Available online at www.sciencedirect.com
ScienceDirect
Behavior Therapy 50 (2019) 828–838
Behavior Therapy
www.elsevier.com/locate/bt

Group Differences in Facial Emotion Expression in Autism: Evidence for the Utility of Machine Classification

Nicole N. Capriola-Hall
University of Alabama

Andrea Trubanova Wieckowski
Deanna Swain
Virginia Tech
Sherin Aly
Alexandria University

Amira Youssef
A. Lynn Abbott
Virginia Tech

Susan W. White
University of Alabama

Effective social communication relies, in part, on accurate nonverbal expression of emotion. To evaluate the nature of facial emotion expression (FEE) deficits in children with autism spectrum disorder (ASD), we compared 20 youths with ASD to a sample of typically developing (TD) youth (*n* = 20) using a machine-based classifier of FEE. Results indicate group differences in FEE for overall accuracy across emotions. In particular, a significant group difference in accuracy of FEE was observed when participants were prompted by a video of a human expressing an emotion, $F(2, 36) = 4.99$, $p = .032$, $\eta^2 = .12$. Specifically, youth with ASD made significantly more errors in FEE relative to TD youth. Findings support continued refinement of machine-based approaches to assess and potentially remediate FEE impairment in youth with ASD.

Keywords: autism spectrum disorder; facial emotion expression; machine learning

Youth with autism spectrum disorder (ASD) exhibit impairments related to social behavior, which often include a diminished ability to interpret and convey facial expressions (Evers, Steyaert, Noens, & Wagemans, 2015; Rump, Giovannelli, Minshew, & Strauss, 2009). The ability to understand nonverbal emotion expression is fundamental for effective social communication (Wang, Dapretto, Hariri, Sigman, & Bookheimer, 2004). Typically, the ability to discriminate certain expressions develops very early in childhood with ability to identify basic emotions using static cues appearing as early as 3 months of age (e.g., Soken & Pick, 1999). However, one of the hallmark characteristics of ASD is a failure to attend to facial expressions (Frazier et al., 2017). In addition to impaired facial emotion recognition (FER), the more limited research on facial emotion

Since study completion, Nicole N. Capriola-Hall and Susan W. White are now at the Department of Psychology, University of Alabama.
This study was funded by NICHD (R03HD081070; Co-PIs Abbott & White).
Address correspondence to Nicole Capriola-Hall, Department of Psychology, University of Alabama, Tuscaloosa, AL 35487; e-mail: nncapriola@crimson.ua.edu.

0005-7894/

expression (FEE) indicates that youth with ASD may demonstrate diminished and perhaps atypical FEE. The current study was undertaken to validate machine-based classification of FEE and to determine if machine-based detection and classification of FEE could detect group differences between youth with ASD and those without.

Although children with ASD do spontaneously express basic emotions and generally do not show impairment in imitation or mimicry of facial expressions (e.g., Deschamps, Coppes, Kenemans, Schutter, & Matthys, 2015; Press, Richardson, & Bird, 2010), some studies have found that they do so less precisely and less frequently than peers without ASD (Kasari, Sigman, Mundy, & Yirmiya, 1990; Shalom et al., 2006). However, social pressure may account for some of the differences in emotional behaviors in children with ASD, as some studies have found no deficits in display of emotion when social pressure is removed (i.e., the paradigm was nonsocial and noninteractive; youth watched emotion-inducing YouTube videos without a social partner; Zane, Neumeyer, Mertens, Chugg, & Grossman, 2018). This study offers important considerations for the current study given that our present research was nonsocial, as defined by Zane and colleagues (2018), yet interactive in nature. Within a social context, however, people with ASD display fewer nonverbal expressions of affect (Yirmiya, Kasari, Sigman, & Mundy, 1989), and their facial expressions are often perceived as incorrect or incongruent relative to the situation by independent raters. In addition, in a recent study on perception of facial expressivity in ASD, researchers found that ASD expressions were rated as more intense and less natural, even when accurate, relative to typically developing (TD) adults (e.g., Faso, Sasson, Pinkham, 2015). Children with ASD were also rated as more socially awkward from short visual and/or auditory stimuli during a story-retelling task, as well as from still images, showing that nonverbal emotion expression influences perception of social awkwardness (Grossman, 2015), which may lead to or exacerbate social difficulty.

Compared to the interventions targeting FER, there has been little clinical research on remediation of deficient FEE (Wieckowski & White, 2017a) and most of the extant research examining difficulties in FEE has relied on human-coded data from video recordings of the expressed emotion (Gordon, Pierce, Bartlett, & Tanaka, 2014; Hobson & Lee, 1998; Loveland et al., 1994). Human coding poses clear limitations due to resource demands (e.g., time to train multiple raters). Novel approaches to assessment and potential modification of FEE that rely less on human coding could prove more viable, especially if the approach could classify expression in real time in order to provide in-situ feedback to the child. Computerized intervention allows for stable, predictable, and consistent training, and has been found to provide a motivational learning environment for individuals with ASD (Goldsmith & LeBlanc, 2004; White et al., 2018). Within the past 10 years, affective computational approaches (e.g., machine learning algorithms) have become increasingly important to the understanding of affective neuroscience and autism (el Kaliouby, Picard, & Baron-Cohen, 2006).

We sought to evaluate machine-classified FEE using a newly developed computer system (White et al., 2018), termed Facial Emotion Expression Training (FEET), which utilizes a noninvasive approach to recognize four primary emotions (i.e., happiness, fear, anger, and neutral expressions) using 3-dimensional (3D) facial expressions alone. The decision to target these emotions in particular was based on prior research indicating that recognition deficits may be most pronounced for negative emotions such as fear (Uljarevic & Hamilton, 2013), and because the initial piloting of the system indicated that these four emotions were the most readily and successfully classified (Aly et al., 2015). Unlike other computational approaches, FEET recognizes the subject's facial expressions automatically and provides real-time responses based on the detected expression. The system is interactive, allowing the child to get immediate feedback. Importantly, the FEET system has been found to be acceptable and enjoyable to youth with ASD (White et al., 2018). FEET was developed by an interdisciplinary research team as a proof of concept and was used to obtain the data presented in this study.

Past research (e.g., Malawski, Kwolek, & Sako, 2014; Seddik, Maâmatou, Gazzah, Chateau, & Ben Amara, 2013) has employed Kinect-based technology to detect FER in adults. To our knowledge, no prior research has examined differences in FEE among youth with ASD and TD youth, using a computer-based system capable of automated recognition of human expressions. Using automatic, computerized detection, the current study has examined differences in FEE across types of stimuli. Specifically, when interacting with the FEET system, the child proceeded through three scaffolded levels, with stimuli becoming increasingly subtle (i.e., signaling of emotions was less obvious) as the levels progressed (Figure 1). Within each level, depicted emotions increased in intensity of expression (e.g., smile grows more pronounced for a happy expression) across three trials. Youth with ASD have been found to more accurately identify emotions when depicted at full intensity but often struggle with recognition of emotions at low or intermediate intensity levels (Doi et al., 2013; Grossman & Tager-Flusberg, 2012). As such, these scaffolded levels

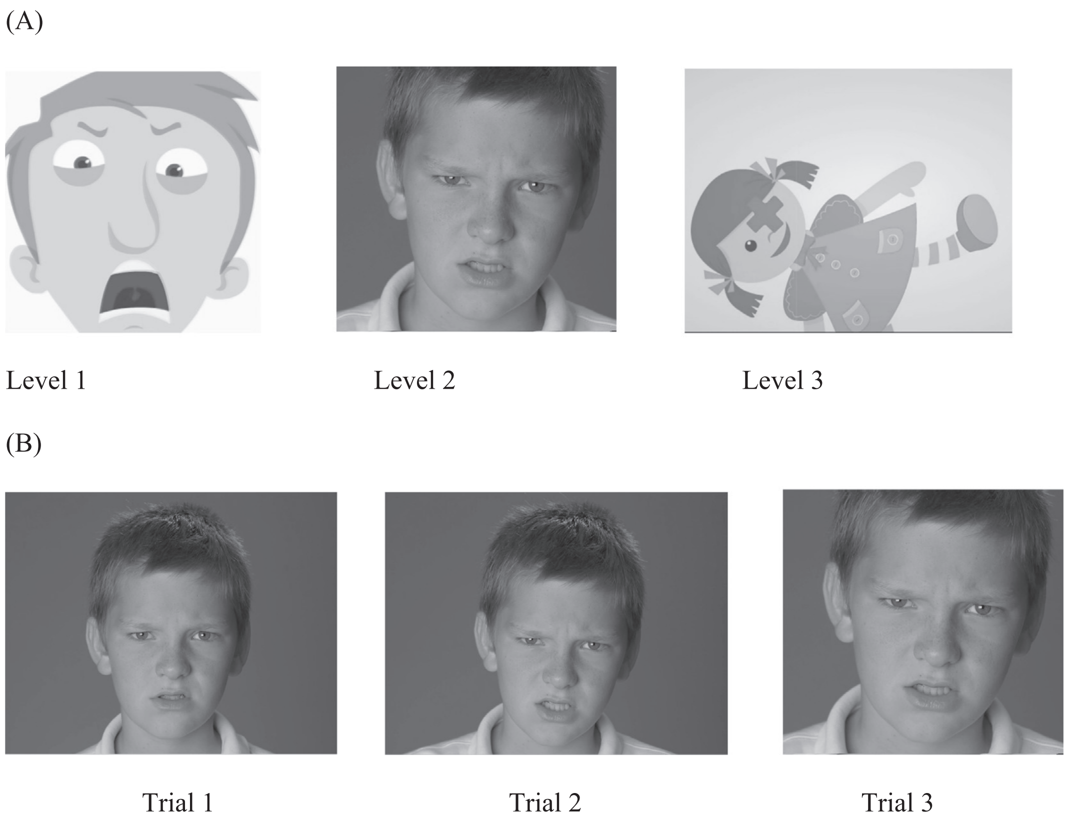

FIGURE 1 (A) Single frames from sample video stimuli for scaffolded levels depicting anger. (B) Single frames from a sample video stimuli for trials demonstrating increase in intensity for anger.

(i.e., different stimuli) and trials (i.e., different intensities) afford the ability to examine how differences in stimuli and intensity might influence FEE.

Analysis of both the scaffolded levels and trials of FEET allows us to account for potential interactions between group and stimulus type. Before examining group differences in FEE, we assessed convergence between the human coding and machine-based classification in order to establish ground truth for the machine-based classification and to determine whether machine-based classification is acceptable to utilize for further analyses. We hypothesized that there would be a positive association between human-coded FEE and FEE determined by the machine classifier (FEET). Support for this hypothesis would indicate that machine-based classification of FEE could serve as a viable supplement to, or perhaps replacement for, more time-intensive human coding. Ground truth is needed for the training of machine-based models (Rudovic, Lee, Dai, Schuller, & Picard, 2018), as it provides an external criterion against which the new system's results are compared. Consistent with past research in this area (e.g., Rudovic et al., 2018), human codes were examined to establish ground truth for machine-based classification.

The central aim of the present study was to determine if machine-based detection and classification of FEE could detect group-level differences between youth with ASD and those without. In order to validate the use of machine-based detection and classification of FEE, we examined the convergence between human coding of FEE and FEET's classification of FEE. Following the validation of FEET and in line with the primary aim of this preliminary study, we predicted that children with ASD would demonstrate overall FEE deficiency, based on FEET's machine-based classification, in comparison to TD peers.

Method

PARTICIPANTS

Forty children were enrolled in a pilot study to determine the feasibility of the FEET system and machine-learning classifier, as well as to examine group differences. All participants were 9–12 years old and free of a co-occurring intellectual disability as determined by the Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II; Wechsler, 2011). Although this is a fairly narrow age range, facial features of children this age are sufficiently mature for automated processing (Huelke, 1998) with the proposed FEE technology (Arbogast et al., 2016), and there is more extreme change in facial dimensions earlier in childhood (Bishara et al., 1984). Further, children of this age with ASD have been shown to display difficulties identifying emotions from static facial expressions (Lindner & Rosén, 2006). Children in the ASD group ($n = 20$; mean age = 10.20 years; 90% boys) met diagnostic criteria for ASD as determined by clinical evaluation and supported by the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2; Lord et al., 2012) scored by research reliable clinicians. Youth with ASD were permitted to participate regardless of co-occurring psychopathology (as determined by parent report). The TD group included 20 participants with no clinical diagnoses or an immediate family member diagnosed with ASD (as determined by parent report; mean age = 10.81 years; 70% boys). Both groups were non-treatment seeking.

PROCEDURE

The study was approved by the university's Institutional Review Board for human subject research. All parents provided informed consent and youth gave assent prior to beginning the study. Participants were recruited through the psychology department's child participant database, university-affiliated clinics, and advertisements in the community and on campus. Each child completed the computerized FEET task while the parent completed forms in a separate room.

MEASURES

Facial Emotion Expression Training (FEET)

FEET is an interactive, automated system developed for the purpose of helping children with ASD learn appropriate FER/FEE (White et al., 2018). The Microsoft Kinect was selected as the interface because it is relatively low in cost, fast, and because it captures 3D information that helps address problems related to variations in illumination and head pose, as compared to traditional video cameras. The FEET system consists of a small, monitor-mounted Kinect sensor to capture 3D representations of the child's face, along with software that presents audiovisual stimuli and automatically detects the emotion being expressed by the child.

We developed the facial emotion recognition component of FEET in order to interpret FEE by the participant using machine-learning techniques trained using the VT-KFER data set (Aly et al., 2015). VT-KFER is a publicly available dataset captured by the Kinect sensor. The development of FEET involved, first, the automatic extraction and selection of 3D geometric facial features that best discriminate the four emotions from the VT-KFER dataset. Then, a training procedure was conducted in which these selected features, for each emotion type, were presented to a machine-learning system, and the system applied a form of kernel discriminant analysis to establish internal rules for distinguishing each emotion type from all others (Li, Gong, & Liddell, 2003). Those rules are in the form of mathematical computations based on geometric features of the subject's face. Our automated FER component of FEET achieved 86.9% accuracy when tested on VT-KFER data using leave-one-subject-out cross-validation. We expected slightly lower accuracy with FEET participants as all training was performed using data from VT-KFER. Indeed, off-line assessment of FEET's recognition performance indicated that the recognition accuracy for the 20 TD children was 72.5%.

In addition to automated recognition of expressions, FEET presented stimuli to participants in order to prompt certain emotions and provided feedback to the participants based on their responses. Stimuli that were presented by FEET became increasingly subtle (i.e., signaling of emotions was more subdued) as the levels progressed (Figure 1), proceeding from simple cartoon faces showing discrete emotions (Level 1), to dynamic short films of a child actor showing emotions (Level 2), to scenes with an audiovisual track indicating the target emotion, but without a child's face (e.g., broken doll with child's voice saying, "*My brother broke my toy, so I'm going to break his toy!*" to target anger; Level 3).

Within each of these levels, the intensity of the depicted emotion increased across the trials (i.e., within-level; e.g., scowl becomes more pronounced, expression becomes clearer). For each of the three levels, the system presented the stimuli on a digital display one at a time, with audible prompts encouraging the subject to show what the emotion currently being displayed was (e.g., "With your face, show what I am feeling"), and then detected the subject's facial expression in response to the prompt. Based on the subject's expression, the system presented either encouragement or corrective feedback (e.g., "Not quite right, try again"). In this way, the system was formulated to be interactive and game-like to facilitate child engagement and enjoyability. The system, however, was not adaptive, as regardless of performance, the FEET system progressed through all levels and trials for all participants, with subsequent attempts of expressing the correct emotion counted as inaccurate for the basis of the accuracy coding. Although feedback was offered by the FEET system, the feedback offered was standard across participants and the system did not coach participants on how to produce the correct expression.

For each stimulus within each trial for each level, FEET automatically captured sensor data from the Microsoft Kinect and determined whether the participant expressed the emotion that was expected. The low-cost, three-dimensional sensors, coupled with inexpensive computers and recently developed machine-learning techniques, offer the potential for advances in the ways that computational devices can be used for interactive therapy to address FEE impairment in children with ASD. The machine learning techniques utilize a computational framework for discovering patterns and represent strong potential for enhancing current intervention practices. Further technical details on the development of the machine-based classifier can be found in Aly, Abbott, Wieckowski, White, and Youssef (submitted for publication).

HUMAN CODING OF FEE

Human coding was performed solely to establish a ground-truth reference for assessing FEET's machine-based classifier. This step was necessary in order to validate the use of the FEET system. After the sessions, two trained undergraduate students viewed recordings and independently coded each participant's expressions for all stimulus presentations (i.e., all levels and trials). Human coders received didactic training, including assigned readings on human emotion expression coding and training on the specific facial features associated with basic emotions (EM-FACS; Ekman & Friesen, 1984). They then co-coded selected videos as a group and discussed codes and stimuli to ensure consistency before proceeding to code independently. Human coders attained reliability by coding a set of training stimuli (i.e., same stimuli and codes used to train the FEET system) and were considered reliable for independent coding when they achieved greater than 80% agreement. In addition to assigning the emotion label, the coders were asked to code the clarity of expression (i.e., "2" for complete confidence in emotion code, and "0" for indiscernible emotion). In order to gauge inter-rater agreement of behavior coding, 50% of the videos were co-coded. Given the focus of the analysis is on the machine-based FEE classification, and the fact that prior research in this area has generally relied on lower percentages for co-coding (e.g., 25% in Wieckowski & White, 2017b; 26.7% in Yoshimura, Sato, Uono, & Toichi, 2015; 25% in Loveland et al., 1994), 50% co-coding was determined sufficient. When disagreement between coders occurred, a third rater (i.e., trained master's-level graduate student or Ph.D.-level psychologist) assigned the final code, which could differ from either rater's code (although there was no instance where it did differ). After approximately half of the videos were coded, all raters and trainers participated in consensus coding for one subject's expressions to recalibrate in order to prevent drift in codes. For the co-coded videos, intraclass correlations (ICCs) were computed to determine the level of agreement between two raters' judgments on what emotion the participant portrayed. According to the calculated ICCs value, there was good agreement (Cicchetti, 1994) between the two independent raters, ICC (1,2) = .638, $p < .001$, for the co-coded videos.

Convergence of Human Coding and FEET

Of note, the number of possible emotions coded between FEET and human coding differed. Coders chose one of the six basic emotions: happiness, sadness, fear, surprise, disgust, or anger, in addition to neutral expression, whereas the FEET classifier had just four options: happiness, anger, fear, and neutral. This approach was taken primarily for pragmatic reasons. First, in FEET development it was determined that the machine classifier had difficulty distinguishing surprise and disgust. Although FEET did not attempt to elicit surprise or disgust, from presented stimuli, it is possible that participants showed these emotions, and the human coders could not be instructed to code incorrectly (i.e., pretend that surprise and disgust do not exist and just code as one of the other four basic emotions). Human coders were shown only stimuli with the four emotions (happiness, anger, fear, and neutral) since those were the only emotions intended to be elicited by FEET. The human coders were not aware that all emotions were not included as stimuli. As such, convergence was calculated for both the same number of emotions coded by FEET and the additional emotions available to the human coders, thereby providing the most stringent test of ground truth. Results demonstrated excellent agreement (Cicchetti, 1994) between the FEET classifier and the full set of human coded emotions (e.g., emotion choices of disgust, surprise, and sadness included), ICC(1,2) = .795[1], $p < .001$. When the two sets of emotions were matched to the emotions based on the machine-based classification (e.g., removing cases of disgust, sadness, and surprise from those used for human coding), the intraclass correlation remained consistent ICC(1,2) = .795[1], $p < .001$. Using human coding as ground truth, these levels of convergence indicate that the FEET classifier results can be relied upon for subsequent analyses. Although

[1] The FEET system only classified expressions as the emotion prompted or "Incorrect." As such, the "Incorrect" codes were deleted from this convergence calculation as human coders could not classify an emotion as "Incorrect." With the "Incorrect" label included, reliability was fair, ICC(1,2) = .501, $p < .001$.

feedback was provided by FEET to the participating youth in real time, data used for the present study was obtained by applying the FEET classifier to recorded data after the sessions. This approach was used for consistency, as small improvements were made to FEET during the several months of conducting interactive sessions.

DATA ANALYSIS

Data were analyzed with IBM SPSS Statistics Version 24. In order to investigate whether group (ASD, TD) differences existed among the questionnaire and demographic variables, we used a series of independent samples *t*-tests for the continuous variables and Mann-Whitney Test for the ordinal variables to determine whether any variables (e.g., IQ, participant sex, family income) needed to be controlled for as possible covariates. The primary aim of the manuscript was to examine the utility of machine-based classification in detection of group-level differences in FEE. This would be supported by significantly lower accuracy scores on overall expression and expression at each level (i.e., modeled, cartoon prompt versus cued human stimulus versus audio-visual scenario) for the youth with ASD compared to the TD youth. Given the preliminary nature of the current study, we conducted exploratory analyses to examine FEE differences across emotion type. We also explored FEE accuracy across stimuli intensity within all FEET levels. Since the current study was a pilot study, directional hypotheses were not offered regarding emotion specificity and differences in stimulus intensity given a lack of prior research to inform us as well as limited statistical power. To test group differences in FEE across FEET's levels and trials as well as emotion type, a one-way between-groups multivariate analysis of variance (MANOVA) was performed to investigate group differences in FEE based on machine-based classification. Group (ASD vs. TD) was the independent variable. Levels of FEET (Level 1, 2, and 3), trials of FEET (Trials 1, 2, and 3) for all Levels, and emotion (4 emotions: happiness, anger, fear, and neutral) were the dependent variables.

Results

Group descriptive statistics (see Table 1) and preliminary analyses indicated no problems with skewness or kurtosis and no significant differences between the TD and ASD youth in terms of participant sex ($\chi^2 = 2.50$, $p = .114$) and age ($t = 1.76$, $p = .087$). However, there was a statistically significant group difference in cognitive ability, with the TD group scoring higher, $t(38) = 4.35$, $p < .001$. Despite this group difference, cognitive ability was not statistically associated with any of the dependent variables for either group, *ps* =.082-.997. Given that family income was an ordinal variable (ranges = < 20,000, 21,000–40,000, 41,000–60,000, 61,000–80,000, > 81,000), a Mann-Whitney Test was used to calculate group-level differences in family income. Family income was significantly greater for the TD group relative to the ASD group, $Z = -2.22$, $p = .033$. Family income was also not significantly associated with our variables of interest for either group, *p*s = .082–.994. Thus, we did not include any covariates in the current analyses.

Table 1
Demographic Data

	ASD (*n* =20)	TD (*n* =20)	X^2 / t
Gender (male)	18	14	2.50
Race			5.03
White	16	17	
Black	2	0	
Latino	1	0	
Asian/Other	1	2	
NA	0	1	
Age (months)	122.50	129.75	1.76^
IQ	100.55 (13.96)	118.15 (11.53)	4.35**

^ $p < .10$; *$p < .05$; **$p < .01$.
Note. TD = typically developing; ASD = autism spectrum disorder.

For the MANOVA, the Box's M test of equality of covariance matrices was significant, demonstrating that the model assumption of homogeneity of covariances was violated. Given the violation, Pillai's trace was used to evaluate whether there were statistically significant differences among the groups on the linear combination of the dependent variables given that it is more robust and not linked to assumptions about the normality or distribution of the data (Tabachnick & Fidell, 2007).

Based on classification by FEET, there was a statistically significant group difference between the ASD and control children on the combined dependent variables, Pillai's Trace = .67, $F(17, 22) = 2.57$, $p = .020$, $\eta_p^2 = .67$. When the dependent variables were considered separately, the univariate *F*-tests showed there was a statistically significant difference between the groups for overall FEE accuracy, $F(1, 38) = 7.93$, $p = .008$, $\eta_p^2 = .17$ (Table 2). Accuracy of FEE for the TD group across emotions was 82.53% based on classification by FEET. For the ASD group, overall accuracy of FEE across emotions was 73.25% based on the machine-based classification. Univariate *F*-tests also suggested that the ASD and TD groups significantly differed on Level 2 of FEET (human faces, please see Figure 2). The ASD youth evinced more errors within Level 2. Accuracy of FEE for the TD group was 87.90%, as compared to 75.93% for the ASD group.

Table 2
Means and Standard Deviations for Accuracy of FEE Across FEET Levels and Emotions, as Determined by FEET

	ASD (n =20)	TD (n =20)	Significant Group Differences
Accuracy across Emotions (%)	73.25 (10.44)	82.53 (10.40)	9.28**
Level 1 Accuracy (%)	72.20 (16.31)	81.65 (16.93)	9.45^
Level 2 Accuracy (%)	75.95 (15.50)	87.90 (11.71)	11.95*
Level 3 Accuracy (%)	71.65 (13.08)	79.10 (14.03)	7.45^
Level 1 X Trial 1 Accuracy (%)	62.50 (25)	78.75 (16.77)	16.25*
Level 1 X Trial 2 Accuracy (%)	73.75 (20.64)	81.25 (17.91)	7.50
Level 1 X Trial 3 Accuracy (%)	80 (19.19)	71.25 (23.33)	-8.75
Level 2 X Trial 1 Accuracy (%)	75 (18.13)	85 (17.01)	10.00
Level 2 X Trial 2 Accuracy (%)	80 (22.36)	87.50 (15.74)	7.50
Level 2 X Trial 3 Accuracy (%)	75 (21.46)	87.50 (17.21)	12.50*
Level 3 X Trial 1 Accuracy (%)	70 (26.41)	76.25 (15.12)	6.25
Level 3 X Trial 2 Accuracy (%)	73.75 (18.98)	81.25 (19.66)	7.50
Level 3 X Trial 3 Accuracy (%)	70 (26.41)	77.50 (19.70)	7.50
Happiness Accuracy (%)	58. 60 (30.01)	73.45 (19.26)	14.85^
Anger Accuracy (%)	68.40 (26.20)	78.35 (23.61)	9.95
Neutral Accuracy (%)	85.85 (16.44)	96.15 (8.94)	10.30*
Fear Accuracy (%)	80.35 (15.90)	82.85 (20.28)	2.50

^ $p < .10$; *$p < .05$; **$p < .01$.
Note. Level 1 =Cartoon stimuli; Level 2= Human stimuli; 3= Audio-visual track stimuli.
TD = typically developing; ASD = autism spectrum disorder.

In terms of group differences for discrete emotions, there was a statistically significant group difference for neutral expressions across the levels (Table 2) as determined by the univariate F-tests, $F(1, 38) = 6.06$, $p = .018$, $\eta_p^2 = .14$. Specifically, the TD sample demonstrated greater FEE accuracy (i.e., correspondence between child's expression and the presented stimuli) than the ASD sample (96.15% vs. 85.85% accurate, respectively).

We also examined group differences in FEE, as determined by FEET, across degrees of intensity of the expressed emotion (i.e., trial) for Levels 1, 2, and 3. For Level 1, there was a statistically significant group difference for the lowest intensity trial, $F(1, 37)= 5.82$, $p =.021$, $\eta^2 = .13$. Mean accuracy for FEE for Level 1, Trial 1 was 62.50% for the ASD group and 78.75% for the TD group. Differences were not statistically significant as the intensity increased for the Level 1 trials. Although not statistically significant, youth with ASD showed a trend toward greater accuracy (80%) relative to the TD youth (71.25%) for Trial 3 of Level 1.

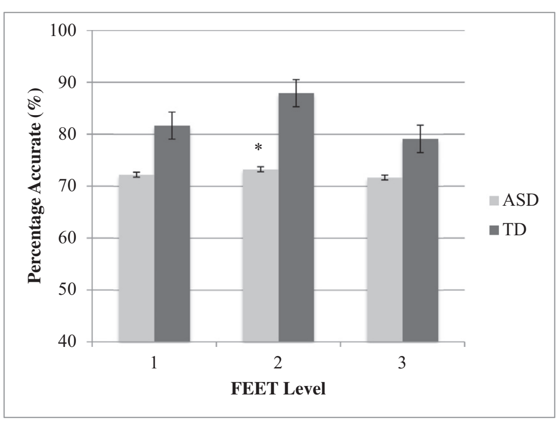

FIGURE 2 Accuracy of emotions expressed across FEET levels, as determined by FEET.Note. Level 1 =Cartoon stimuli; Level 2= Human stimuli; 3= Audio track stimuliAn asterisk indicates significant between-group difference (p < .05). TD = typically developing; ASD = autism spectrum disorder.

For Level 2 (human stimuli), Trial 3 (highest intensity stimuli) youth with ASD demonstrated poorer FEE accuracy (75 %) relative to the TD group (87.5%) as determined by FEET, $F(1, 37)= 4.13$, $p =.049$, $\eta^2 = .10$. No other significant differences were observed at the trial level for Level 2. In addition, there were no statistically significant group differences in FEE accuracy for any of the trials for Level 3, suggesting that the ASD and TD groups did not differ significantly in accuracy of FEE as the stimuli in Level 3 increased in intensity.

Discussion

This preliminary study sought to validate the use of machine-based classification of FEE (convergence between human coders and machine-based classification) and to explore group differences in FEE, as determined by FEET. Group differences in FEE for overall accuracy as well as across the scaffolded levels and trials of FEET were explored in order to account for potential interactions between group, stimulus type, and intensity. There was a positive association between human coding and FEET in identifying the emotions expressed across the sample. As noted by others in the field (Gordon et al., 2014), human coding can be cumbersome (i.e., time demands, the need for training, and lack of ability to be used in real time) and therefore automated approaches to FEE classification should be considered. Convergence findings suggest that the use of the FEET system, which is both non-invasive and employs machine-based classification to provide corrective feedback, has promise given the time demands associated with the human coding of emotions.

In support of our primary hypothesis, there was a significant difference between the youth with ASD and TD youth in terms of accuracy of FEE. As determined by the machine-based classification, youth with ASD made significantly different faces, classified as errors in typical expression, compared to TD youth, in response to human face stimuli. In typical human social interaction, emotional expressions are likely to be of low intensity and fleeting, changing from moment to moment (Matsumoto & Hwang, 2014). This observation is an important consideration in interpreting these results; youth with ASD may need more training with human stimuli for FEE deficit remediation, if deficits are not so apparent with less realistic stimuli such as cartoons. Our findings are in contrast to those reported by Zane and colleagues (2018), who suggested that youth with ASD demonstrated more intensive and varied facial expressions relative to TD youth, as coded by researchers, when the paradigm was nonsocial and noninteractive. This group level difference observed in our study might be a function of the FEET paradigm. Specifically, although FEET was not social in nature, as defined by Zane and colleagues, the paradigm was indeed interactive.

In terms of group-level differences in FEE for discrete emotions, results revealed significant group-level differences in FEE for only neutral facial expressions. Specifically, youth with ASD demonstrated less accurate FEE when prompted to display neutral expressions. Prior research has suggested that neutral faces are often interpreted as ambiguous social information, resulting in less sustained social attention (Cooney, Atlas, Joormann, Eugène, & Gotlib, 2006). Perhaps neutral faces were perceived as more ambiguous by the ASD group, which affected facial responses. Past research by Eack and colleagues (2015) and Dalton and colleagues (2005) has suggested that individuals with ASD demonstrated significantly greater difficulties in distinguishing neutral facial expressions from emotional facial expressions, which might explain the difficulties in expressing these emotions, at least within the context of FEET.

Our findings suggest that youth with ASD demonstrated poorer accuracy in FEE for cartoon-like stimuli at the lowest intensity. These results suggest that youth with ASD improved in the FEE accuracy as the intensity of the cartoon stimuli increased. The increase was such that their accuracy in FEE, although not statistically significant, was higher than the TD group for the highest intensity stimuli at the cartoon level. Even within the same stimulus category, intensity is an important consideration for future research. Regarding trial level differences, our findings also demonstrated that youth with ASD demonstrated significantly poorer accuracy in FEE for the highest intensity stimuli of the human level. Youth with ASD demonstrate behavioral abnormalities in processing faces (Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Spezio, Adolphs, Hurley, & Piven, 2007), often taking a piecemeal processing approach (i.e., lack of integration of all facial information to recognize and in turn express the emotion prompted; Behrmann, Thomas, & Humphreys, 2006). In cartoons, single components can change (e.g., big smile) more radically, while other components of the face do not. In contrast, human emotional expression involves the whole face. Though speculative, it is possible that the youth with ASD were focused on a particular facial feature (e.g., mouth)—an approach that can help with reading emotions in cartoons, but is perhaps detrimental when interpreting the emotions expressed by other people.

Overall, our findings are consistent with prior research showing that FEE in youth with ASD differs from what is seen in typically developing youth (Kasari et al., 1990; Shalom et al., 2006), and contributes in a novel way by demonstrating that machine-based classification was able to detect group-level differences across stimulus type. Atypical or deficient FEE can impede social discourse, which in turn can give rise to the development and exacerbation of social difficulties for youth with ASD. Nonverbal communication, including FEE, is fundamental for effective social discourse (Wang et al., 2004). Impaired FEE can lead to limited or negative social interactions and difficulties with peer relationships. Findings from this study suggest that for youth with ASD, FEE may be impaired in fairly subtle and nuanced ways, and that detection of FEE impairment is dependent on the nature of the stimulus.

Results of this study must be considered in light of several methodological limitations. First, the composition of the sample limits our ability to generalize findings. The participants were predominantly

Caucasian and, by study design, were all relatively cognitively unimpaired. Although IQ was not a statistically significant covariate in the analyses, future research should also examine FEE within well-matched groups. Despite all participants demonstrating average verbal abilities, future studies should include language testing to determine if all participants were equally able to follow the instructions prompted by the FEET system. In addition, youth with ASD could have comorbid psychopathology, which may have influenced our findings. Unfortunately, due to the small sample size, we could not examine whether these individual differences influenced our findings. Although our sample composition was not unlike other ASD research samples that examined displays of emotion (Zane et al., 2018), the gender composition within our sample was likely not representative of the more general ASD population. Within the bounds of the current study design, we were unable to explicitly disentangle whether both FER and FEE were targeted as these processes are strongly correlated (Ricciardi et al., 2017; Wieckowski & White, 2017a) and may be partially codependent. Within FEET, some youth might rely on valence, or might mimic the expression, or they might make an expression based on genuine feeling. The automated expression-recognition component of FEET was not developed specifically for children (White et al., 2018), and therefore it is possible that FEET's performance could be improved through further development that targets children. Participants younger than 9 were not included in part because of concern that FEET would not recognize expressions as accurately as for the selected age group of 9 to 12 years of age. As such, this is an important area for future research. Despite the fact that the feedback offered by the FEET system was standard across participants and the system did not coach participants on how to produce the correct expression, we are unable to account for whether feedback from FEET functioned as an intervention that might have influenced the actions of the participants. The FEET system was designed for eventual use as an intervention; however, the present study solely focused on group differences in FEE and whether FEET was able to detect said differences. A priori power calculations indicated that our study design had enough statistical power to detect medium effects for our primary analyses; nonetheless, the sample sizes were small across the two groups. Concerning the technological approach because the number of possible emotions coded between FEET and human coding differed, the relationship between them cannot be directly tested. FEET was constrained to a four option coding system whereas human coding allowed for six possible options; thus, the lower ICC calculation (see Footnote1) likely reflects this difference. In addition, FEET classification data analyzed here may have differed in a few cases from the responses that were presented to participants. Small improvements to the FEET classifier software were implemented during the study in order to provide greater accuracy. For consistency, the results for the final FEET classifier were used in this work. Finally, the classification reliability of the system should be improved for FEET to be used within future clinical and research settings. The accuracy for TD participants was about 72.5% (Aly et al., submitted for publication), which may seem low. However, this accuracy level is only slightly lower than the threshold of 80% agreement that was required for human coders to be considered proficient. This study therefore raises questions concerning agreement between human coders of emotion, and the extent to which data provided by human coders should be treated as ground truth during the development of automated FER systems. Refinement of the machine-learning techniques and migration to next-generation 3D sensing technology are under way to improve the system's reliability. FEET was developed using the first-generation Kinect system, and we are actively investigating the second-generation Kinect system. Newer sensing platforms may also be considered, such as the 3D component of Apple's Face ID system.

These limitations notwithstanding, this is a preliminary study with both a well-characterized clinical sample as well as nonclinical sample, which allows for an in-depth examination of group differences in FEE using an innovative machine-based classification approach. The present study is the first to determine that differences in FEE, as identified through the use of machine-based classification, between youth with ASD and TD youth do indeed exist. The prototype FEET system has been found to be feasible to implement and generally acceptable to participants (White et al., 2018). Moreover, computerized systems, such as FEET, are low-cost and portable. A further advantage of the FEET system is that it can be packaged for use at home. Unlike robotic systems or traditional videos that are currently limited to research facilities or to clinics, we envision a future version of FEET providing interaction in other settings. Similar to today's interactive video games, such a system would be available at any hour, and as frequently as needed by the subject. Although the proposed system would not replace more traditional therapy, it could be used as an aid at home, automatically transferring information that it collects to therapists through secure channels over the Internet. The use of computerized systems like FEET also allows for classification and expression training in real time. Automated recognition of facial expressions is an active area of research, and offer the potential to perform both more efficiently and more accurately that human coding. In conclusion, group differences in FEE emerged using a novel, computerized system which made classifications of the dynamic expressions of youth with ASD and TD controls. These findings support the need for continued development of technologies for remediation of difficulties in FEE expressed by youth with ASD.

Conflict of Interest Statement

The authors have no conflicts of interest.

Human and Animal Rights and Informed Consent

All study procedures were approved by the institutional review board for human subject research. All participants provided informed consent.

References

Aly, S., Abbott, A. L., Wieckowski, A. T., White, S. W., & Youssef, A. E. (2019). Fast 3D facial expression recognition. Manuscript submitted for publication.

Aly, S., Trubanova, A., Abbott, L., White, S., & Youssef, A. (2015, May). VT-KFER: A Kinect-based RGBD+ time dataset for spontaneous and non-spontaneous facial expression recognition. *In Biometrics (ICB), 2015 International Conference on* (pp. 90–97). IEEE. https://doi.org/10.1109/ICB.2015.7139081

Arbogast, K. B., Kim, J., Loeb, H., Kuo, J., Koppel, S., Bohman, K., & Charlton, J. L. (2016). Naturalistic driving study of rear seat child occupants: Quantification of head position using a Kinect™ sensor. *Traffic Injury Prevention*, *17*(suppl), 168–174.

Behrmann, M., Thomas, C., & Humphreys, K. (2006). Seeing it differently: Visual processing in autism. *Trends in Cognitive Sciences*, *10*, 258–264. https://doi.org/10.1016/j.tics.2006.05.001

Bishara, S. E., Peterson, L. C., & Bishara, E. C. (1984). Changes in facial dimensions and relationships between the ages of 5 and 25 years. *American Journal of Orthodontics*, *85*, 238–252. https://doi.org/10.1016/0002-9416(84)90063-0

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cooney, R. E., Atlas, L. Y., Joormann, J., Eugène, F., & Gotlib, I. H. (2006). Amygdala activation in the processing of neutral faces in social anxiety disorder: Is neutral really neutral? *Psychiatry Research: Neuroimaging*, *148*, 55–59. https://doi.org/10.1016/j.pscychresns.2006.05.003

Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., . . . Davidson, R. J. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature Neuroscience*, *8*, 519–526. https://doi.org/10.1038/nn1421

Deschamps, P. K., Coppes, L., Kenemans, J. L., Schutter, D. J. L. G., & Matthys, W. C. H. J. (2015). Electromyographic responses to emotional facial expressions in 6–7 year olds with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *45*, 354–362. https://doi.org/10.1007/s10803-013-1890-z

Doi, H., Fujisawa, T. X., Kanai, C., Ohta, H., Yokoi, H., Iwanami, A., & Shinohara, K. (2013). Recognition of facial expressions and prosodic cues with graded emotional intensities in adults with Asperger syndrome. *Journal of Autism and Developmental Disorders*, *43*, 2099–2113. https://doi.org/10.1007/s10803-013-1760-8

Eack, S. M., Mazefsky, C. A., & Minshew, N. J. (2015). Misinterpretation of facial expressions of emotion in verbal adults with autism spectrum disorder. *Autism*, *19*, 308–315. https://doi.org/10.1177/1362361314520755

el Kaliouby, R., Picard, R., & Baron-Cohen, S. (2006). Affective computing and autism. *Annals of the New York Academy of Sciences*, *1093*, 228–248. https://doi.org/10.1196/annals.1382.016

Ekman, P., & Friesen, W. V. (1984). *Emotion facial action coding system (EM-FACS)*. San Francisco: University of California.

Evers, K., Steyaert, J., Noens, I., & Wagemans, J. (2015). Reduced recognition of dynamic facial emotional expression and emotion-specific response bias in children with an autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *45*, 1774–1784. https://doi.org/10.1007/s10803-014-2337-x

Faso, D. J., Sasson, N. J., & Pinkham, A. E. (2015). Evaluating posed and evoked facial expressions of emotion from adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *45*, 75–89. https://doi.org/10.1007/s10803-014-2194-7

Frazier, T. W., Strauss, M., Klingemier, E. W., Zetzer, E. E., Hardan, A. Y., Eng, C., & Youngstrom, E. A. (2017). A meta-analysis of gaze differences to social and nonsocial information between individuals with and without Autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*, 546–555. https://doi.org/10.1016/j.jaac.2017.05.005

Goldsmith, T. R., & LeBlanc, L. A. (2004). Use of technology in interventions for children with autism. *Journal of Early and Intensive Behavior Intervention*, *1*, 166–178. https://doi.org/10.1037/h0100287

Gordon, I., Pierce, M. D., Bartlett, M. S., & Tanaka, J. W. (2014). Training facial expression production in children on the autism spectrum. *Journal of Autism and Developmental Disorders*, *44*, 2486–2498. https://doi.org/10.1007/s10803-014-2118-6

Grossman, R. B. (2015). Judgments of social awkwardness from brief exposure to children with and without high-functioning autism. *Autism*, *19*, 580–587. https://doi.org/10.1177/1362361314536937

Grossman, R. B., & Tager-Flusberg, H. (2012). "Who said that?" Matching of low-and high-intensity emotional prosody to facial expressions by adolescents with ASD. *Journal of Autism and Developmental Disorders*, *42*, 2546–2557. https://doi.org/10.1007/s10803-012-1511-2

Huelke, D. F. (1998). An overview of anatomical considerations of infants and children in the adult world of automobile safety design. *Annual Proceedings/Association for the Advancement of Automotive Medicine*, *42*, 93–113.

Hobson, R. P., & Lee, A. (1998). Hello and goodbye: A study of social engagement in autism. *Journal of Autism and Developmental Disorders*, *28*, 117–126. https://doi.org/10.1023/A:1026088531558

Kasari, C., Sigman, M., Mundy, P., & Yirmiya, N. (1990). Affective sharing in the context of joint attention interactions of normal, autistic, and mentally retarded children. *Journal of Autism and Developmental Disorders*, *20*, 87–100. https://doi.org/10.1007/BF02206859

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, *59*, 809–816. https://doi.org/10.1001/archpsyc.59.9.809

Li, Y., Gong, S., & Liddell, H. (2003). Recognising trajectories of facial identities using kernel discriminant analysis. *Image and Vision Computing*, *21*, 1077–1086. https://doi.org/10.1016/j.imavis.2003.08.010

Lindner, J. L., & Rosén, L. A. (2006). Decoding of emotion through facial expression, prosody and verbal content in children and adolescents with Asperger's syndrome. *Journal of Autism and Developmental Disorders*, *36*, 769–777. https://doi.org/10.1007/s10803-006-0105-2

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012). *Autism diagnostic observation schedule (ADOS-2)* (2nd ed.). Torrence, CA: Western Psychological Services.

Loveland, K. A., Tunali-Kotoski, B., Pearson, D. A., Brelsford, K. A., Ortegon, J., & Chen, R. (1994). Imitation and expression of facial affect in autism. *Development and Psychopathology*, *6*, 433–444. https://doi.org/10.1017/S0954579400006039

Malawski, F., Kwolek, B., & Sako, S. (2014). Using Kinect for facial expression recognition under varying poses and illumination. *International Conference on Active Media Technology* (pp. 395–406). Springer, Cham. https://doi.org/10.1007/978-3-319-09912-5_33

Matsumoto, D., & Hwang, H. C. (2014). Judgments of subtle facial expressions of emotion. *Emotion*, *14*, 349–357. https://doi.org/10.1037/a0024356

Press, C., Richardson, D., & Bird, G. (2010). Intact imitation of emotional facial actions in autism spectrum conditions. *Neuropsychologia*, *48*, 3291–3297. https://doi.org/10.1016/j.neuropsychologia.2010.07.012

Ricciardi, L., Visco-Comandini, F., Erro, R., Morgante, F., Bologna, M., Fasano, A., ... Kilner, J. (2017). Facial emotion recognition and expression in Parkinson's Disease: An emotional mirror mechanism? *PLoS One*, *12*, e0169110

Rudovic, O., Lee, J., Dai, M., Schuller, B., & Picard, R. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *arXiv preprint arXiv:1802.01186*.

Rump, K. M., Giovannelli, J. L., Minshew, N. J., & Strauss, M. S. (2009). The development of emotion recognition in individuals with autism. *Child Development*, *80*, 1434–1447. https://doi.org/10.1111/j.1467-8624.2009.01343.x

Seddik, B., Maâmatou, H., Gazzah, S., Chateau, T., & Ben Amara, N. E. (2013). Unsupervised facial expressions recognition and avatar reconstruction from kinect. *Proceedings: 10th International Multi-Conference on Systems, Signals & Devices (SSD)*, Hammamet.

Shalom, D. B., Mostofsky, S. H., Hazlett, R. L., Goldberg, M. C., Landa, R. J., Faran, Y., ... Hoehn-Saric, R. (2006). Normal physiological emotions but differences in expression of conscious feelings in children with high-functioning autism. *Journal of Autism and Developmental Disorders*, *36*, 395–400. https://doi.org/10.1007/s10803-006-0077-2

Soken, N. H., & Pick, A. D. (1999). Infants' perception of dynamic affective expressions: do infants distinguish specific expressions? *Child Development*, *70*, 1275–1282. https://doi.org/10.1111/1467-8624.00093

Spezio, M. L., Adolphs, R., Hurley, R. S., & Piven, J. (2007). Abnormal use of facial information in high-functioning autism. *Journal of Autism and Developmental Disorders*, *37*, 929–939. https://doi.org/10.1007/s10803-006-0232-9

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Allyn & Bacon/Pearson Education.

Uljarevic, M., & Hamilton, A. (2013). Recognition of emotions in autism: A formal meta-analysis. *Journal of Autism and Developmental Disorders*, *43*, 1517–1526. https://doi.org/10.1007/s10803-012-1695-5

Wang, A. T., Dapretto, M., Hariri, A. R., Sigman, M., & Bookheimer, S. Y. (2004). Neural correlates of facial affect processing in children and adolescents with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*, 481–490. https://doi.org/10.1097/00004583-200404000-00015

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II)*. San Antonio, TX: NCS Pearson.

White, S. W., Abbott, L., Wieckowski, A. T., Capriola-Hall, N. N., Aly, S., & Youssef, A. (2018). Feasibility of automated training for facial emotion expression and recognition in Autism. *Behavior Therapy*, *49*(6), 881–888.

Wieckowski, A. T., & White, S. W. (2017a). Application of technology to social communication impairment in childhood and adolescence. *Neuroscience & Biobehavioral Reviews*, *74*, 98–114. https://doi.org/10.1016/j.neubiorev.2016.12.030

Wieckowski, A. T., & White, S. W. (2017b). Eye-gaze analysis of facial emotion recognition and expression in adolescents with ASD. *Journal of Clinical Child & Adolescent Psychology*, *46*, 110–124. https://doi.org/10.1080/15374416.2016.1204924

Yirmiya, N., Kasari, C., Sigman, M., & Mundy, P. (1989). Facial expressions of affect in autistic, mentally retarded and normal children. *Journal of Child Psychology and Psychiatry*, *30*, 725–735. https://doi.org/10.1111/j.1469-7610.1989.tb00785.x

Yoshimura, S., Sato, W., Uono, S., & Toichi, M. (2015). Impaired overt facial mimicry in response to dynamic facial expressions in high-functioning autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *45*, 1318–1328. https://doi.org/10.1007/s10803-014-2291-7

Zane, E., Neumeyer, K., Mertens, J., Chugg, A., & Grossman, R. B. (2018). I think we're alone now: Solitary social behaviors in adolescents with autism spectrum disorder. *Journal of Abnormal Child Psychology*, *46*, 1111–1120. https://doi.org/10.1007/s10802-017-0351-0

Received: May 21, 2018
Accepted: December 18, 2018
Available online: 21 December 2018