant correlations with the Videotest indices. The sensitivity index positively correlated with the scale of “Interpersonal Emotional Intelligence” from the EmIn Questionnaire ($r = .13$, $p < .05$) and the Questionnaire Measure of Emotional Empathy ($r = .27$, $p < .05$). It also produced a marginally significant correlation with the “Openness” factor of the NEO Five-Factor Inventory ($r = .21$, $p < .10$). No significant correlations were found between the *A* index and the questionnaire scales.

5. Discussion

The study provided evidence to support the possibility and necessity of distinguishing between two different indices of emotion recognition,

Table 4
Sex differences for the Videotest indices.

	Accuracy index (*A*)	Sensitivity index (*S*)
Men (N = 221)	1.19	0.23
Women (N = 463)	1.16	0.20

Table 5
Spearman's correlations between Videotest indices and cognitive and personality variables.

		n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Accuracy Index (A)	684														
2	Sensitivity Index (S)	684	$-.39^{**}$													
3	MSCEIT Emotion Perception: Total Score	45	$.38^{**}$	$-.31^{*}$												
4	EmIn Questionnaire: Interpersonal EI	274	.03	$.13^{*}$	–											
5	EmIn Questionnaire: Intrapersonal EI	274	.05	.03	–	$.45^{**}$										
6	EmIn Questionnaire: Emotion Comprehension	274	.04	.10	–	$.78^{**}$	$.69^{**}$									
7	EmIn Questionnaire: Emotion Management	274	.04	.06	–	$.65^{**}$	$.82^{**}$	$.53^{**}$								
8	Raven's APM: Total score	230	−.02	.08	−.09	−.06	−.03	−.04	−.05							
9	NEO FFI: Neuroticism	66	.03	−.02	.38	$-.54^{**}$	$-.76^{**}$	$-.54^{**}$	$-.79^{**}$	−.23						
10	NEO FFI: Extraversion	66	−.05	.07	−.38	.16	.07	.16	.06	−.15	$-.34^{**}$					
11	NEO FFI: Openness	66	−.15	$.21^{\dagger}$	.22	.20	.11	$.25^{\dagger}$	.10	−.04	−.02	.03				
12	NEO FFI: Agreeableness	66	.20	−.18	.02	$.33^{*}$	$.27^{\dagger}$	$.31^{*}$	$.28^{\dagger}$	−.03	$-.23^{\dagger}$	.20	−.05			
13	NEO FFI: Conscientiousness	66	−.08	.00	$.68^{*}$	$.29^{\dagger}$	$.53^{**}$	$.40^{**}$	$.49^{**}$	.11	$-.37^{**}$	$.22^{\dagger}$	.06	.14		
14	Questionnaire Measure of Emotional Empathy	55	.12	$.26^{\dagger}$	–	.15	$-.26^{\dagger}$	.04	$-.23^{\dagger}$	$.05^{\dagger}$	–	–	–	–	–	

† p < .10.
* p < .05.
** p < .01.

namely accuracy and sensitivity. On the one hand, the reliability coefficients of these two indices are quite satisfactory. On the other hand, accuracy and sensitivity indices can be regarded as measuring two different aspects of emotion recognition. This latter claim is confirmed by two facts. First, the correlation between these indices is not high (r = −.39), although it is statistically significant. Secondly, the accuracy and sensitivity indices gave different correlation patterns for other cognitive and personality variables. The most dramatic difference was found in correlations with the Emotion Perception subtest of the MSCEIT. The *A* index gives a positive correlation (r = .38), which confirms its validity as an accuracy measure, whereas the *S* index gives a negative correlation (r = −.32).

Only the *S* index correlates with some of the questionnaire scales. It seems quite reasonable to suggest that sensitivity is more of a personality construct, whereas accuracy is a purely cognitive construct reflecting the abilities of emotional information processing. However, no significant correlation between the *A* index and general intelligence has been found. It might be valuable to compare these results with the evidence obtained in other studies of emotion recognition ability. Most of the extant literature tends to focus on the MSCEIT. The subtest scores and the total score of the MSCEIT give low or moderate correlations with intellectual abilities scores. The relationships between the Emotion Understanding subtest and crystallized intelligence (in particular, verbal intelligence) are the most stable (Roberts et al., 2006). Many studies report positive, albeit low correlations between the total score of the MSCEIT and GPA (e.g., r = .16, p < .05, see Brackett & Mayer, 2003).

The main discrepancy between our result and those described above can be explained by the suggestion that those cognitive processes that account for the level of general intelligence do not play an essential role in emotion recognition. This suggestion is indirectly supported by the evidence that the relationship between the MSCEIT subtest ‘Emotion Understanding’ and general intelligence is widely replicated in different studies. The material of this subtest is entirely verbal; therefore, it mostly uses participants' verbal abilities. However, the Emotion Perception subtest (the least verbal in the MSCEIT and the most similar to the Videotest in this sense) does not provide any stable relationships with general intelligence.

Another possible explanation of the absence of the relationships between the *A* index and intelligence scores could be the response format of test items. According to MacCann and Roberts (2008), correlations between emotional intelligence and general intelligence depend on the response format of the emotional intelligence test items. Items with the same content give higher correlations with general intelligence if a multiple-choice response format is used, instead of a Likert-scale format. This regularity holds true for any emotional abilities including emotion understanding and emotion management. A version of the Videotest with multiple-choice items may give higher correlations with intelligence tests.

It is particularly interesting that no significant sex differences in the Accuracy and Sensitivity scores were found. Women had a slightly better accuracy of emotion recognition, but this difference did not reach statistical significance. The most recent meta-analysis of emotion recognition (Thompson & Voyer, 2014) showed a small overall advantage in favor of women (Cohen's *d* = .19). This study analyzed various factors that can moderate sex differences including the intention of the actors, posed vs. spontaneous portrayals of emotions. The mean effect size for posed emotions was .78, whereas the mean effect size for spontaneous emotions was only .06. Although the moderator analysis revealed no significant contribution of this factor to the variability of effect sizes, we find the observed difference between posed and spontaneous emotion to be particularly relevant for our study. Most studies on emotion recognition use stimuli with posed emotion, whereas our Videotest used videos with spontaneous real-life emotions. This feature of our stimuli can probably explain the absence of sex differences in the results.

The Videotest of Emotion Recognition, as described in this paper, can be developed further and improved in various ways. The selection of a larger set of video recordings would represent a more diverse array of emotional states. The next step in analyzing the validity of the accuracy and sensitivity indices would be searching for their connections with some real-life achievements. Future research could also examine the use of the two suggested indices of emotion recognition in other emotion abilities measures.

Acknowledgment

The authors would like to thank Martin Lynch for his helpful comments on the text during manuscript preparation.

References

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

Brackett, M.A., & Mayer, J.D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin*, *29*, 1147–1158.

Costa, P.T., & McCrae, R.R. (1989). *The NEO-PII/NEO-FFI manual supplement*. Odessa, FL: Psychological Assessment Resources.

Feldman Barrett, L., & Russell, J.A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science*, *8*, 10–14.

Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., & Ellsworth, P. (2007). The world of emotions is not two-dimensional. *Psychological Science*, *18*, 1050–1057.

Gunes, H., & Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, *1*, 68–99.

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

Kuppens, P., Tuerlinckx, F., Russell, J.A., & Feldman Barrett, L. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, *139*, 917–940.

Legree, P.J., Psotka, J., Tremble, T., & Bourne, D.R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze, & R.D. Roberts (Eds.), *Emotional intelligence: An international handbook* (pp. 155–179). Cambridge, MA: Hogrefe.

Lyusin, D. (2006a). Emotional intelligence as a mixed construct: Its relation to personality and gender. *Journal of Russian and East European Psychology*, *44*, 54–68.

Lyusin, D. (2006b). A new measure for emotional intelligence: EmIn Questionnaire. *Psikhologicheskaya Diagnostika*, 3–22 (In Russian).

MacCann, C., & Roberts, R.D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, *8*, 540–551.

Matsumoto, D., LeRoux, J.A., 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.

Mayer, J.D., Salovey, P., & Caruso, D. (2000). Competing models of emotional intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 396–420) (2nd ed.). New York: Cambridge University Press.

Mayer, J.D., Salovey, P., & Caruso, D.R. (2002). *The Mayer, Salovey, and Caruso Emotional Intelligence Test: Technical manual.* Toronto: Multi-Health Systems.

Mayer, J.D., Salovey, P., Caruso, D.R., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion*, *1*, 232–242.

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

Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality*, *40*, 525–543.

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.

Ovsyannikova, V. (2007). *Role of cognitive factors in emotional states recognition.* (Doctoral thesis) Moscow: Institute of Psychology of the Russian Academy of Sciences (In Russian).

Raven, J., Raven, J.C., & Court, J.H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 4. The Advanced Progressive Matrices.* San Antonio, TX: Harcourt Assessment.

Reis, H.T., & Judd, C.M. (Eds.). (2000). *Handbook of research methods in social and personality psychology*. New York: Cambridge University Press.

Reisenzein, R. (1994). Pleasure-activation theory and the intensity of emotions. *Journal of Personality and Social Psychology*, *67*, 525–539.

Roberts, R.D., Schulze, R., O'Brien, K., MacCann, C., Reid, J., & Maul, A. (2006). Exploring the validity of the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) with established emotions measures. *Emotion*, *6*, 663–669.

Roseman, I.J. (2001). A model of appraisal in the emotion system: Integrating theory, research, and applications. In K.R. Scherer, & A. Schorr (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 68–91). New York: Oxford University Press.

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.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178.

Salovey, P., & Grewal, D. (2005). The science of emotional intelligence. *Current Directions in Psychological Science*, *14*, 281–285.

Scherer, K.R., & Scherer, U. (2011). Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the emotion recognition index. *Journal of Nonverbal Behavior*, *35*, 305–326.

Sergienko, E.A., & Vetrova, I.I. (2010). *J. Mayer, P. Salovey, and D. Caruso's Emotional Intelligence Test (MSCEIT v. 2.0). Manual.* Moscow: Institute of Psychology of the Russian Academy of Sciences (In Russian).

Thompson, A.E., & Voyer, D. (2014). Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis. *Cognition and Emotion*, *28*, 1164–1195.

Tutushkina, M.K. (Ed.). (1996). *Practical psychology for managers*. Moscow: Filin (In Russian).

Warwick, J., Nettelbeck, T., & Ward, L. (2010). AEIM: A new measure and method of scoring abilities-based emotional intelligence. *Personality and Individual Differences*, *48*, 66–71.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., & Rigoll, G. (2013). LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, *31*, 153–163.