ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/326503071

# Development and Validation of the Facial Expression Recognition Test (FERT)

**Article** *in* Psychological Assessment · July 2018

DOI: 10.1037/pas0000595

CITATIONS
0

READS
596

**5 authors**, including:

Marcello Passarelli
Italian National Research Council
**20** PUBLICATIONS **9** CITATIONS
SEE PROFILE

Michele Masini
Università degli Studi di Genova
**7** PUBLICATIONS **4** CITATIONS
SEE PROFILE

Fabrizio Bracco
Università degli Studi di Genova
**39** PUBLICATIONS **110** CITATIONS
SEE PROFILE

Carlo Chiorri
Università degli Studi di Genova
**114** PUBLICATIONS **655** CITATIONS
SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**



Gay, Lesbian and Bisexual People's Attitudes towards Sexual Minority People: The Adherence to Traditional Gender Roles View project



Development of an integrated methodology for the assessment and development of non-technical skills in simulated scenarios. View project

All content following this page was uploaded by Marcello Passarelli on 21 March 2019.

The user has requested enhancement of the downloaded file.

## Abstract

Detecting the emotional state of others from facial expressions is a key ability in emotional competence and several instruments have been developed to assess it. Typical emotion recognition tests are assumed to be unidimensional, use pictures or videos of emotional portrayals as stimuli, and ask the participant which emotion is depicted in each stimulus. However, using actor portrayals adds a layer of difficulty in developing such a test: the portrayals may fail to be convincing and may convey a different emotion than intended. For this reason, evaluating and selecting stimuli is of crucial importance. Existing tests typically base item evaluation on consensus or expert judgement, but these methods could favour items with high agreement over items that better differentiate ability levels and they could not formally test the item pool for unidimensionality. In order to address these issues we propose a new test, named Facial Expression Recognition Test (FERT), developed using an IRT 2PL model. Data from 1002 online participants were analysed using both a unidimensional and a bifactor model, and showed that the item pool could be considered unidimensional. The selection was based on the items' discrimination parameters, retaining only the most informative items to investigate the latent ability. The resulting 36-item test was reliable and quick to administer. We found both a gender difference in the ability to recognize emotions and a decline of such ability with age. The PsychoPy implementation of the test and the scoring script are available on a Github repository.



*Keywords:* facial expression recognition, emotion detection, Item Response Theory, emotional competence, Bayesian 2PL.


*Public Significance Statement:* This study presents the development and validation of a test measuring the ability to recognize emotions from facial expressions. We present evidence that the emotion recognition ability could be considered a single ability, rather than several emotion-specific abilities. Furthermore, great care was taken in devising a strategy to select only the most informative items for inclusion in the final test.

# Development and Validation of the Facial Expression Recognition Test (FERT)

The ability to accurately perceive the emotional state of others is a crucial component of emotional competence (Scherer, 2007) that comes into play in many social situations (Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979) and varies across individuals (Lyusin & Ovsyannikova, 2015). Information on emotional states can be conveyed through many verbal and nonverbal channels, among which facial expressions are one of the most studied (Bänziger, Grandjean, & Scherer, 2009). Specific training programs have been devised to teach explicit strategies to detect emotions on the basis of facial expressions alone (e.g. Bölte et al., 2002). Several instruments have also been designed to measure individual differences in this ability; however, some authors noted a surprising lack of concern in developing measures that are psychometrically sound (Bänziger et al., 2009). Table 1 offers a comparison of existing measures, listing their strengths and shortcomings.

[Please insert Table 1 about here]

It can be observed that the typical emotion recognition test is based on videos, pictures, or recordings of actors interpreting a specific emotion. The participant has to recognize the emotion in each item, and the number or proportion of correct responses is used as a measure of general emotion recognition ability (e.g.: Bänziger et al., 2009; Nowicki & Duke, 1994; Scherer & Scherer, 2011). Some of these instruments are designed to inspect this ability by using different channels of communication (e.g. voice or body movements; see Bänziger et al., 2009; Rosenthal et al., 1979), while others specifically focus on the use of facial expressions (e.g. Lyusin & Ovsyannikova, 2015; Mayer, Salovey, Caruso, & Sitarenios, 2003). However, a critical challenge in devising these measures lies in the process used for developing and selecting the test material. Emotional portrayals by actors may provide inadequate items: as the actor may fail to convincingly portray an emotion, an item may present an ambiguous, or even incorrect, scoring. In order to guide item selection and exclude problematic stimuli, test designers have relied on consensus by participants and/or experts. However, using a consensus-based method might drive the selection towards

relatively easy items (i.e., items for which the correct answer is the most frequently chosen alternative), without considering which items are more informative about the latent ability. On the other hand, if a direct measure of item informativeness can be obtained, it can be used to weigh item responses for scoring, thus obtaining a more precise measure.

*The issue of test unidimensionality*

A second challenge in assessing the ability to recognize emotions on the basis of facial expressions pertains to the dimensional structure of the construct itself. Typical emotion recognition tests measure emotional recognition ability as a general, unidimensional construct, meaning that a highly-skilled individual has a higher recognition ability for *all* emotions. However, several studies report impairments in the recognition of *specific* emotions: lesions to the amygdala lead to reduced recognition of fear (Adolphs, Tranel, Damasio, & Damasio, 1994; Calder et al., 1996; Calder, Lawrence, & Young, 2001; Sprengelmeyer et al., 1999); recognition of disgust is impaired in patients with Huntington's disease (Sprengelmeyer et al., 1996), lesions to insular cortex or putamen (Calder, Keane, Manes, Antoun, & Young, 2000), or obsessive-compulsive disorder (Sprengelmeyer et al., 1997); lesions to ventral striatum or alterations of the dopaminergic system seem to impair recognition of anger (Calder, Keane, Lawrence, & Manes, 2004; Lawrence, Calder, McGowan, & Grasby, 2002). The existence of emotion-specific impairments, as well as the presence of distinct patterns of cerebral activation to different facial expressions (Breiter et al., 1996; Krolak-Salmon et al., 2003; Morris et al., 1996; Phillips et al., 1997), casts doubt on the unidimensionality of the emotion recognition ability.

A promising approach for addressing the item selection issue is represented by the Item Response Theory (IRT) framework. Although recently it has moved beyond the confines of maximum performance tests into assessment domains such as personality, psychopathology, and patient-reported outcomes (Reise & Revicki, 2015), IRT has traditionally been applied in contexts

in which there is no doubt about which answer is correct (e.g., a mathematical test with unambiguous solutions).

The most common IRT measurement model assumes that a single continuous latent variable (usually labelled as $\theta$) can represent individual differences on a psychological construct, either an ability or a trait. For dichotomous item responses, a possible item response curve can be:

$$P(x = 1 \mid \theta) = \exp [1.7\alpha(\theta - \beta)] / (1 + \exp [1.7\alpha(\theta - \beta)])$$

where $P$ is the probability to provide a correct answer to the item, $\alpha$ is the item discrimination (or slope) and corresponds to the slope of the item response curve at $P = .50$, $\beta$ is the item location (or difficulty, i.e., the amount of ability required for having a .50 probability to provide a correct answer), and 1.7 is a scaling factor that makes the value of the item discrimination parameter in logistic models comparable to a normal-ogive model.

This model is known as the two-parameter logistic (2PL) model, and its main feature is that items with higher discrimination count more towards $\theta$ than items with lower discrimination. In other words, it is not only a matter of *how many* items a participant gets correct, but also *which* ones.

As an item's *discrimination* is the slope of the function linking the probability of correctly answering the item with the latent ability of the respondent, higher discrimination means that the latent ability is more strongly associated with the probability of answering correctly to that specific item. High discrimination also provides evidence that an item has a correct and unambiguous scoring: sub-optimal items will have a low, or even negative (if incorrectly labelled) discrimination. This can happen, for instance, if the actor has failed to produce a facial expression representative of the emotion requested (e.g., the experimenter requested an expression of anger, but the expression produced is more representative of disgust). In this case, item discrimination will be negative, as individuals with higher emotion recognition ability will select an emotional label (disgust) different from the one deemed correct (anger) with higher probability than individuals with low emotion recognition ability. Maximizing for item discrimination can also help to obtain more information on

the latent ability for a given number of items.

The estimation of IRT item parameters depends on the degree to which item response data meet the unidimensionality (or multidimensionality) assumption. Nevertheless, in applied research data are rarely *strictly* unidimensional and some rules of thumb (e.g., a combination of fit indices from the structural equation modeling [SEM] framework, residual values, and eigenvalue ratios) have been developed to decide whether data are "unidimensional enough" for such models (for a review, see Reise, Cook, & Moore, 2015). Reise and coworkers (Reise, 2012; Reise, Bonifay, & Haviland, 2013; Reise, Cook, & Moore, 2015; Reise, Moore, & Haviland, 2010; Reise, Morizot, & Hays, 2007; Reise, Scheines, Widaman, & Haviland, 2013; Reise, Moore, & Maydeu-Olivares, 2011; Rodriguez, Reise, & Haviland, 2016a, 2016b) have recently pointed out that event when the commonly applied procedures to check for unidimensionality provide evidence that support unidimensionality, the researcher cannot be confident that the common target latent trait is identified correctly or that the estimated item parameters properly reflect the relation between item responses and the common latent trait. Moreover, an adequately fitting unidimensional model according to common SEM fit indices can still yield item parameter estimates biased by multidimensionality, due, e.g., to a single correlated residual. Conversely, it is also possible that even when a unidimensional model shows a poor fit according to SEM fit indices, and/or a multidimensional solution yields improved statistical fit, the application of IRT may still be viable

Hence, Reise et al. (2015) proposed a different approach, that does not address the issue of whether the data are "unidimensional enough", but rather the degree to which multidimensionality impacts or distorts the estimation of item parameters. This criterion is based on the equivalence of IRT and item-level factor analysis (Takane & de Leeuw, 1987) and on the application of exploratory bifactor analyses (Jennrich & Bentler, 2011; Schmid & Leiman, 1957) and targeted factor rotations (Browne, 2001) to directly model and assess the impact of multidimensionality on IRT item parameter estimates.

Relevant to this study, none of the measures presented in Table 1 has used factor loadings for

selecting the items to be included in the test and, to the best of our knowledge, the only instrument to measure facial expression recognition built using an IRT framework is the Geneva Expression Recognition Test (GERT; Schlegel, Grandjean & Scherer, 2014). However, due to a relatively small sample size, the GERT employed a 1PL model, which estimated items' difficulty but not their discrimination. Estimating item difficulty is useful in order to select items that discriminate well for all ability levels of interest. However, in an emotion recognition test difficulty alone cannot guarantee that the item is correctly scored and is therefore a valid indicator of the latent variable of interest. The 2PL model, while requiring larger sample sizes, estimates both difficulty and discrimination parameters, giving an additional measure of item quality.

The aim of the current study was therefore to develop a new test to measure an individual's ability to detect the six basic emotions (happiness, sadness, anger, disgust, fear, and surprise; see Ekman, 1992 for an evolutionary account of the fundamental importance of these emotions) on the basis of facial expressions using a 2PL IRT model. The test will focus on facial expression recognition alone, minimizing other cues individuals may rely on for emotion recognition in daily life (e.g. posture, tone of voice, speech). The resulting test would therefore be useful for experimental research on facial expression recognition or for evaluating training programs focusing on it; any deficit in emotion recognition detected by the test would likely be compensated in daily life through the use of channels other than facial expressions.

In order to evaluate the possibility that facial expression recognition ability is not a unidimensional construct, we followed the procedure suggested by Reise et al. (2015) and labelled by the authors as the "comparison modeling" method. As a first step, a unidimensional model is estimated (referred to as the "restricted" model). In this model, items are considered reflective indicators of a single, general latent dimension (in this case, the ability to recognize emotions). Then an "unrestricted" bifactor model that could better represent the multidimensional (i.e., bifactor) data structure is estimated. In this model, it is assumed that one common, general factor underlies the variance of all the scale items and a set of orthogonal group factors are specified that

account for additional variation, usually assumed to arise because of item parcels with similar content (in this case, the ability to recognize specific emotions). Item slope parameter estimates on the restricted model are compared to item slope parameter estimates on the general factor in the unrestricted model. Grounding on the assumption that the unrestricted model is a more accurate representation of the relationship between the items and the common trait that is measured by the scale, the comparison of the two sets of parameter estimates provides a direct index of the degree to which item slope parameters are distorted because of forcing multidimensional data into a unidimensional model.

To the best of our knowledge, no existing instrument was developed in a multidimensional framework, taking into account the possibility of the existence of both a general recognition ability and emotion-specific recognition abilities. However, there are studies in which responses for different emotions have been evaluated separately (e.g. Bänziger et al., 2011; Matsumoto et al., 2000). The high average correlation between emotion-specific scores suggests the existence of a general recognition ability.

A relatively large pool of original items (N=108) was developed. The development of new test material allowed us to address other shortcomings of existing measures, such as low picture quality (e.g. Bänziger et al., 2009; Matsumoto et al., 2000; Nowicki & Duke, 1994; Warwick et al., 2010), long administration time (e.g. Bänziger et al., 2009; Rosenthal et al., 1979; Schlegel, Grandjean & Scherer, 2014 — which are, however, multimodal tests), lack of balance for actor gender (e.g. Rosenthal et al., 1979), or use of non-professional actors (e.g. Lyusin & Ovsyannikova, 2015; Herzmann et al., 2008). Additionally, we took care not to include extremely attractive or unattractive actors, since attractiveness can influence the recognition of some emotions (Limbrecht et al., 2012).

Previous studies evidenced a gender difference in emotion recognition ability. Women seem to be more capable of recognizing emotions from facial expressions (Hall, 1978; Schlegel et al., 2013). Hall (1978) presents three possible explanations for this gender difference in facial

expression recognition. The first has to do with gender socialization: women are believed to be more capable than men at decoding nonverbal cues (Broverman, Vogel, Broverman, Clarkson, & Rosenkrantz, 1972), and will tend to be more attentive to facial expressions in order to conform to this role. The second is that facial expression recognition may be especially socially adaptive for women, due to their status as an oppressed minority (Weitz, 1974). The third is that women are 'wired' to be especially good at facial expression recognition due to evolutionary advantages in being able to detect detect distress in their children and/or threatening signals in other individuals (Hall, 1978).

Ruffman, Henry, Livingstone, & Phillips (2008) also found a decline of emotion recognition ability with age, a finding consistent with age-related changes in the volume of frontal and temporal lobes. We expected to replicate both the gender and age effects on facial expression recognition.

## Method

*Material production*

Six professional actors (3 males) were asked to interpret the six basic emotions of happiness, sadness, anger, disgust, fear, and surprise. Each actor interpreted the set of emotions six times: the first two sessions were completely unguided; the third and fourth sessions were guided by an expert Facial Action Coding System (FACS - Ekman & Friesen, 1978) coder; the fifth and sixth sessions were based on imitation of emotional portraits included in the Picture of Facial Affect (Ekman, 1976). The order of emotion portrayals within sessions was randomized.

Emotional interpretations were video-recorded against a black background using a frontal camera focused on the upper body (see Figure 1 for an example). The actors were asked to represent each emotion only by means of their faces, not using their shoulders, arms, or hands. For each combination of actor and expression a triplet of colour still frames was selected by a FACS coder for the initial item pool, which consisted of 108 frontal emotional portrayals (6 actors x 6 expressions x 3 selected frames).

[Please insert Figure 1 about here]

*Data collection*

Data on emotional recognition were collected using the LimeSurvey platform 2.05+ Build 150211 through snowball sampling on social networks. For each of the 108 items, participants were shown a neutral (i.e., non-expressive) picture of the same actor as reference, along with the emotional portrayal to be recognized. Participants had to select which of the six emotions was being portrayed by the actor. As the complete test was deemed too long for voluntary participation, only 54 out of 108 items were randomly presented to each participant. A total of 1151 Italian speakers took part in the study, and 794 answered all the 54 items they were shown (63.8% Females, mean age 36.13 ± 13.79; no demographic data was available for participants who did not complete the survey) while 1002 provided at least 10 responses. At the end of the emotion detection task, participants judged each actor's attractiveness on a scale of 1 to 10, using the neutral picture. Lastly, the participants' age and gender were collected. The research protocol was approved by the ethical board of the University of Genoa.

*Analysis*

The first step of analysis entailed a check of unidimensionality of the 108 item pool using 5 datasets imputed through the *imputeMissing* function in the R package *mirt* (Chalmers, 2012)[1]. The analyses were performed using Mplus 7 (Muthén & Muthén, 1998-2012) using the weighted least square mean and variance adjusted (WLSMV) estimator for item-level factor analyses and the maximum likelihood (ML) estimator for the 2PL. For both item-level factor analyses and 2PL models we considered both a unidimensional model and a bifactor model. For the bifactor model, a tetrachoric correlation matrix was computed, and a specified number of primary factors (in this case 7: a general ability factor and six emotion-specific factors) were extracted. An oblique factor

1 Computer memory constraints prevented us from conducting the unidimensionality check using a Bayesian approach.

rotation (in this case a bi-geomin oblique rotation, as implemented in Mplus 7) was performed and a higher-order factor from the primary factor correlation matrix was extracted. Finally, a Schmid-Leiman (SL, Schmid & Leiman, 1957) orthogonalization to obtain the loadings for each item on the general and group factors was performed. If items were to present simple loading patterns (i.e., no cross-loadings) on the oblique factors, they would tend to load on one and only one group factor. Reise et al. (2015) suggest to inspect the pattern of loadings on the group factors to specify a target rotation matrix in which, if in the SL a loading is greater than or equal to |.15|, then the corresponding element of the target matrix is unspecified (?) and if it is less than |.15| it is specified (0).

After providing evidence for the unidimensionality of the construct, the model was fitted using a Bayesian approach. Unlike the ML estimation, the Bayesian approach can take into consideration uncertainty in the parameter estimates, a feature that is especially useful when modelling data with a high amount of missing observations. Additionally, Bayesian modelling using (weakly) informative priors provides finite – if highly uncertain – parameter estimates when ML cannot (e.g., if a participant answers correctly to all items). The unidimensional model and its priors are summarized in Figure 2. All priors were weakly informative. Weakly informative priors give slightly higher prior probability to parameter values most commonly seen in IRT (e.g., a discrimination > 5 would be highly improbable), and have a lower risk of distorting results than strongly informative priors. Discriminations were constrained to be positive in order to avoid sign-switching between iterations; this was necessary to achieve convergence, as the possibility of sign-switching would allow two equally likely solutions, and lead to bimodal distributions for discrimination parameters. However, since some discriminations were expected to be negative (due to possible incorrect item coding), they were expected to show values close to zero when constrained to be positive.

[Please insert Figure 2 about here]

The model was fitted using Stan 2.12.1 (Stan Development Team, 2015) through R 3.3.1 (R Core Team, 2016) using Stan's built-in Hamiltonian Monte Carlo sampler. Four parallel chains were run for 20,000 iterations (10,000 burn-in), resulting in a total of 40,000 samples for each parameter. The number of chain and iterations were chosen following the suggestions made by Depaoli & van de Schoot (2015) and checking for convergence post running using the R hat statistic (see the Supplementary Materials for R and Stan codes, convergence statistics, and averages and standard deviations for estimated parameters of the 108-item model).

The third step of analysis aimed at reducing the number of items in the final version of the test. The original pool of 108 items could be divided into triplets, as there were three items for each combination of actor and emotion portrayed. For each triplet, only one item was retained, following a single criterion: item discriminations for each triplet were directly compared using the mean of the parameters' posteriors as a summary statistic, and the most highly discriminative item for each triplet was selected for inclusion in the final version of the test. Once items were selected, the model was re-run considering only the 36 selected items. Parameters estimated using this model were analysed to determine the influence of gender and age on facial expression recognition ability (or abilities), and to check whether or not actor gender and emotion portrayed were associated with item discrimination and/or difficulty.

## Results

Descriptive statistics are reported in the Supplementary Materials. As actors' attractiveness judgements means ranged from 4.77 to 6.06 — very close to the median point of the scale — analyses were conducted considering all six actors.

The unidimensionality check was performed twice: one considering only 794 participants that completed all the 54 items they were shown, and the second time including all 1002 participants that provided at least 10 valid answers. Results were overlapping, hence we have reported in the Supplementary Materials (Table SM3) only results on the latter sample.

Table SM3 reports the factor loadings and the discrimination parameters for the restricted unidimensional model and for the bifactor model. As shown in Table SM3, however, we found very little evidence of item loadings on the group factors exceeding the |.15| threshold, hence almost all the cells of the target matrix would be specified as 0. Moreover, both the factor loadings and the discrimination parameters in the restricted and unrestricted models were substantially the same ($r$ = .998 in both cases), suggesting that the unidimensionality of the 108 items could be reasonably assumed. Notably, commonly used fit indices from factor analytic and IRT models also supported the unidimensionality of the item pool (see Table SM4 in the Supplementary Materials). Residual correlations for the unidimensional model exceeded |.20| in 75 out of 5778 cases (1.30%, highest residual correlation = |.42|). Of these relatively higher residual correlations, 11 involved items related to the same actor (regardless of the emotion), 22 the same emotion (regardless of the actor), and 1 the same emotion in the same actor. Note that at this stage of the analyses we did not consider the loadings on the general ability factor (Reise et al., 2015 suggest to drop items with a loading smaller than |.30| on the general factor), since the final item selection was to be guided by the Bayesian estimations.

We therefore proceeded with the Bayesian analyses considering only a unidimensional model. Parameters estimated for the 108-items model are available in the Supplementary Materials. The estimates of discrimination parameters were used to guide the selections of the final 36 items retained, according to the procedure detailed in the methods section. Parameters estimated for the resulting 36-items model are shown in Table 2.

[Please insert Table 2 about here]

The resulting test can be considered easy, as average item difficulty was low (-2.95; see Figure 3 for the Test Information Function). Therefore, the test is more accurate when the estimated latent ability is relatively low.

[Please insert Figure 3 about here]

The test score reliability, computed as described in Raykov, Dimitrov, and Asparouhov (2010), was .92. Considering the ability score estimated using the 36-item unidimensional model, we found a small gender difference (higher scores in females: $t(566.78) = -3.71$, $p < .001$, *Cohen's d* = 0.28) and a slight decline of ability with age ($r = -.17$, $p < .001$).

Estimated item discrimination and difficulty were analysed using factorial ANOVA to test for differences by actor gender, emotion portrayed, and interaction of actor gender and emotion (see Table 3 for estimated marginal means). For item discrimination, no difference was found on the basis of gender ($F(1, 24) = 0.003$, $p = .956$, partial $\eta^2 < .001$). Difference on the basis of emotion depicted was non-significant ($F(5, 24) = 2.33$, $p = .074$, partial $\eta^2 = .326$). Post-hoc tests performed using the Tukey's Honest Significant Difference procedure did not reveal any significant difference; the highest difference was found between disgust and anger (anger was more discriminative; $p = .116$). The interaction of gender and emotion was non-significant ($F(5, 24) = 0.39$, $p = .854$, partial $\eta^2 = .074$). Regarding difficulty, no difference was found for actor gender ($F(1, 24) = 0.04$, $p = .840$, partial $\eta^2 = .002$). A significant effect was found for emotion portrayed ($F(5, 24) = 6.58$, $p < .001$, partial $\eta^2 = .578$), and post-hoc tests revealed that fear was significantly harder to recognize than disgust ($p = .030$), happiness ($p = .001$), and surprise ($p = .004$). Sadness was also harder to recognize than happiness ($p = .022$). No significant effect was found for the interaction of actor gender and emotion ($F(5, 24) = 1.90$, $p = .132$, partial $\eta^2 = .283$). Table 3 reports estimated marginal means for each combination of actor gender and emotion portrayed.

[Please insert Table 3 about here]

Of the final 36 items, out of 630 residual correlations, 30 (4.76%) exceeded |.20| (highest residual correlation = |.30|). Of these, 3 involved items related to the same actor (regardless of the emotion), 6 the same emotion (regardless of the actor), and none the same emotion in the same actor.

## Discussion

Several measures of emotional recognition ability have been developed. Some of them present minor issues (e.g. dated stimuli, gender imbalance, significant length). Two major issues, however, pertain to the assessment of stimulus quality and to the possible multidimensionality of the emotion recognition ability.

Regarding the first issue, the measures listed in Table 1 determine the 'correct' emotion depicted in each stimulus using consensus among participants, expert judgment, and/or the actor's intention, and weigh all items equally when estimating a participant's ability; some of them retained all items with sufficient consensus. A step forward in item evaluation is represented by the GERT, which was developed using a Rasch model (Schlegel, et al., 2014). However, a Rasch model estimates only item difficulty (and not item discrimination), and can be used to select items that are informative on the ability range of interest.

Our aim was to build upon Schlegel et al.'s (2014) recent increase in methodological rigor, using a larger sample and a more complex (2PL) model in order to evaluate stimuli on their quality and informativeness, while still obtaining estimates of their difficulty. Stimuli selection was directly guided by the estimate of discrimination for each item (whereas previous measures typically based item selection on consensus and/or expert judgement). This strategy allowed us to reliably detect scoring errors, which may go unnoticed using other approaches. The 2PL scoring, additionally, does not weigh items equally and takes into account both their difficulty and discrimination when estimating the participant's latent ability. Additionally, the use of IRT models allowed us to compute the test information function, and detect which levels of ability would be more reliably measured by the FERT.

As for the possible multidimensionality of the emotion recognition ability, most existing measures treat emotion recognition ability as a unidimensional construct, computing a single ability score for each individual. There is compelling evidence that neurological and psychological disorders may selectively impair recognition of a single emotion, suggesting that emotion

recognition ability can be emotion-specific. This would imply a multidimensionality of the construct in a patient population, but similar evidence is not available for the non-clinical population. Therefore, we used a recently developed method for assessing the unidimensionality of the item pool, i.e., the Reise et al. (2015) comparison method, that allows the comparison of item factor loadings/discrimination parameters between a restricted, unidimensional model (i.e., a single emotion recognition ability factor) and an unrestricted bi-factor model, in which group factors (i.e., six emotion-specific abilities) are also considered. The results supported the unidimensionality of the item pool, suggesting that the estimation of FERT item parameters with the 2PL model was negligibly, if ever, impacted or distorted by any sort of multidimensionality. However, since we used a convenience sample of online participants, further research is needed to test whether emotion recognition ability is not emotion-specific in both patient and non-clinical populations.

We also aimed to address some of the minor issues highlighted in the introduction. Great care was taken in obtaining stimuli with high picture quality in controlled lighting conditions. Stimuli selection sought to ensure that actor gender and emotions depicted would be balanced across retained stimuli. Retaining only 36 items ensured that administration time would be short (the median time for taking the full 108-items test was 10.1 minutes; the median time for completing the 36-item version, as measured in follow up studies, was 5.15 minutes), and selecting the items with highest discrimination allowed us to maximize the informativeness of the resulting measure. The involvement of professional actors only may have helped in obtaining convincing emotional portrayals (Bänziger, Mortillaro, & Scherer, 2011), and the decision to use three different methods of expression elicitation (free interpretation, guided interpretation, and imitation) allowed us to select the stimulus material from a wide pool of varied portrayals. The crucial decision to limit the test material to Ekman's six basic emotions, while restricting the range of emotions considered, ensured that the facial expressions included in the test would be cross-culturally valid, at least in Western societies. Lastly, the test material is in colour (potentially adding to ecological validity), and stimuli have been controlled for perceived attractiveness of the actor, a characteristic that

influences expression recognition (Limbrecht et al., 2012). Taken together, these features represent important, albeit small, steps towards ecological validity of the test material with respect to previous measures.

The results presented here replicate previous results on emotion recognition ability, showing a small gender effect (females perform better than males) and a decline in ability with age, consistent with results reported by Hall (1978), Ruffman et al. (2008), and Schlegel et al. (2013).

While gender differences are not the focus of the present study, it should be noted that Hall (1978) theorized that, if the gender effect were the result of gender socialization, its effect size should diminish over time due to rapid changes in gender stereotypes and gender inequality in Western societies. However, almost 40 years later, the effect size we found (d = .28 [.42, .13]) is not significantly different from the one reported by Hall (1978; d = .40).

Results on item difficulty are consistent with those reported by Biehl et al. (1997), who argued that happiness and surprise are relatively easy to recognize, and fear comparatively harder, as well as with Matsumoto et al. (2000), who identify happiness, disgust, and surprise as the most easily recognized emotions and fear and sadness as more difficult to detect.

### *Limitations*

The FERT presents a few shortcomings that could not be overcome. The use of actor portrayals instead of spontaneous emotions is a widespread problem in emotional recognition measures, which cannot be easily sidestepped due to ethical concerns on recording private, substantially unpredictable, and fleeting episodes for a number of individuals (Bänziger et al., 2011).

Other limitations stem from the sampling strategy: data was collected online, with no incentives for finishing the questionnaire; the resulting sample over-represents females, and there is no guarantee that participants took the test in a quiet environment.

Furthermore, the decision to restrict the stimuli to still images, while leading to a very short measure, limits – by design – the assessment of the explored construct to a single modality (still

facial expressions). The construct explored may be narrow compared to the more general emotion detection ability, which may include the capacity to detect emotions from dynamics, speech, and body movements. Therefore, possible future extensions of the test could take into account a wider range of emotional cues.

A yet more practical issue concerns the ease of use of the final test: the scoring procedure for 2PL models can be quite complex. We provide an R script that can be used to compute the ability score for new participants on the basis of their response pattern, using the PsychoPy output file as input (see Supplementary Materials and Github repository https://github.com/M-Pass/FERT, where the final test is available for download). The script computes uncertainty estimates for $\theta$. We hope this endeavour will facilitate scoring enough to encourage use of the test, but any possible countermeasure will take longer than simply computing the proportion of correct responses.

Lastly, due to the low mean difficulty of the items, the FERT is more precise when estimating low ability scores than when administered to highly skilled individuals. On the other hand, the high amount of correct responses observed suggests that test instructions are clear and that participants had no trouble understanding the task.

*Future work*

Future research should focus on further testing the psychometric properties of the test: data should be collected to test generalization of results to other cultures[2], test-retest reliability, and predictive validity. Parallel versions of the test can be built by matching 2PL item parameters in order to mitigate training effect in test-retest (see, e.g., Chen, Chang, & Wu, 2012). Adding new test items with both high difficulty and high discrimination could make the test more accurate in estimating highly-skilled individuals' ability.

The development of the test opens up new research possibilities, as it can be used to explore to what extent facial expression recognition ability evolves over time and can be actively trained (or, on the other hand, if it is a trait ability that cannot be trained at all); whether specific

populations (e.g. patients with mental disorders) significantly differ from the general population; which strategies (e.g. in visual exploration) lead to correct or erroneous emotion detection; whether or not the unidimensional model holds in clinical populations (especially in those populations which exhibit specific deficits – see the introduction). While some of these questions have already been explored by previous studies, we believe the use of a psychometrically-sound measure, whose items have been specifically tested for the informativeness on the underlying construct, can be of help in shedding some more light on ongoing research issues.

References

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature, 372*(6507), 669-672. http://doi.org/10.1038/372669a0

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion, 9*(5), 691–704. https://doi.org/10.1037/a0017088

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2011). Introducing the Geneva multimodal expression corpus of experimental research on emotion perception. *Emotion, 12*(5), 1161-1179. https://doi.org/10.1037/a0025827.

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21*(1), 3-21. https://doi.org/10.1023/A:1024902500935

Bölte, S., Feineis-Matthews, S., Leber, S., Dierks, T., Hubl, D., & Poustka, F. (2002). The development and evaluation of a computer-based program to test and to teach the recognition of facial affect. *International Journal of Circumpolar Health, 61*(2), 61-68. https://doi.org/10.3402/ijch.v61i0.17503

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., ... & Rosen, B. R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron, 17*(5), 875-887. http://doi.org/10.1016/S0896-6273(00)80219-6

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal. *Journal of Social Issues, 28*(2), 59-78. https://doi.org/10.1111/j.1540-4560.1972.tb00018.x

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate*

*Behavioral Research, 35*(1), 111–150. http://doi.org/10.1207/S15327906MBR3601

Calder, A. J., Keane, J., Lawrence, A. D., & Manes, F. (2004). Impaired recognition of anger following damage to the ventral striatum. *Brain, 127*(9), 1958-1969. http://doi.org/10.1093/brain/awh214

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience, 3*(11), 1077-1078. http://doi.org/10.1038/80586

Calder, A. J., Lawrence, A. D., & Young, A. W. (2001). Neuropsychology of Fear and Loathing. *Nature Reviews Neuroscience, 2*(5), 352–363. http://doi.org/10.1038/35072584

Calder, A. J., Young, A. W., Rowland, D., Perrett, D. I., Hodges, J. R., & Etcoff, N. L. (1996). Facial emotion recognition after bilateral amygdala damage: Differentially severe impairment of fear. *Cognitive Neuropsychology, 13*(5), 699–745. http://doi.org/10.1080/026432996381890

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29. URL http://www.jstatsoft.org/v48/i06/.

Chen, P.-H., Chang, H.-H., & Wu, H. (2012). Item selection for the development of parallel forms from an IRT-based seed test using a sampling and classification approach. *Educational and Psychological Measurement, 72*(6), 933–953. https://doi.org/10.1177/0013164412443688

Depaoli, S., & van de Schoot, R. (2015). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods, 22*(2), 240-261. https://doi.org/10.1037/met0000065

Ekman, P., & Friesen, W. V. (1974). Nonverbal behavior and psychopathology. In R. J. Friedman & M. Katz (Eds.), *The psychology of depression: Contemporary theory and research* (pp. 3-31). Washington, DC: Winston & Sons

Ekman, P. (1976) *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists.

Ekman, P., & Friesen, W. V. (1978). *Manual for the facial action coding system.* Palo Alto, CA: Consulting Psychologist Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3), 169–200. https://doi.org/10.1080/02699939208411068

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*(4), 845–857. http://doi.org/10.1037/0033-2909.85.4.845

Herzmann, G., Danthiir, V., Schacht, A., Sommer, W., Wilhelm, O., Scientific, C., … Wilhelm, O. (2008). Toward a comprehensive test battery for face cognition: Assessment of the tasks. *Behavior Research Methods, 40*(3), 840–857. https://doi.org/10.3758/BRM.40.3.840

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76*(4), 537–549. http://doi.org/10.1007/s11336-011-9218-4

Kay, T. (1984). *Individual differences in children's abilities to discriminate positive and negative affect from facial cues*. Unpublished doctoral dissertation, Emory University, Atlanta.

Krolak-Salmon, P., Hénaff, M. A., Isnard, J., Tallon-Baudry, C., Guénot, M., Vighetto, A., ... & Mauguiére, F. (2003). A specific response to disgust modulated by attention in human ventral anterior insula. *Annals of Neurology, 53*(4), 446-453. http://doi.org/10.1002/ana.10502

Lawrence, A. D., Calder, A. J., McGowan, S. W., & Grasby, P. M. (2002). Selective disruption of the recognition of facial expressions of anger. *Neuroreport, 13*(6), 881-884. http://doi.org/10.1097/00001756-200205070-0002

Limbrecht, K., Rukavina, S., Scheck, A., Walter, S., Hoffmann, H., & Traue, H. C. (2012). The influence of naturalness, attractiveness and intensity on facial emotion recognition. *Psychology Research, 2*(3), 166-176. https://doi.org/10.17265/2159-5542/2012.03.004

Lyusin, D., & Ovsyannikova, V. (2014). Measuring two aspects of emotion recognition ability: Accuracy vs. sensitivity. *Learning and Individual Differences, 52*, 129-136. https://doi.org/10.1016/j.lindif.2015.04.010

Martinez, A. M., & Benavente, R. (1998). *The AR Face Database (Tech. Rep. 24)*. Barcelona, Spain: Universitat Autònoma de Barcelona, Computer Vision Center

Matsumoto, D., LeRoux, J., Wilson‐Cohn, C., Raroque, J., Kooken, K., Ekman, P., ⋯ Goh, A. (2000). A New Test to Measure Emotion Recognition Ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior*, *24*(3), 179–209. https://doi.org/10.1023/A:1006668120583

Mayer, J.D., Salovey, P., Caruso, D.R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*(1), 97–105. https://doi.org/10.1037/1528-3542.3.1.97

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature, 383*(6603), 812-815. http://doi.org/10.1038/383812a0

Muthén, B., & Muthén, L. (1998-2012). *Mplus user's guide (7th ed.)*. Los Angeles, CA: Muthén & Muthén.

Nowicki Jr., S., & Carton, J. (1993). The Measurement of Emotional Intensity From Facial Expressions. *Journal of Social Psychology*, *133*(5), 749–750. https://doi.org/10.1080/00224545.1993.9713934

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior*, *18*(1), 9–35. http://doi.org/10.1007/BF02169077

Passarelli, M., Masini, M., Bracco, F., Petrosino, M., & Chiorri, C. (2018, January 10). FERT data file. Retrieved from https://osf.io/wd6vt/

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., ... & Gray, J. A. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389*(6650), 495-498. http://doi.org/10.1038/39051

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling*, *17*(2), 265–279. http://doi.org/10.1080/10705511003659417

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., & Revicki, D. A. (Eds.) (2015). *Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment.* New York: Routledge.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95,* 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling. Applications to typical performance assessment* (pp. 13–40). New York, NY: Routledge.

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling. Applications to typical performance assessment* (pp. 13–40). New York, NY: Routledge.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., Moore, T., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement*, *71*(4), 684–711. https://doi.org/10.1177/0013164410378690

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19–31. https://doi.org/10.1007/s11136-007-9183-7

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*(1), 5–26. https://doi.org/10.1177/0013164412449831

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

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

Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience and Biobehavioral Reviews*, *32*(4), 863-881 http://doi.org/10.1016/j.neubiorev.2008.01.001

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

Scherer, K. R., & Scherer, U. (2011). Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the emotion recognition index. *Journal of Nonverbal Behavior*, *35*(4), 305–326. https://doi.org/10.1007/s10919-011-0115-4

Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the Geneva Emotion Recognition Test: An example of Rasch-based test development. *Psychological Assessment, 26*(2), 666–672. https://doi.org/10.1037/a0035246

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*(1), 53–61. http://doi.org/10.1007/BF02289209

Sprengelmeyer, R., Young, A. W., Calder, A. J., Karnat, A., Lange, H., Hömberg, V., ... & Rowland, D. (1996). Loss of disgust. *Brain, 119*(5), 1647-1665. https://doi.org/10.1093/brain/119.5.1647

Sprengelmeyer, R., Young, A. W., Pundt, I., Sprengelmeyer, A., Calder, A. J., Berrios, G., ... & Przuntek, H. (1997). Disgust implicated in obsessive–compulsive disorder. *Proceedings of the Royal Society of London B: Biological Sciences, 264*(1389), 1767-1773. https://doi.org/10.1098/rspb.1997.0245

Sprengelmeyer, R., Young, A. W., Schroeder, U., Grossenbacher, P. G., Federlein, J., Buttner, T., & Przuntek, H. (1999). Knowing no fear. *Proceedings of the Royal Society of London B: Biological Sciences, 266*(1437), 2451-2456. http://doi.org/10.1098/rspb.1999.0945

Stan Development Team (2015). *Stan Modeling Language User's Guide and Reference Manual, Version 2.10.0*. URL http://mc-stan.org/.

Takane, Y., & Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408. https://doi.org/10.1007/

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., … Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research, 168*(3), 242–9. https://doi.org/10.1016/j.psychres.2008.05.006

Warwick, J., Nettelbeck, T., & Ward, L. (2010). AEIM: A new measure and method of scoring abilities-based emotional intelligence. *Personality and Individual Differences, 48*(1), 66–71. https://doi.org//10.1016/j.paid.2009.08.018

Weitz, S. (1974). *Nonverbal communication: Readings with commentary*. New York, NY: Oxford University Press.

Footnotes

1. Note that at the time we performed this study the Amazon Mechanical Turk was not available for non-US residents and the recently developed Prolific was not yet online.

2. Data collection is ongoing, and the test can be accessed at http://130.251.230.192/limesdf_new/index.php/883421/lang-en in Italian, English, French, Russian, Polish, Spanish, Dutch, Brazilian Portuguese, Turkish, Romanian, Swedish, and Norwegian (Bokmål and Nynorsk).

Figure Captions

Figure 1: Example of an emotional portrayal (disgust) included in the test

Figure 2: Bayesian 2PL model specification

Figure 3: Test information function for the final (36-items) FERT

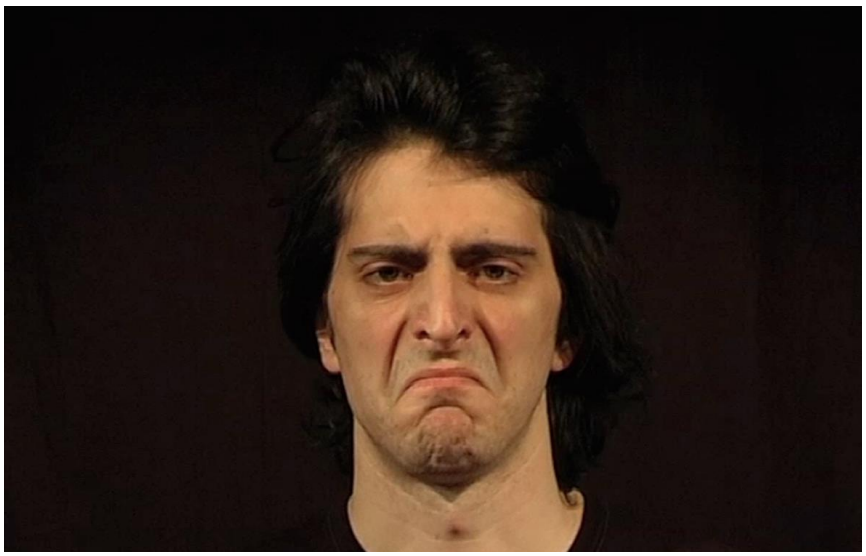

Figure 1

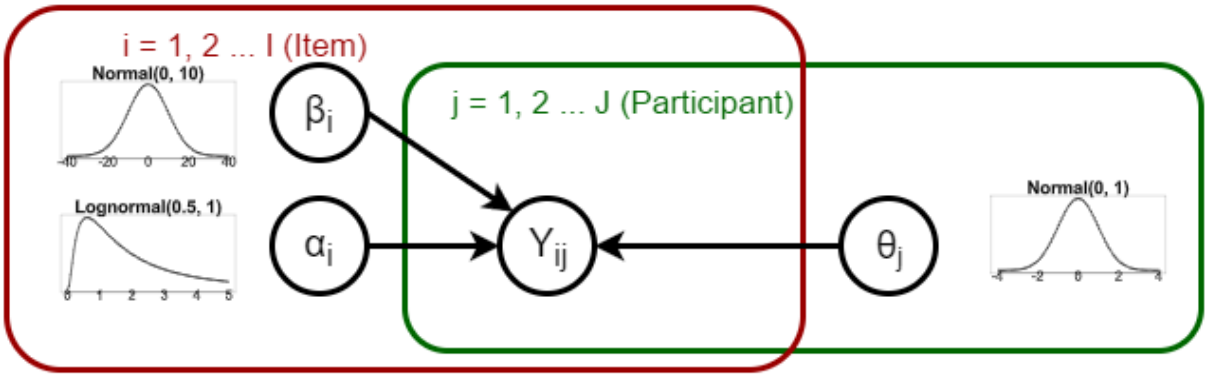
i = 1, 2 ... I (Item)
Normal(0, 10)
-40 -20 0 20 40
Lognormal(0.5, 1)
0 1 2 3 4 5
βi
αi
j = 1, 2 ... J (Participant)
Yij
θj
Normal(0, 1)
-4 -2 0 2 4

Figure 2

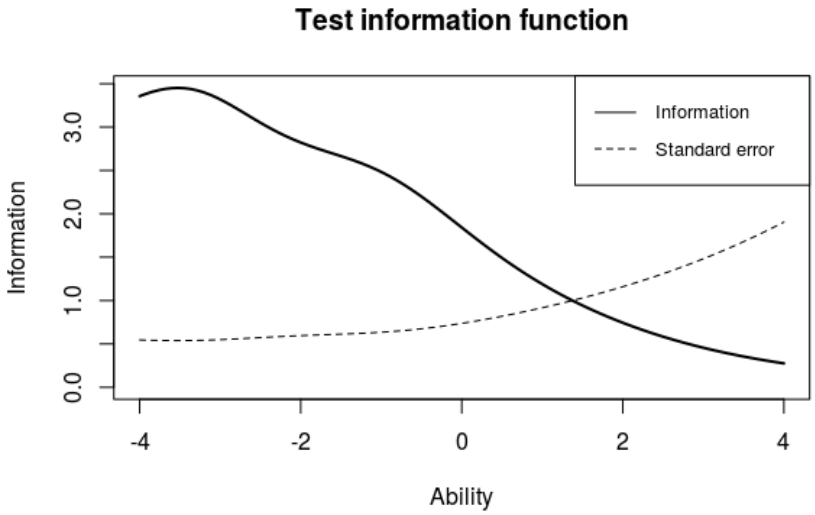
Test information function
Information
Standard error
Information
0.0
1.0
2.0
3.0
-4
-2
0
2
4
Ability

Figure 3

Table 1
*Main Features of Facial Expression Recognition Ability Tests*

| Test | Reference | Stimuli database | Professional actors | Scoring system | Stimuli evaluation |
|---|---|---|---|---|---|
| Videotest of Emotion Recognition | Lyusin & Ovsyannikova (2015) | Original | No | 2 measures of deviation from expert judgements | Experts |
| GERT | Schlegel, Grandjean, & Scherer (2014) | GEMEP – Bänziger, Mortillaro, & Scherer (2011) | Yes | 1PL ability score | Model fit |
| ERI (FACIAL-INDEX subtest) | Scherer & Scherer (2011) | POFA (Ekman & Friesen, 1976) | FACS-trained | Proportion of correct responses | Consensus |
| AEIM (perception subscale)* | Warwick, Nettelbeck, & Ward (2010) | FACS Manual (Ekman & Friesen, 1978) | FACS-trained | Deviation experts, self-reported confidence | Consensus |
| MERT | Bänziger, Grandjean, & Scherer (2009) | GEMEP – Bänziger, Mortillaro, & Scherer (2011) | Yes | Proportion of correct responses | Experts |
| Battery of face cognition measures (subtest – Facially Expressed Emotion Labeling) | Herzmann et al. (2008) | AR Face Database – Martinez & Benavente (1999); NimStim Face Stimulus Set – Tottenham et al. (2009) | No (AR); Yes (NimStim) | Proportion of correct responses | Unknown |
| MSCEIT 2.0 (faces task) | Mayer, Salovey, Caruso, & Sitarenios (2003) | Unknown (likely to be original) | Unknown | Deviation from consensus and experts | Experts and consensus |
| JACBART* | Matsumoto et al. (2000) | JACFEE and JACNeuF – Biehl et al. (1997) | FACS-trained | Proportion of correct responses | Experts and consensus |
| DANVA (Receptive Facial Expression Subtest)* | Nowicki & Duke (1994) | Children's Affect Test – Kay (1984); BART – Ekman & Friesen (1974) | FACS-trained | Proportion of correct responses | None |
| DANVA FACES 2 | Nowicki & Carton (1993) | Original | Unknown | Proportion of correct responses | Consensus |
| PONS | Rosenthal, Hall, DiMatteo, Rogers, & Archer (1979) | Original | No | Proportion of correct responses | Experts |

| Test | Item | Response | Stimuli | Expression elicitation method | Administration time |
|---|---|---|---|---|---|
| Videotest of Emotion Recognition | 7 | Dimensional | Video recordings | Spontaneous | Unknown (likely to be brief) |
| GERT | 83 | Categorical | Video and audio recordings | Scenarios | ~ 30’ |
| ERI (FACIAL-INDEX subtest) | 30 | Dimensional | Still pictures | FACS-based | < 20’ |
| AEIM (perception subscale)* | 20 | Categorical | Still pictures | FACS-based | Unknown (likely to be brief) |
| MERT | 120 items from 30 portrayals (90 with faces) | Categorical | Video and audio recordings, still pictures | Scenarios | ~ 45’ |
| Battery of face cognition measures (subtest – Facially Expressed Emotion Labeling) | 30 | Categorical | Still pictures | Unknown | ~ 3’ |
| MSCEIT 2.0 (faces task) | 20 | Categorical | Still pictures | Unknown | Unknown (likely to be brief) |

| | | | | | |
|---|---|---|---|---|---|
| JACBART* | 56 | Categorical | Still pictures | FACS-based | Unknown (likely to be brief) |
| DANVA (Receptive Facial Expression Subtest)* | 20 | Dimensional | Still pictures | FACS-based | ~ 2’ |
| DANVA FACES 2 | 96 | Dimensional | Still pictures | Scenarios | Unknown (likely to be brief) |
| PONS | 220 (120 with faces) | Categorical | Video and audio recordings | Free interpretation | > 47' |

| **Test** | **Actor attractiveness evaluation** | **Sample size** | **Color / greyscale** | **Actor number and gender** | **Emotions** |
|---|---|---|---|---|---|
| Videotest of Emotion Recognition | No | 684 | Unknown | Unknown (F and M) | Anger, Anxiety, Arousal, Calmness, Contempt, Disgust, Displeasure, Fear, Guilt, Happiness, Interest, Relaxation, Shame, Suffering, Surprise |
| GERT | No | 295 (82 males) | Unknown | 5F + 5M | Amusement, Anger, Anxiety, Despair, Disgust, Fear, Interest, Irritation, Joy, Pleasure, Pride, Relief, Sadness, Surprise |
| ERI (FACIAL-INDEX subtest) | No | 4755 | Greyscale | Unknown | Amusement, Anger, Anxiety, Despair, Disgust, Fear, Interest, Irritation, Joy, Pleasure, Pride, Relief, Sadness, Surprise |
| AEIM (perception subscale)* | No | 272 (psychology students) | Greyscale | 2F + 2M | Anger, Happiness, Sadness, Surprise |
| MERT | No | 62 (psychology students) | Greyscale | 5F + 5M | Anxiety, Boredom, Cold anger, Contempt, Disgust, Despair, Elation, Fear, Happiness, Hot anger, Interest, Panic, Pride, Sadness, Shame |
| Battery of face cognition measures (subtest – Facially Expressed Emotion Labeling) | No | 153 | Greyscale | Unknown | Anger, Disgust, Fear, Happiness, Sadness, Surprise |
| MSCEIT 2.0 (faces task) | No | 2112 | Unknown | Unknown | Unknown |
| JACBART* | No | 579 (multiple psychology students samples) | Greyscale | 4F + 4M | Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise |
| DANVA (Receptive Facial Expression Subtest)* | No | >2300 children (multiple samples) | Greyscale | Unknown | Anger, Fear, Happiness, Sadness |
| DANVA FACES 2 | No | Unclear (multiple samples of students and children ~ 500) | Greyscale | Unknown | Anger, Fear, Happiness, Sadness |
| PONS | No | 2615 | Greyscale | 1F | No direct measure of emotion |

Table 2
*Item Parameters for the 36-Items Model*

| Item Code | Mean (α) | SD (α) | Mean (β) | SD (β) |
|---|---|---|---|---|
| ARANG2 | 0.47 | 0.23 | -3.33 | 2.18 |
| ARDIS3 | 0.30 | 0.11 | -3.93 | 1.74 |
| ARFEA2 | 0.72 | 0.33 | -3.83 | 1.95 |
| ARHAP2 | 0.88 | 0.40 | -4.10 | 2.13 |
| ARSAD3 | 0.55 | 0.26 | -1.47 | 1.02 |
| ARSUR2 | 0.61 | 0.18 | -3.73 | 1.24 |
| EBANG1 | 0.61 | 0.25 | -1.20 | 0.69 |
| EBDIS2 | 0.21 | 0.12 | 0.38 | 1.10 |
| EBFEA3 | 0.78 | 0.38 | 0.39 | 0.51 |
| EBHAP3 | 1.43 | 1.13 | -4.75 | 2.87 |
| EBSAD3 | 0.55 | 0.17 | -4.77 | 1.62 |
| EBSUR3 | 0.45 | 0.16 | -6.29 | 2.47 |
| FFANG3 | 1.70 | 1.38 | -3.35 | 1.74 |
| FFDIS2 | 0.48 | 0.26 | -6.43 | 3.33 |
| FFFEA2 | 0.59 | 0.29 | 0.49 | 0.43 |
| FFHAP3 | 0.88 | 0.54 | -5.83 | 3.10 |
| FFSAD1 | 0.55 | 0.23 | -1.71 | 0.94 |
| FFSUR2 | 0.89 | 0.48 | -3.86 | 2.10 |
| FGANG2 | 0.83 | 0.37 | -2.87 | 1.39 |
| FGDIS2 | 0.88 | 0.47 | -4.39 | 2.40 |
| FGFEA2 | 0.28 | 0.11 | 0.11 | 0.34 |
| FGHAP3 | 0.63 | 0.32 | -6.20 | 3.23 |
| FGSAD1 | 0.61 | 0.17 | -3.46 | 1.15 |
| FGSUR3 | 0.46 | 0.23 | -3.05 | 1.84 |
| LDANG3 | 0.78 | 0.19 | -2.80 | 0.71 |
| LDDIS3 | 0.53 | 0.27 | -4.69 | 2.73 |
| LDFEA3 | 0.85 | 0.33 | -0.36 | 0.32 |
| LDHAP2 | 0.63 | 0.30 | -3.41 | 1.95 |
| LDSAD3 | 0.49 | 0.22 | 0.29 | 0.44 |
| LDSUR2 | 0.60 | 0.25 | -3.70 | 1.81 |
| MGANG1 | 1.41 | 0.60 | -0.93 | 0.32 |
| MGDIS2 | 0.52 | 0.16 | -3.58 | 1.28 |
| MGFEA3 | 1.10 | 0.38 | -1.48 | 0.48 |
| MGHAP3 | 0.73 | 0.25 | -5.41 | 2.00 |
| MGSAD2 | 0.60 | 0.28 | 0.20 | 0.40 |
| MGSUR1 | 0.31 | 0.12 | -6.54 | 2.73 |

Note: SD = standard deviation; α = item discrimination parameter; β = item difficulty parameter. The item code comprises the initials of the actor (first and second letter), the emotion displayed (third, fourth, and fifth letter; ANG = anger; DIS = disgust, FEA = fear; HAP = Happiness; SAD = sadness; SUR = surprise), and the frame number (see text).

Table 3

*Marginal Means for Item Parameters According to Actor Gender and Emotion Portrayed. Numbers in Brackets Indicate 95% Confidence Intervals.*

| Actor gender | Emotion portrayed | Discrimination | Difficulty |
|---|---|---|---|
| Male | Anger | 0.92 [0.60, 1.23] | -3.03 [-5.13, -0.93] |
| Male | Disgust | 0.40 [0.09, 0.71] | -5.33 [-7.43, -3.23] |
| Male | Fear | 0.65 [0.34, 0.96] | -1.36 [-3.46, 0.74] |
| Male | Happiness | 0.75 [0.43, 1.06] | -4.54 [-6.64, -2.43] |
| Male | Sadness | 0.45 [0.14, 0.76] | 0.68 [-1.42, 2.79] |
| Male | Surprise | 0.67 [0.36, 0.98] | -3.96 [-6.07, -1.86] |
| Female | Anger | 0.93 [0.62, 1.25] | -1.62 [-3.73, 0.48] |
| Female | Disgust | 0.61 [0.30, 0.93] | -1.53 [-3.63, 0.57] |
| Female | Fear | 0.80 [0.48, 1.11] | -0.22 [-2.32, 1.89] |
| Female | Happiness | 0.63 [0.41, 1.04] | -6.30 [-8.40, -4.20] |
| Female | Sadness | 0.56 [0.24, 0.87] | -2.48 [-4.59, -0.38] |
| Female | Surprise | 0.38 [0.06, 0.69] | -5.51 [-7.61, -3.40] |

View publication stats