Psychological Assessment

Test–Retest Reliability of the Facial Expression Labeling Task

Jennifer L. Cecilione, Lance M. Rappaport, Brad Verhulst, Dever M. Carney, R. J. R. Blair, Melissa A. Brotman, Ellen Leibenluft, Daniel S. Pine, Roxann Roberson-Nay, and John M. Hettema

Online First Publication, February 23, 2017. http://dx.doi.org/10.1037/pas0000439

Psychological Assessment

1040-3590/17/$12.00 http://dx.doi.org/10.1037/pas0000439

BRIEF REPORT

Test–Retest Reliability of the Facial Expression Labeling Task

Jennifer L. Cecilione, Lance M. Rappaport, Brad Verhulst, and Dever M. Carney
Virginia Commonwealth University

R. J. R. Blair, Melissa A. Brotman, Ellen Leibenluft, and Daniel S. Pine
National Institutes of Health, Bethesda, Maryland

Roxann Roberson-Nay and John M. Hettema
Virginia Commonwealth University

Recognizing others' emotional expressions is vital for socioemotional development; impairments in this ability occur in several psychiatric disorders. Further study is needed to map the development of this ability and to evaluate its components as potential transdiagnostic endophenotypes. Before doing so, however, research is required to substantiate the test–retest reliability of scores of the face emotion identification tasks linked to developmental psychopathology. The current study estimated test–retest reliability of scores of one such task, the facial expression labeling task (FELT) among a sample of twin children ($N = 157$; ages 9–14). Participants completed the FELT at two visits two to five weeks apart. Participants discerned the emotion presented of faces depicting six emotions (i.e., happiness, anger, sadness, fear, surprise, and disgust) morphed with a neutral face to provide 10 levels of increasing emotional expressivity. The present study found strong test–retest reliability (Pearson r) of the FELT scores across all emotions. Results suggested that data from this task may be effectively analyzed using a latent growth curve model to estimate overall ability (i.e., intercept; r's = 0.76—0.85) and improvement as emotions become clearer (i.e., linear slope; r's = 0.69—0.83). Evidence of high test–retest reliability of this task's scores informs future developmental research and the potential identification of transdiagnostic endophenotypes for child psychopathology.

Public Significance Statement
This study demonstrated strong test–retest reliability of the facial emotion labeling task scores in a sample of child twins aged 9 to 14. Evidence from this study of strong test–retest reliability supports use of this task in future research on the psychobiological etiology and longitudinal development of face emotion identification ability.

Keywords: children, face emotion identification, facial expression labeling task, test–retest reliability

Supplemental materials: http://dx.doi.org/10.1037/pas0000439.supp

The ability to identify facial expressions is crucial for healthy socioemotional development (Marsh & Blair, 2008; Trentacosta & Fine, 2010). Deficits in face emotion identification have been associated with several psychiatric disorders (Brotman et al., 2008; Shaw, Stringaris, Nigg, & Leibenluft, 2016; Blair, Colledge, Murray, & Mitchell, 2001; Trentacosta & Fine, 2010). Evidence that face emotion identification deficits are implicated in a myriad of psychiatric disorders suggests that components of face emotion

Jennifer L. Cecilione, Lance M. Rappaport, Brad Verhulst, and Dever M. Carney, Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University; R. J. R. Blair, Section on Affective Cognitive Neuroscience, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland; Melissa A. Brotman, Ellen Leibenluft, and Daniel S. Pine, Emotion and Development Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland; Roxann Roberson-Nay and John M. Hettema, Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University.

The material contained in this article has not been published nor is it under consideration elsewhere. There are no previous publications based on this data. The authors do not have any financial interests that might influence this research. This study was supported by the National Institutes of Health (R01MH098055 to J.M.H, NIMH- T32MH020030 to L.M.R., and NIMH-IRP-ziamh002781 to D.S.P.).

Correspondence concerning this article should be addressed to Jennifer L. Cecilione, Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, P.O. Box 980489, Richmond, VA 23298. E-mail: jennifer.cecilione@vcuhealth.org

identification may be transdiagnostic endophenotypes of psychiatric illnesses.

Research regarding face emotion identification deficits in internalizing disorders has resulted in mixed findings (e.g., Easter et al., 2005, but see Guyer et al., 2007). This may be because different paradigms are used to assess face emotion identification ability, which hinders the ability to reconcile discordant findings in this field (Bourke et al., 2010). Some of the various face emotion identification tasks include the Child and Adult Facial Expressions subtests of the Diagnostic Analysis of Nonverbal Accuracy (Nowicki & Duke, 1994), the Penn emotion recognition test (Kohler et al., 2003), and the Bell Lysaker emotion recognition task (Pinkham, Penn, Green, & Harvey, 2016). Other studies use variations of the facial expression labeling task (FELT), such as the dynamic emotion identification task (Kirsh & Mounts, 2007) and the emotional expression multimorph task (Blair et al., 2001).

Because task heterogeneity creates a methodological confound in aggregating results to draw reliable conclusions about face emotion identification ability, the current study sought to evaluate a promising version of the task by determining the test–retest reliability of FELT scores; this task has been used in several previous studies in varying forms (e.g., Averbeck, Bobin, Evans, & Shergill, 2012; Kirkpatrick, Lee, Wardle, Jacob, & de Wit, 2014; Marsh, Yu, Pine, & Blair, 2010). The present study's version of the FELT may be a more sensitive and preferred measure of face emotion identification ability than prior variations for several reasons.

First, this version uses all six of Ekman's exemplar emotions, which is particularly important because prior research suggests that certain psychiatric disorders (e.g., major depressive disorder) may be associated with difficulty or a predilection to identifying particular emotions (e.g., sadness; Bourke et al., 2010). Including fewer emotions (e.g., Nowicki & Duke, 1994) may preclude establishing emotion-specific deficits. Second, this task presents each emotion at 10% to 100% emotional expressivity in 10% increments (Harmer, Rogers, Tunbridge, Cowen, & Goodwin, 2003; Marsh et al., 2010); this may provide a more realistic, nuanced display of facial expressions than other tasks (e.g., Nowicki & Duke, 1994). Third, in the present task, participants select a response choice after seeing each of the 360 static pictures of faces; this is preferred to previous tasks wherein only one data point is collected when participants press a key to identify an emotion in a continuously morphing image (e.g., Kirsh & Mounts, 2007). The current paradigm provides information on face emotion identification ability at each level of emotional expressivity. Finally, in the present task, trials are presented in random order. When trials are presented in sequential order (e.g., Blair et al., 2001), responses at various levels of expressivity become nonindependent, such that a participant's response at one level may inform their response at the next.

Moreover, the analytic strategy used to assess accuracy often varies over studies of face emotion identification. One approach is to compute an individual's raw accuracy, which conflates an individual's ability to correctly identify an emotion with his or her tendency to endorse the emotion. The current study utilized the "unbiased hit rate." This approach adjusts an individual's *raw accuracy* (i.e., proportion of trials correctly identified) for an individual's *differential accuracy* (i.e., proportion of correct uses of an emotion). The resulting score was adjusted for guessing and non-normality (Marsh et al., 2010; Wagner, 1993). Additionally, in all variations of face emotion identification tasks, levels of emotional expressivity are nested within individual and present a gradient over which one can estimate an intercept and rate of change as the expressed emotion becomes clearer (i.e., linear slope). At a minimum, the analysis of data from this and related tasks should account for the nonindependence of trials within person. However, existing research has not examined the test–retest reliability of face emotion identification tasks' scores through analytic approaches that account for the structure of this data (e.g., multilevel/hierarchical modeling or latent growth curve modeling).

The present study provides the first examination of test–retest reliability of change in scores over increasing expressivity of emotional expression in preadolescent children. Prior research has reported adequate test–retest reliability of FELT scores among an adult community sample (Adams et al., 2016). Because the test–retest reliability of scores of any face emotion identification task using a child sample has not yet been established, the current study's approach is novel in that it assesses test–retest reliability of FELT scores in a genetic epidemiological sample of preadolescent children. It is important to study face emotion identification ability in this demographic, as preadolescent children are learning, developing, and solidifying various psychological characteristics that will affect their developmental trajectories into adolescence and adulthood (Steinberg & Morris, 2001). Given that deficits in face emotion identification ability are correlated with a variety of psychopathology, intervening at this early stage in development is critical. Similar to previous research (Adams et al., 2016), the present study hypothesized that test–retest reliability would be high. Because findings from research using adult participants (i.e., Adams et al., 2016) cannot be assumed to apply to children, demonstrating high test–retest reliability in the current study is necessary to inform future developmental research on face emotion identification. In particular, high test–retest reliability of scores on a face emotion identification task is essential for future research to examine face emotion identification deficits in childhood as a marker for abnormal development and psychopathology. Additionally, this study used an analytic approach that is more consistent with the nature of the task (i.e., a latent growth curve model), in which performance can be examined as a function of decreasing task difficulty. This analytic approach permitted examining reliability of level, which is similar to research described by Adams et al. (2016), and change as difficulty decreases. In this manner, the present study sought to extend research by Adams et al. (2016) to demonstrate test–retest reliability in a child sample and sought to examine the test–retest reliability of change during the task.

Method

Participants and Procedure

Participants ($N = 796$ individuals; 52.4% female) were recruited as twin pairs from the Mid-Atlantic Twin Registry (Lilley & Silberg, 2013) to participate in a study on juvenile anxiety risk. This genetic epidemiological sample was comprised of Caucasian twin children (aged 9 to 14, $M = 10.78$ years) who lived in the mid-Atlantic region. This study's protocol was approved by Vir-

ginia Commonwealth University's institutional review board. Consent was obtained from participants' legal guardian and assent was obtained from participants. Participants attended two laboratory-based sessions that were two to five weeks apart. At visit 1 (V1), 393 twin pairs (786 individuals) completed the study's full protocol. Of those who participated in visit 2 (V2), 157 individuals were assigned to a randomization order that included the FELT (i.e., a planned missingness design). The study's full protocol is detailed elsewhere (Carney et al., 2016).

Measures

To measure zygosity, a parent or legal guardian of participants completed questionnaires about twins that included questions regarding physical similarities between them; these items have previously been used to determine the zygosity of twin participants, and the interpretations of the questionnaire's results have shown good validity compared to blood testing (Kasriel & Eaves, 1976) and DNA evaluation of zygosity (Jackson, Snieder, Davis, & Treiber, 2001).

The task used herein is most similar to that described by Marsh et al. (2010). Participants sat with an experimenter in front of a computer and were given these instructions: "In this task, you will be shown pictures of people with various emotional expressions. After viewing the faces, please choose the correct emotion from the choices provided on the next screen." Participants then completed one practice trial during which they were shown a picture of a man making a happy face and asked to identify the emotion from the list of Ekman's six emotions.

Pictures were created by morphing a static photo of a person expressing the target emotion with a photo of the same person making a neutral expression. The lowest level of emotional expressivity was 10%, and the highest level was 100%; each emotion had 10 levels of emotional expressivity. Emotions reflected Ekman's six basic emotions (i.e., happiness, sadness, disgust, fear, anger and surprise; Ekman & Friesen, 1976). All faces were Caucasian adults (50% female). Participants were presented with six trials of each emotion at each expressivity level for a total of 360 trials (six trials by six emotions by 10 expressivity levels). Trials were presented in random order and consisted of a fixation cross for 250 ms, then the target face for 500 ms. Following each trial, participants were asked to choose (at their own pace) the emotion displayed from a list of the six possible emotions.

Data Analysis

Raw data were collected using E-Prime 2.0 software, scored in E-Merge software (Schneider, Eschman, & Zuccolotto, 2002) and exported to R for analysis. An unbiased hit rate was computed at each expressivity level for each emotion as the product of *raw accuracy* (i.e., number of trials correct/number of trials of an emotion) and *differential accuracy* (i.e., number of trials correct/number of times an emotion was endorsed; Marsh et al., 2010; Wagner, 1993). To account for accuracy attributable to guessing, the result was adjusted by subtracting 1/6 and then transformed using an arcsin transformation to improve normality over participants. Experimenters noted that 33 participants from V1 and 17 participants from V2 stopped engaging with the task (e.g., rapidly and repeatedly pressed the same number) before its completion. Data from these cases were removed due to concerns about data quality.

Test–retest reliability of FELT scores (i.e., unbiased hit rates) was assessed in two approaches. First, test–retest reliability was assessed as the correlation between unbiased hit rates at V1 and V2, calculated for each emotion at each expressivity level. Nonindependence of twins due to familial aggregation was accounted for by estimating correlations within the biometrical model; a multivariate biometrical model was fit using structural equation modeling for V1 and V2 nested within person, nested within family. The correlation between V1 and V2 was taken from the resulting estimated variance-covariance matrix. This model provided an appropriate adjustment for nonindependence due to nesting twins within families and an adjustment for greater nonindependence among monozygotic twins.

Clinical and neurological correlates of the FELT may be clarified by analyzing data using a latent growth curve model or similar multilevel modeling to estimate the rate of change in accuracy as emotional expression becomes clearer and to account for the nesting of levels of emotional expressivity within person. Therefore, test–retest reliability of FELT scores (i.e., unbiased hit rates) was also assessed using a latent growth curve model to estimate the intercept, linear slope, and quadratic slope over difficulty, which describes a participant's improvement in detecting an emotion with increasing emotional expressivity. Test–retest reliability of FELT scores was computed using structural equation modeling to fit a latent growth curve to the 10 expressivity levels for each person at each visit. Six separate structural equation models were fit to estimate the test–retest reliability of participants' accuracy at detecting each of the six emotions. Similar to estimating the correlation between V1 and V2 for each emotion at each expressivity level, test–retest reliability for intercept, linear slope, and quadratic slope over difficulty were estimated within a biometrical model to account for nonindependence due to nesting of twins within family and greater nonindependence among monozygotic twins. A latent growth curve was estimated for each visit such that four latent growth curves were fit for each model (two per twin by two twins). Because the factor loadings in a latent growth curve model can influence the correlations between the latent parameters, orthogonal loadings were used to minimize the impact of nonessential multicollinearity on the reliability estimates (Cohen, Cohen, West, & Aiken, 2003). Orthogonal loadings were centered so that the intercept for each model was centered at 55%. The correlation between V1 and V2 for intercept, linear slope, and quadratic slope were extracted from the resulting estimated variance-covariance matrix.

Results and Discussion

Results revealed strong test–retest reliability of FELT scores as a measure of face emotion identification ability among juveniles. Reliability was higher for emotion presentations at higher levels of expressivity (see Table 1). At 100% expressivity of emotion, V1 and V2 unbiased hit rate scores were correlated for each emotion: anger ($r = .409$), fear ($r = .432$), happiness ($r = .542$), sadness ($r = .467$), disgust ($r = .462$), and surprise ($r = .390$). Reliability of task performance scores (i.e., unbiased hit rates) improved as emotions became easier to recognize and approached an asymptote at approximately 60% expressivity (see Figure 1). That test–retest

Table 1
Test–Retest Reliability of Unbiased Hit Rates at Each Level of Emotional Expressivity

	Anger		Fear		Happiness	
Expressivity	Estimate	CI	Estimate	CI	Estimate	CI
10	−.082	(−.248, .088)	.146	(−.058, .338)	−.016	(−.187, .157)
20	.236	(.072, .386)	−.101	(−.295, .102)	−.009	(−.177, .159)
30	.441	(.293, .567)	.085	(−.102, .264)	.156	(−.009, .313)
40	.069	(−.096, .230)	.264	(.098, .416)	.429	(.279, .560)
50	.275	(.113, .423)	.263	(−.728, .88)	.431	(.284, .561)
60	.411	(.263, .541)	.240	(−.560, .928)	.516	(.380, .631)
70	.475	(.333, .598)	.465	(.318, .592)	.581	(.457, .685)
80	.315	(.153, .464)	.374	(.210, .520)	.587	(.461, .691)
90	.508	(.375, .621)	.388	(.236, .524)	.546	(.412, .659)
100	.409	(.259, .541)	.432	(.282, .563)	.542	(.408, .655)
	Sadness		Disgust		Surprise	
Expressivity	Estimate	CI	Estimate	CI	Estimate	CI
10	.152	(−.017, .313)	.320	(.156, .464)	−.116	(−.332, .114)
20	.040	(−.131, .209)	.120	(−.049, .283)	.230	(.020, .417)
30	.020	(−.153, .193)	.072	(−.100, .241)	.022	(−.170, .214)
40	.107	(−.068, .278)	.164	(−.010, .327)	.178	(.006, .338)
50	.358	(.209, .495)	.499	(.361, .615)	.284	(.123, .429)
60	.397	(.244, .532)	.574	(.447, .680)	.440	(.291, .568)
70	.352	(.194, .493)	.600	(.480, .698)	.428	(.280, .555)
80	.442	(.293, .572)	.555	(.425, .662)	.388	(.235, .525)
90	.414	(.261, .548)	.555	(.428, .662)	.364	(.209, .501)
100	.467	(.323, .590)	.462	(.314, .588)	.390	(.237, .525)

Note. CI refers to a 95% confidence interval.

reliability approaches an asymptote is further supported by overlapping confidence intervals, which are consistent with no improvement in test–retest reliability of scores for each emotion from 60% to 100% expressivity (see Table 1). Reliability of scores is similar in the detection of all six emotions, with estimates of reliability for each falling within the confidence intervals for most scores at other emotions at 100% expressivity.

Findings demonstrate high test–retest reliability for FELT scores when data are analyzed using latent growth curve modeling that aggregates over levels of expressivity to estimate performance as both an intercept and change within person. Intercept scores represent participant median performance for each emotion; the correlation between intercepts represents aggregated scores for performance on each emotion (as opposed to considering scores at each expressivity level independently). In contrast to a participant's mean performance, the linear slope indexes the extent to which a participant's performance improved as the presentation of emotion became clearer. The correlations between intercepts at V1 and V2 (anger: $r = .811$, fear: $r = .796$, happiness: $r = .831$, sadness: $r = .763$, disgust: $r = .847$, surprise: $r = .789$) as well as correlations between linear slopes at V1 and V2 (anger: $r = .759$, fear: $r = .688$, happiness: $r = .765$, sadness: $r = .734$, disgust: $r = .831$, surprise: $r = .692$) are high and similar for all emotions (see Table 2). Notably, estimated reliability is substantially higher for unbiased hits rates aggregated over expressivity levels, lending support for this analytic approach in future research. There were no consistent differences in reliability of scores based on gender (see supplemental Table 1) or age (see supplemental Table 2). The estimated growth parameters (i.e., intercept, linear slope, and quadratic slope) for V1 and V2 were examined to evaluate potential practice effects. There is evidence that performance on the task followed a nonlinear trajectory (see supplemental Table 3). There was little interindividual variance in the quadratic slope, which precludes estimating test–retest reliability. Although there is some improvement of performance at V2 (see supplemental Table 3), consistent improvement in performance may reduce estimated correlations describing the test–retest reliability of performance on the FELT.

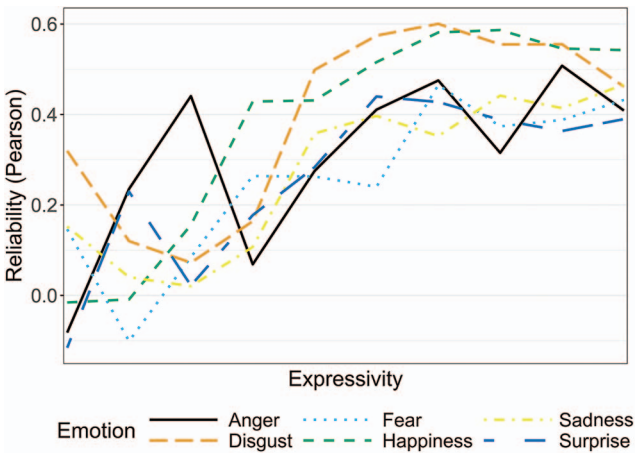

Figure 1. Test–retest reliability for unbiased hit rates of each emotion at each level of expressivity. See the online article for the color version of this figure.

Table 2
Reliability of Intercept, Linear Slope, and Quadratic Slope Over Difficulty

	Intercept		Linear slope		Quadratic slope	
Emotion	Estimate	CI	Estimate	CI	Estimate	CI
Anger	.811	(.661, .829)	.759	(.535, .793)	−.068	(−.430, .627)
Fear	.796	(.652, .897)	.688	(.429, .695)	.652	(−.050, .651)
Happiness	.831	(.741, .899)	.765	(.629, .860)	.711	(.155, .873)
Sadness	.763	(.619, .864)	.734	(.578, .892)	.136	(−.613, .267)
Disgust	.847	(.715, .934)	.831	(.679, .927)	.686	(−.044, .887)
Surprise	.789	(.606, .904)	.692	(.431, .692)	.874	(−.133, .874)

Note. CI refers to a 95% confidence interval.

While this study demonstrates high test–retest reliability of FELT scores, some limitations should be mentioned. First, the FELT requires participants to attend for the entire duration of task (~ 25 min), and lack of attention was evident for some participants in this study. It is unclear whether a shorter task with fewer trials would provide an acceptable level of signal-to-noise. Second, prior studies have found effects of IQ on participants' ability to identify emotional facial expressions (Lawrence et al., 2015); however, those effects were not considered in these reliability estimates. Finally, this sample is limited to Caucasian families because of its primary genetic aims, so these findings might not be generalizable to other races.

Despite these limitations, evidence of high test–retest reliability of FELT scores supports the use of this task in longitudinal research on children's socioemotional developmental trajectories; test–retest reliability of a task's scores is necessary to use that task to study change in a process over time (Khoo, West, Wu, & Kwok, 2006). This task and the analytic methods of the current study may also be used in future research on the development and clinical correlates of face emotion identification in children. Test–retest reliability of performance on the FELT is needed to clarify the association of face emotion identification deficits with psychopathology. Reliability is important for generalizability and to determine that the variance in results is not attributable to noise, but to variance caused by psychopathology. This task may also be used in future research on emotion regulation. Since recruiting assistance from others in coping with distress and providing assistance to others in need are key components of interpersonal emotion regulation (Hofmann, 2014), face emotion identification seems to play a crucial role in this process as well.

Moreover, because deficits in face emotion identification ability are associated with a variety of psychiatric disorders, deficits in the detection of some emotions may represent a transdiagnostic endophenotype (i.e., a broad, intermediate risk factor for psychiatric illness). This might be possible given that some emotions activate a common group of brain regions while others are region-specific (Fusar-Poli et al., 2009). Endophenotypes may be less genetically complex and thus provide a link between genetic code and symptoms indicative of psychiatric disorders (Cannon & Keller, 2006; Insel et al., 2010). Because one of the criteria of an endophenotype is that it is heritable (Gottesman & Gould, 2003), and test–retest reliability is needed to assess heritability based on genetic epidemiological research (Cannon & Keller, 2006), this study's findings recommend this variation of the FELT for future research on the heritability of face emotion identification; this future work will utilize twin models. This research, using the FELT and corresponding analyses demonstrated in the current study, will be critical to evaluate whether components of face emotion identification are transdiagnostic endophenotypes.

The findings of the current study establish the reliability of the FELT to assess children's face emotion identification ability, which is necessary to support the use of this task in future studies of children's face emotion identification ability. Additionally, the analytic approach utilized in the current study (i.e., latent growth curve analysis using unbiased hit rates) is preferred to other methods (e.g., bivariate correlations using hit rates), as the analytic strategies of this study take into account change in accuracy as the emotion becomes clearer and adjust for accuracy attributable to guessing and non-normality when determining participants' overall ability to identify emotions.

References

Adams, T., Pounder, Z., Preston, S., Hanson, A., Gallagher, P., Harmer, C. J., & McAllister-Williams, R. H. (2016). Test–retest reliability and task order effects of emotional cognitive tests in healthy subjects. *Cognition and Emotion, 30,* 1247–1259.

Averbeck, B. B., Bobin, T., Evans, S., & Shergill, S. S. (2012). Emotion recognition and oxytocin in patients with schizophrenia. *Psychological Medicine, 42,* 259–266. http://dx.doi.org/10.1017/S0033291711001413

Blair, R. J. R., Colledge, E., Murray, L., & Mitchell, D. G. V. (2001). A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies. *Journal of Abnormal Child Psychology, 29,* 491–498. http://dx.doi.org/10.1023/A:1012225108281

Bourke, C., Douglas, K., & Porter, R. (2010). Processing of facial emotion expression in major depression: A review. *Australian and New Zealand Journal of Psychiatry, 44,* 681–696. http://dx.doi.org/10.3109/00048674.2010.496359

Brotman, M. A., Guyer, A. E., Lawson, E. S., Horsey, S. E., Rich, B. A., . . . Leibenluft, E. (2008). Facial emotion labeling deficits in children and adolescents at risk for bipolar disorder. *American Journal of Psychiatry, 165,* 385–389. Retrieved from http://ajp.psychiatryonline.org/doi/pdf/10.1176/appi.ajp.2007.06122050

Cannon, T. D., & Keller, M. C. (2006). Endophenotypes in the genetic analyses of mental disorders. *Annual Review of Clinical Psychology, 2,* 267–290. http://dx.doi.org/10.1146/annurev.clinpsy.2.022305.095232

Carney, D. M., Moroney, E., Machlin, L., Hahn, S., Savage, J. E., Lee, M., . . . Hettema, J. M. (2016). The twin study of negative valence emotional constructs. *Twin Research and Human Genetics, 19,* 456–464. http://dx.doi.org/10.1017/thg.2016.59

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Easter, J., McClure, E. B., Monk, C. S., Dhanani, M., Hodgdon, H., Leibenluft, E., . . . Ernst, M. (2005). Emotion recognition deficits in pediatric anxiety disorders: Implications for amygdala research. *Journal of Child and Adolescent Psychopharmacology, 15,* 563–570. http://dx.doi.org/10.1089/cap.2005.15.563

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Consulting Psychologists Press.

Fusar-Poli, P., Placentino, A., Carletti, F., Landi, P., Allen, P., Surguladze, S., . . . Politi, P. (2009). Functional atlas of emotional faces processing: A voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *Journal of Psychiatry & Neuroscience, 34,* 418–432.

Gottesman, I. I., & Gould, T. D. (2003). The endophenotype concept in psychiatry: Etymology and strategic intentions. *The American Journal of Psychiatry, 160,* 636–645. http://dx.doi.org/10.1176/appi.ajp.160.4.636

Guyer, A. E., McClure, E. B., Adler, A. D., Brotman, M. A., Rich, B. A., Kimes, A. S., . . . Leibenluft, E. (2007). Specificity of facial expression labeling deficits in childhood psychopathology. *Journal of Child Psychology and Psychiatry, 48,* 863–871. http://dx.doi.org/10.1111/j.1469-7610.2007.01758.x

Harmer, C. J., Rogers, R. D., Tunbridge, E., Cowen, P. J., & Goodwin, G. M. (2003). Tryptophan depletion decreases the recognition of fear in female volunteers. *Psychopharmacology, 167,* 411–417. http://dx.doi.org/10.1007/s00213-003-1401-6

Hofmann, S. G. (2014). Interpersonal emotion regulation model of mood and anxiety disorders. *Cognitive Therapy and Research, 38,* 483–492. http://dx.doi.org/10.1007/s10608-014-9620-1

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry, 167,* 748–751. http://dx.doi.org/10.1176/appi.ajp.2010.09091379

Jackson, R. W., Snieder, H., Davis, H., & Treiber, F. A. (2001). Determination of twin zygosity: A comparison of DNA with various questionnaire indices. *Twin Research, 4,* 12–18. http://dx.doi.org/10.1375/1369052012092

Kasriel, J., & Eaves, L. (1976). The zygosity of twins: Further evidence on the agreement between diagnosis by blood groups and written questionnaires. *Journal of Biosocial Science, 8,* 263–266. http://dx.doi.org/10.1017/S0021932000010737

Khoo, S.-T., West, S., Wu, W., & Kwok, O.-M. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 301–317). Washington, DC: American Psychological Association.

Kirkpatrick, M. G., Lee, R., Wardle, M. C., Jacob, S., & de Wit, H. (2014). Effects of MDMA and Intranasal oxytocin on social and emotional processing. *Neuropsychopharmacology, 39,* 1654–1663. http://dx.doi.org/10.1038/npp.2014.12

Kirsh, S. J., & Mounts, J. R. W. (2007). Violent video game play impacts facial emotion recognition. *Aggressive Behavior, 33,* 353–358. http://dx.doi.org/10.1002/ab.20191

Kohler, C. G., Turner, T. H., Bilker, W. B., Brensinger, C. M., Siegel, S. J., Kanes, S. J., . . . Gur, R. C. (2003). Facial emotion recognition in schizophrenia: intensity effects and error pattern. *American Journal of Psychiatry, 160,* 1768–1774. http://dx.doi.org/10.1176/appi.ajp.160.10.1768

Lawrence, K., Campbell, R., & Skuse, D. (2015). Age, gender, and puberty influence the development of facial emotion recognition. *Frontiers in Psychology, 6,* 761. http://dx.doi.org/10.3389/fpsyg.2015.00761

Lilley, E. C. H., & Silberg, J. L. (2013). The Mid-Atlantic Twin Registry, revisited. *Twin Research and Human Genetics, 16,* 424–428. http://dx.doi.org/10.1017/thg.2012.125

Marsh, A. A., & Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: A meta-analysis. *Neuroscience and Biobehavioral Reviews, 32,* 454–465. http://dx.doi.org/10.1016/j.neubiorev.2007.08.003

Marsh, A. A., Yu, H. H., Pine, D. S., & Blair, R. J. R. (2010). Oxytocin improves specific recognition of positive facial expressions. *Psychopharmacology, 209,* 225–232. http://dx.doi.org/10.1007/s00213-010-1780-4

Nowicki, S., Jr., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior, 18,* 9–35. http://dx.doi.org/10.1007/BF02169077

Pinkham, A. E., Penn, D. L., Green, M. F., & Harvey, P. D. (2016). Social cognition psychometric evaluation: Results of the Initial Psychometric Study. *Schizophrenia Bulletin, 42,* 494–504. http://dx.doi.org/10.1093/schbul/sbv056

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime computer software and manual*. Pittsburgh, PA: Psychology Software Tools Inc.

Shaw, P., Stringaris, A., Nigg, J., & Leibenluft, E. (2016). Emotion dysregulation in attention deficit hyperactivity disorder. *Focus, 14,* 127–144. http://dx.doi.org/10.1176/appi.focus.140102

Steinberg, L., & Morris, A. S. (2001). Adolescent development. *Journal of Annual Review of Psychology, 52,* 83–110. http://dx.doi.org/10.1146/annurev.psych.52.1.83

Trentacosta, C. J., & Fine, S. E. (2010). Emotion knowledge, social competence, and behavior problems in childhood and adolescence: A meta-analytic review. *Social Development, 19,* 1–29. http://dx.doi.org/10.1111/j.1467-9507.2009.00543.x

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17,* 3–28. http://dx.doi.org/10.1007/BF00987006

Received June 16, 2016
Revision received November 28, 2016
Accepted December 5, 2016 ■