

METHODS AND MEASURES

Non-expert ratings of infant and parent emotion: Concordance with expert coding and relevance to early autism risk

Jason K. Baker^a, John D. Haltigan^a, Ryan Brewster^a, James Jaccard^b, and Daniel Messinger^a

This study investigated a novel approach to obtaining data on parent and infant emotion during the Face-to-Face/Still-Face paradigm, and examined these data in light of previous findings regarding early autism risk. One-hundred and eighty eight non-expert students rated 38 parents and infant siblings of children who did (20) or did not (18) have autism spectrum disorders. Ratings averaged across 10 non-experts exhibited high concordance with expert facial-action codes for infant emotion, and 20 non-experts were required for reliable parent ratings. Findings replicated the well-established still-face effect and identified subtle risk associations consonant with results from previous investigations. The unique information offered by intuitive non-expert ratings is discussed as an alternative to complex and costly behavioral coding systems.

Keywords: autism; emotion; infancy; measurement; observation; parent–child interaction; rating; still-face

Much of children's social and emotional adjustment takes root in the context of early parent–infant interaction (Feldman, 2007). The emotional expression of both parties plays a central role in these exchanges – infant positive and negative expressions soliciting engagement and/or assistance from parents, and parent expression providing comfort to infants and scaffolding their regulatory efforts (see Messinger & Fogel, 2007; Weinberg, Tronick, Cohn, & Olson, 1999). Data on infant and parent emotional expression during face-to-face interactions have informed our understanding of the emergence of children's attachment (Braungart-Rieker, Garwood, Powers, & Wang, 2001; Cohn, Campbell, & Ross, 1991), self-regulation (Feldman, Greenbaum, & Yirmiya, 1999), cognitive development (Feldman, Greenbaum, Yirmiya, & Mayes, 1996), prosociality (Kochanska, Aksan, & Carlson, 2005), empathy (Kochanska, Forman, & Coy, 1999), and behavior problems (Moore, Cohn, & Campbell, 2001). The present study evaluated a novel method for obtaining information on parent and infant emotion using ratings generated by non-experts. These data were then used to replicate findings relating to early autism risk.

Coding of infant and parent emotion

Obtaining efficient, replicable measurement of ongoing emotion-related behavior is a chronic difficulty for students of interaction. Attempts to operationalize the measurement of

emotion have led to the development of complex behavioral coding systems. The premise of these systems is that facial expressions and related actions provide clues to specific emotional states. The Facial Action Coding System (FACS; Ekman & Friesen, 1978) is a system for coding facial behavior that was created for adults, and has been adapted for use with infants (BabyFACS; Oster, 2000). FACS allows for the identification of facial Action Units (AUs) that index smiles and cry faces – prototypical expressions of positive and negative infant emotion (Camras, 1992; Messinger & Fogel, 2007). Information generated by FACS and similar behavioral coding systems (e.g., AFFEX – Izard & Dougherty, 1980) represents a significant contribution to the literature on infant emotion and communication (see Messinger & Fogel, 2007). However, the substantial training and certification required to use these systems, and the complexity of the coding itself, require a significant commitment of time and effort. These costs may reduce wide-spread adoption of these systems across laboratories, increasing reliance upon investigator-specific systems and limiting research performed in this important area.

Non-expert ratings as an alternative to expert coding

In creating observational measurement systems, it is generally assumed that coders and raters must learn to ignore subjective or “intuitive” information in favor of concrete observations that can be clearly agreed upon by multiple observers. However, an important line of research suggests that non-expert ratings can contribute greatly to the field of observational measurement

^a University of Miami, USA. ^b Florida International University, USA.

(Ambady & Rosenthal, 1992; Gottman & Levenson, 1985), with the predictive ability of these ratings leading some to suggest that non-experts possess “intuitive expertise” in certain domains (Waldinger, Schulz, Hauser, Allen, & Crowell, 2004). Few areas would seem to benefit more from human intuition than the assessment of emotion, as much of the functional significance of emotional expression lies in how it is perceived by others (Campos, Mumme, Kermoian, & Campos, 1994; Dinehart et al., 2005). The option of obtaining reliable, valid ratings of complex parent and infant emotional behavior from non-experts would simultaneously circumvent the need for intensive coder training while also capitalizing on the potential benefits of intuitive emotion expertise.

Although it is unlikely that individual ratings from a single non-expert could adequately replace expert coding, aggregating across multiple raters minimizes error, increasing the accuracy of measurement (Larrick & Soll, 2006). Indeed, Waldinger et al. (2004) found that ratings of marital interaction averaged across a small group of non-experts showed correlations with expert ratings at values as high as .95 (e.g., for marital hostility). The present study attempted to tap the so-called *wisdom of crowds* phenomenon (Surowiecki, 2004) by averaging across multiple non-expert ratings.

The Face-to-Face Still-Face as a measurement paradigm

Merging naturalistic and experimental methods, the Face-to-Face/Still-Face (FFSF – Tronick, Als, Adamson, Wise, & Brazelton, 1978) has emerged as one of the most widely used paradigms in the study of early parent–infant interaction (Adamson & Frick, 2003). In the FFSF, parents are instructed to play with their infants as they typically would, and then to hold a motionless face and to refrain from any communicative acts. Across this transition, infants generally become less positive and more negative – the well-established “still-face effect” (Adamson & Frick, 2003; Mesman, van IJzendoorn, & Bakermans-Kranenburg, 2009; Tronick et al., 1978). Following this period of unresponsiveness, the parent is asked to re-engage with the infant, at which time infants typically exhibit a partial rebound of positive emotional expression and carry-over of negative emotion, often labeled the “reunion effect” (Adamson & Frick, 2003; Mesman et al., 2009; Weinberg & Tronick, 1996). Use of the FFSF has been wide-spread, informing issues related to group differences in children differing in developmental risk (e.g., Bendersky & Lewis, 1998) and predicting important child outcomes into the preschool period (Bates, Maslin, & Frankel, 1985; Braungart-Rieker et al., 2001; Cohn et al., 1991; Moore et al., 2001).

Continuous measurement system and the current study

The current study examined the validity of ratings of parent and child emotional valence during the FFSF as performed by non-expert college students who rated using the Continuous Measurement System (CMS). The primary aim of the current study was to compare non-expert ratings of infant and parent emotion in the FFSF with the measurements of certified FACS coders. We addressed validity further by examining the potential of non-expert ratings to replicate findings from studies that utilized relevant expert coding. First, we examined whether non-expert ratings produced the well-established and aforementioned “still-face” and “reunion effects”. We then used this

same dataset to examine the ability of non-expert ratings to distinguish between two groups of children differing in a form of developmental risk particularly relevant to the construct of interest, as discussed below.

Children at-risk for autism and early emotional communication

Autism spectrum disorders are neurodevelopmental disorders involving core impairment in social functioning and communication (Landa, Holman, & Garret-Mayer, 2007). Social and emotional communication deficits during the first year of life in children later diagnosed with autism spectrum disorders have been noted through retrospective and, more recently, prospective methods (Osterling, Dawson, & Munson, 2002; Zwaigenbaum et al., 2005). Prospective, longitudinal studies include younger siblings of children diagnosed with autism spectrum disorders as a high-risk group with whom to study early autism-related deficits (see Zwaigenbaum et al., 2007). Compared to infant siblings of typically developing controls, infant siblings of individuals with autism spectrum disorders are at increased risk both for developing autism spectrum disorders and also, some suggest, for milder deficits in social-communicative functioning (Yirmiya & Ozonoff, 2007). The presence of subtle emotional communication deficits during the FFSF has been identified in infant siblings of children with autism spectrum disorders as a group (Yirmiya et al., 2006), but another study reported no differences during a similar but abbreviated paradigm (Merin, Young, Ozonoff, & Rogers, 2007).

A previous report from our laboratory, utilizing a large portion (82%) of the present sample, examined FACS-coded smiling and cry-faces during the FFSF and found a tendency for 6-month-old siblings of children with autism spectrum disorders to smile for a lower proportion of the overall task than siblings of typically-developing children, although the significance of this effect differed by the analytical approach used (Cassel et al., 2007). The present study attempted to replicate autism risk group findings in our expanded sample using non-expert ratings rather than the aforementioned FACS coding.

Hypotheses

It was predicted that non-expert ratings of positive emotion would correlate with expert FACS-based coding of smiling for infants and parents in each episode of the FFSF, despite the fact that these systems measure somewhat different constructs. Similarly, we expected that non-expert ratings of negative emotion would relate strongly to cry-faces for infants. Non-expert ratings were predicted to exhibit patterns reflecting still-face and reunion effects. Finally, it was anticipated that group differences would emerge for non-expert ratings in which at-risk siblings would exhibit subtle deficits in adaptive emotional communication.

Method

Participants

Our initial sample consisted of 39 parents and their 6-month-old infants. However, data were dropped for one child (see below), resulting in a final sample of 38. Comparison siblings

(Comparison-sibs) were infants whose older sibling(s) had no diagnosis of an autism spectrum disorder and showed no evidence of heightened autism symptomatology. Autism-sibs had at least one sibling who was diagnosed with Autism, Asperger's Disorder, or Pervasive Developmental Disorder – Not Otherwise Specified. Independent community diagnoses for the older siblings with autism spectrum disorders were confirmed for this study via record review by a clinical psychologist and performance of the Autism Diagnostic Observation Schedule (ADOS; Lord, Rutter, DiLavore, & Risi, 1999). Complete data regarding any eventual autism spectrum diagnoses in the target (younger) siblings are not yet available. Infants were included in this sample if they were at least 36 weeks gestation at birth, and had a birth weight above 2500g. Due to persistent distress, data collection was terminated for one of the Autism-sibs (male) during the FFSF. Excluding the data from this dyad yielded a final sample of $N = 38$ dyads, with 18 Comparison-sibs and 20 Autism-sibs.

The mean age of the 38 infants at the 6-month assessment was 6.1 months ($SD = .3$; range 5.1 to 6.9 months) and did not differ by group ($p > .71$). Nine of the Comparison-sibs, and seven of the Autism-sibs were female, and ethnicity of the total group of infants as reported by parents was 39% Caucasian, 39% Hispanic, 6% Asian, 4% Black, and 12% Other. Mean parent age was 34.92 years ($SD = 4.58$; range = 23–47 years). A total of 58% of the mothers and 42% of the fathers reported earning an advanced or professional degree and another 21% of mothers and 32% of fathers reported completing a 4-year college degree. There were no group differences with regard to parent age, parent education, child ethnicity, family income, or child gender between the Autism-sib and the Comparison-sibs samples. Of the 38 parents interacting with their infants, two were fathers and 36 were mothers. The families of the two fathers identified these men as primary caregivers for their infants.

Procedure

At the 6-month assessment, infant-parent dyads completed the Face-to-Face/Still-Face Protocol (Tronick et al., 1978). Parents were asked to play with their baby without toys for three minutes (Face-to-Face), stop playing and maintain a still face with no emotional expression for two minutes (Still-Face), and resume play for another three minutes (Reunion). Infants were placed in an elevated car seat and their parents were positioned on a small chair opposite to them. One time-synched camera was used to record the face and upper body of the infant; the second time-synched camera recorded the face and upper-body of his or her parent. Each episode of the FFSF for both infant and parent was then separately exported to a video file for rating.

Non-expert student raters

Ratings of the 38 infants and parents were completed by 188 non-expert student raters at a large urban university in the Southeast in fulfillment of the research component of an introductory psychology course. The raters were non-experts in that they had no specialized training in coding emotion. A given student rated either infants or parents (not both). The mean age of raters was 19.56 years ($SD = 2.51$; range = 17–42 years), 28% were male, and ethnicity was reported as 54% Caucasian, 24% Hispanic, 11% Black, 7% Asian, and 4% Other.

Continuous Measurement System

The Continuous Measurement System (available for download, free of charge, at <http://measurement.psy.miami.edu/cms.phtml>; Messinger, Mahoor, Chow, & Cohn, 2009) enables the direct, real-time measurement of individual perceptions of human behavior. Video files were presented to raters in a random sequence and raters used a joystick to move the cursor up or down along a graduated color bar adjoining the right margin of the picture frame where the video was shown. Videos were presented without audio because the parent and infant videos contained the same audio track. Raters used the joystick to measure emotional valence as a continuous dimension from positive emotion to negative emotion in response to the following instructions (the only training they received): "Please use this joystick to rate emotion. Ratings above the tic mark indicate positive emotion (joy, happiness, pleasure). Ratings below the mark indicate negative emotion (distress, sadness, anger)." Upon playback of the video file, the Continuous Measurement System captured rating data for every frame of the video. Final emotion ratings for each episode, for each rater, were represented by the average rating above the "neutral" tic mark (positive emotion mean) or below it (negative emotion mean) for each episode. When averaging, frames that included a rating of the opposite valence were treated as zero. Though this procedure truncated the variability in the original emotion ratings and may have increased reliability among non-experts, conversion to two scales allowed for more construct similarity between non-expert ratings and FACS coding.

Each non-expert rated a batch of approximately eight infants or parents (mean = 7.60, $SD = 3.51$; range = 4–11), containing separate video clips for each episode of the FFSF (i.e., Face-to-Face, Still-Face, and Reunion). Each episode for each infant was rated by approximately 18 non-expert raters (mean = 18.24, $SD = 2.44$; range 13–20), and each parent episode was rated by approximately 19 raters (mean = 19.17, $SD = 2.49$; range 13–21). All rating was done in real-time, requiring two to three minutes per episode.

FACS coding of infants and parents

The onset and offset of infants' smiles and cry-faces in the FFSF were coded by graduate students certified in the FACS system (Ekman & Friesen, 1978) and trained in its application to infants, BabyFACS (Oster, 2000). Training included study of the FACS manual (>500 pages) and the BabyFACS manual, several months of practice toward the FACS final reliability test, and attendance at a BabyFACS training workshop. Coding time averaged approximately 30 minutes for each infant episode, and 20 minutes for each parent episode. In FACS, coding of smiles occurs when the lip corners are pulled diagonally upward by the zygomaticus major (AU12). In cry-faces, the lips are stretched laterally by the risorius muscle (AU20) and the brows are lowered by the corrugator muscle (AU4). Twenty-nine percent of infants were coded by separate certified FACS coders for reliability with mean 86% agreement between coders (Cohen's Kappa [κ], which corrects for chance agreement = .69). Parent smiles (AU12) were also coded using FACS. Approximately 16% of the parent video clips were randomly selected and coded by two coders with a mean agreement of 84% ($\kappa = .68$). Reliability was calculated in time, reflecting agreement and disagreement for each moment of

observation. The proportion of each episode that the individual exhibited each of the facial configurations (a proportional duration measure) was calculated for use in the analyses.

Results

This section begins with an examination of inter-rater reliability among the non-expert raters followed by the central analyses investigating the concordance between mean non-expert ratings and measurements by the FACS-certified coders (examined with Pearson's correlations). Three separate 2 (Sibling Group) \times 3 (FFSF Episode) repeated-measures designs (one for positive infant emotion, one for negative infant emotion, and one for positive parent emotion) were then examined in order to investigate hypotheses related to the still-face and reunion effects and to examine differences in the FFSF between infants of siblings with and without autism spectrum disorders. Following procedures outlined by Jaccard (1998), planned single degree-of-freedom contrasts were prioritized over the examination of omnibus F tests. We conclude the section by asking whether smaller numbers of non-expert raters produce effects similar to the full complement of raters.

Reliability among non-expert raters

Average-measures intra-class correlations were conducted for the non-expert ratings. These reflect the reliability of the raters if the mean of all the relevant raters were to be used as data – as was the case for the present study. Average reliability (weighted by the number of episodes rated in a given batch) for the infant ratings were .91 for Face-to-Face, .92 for the Still-Face, and .98 for the Reunion, all reflecting very high reliability. Reliability for parent ratings were .70 for the Face-to-Face, .76 for the Still-Face, and .75 for the Reunion, reflecting acceptable reliability.

Concordance between non-expert ratings and expert FACS codes

As seen in Table 1, high correlations existed between non-expert raters and expert coders for variables relating to infant negative emotion across all three episodes, and for those relating to positive emotion in the Face-to-Face and Reunion. Concordance was slightly lower (at the high end of “moderate”) between non-expert-rated infant positivity and expert-coded smiles in the Still-Face. Associations at the moderate to high range were found between non-expert-rated parent positive emotion and expert-coded parent smiles.

Table 1

Correlations between non-expert ratings and relevant Facial Action Coding System (FACS) expert codes

	<i>Face-to-Face</i>	<i>Still-Face</i>	<i>Reunion</i>	<i>Across all episodes</i>
Non-expert infant positive emotion and FACS smiling	.83**	.59**	.84**	.85**
Non-expert infant negative emotion and FACS cry faces	.99**	.86**	.80**	.86**
Non-expert parent positive emotion and FACS smiling	.64**	.64**	.70**	.77**

Note. Correlations were based on 38 infants and 38 parents. ** $p < .01$.

Examination of the “Still Face” and “Reunion Effects”

It was hypothesized that positive infant emotion would be highest in the Face-to-Face, decrease in the Still-Face, and rebound to intermediate levels in the Reunion. Negative infant emotion was expected to be lowest in the Face-to-Face as compared to the Still-Face and Reunion episodes. As discussed above, two infant 2 \times 3 repeated-measures designs were examined in which sibling group status was the between-subjects factor, and episode (Face-to-Face, Still-Face, Reunion) served as the within-subjects factor. Infant gender did not significantly differ by status group and did not interact with variables of interest in preliminary analyses, so it was not controlled for in these analyses.

Planned contrasts supported predictions that positive emotion would be lower in the Still-Face (mean = 14.71, $SD = 16.11$) than in the Face-to-Face (mean = 56.67, $SD = 28.91$), $t(37) = 10.60$, $p < .01$, $\eta_p^2 = .75$, and higher in the Reunion (mean = 35.84, $SD = 29.74$) than in the Still-Face, $t(37) = 5.86$, $p < .01$, $\eta_p^2 = .48$. As predicted, this rebound was only partial, with positive emotion remaining lower in the Reunion than in the Face-to-Face, $t(37) = 4.52$, $p < .01$, $\eta_p^2 = .36$. Contrasts supported the hypothesis that negative emotion would be higher in the Still-Face (mean = 41.66, $SD = 82.16$) and in the Reunion ($M = 46.07$, $SD = 91.35$) than in the Face-to-Face (mean = 11.89, $SD = 57.32$), $t(37) = 2.92$, $p < .01$, $\eta_p^2 = .19$, and $t(37) = 2.86$, $p < .01$, $\eta_p^2 = .18$, respectively.

Examination of risk differences

No significant group-episode interaction effects were present for the above infant still-face and reunion-effect analyses, and groups did not differ significantly on indices of infant emotion for the FFSF as a whole. Interestingly, unequal variances were observed as a function of status group. Autism-sibs as a group exhibited more negative variability in the Still-Face, Levene's $F(1,36) = 9.35$, $p < .01$, and less positive variability in both the Still-Face $F(1,36) = 9.91$, $p < .01$, and the Reunion $F(1,36) = 8.97$, $p < .01$, than did Comparison-sibs (see Figure 1). Planned contrasts utilizing non-pooled error terms (with equal variances not assumed) revealed no significant group differences in positive emotion during the Face-to-Face or the Reunion but indicated that Autism-sibs (mean = 8.54, $SD = 10.37$) were rated significantly lower in positive emotion than were Comparison-sibs (mean = 21.58, $SD = 18.69$) in the Still-Face episode, $t(25.94) = 2.62$, $p = .01$, Cohen's $d = .86$.

Parent emotion in the Face-to-Face/Still-Face

Parent ratings were also examined in order to provide a manipulation check of the FFSF and to examine whether parent

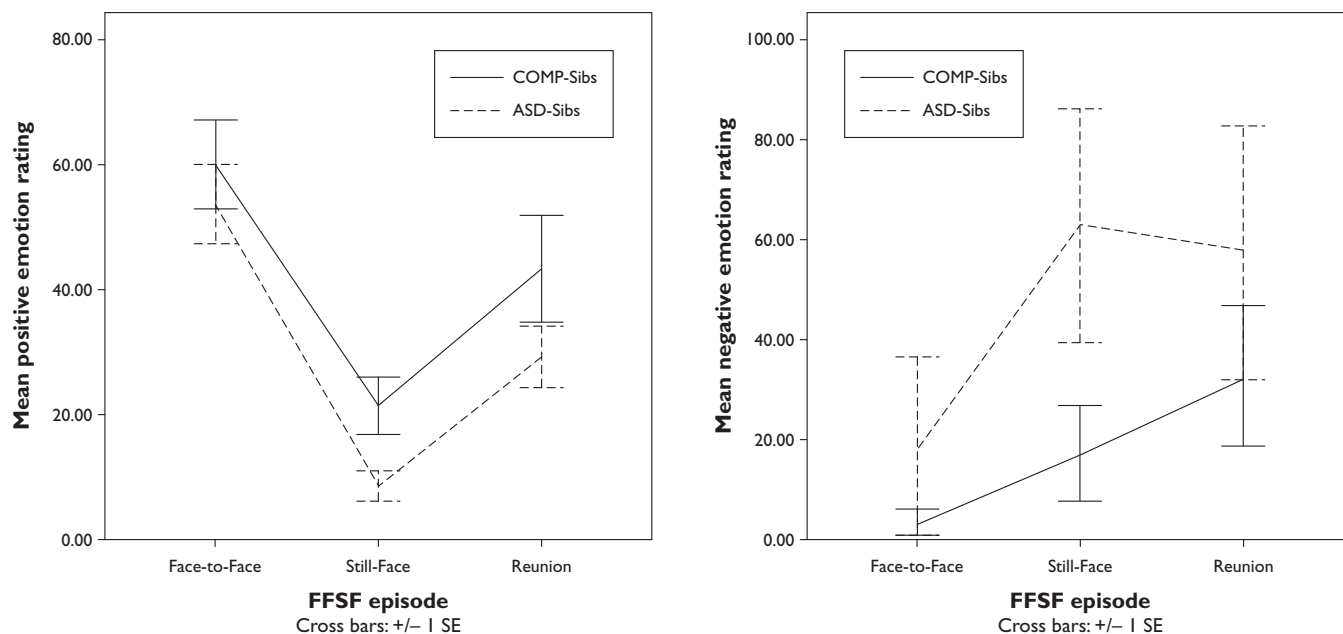


Figure 1. Non-expert positive and negative emotion ratings across Face-to-Face/Still-Face (FFSF) episodes, by risk group.

positive emotion during the overall FFSF differed by Sibling-Group. Planned contrasts supported the validity of the non-expert ratings of the Still-Face manipulation in that parents were rated lower in the Still-Face than in the Face-to-Face, $t(37) = 10.86, p < .01, \eta_p^2 = .76$, and in the Reunion, $t(37) = 14.98, p < .01, \eta_p^2 = .59$. No significant differences existed on parent emotion during the overall FFSF as a function of Sibling-Group.

Reliability and concordance with expert codes for subsamples of non-expert raters

To further examine the potential efficiency of non-expert measurement, infant data were analyzed using random subsamples of 15, 10, and 5 raters. Reliability remained acceptable for all subsamples ($>.70$). Correlations of non-expert infant negative emotion with FACS coding were all within .02 points of the original correlations. Correlations of non-expert infant positive emotion with FACS coding were within .07 points of the original correlations using the 15- and 10-rater subsamples, but dropped substantially when only five raters were utilized (i.e., Face-to-Face = .75, Still-Face = .30, Reunion = .73). Findings derived from all subsamples of raters relating to episode effects and risk group differences were essentially identical to those generated from the full sample (e.g., risk group difference in positive emotion during the Still-Face using five raters, $t(36) = 2.37, p < .05, d = .77$).

For ratings of parent emotion, reliability with a set of 15 non-expert raters was at the threshold of acceptability (i.e., .70), with average intra-class correlations ranging from .68 for the Face-to-Face to .72 for the Still-Face (across all episodes = .70). Concordance with expert codes were somewhat lower than those obtained with the full set of 20 non-expert raters (range of correlations = .63 to .69). Reliability for a set of 10 non-expert raters of parent emotion fell below the level of acceptability (average intra-class correlation = .59), and correlations with expert codes ranged from .60 to .67. Nevertheless, episode effects (and the lack of parent risk group differences)

examined with scores obtained from 15 or 10 raters were essentially identical to those generated from the full sample.

Discussion

Early parent-infant emotional communication has been linked to a host of important child outcomes (Braungart-Rieker et al., 2001; Cohn et al., 1991; Feldman et al., 1996, 1999; Kochanska et al., 1999, 2005). Complex systems for coding emotional communication are effective and well validated, but certification in these systems, maintenance of reliability, and the coding itself can be costly and time consuming. Findings from the present study suggest one alternative to expert behavioral coding – the aggregated ratings of multiple non-expert raters.

Non-experts in the present study appeared particularly adept at rating infant emotion during the FFSF. Mean non-expert ratings of infant emotion showed high reliability, good concurrent validity with expert codes, and resulted in findings consistent with relevant literatures. This suggests that valid assessment of infant emotion can be obtained by aggregating the intuitive ratings of multiple non-experts. Ultimately, the success of non-experts in rating infant emotion also underscores the impressive capacity of infants to effectively communicate emotion to adults (Messinger & Fogel, 2007; Oster, 2005). Sensitive care depends upon the ability of caregivers to comprehend infant cues – many of which are expressed emotionally and understood intuitively.

Assessment of parent emotion does not appear to be as straightforward as assessment of infants. Although correlations between non-expert parent ratings and expert coding fell within the high end of the “moderate” range, associations were not as high as those for infant emotion. Indeed, while approximately 10 raters were able to effectively rate infant emotion, approximately 15 to 20 raters were necessary to achieve adequate reliability for parent ratings – the exact number depending upon whether or not individual scores by episode were considered. One obvious explanation is that adult

emotion is often less intense and more nuanced than that of infants. Raters may intuit “positive emotion” in parents not only from simple smiles or laughs but also from warmth, sensitivity, a loving glance, and/or subtle positive regard. These more subjective constructs may contribute important information, but there are also instances when more objective measurement of parent and infant emotion-related expressions is desired. The use of expert systems such as FACS for measuring facial expression may also be particularly useful when attempting to answer questions regarding the complex interplay of multiple events in time (e.g., parent–child interactive responsiveness, and associations between facial actions, other behaviors, and physiological measures).

Evidence for the effectiveness of non-expert ratings of emotion highlights the potential for application of this type of rating to additional constructs. Expert behavioral coding is used to measure a multitude of personal and interpersonal factors in the psychological sciences, but many coding systems require a great deal of training and expertise. The Continuous Measurement System is flexible, allowing researchers to select their own constructs for measurement and to enter relevant anchors for rating. The range of constructs for which non-expert ratings might prove useful is limited primarily by the degree to which factors can be recognized intuitively and efficiently. Accurate recognition likely depends upon the nature and complexity of the construct in question. Future studies might determine, for example, whether measurement is more challenging when a rater must simultaneously consider the behavior of multiple people (e.g., family conflict) or behaviors expressed through multiple modalities (e.g., parental sensitivity).

Although the present study was focused primarily on measurement, findings regarding autism risk and the Face-to-Face/Still-Face paradigm are noteworthy. The patterned “still-face” and “reunion” effects that provided evidence for validity in the current study have been described extensively in previous investigations with low-risk (see Mesman et al., 2009) and Autism-sib infants (Yirmiya et al., 2006), and are not re-interpreted here. Risk-group findings in the present study were largely consistent with the findings of Cassel et al. (2007), in that Autism-sibs were rated significantly less positive in the Still-Face than were Comparison-sibs. It has been proposed that infant positive expression in the Still-Face, particularly during the initial portion of the episode, may function as an attempt to re-engage the unresponsive caregiver (Adamson & Frick, 2003; Mesman et al., 2009; Tronick, Ricks, & Cohn, 1982). The present findings suggest that Autism-sibs as a group may engage in less positive bidding of this sort as compared to Comparison-sibs. It is important to note that we did not correct for Type I error and that, despite its large effect size, the risk-group difference would no longer be considered significant ($p < .006$) if family-wise modified Bonferroni corrections were applied (Jaccard & Guilamo-Ramos, 2002). Given the focus in the present study on methodology and replication, it is important to note that Cassel et al. (2007) did not apply a correction for Type I error, instead prioritizing the reduction of Type II error so that potential early-identifiers of autism were not missed. Findings from the present study using non-expert ratings are therefore quite similar to those derived from expert facial-action coding.

The present study’s results also suggest that Autism-sibs as a group may exhibit a truncated range of positive expressivity, and more variability in negative emotion than Comparison-sibs

within the context of the still-face stressor. These differences in variances suggest that Autism-sibs as a group may be fairly heterogeneous in negative reactivity and/or regulation ability (as might be expected based on the assumption that only a portion of these children will go on to autism spectrum diagnoses), but more *homogeneous* in their (lower) expression of positive emotion – suggesting the presence of subclinical affective deficits in the Autism-sib group as a whole. No meaningful differences were identified in parent emotion between risk groups, suggesting that, in brief interactive protocols, the emotional behavior of parents of children with autism may be indistinguishable from other parents as they interact with a younger infant sibling. This is consistent with evidence that parents interacting with their children with autism exhibit levels of synchrony (Siller & Sigman, 2002) and sensitivity (van IJzendoorn et al., 2007) similar to those of parents of developmentally-matched children who do not have autism.

Limitations, future directions, and conclusions

Being one of the first of its kind, there are several ways the current study could be improved and expanded. First, the fact that expert and non-expert raters were asked to assess somewhat different indicators of “emotion” is both a limitation and a strength. That is, FACS is not designed to measure emotion per se, and the non-expert raters were not instructed to record smiles and cry faces. Although concordance for infants in particular was nonetheless very high, it would be interesting to know the degree to which convergence would be higher if both sets of individuals were asked to measure identical behaviors. Indeed, it might be interesting to ask experts to attempt to code in real-time using the Continuous Measurement System – although that would not eliminate the extensive training requirements required of FACS coders. Second, although non-expert infant ratings were transformed into separate positive and negative scores, raters actually recorded emotional valence on a single continuum. It is possible that instances of mixed and/or rapidly shifting emotion (e.g., infant frustration while positively bidding in the Still-Face – Adamson & Frick, 2003; parent empathy while attempting to infuse positivity), may have been more difficult to capture with the Continuous Measurement System. Future studies might employ independent rating of positive and negative emotion. Finally, rating in the absence of audio information was necessary to separate parent from infant data, so future studies that record separate audio tracks for each member of the dyad might be of interest.

The present study utilized overall episode means for non-expert ratings. Nevertheless, a powerful contribution of the Continuous Measurement System is the ability to document ratings in real time. Separate time-based data for parents and infants can be combined in order to answer complex questions about dyadic transactions and interactive influence (Chow, Haltigan, & Messinger, in press). The use of the system in this manner was beyond the scope of this initial study, but represents a fruitful avenue for future research. Indeed, the Continuous Measurement System is proving useful in facilitating *expert* rating and coding because of its capacity to record and manage time-based data (Messinger et al., 2009).

On a substantive level, eventual diagnostic data on the Autism-sibs will likely elucidate the findings of the current investigation and similar Autism-sib studies (Zwaigenbaum et

al., 2007). For example, it is possible that risk-group differences might be completely accounted for by those infants who go on to autism spectrum diagnoses. However, the relatively truncated variation of the Autism-sibs' positive emotion scores in the Still-Face suggests less heterogeneity in this group as compared to Comparison-sibs – arguing against the possibility of outliers driving group differences in positive emotion. Finally, collection of more detailed background information on non-expert raters (e.g., how many of them are parents, have experience with children, and/or are familiar with autism), and replication with a non-university sample, would be beneficial.

The present study provides support for the notion that non-expert ratings can contribute to the observational study of parents and children and introduces a method for harnessing this capacity. In addition to its contribution to research reliability and efficiency, the use of non-expert ratings can also teach us important things about the nature of the specific constructs under consideration. For example, non-expert data in the present study informs our understanding of emotion recognition and contributes to an important and challenging line of research focused on the developmental understanding of autism risk and resilience. We hope that researchers from diverse areas of study will consider obtaining non-expert ratings both as a means of assessing the face validity of chosen constructs, and as a substantive tool for understanding those constructs more completely.

References

- Adamson, L.B., & Frick, J.E. (2003). The still face: A history of a shared experimental paradigm. *Infancy, 4*, 451–473.
- Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 2*, 256–274.
- Bates, J.E., Maslin, C.A., & Frankel, K.A. (1985). Attachment security, mother-child interaction, and temperament as predictors of behavior-problem ratings at age three years. *Monographs of the Society for Research in Child Development, 50*, 167–193.
- Bendersky, M., & Lewis, M. (1998). Arousal modulation in cocaine-exposed infants. *Developmental Psychology, 34*, 555–564.
- Braungart-Rieker, J., Garwood, M.M., Powers, B.P., & Wang, X. (2001). Parental sensitivity, infant affect, and affect regulation: Predictors of later attachment. *Child Development, 72*, 252–270.
- Campos, J.J., Mumme, D.L., Kermoian, R., & Campos, R.G. (1994). A functionalist perspective on the nature of emotion. *Monographs of the Society for Research in Child Development, 59*, 284–303.
- Camras, L.A. (1992). Expressive development and basic emotions. *Cognition & Emotion, 6*, 269–283.
- Cassel, T., Messinger, D.S., Ibanez, L., Haltigan, J.D., Acosta, S., & Buchman, A. (2007). Early social and emotional communication in the infant siblings of children with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders, 37*, 122–132.
- Chow, S., Haltigan, J.D., Messinger, D.S. (in press). Dynamic infant-parent affect coupling during the Face-to-Face/Still-Face. *Emotion*.
- Cohn, J.F., Campbell, S.B., & Ross, S. (1991). Infant response in the still-face paradigm at 6 months predicts avoidant and secure attachment at 12 months. *Development and Psychopathology, 3*, 367–376.
- Dinehart, L.H.B., Messinger, D.S., Acosta, S.L., Cassel, T., Ambadar, Z., & Cohn, J. (2005). Adult perceptions of positive and negative infant emotional expressions. *Infancy, 8*, 279–303.
- Ekman, P., & Friesen, W. (1978). *The facial action coding system*. Palo Alto, CA: Consulting Psychologists Press.
- Feldman, R. (2007). Parent-infant synchrony: Biological foundations and developmental outcomes. *Current Directions in Psychological Science, 16*, 340–345.
- Feldman, R., Greenbaum, C.W., & Yirmiya, N. (1999). Mother-infant affect synchrony as an antecedent of the emergence of self-control. *Developmental Psychology, 35*, 223–231.
- Feldman, R., Greenbaum, C.W., Yirmiya, N., & Mayes, L.C. (1996). Relations between cyclicity and regulation in mother-infant interaction at 3 and 9 months and cognition at 2 years. *Journal of Applied Developmental Psychology, 17*, 347–365.
- Gottman, J.M., & Levenson, R.W. (1985). A valid measure for obtaining self-report of affect. *Journal of Consulting and Clinical Psychology, 53*, 151–160.
- Izard, C., & Dougherty, L. (1980). A system for identifying affect expressions by holistic judgements (AFFEX). University of Delaware, Newark, DE.
- Jaccard, J. (1998). *Interaction effects in factorial analysis of variance*. Thousand Oaks, CA: Sage.
- Jaccard, J., & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child and Adolescent Psychology, 31*, 130–146.
- Kochanska, G., Aksan, N., & Carlson, J.J. (2005). Temperament, relationships, and young children's receptive cooperation with their parents. *Developmental Psychology, 41*, 648–660.
- Kochanska, G., Forman, D.R., & Coy, K.C. (1999). Implications of the mother-child relationship in infancy socialization in the second year of life. *Infant Behavior & Development, 22*, 249–265.
- Landa, R., Holman, K.C., & Garret-Mayer, E. (2007). Social and communication development in toddlers with early and later diagnosis of autism spectrum disorders. *Archives of General Psychiatry, 64*, 853–864.
- Larrick, R.P., & Soll, J.B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*, 111–127.
- Lord, C., Rutter, M., DiLavore, P.C., & Risi, S. (1999). *Autism Diagnostic Observation System*. Los Angeles, CA: Western Psychological Services.
- Merin, N., Young, G.S., Ozonoff, S., & Rogers, S. (2007). Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at-risk for autism from comparison infants. *Journal of Autism and Developmental Disorders, 37*, 108–121.
- Mesman, J., van IJzendoorn, M.H., & Bakermans-Kranenburg, M.J. (2009). The many faces of the Still-Face Paradigm: A review and meta-analysis. *Developmental Review, 29*, 120–162.
- Messinger, D., & Fogel, A. (2007). The interactive development of social smiling. In R. Kail (Ed.), *Advances in Child Development and Behavior, 35*, 327–366. Oxford: Elsevier.
- Messinger, D.S., Mahoor, M., Chow, S., & Cohn, J.F. (2009). Automated Measurement of Facial Expression in Infant-Mother Interaction: A Pilot Study. *Infancy, 14*, 285–305.
- Moore, G.A., Cohn, J.F., & Campbell, S.B. (2001). Infant affective responses to mother's still face at 6 months differentially predict externalizing and internalizing behaviors at 18 months. *Developmental Psychology, 37*, 706–714.
- Oster, H. (2000). *Baby FACS: Facial Action Coding System for Infants and Young Children*. Unpublished monograph and coding manual. New York University.
- Oster, H. (2005). The repertoire of infant facial expressions: An ontogenetic perspective. In J. Nadel, D. Muir, J. Nadel, & D. Muir (Eds.), *Emotional development: Recent research advances*. (pp. 261–292). New York, NY: Oxford University Press.
- Osterling, J., Dawson, G., & Munson, J. (2002). Early recognition of one year old infants with autism spectrum disorder versus mental retardation: A study of first birthday party home videotapes. *Development and Psychopathology, 14*, 239–252.
- Siller, M., & Sigman, M. (2002). The behaviors of parents of children with autism predict the subsequent development of their children's communication. *Journal of Autism and Developmental Disorders, 32*, 77–89.
- Surowiecki, J. (2004). *The wisdom of crowds*. New York, NY: Doubleday.
- Tronick, H., Als, H., Adamson, L., Wise, S., & Brazelton, B. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *American Academy of Child Psychiatry, 17*, 1–13.
- Tronick, E., Ricks, M., & Cohn, J.F. (1982). Maternal and infant affective exchange: Patterns of adaptation. In T. Field & A. Fogel (Eds.), *Emotion and early interaction* (pp. 83–100). Hillsdale, NJ: Lawrence Erlbaum.
- van IJzendoorn, M., Rutgers, A., Bakermans-Kranenburg, M., van Daalen, E., Dietz, C., Buitelaar, J., & van Engeland, H. (2007). Parental sensitivity and attachment in children with autism spectrum disorder: Comparison with children with mental retardation, language delays, and typical development. *Child Development, 78*, 597–608.
- Waldinger, R.J., Schulz, M.S., Hauser, S.T., Allen, J.P., & Crowell, J.A. (2004). Reading others' emotions: The role of intuitive judgments in predicting marital satisfaction, quality, and stability. *Journal of Family Psychology, 18*, 58–71.
- Weinberg, M.K., & Tronick, E.Z. (1996). Infant affective reactions to the resumption of maternal interaction after the still-face. *Child Development, 67*, 905–914.
- Weinberg, M.K., Tronick, E.Z., Cohn, J., & Olson, K.L. (1999). Gender differences in emotional expressivity and self-regulation during early infancy. *Developmental Psychology, 35*, 175–188.
- Yirmiya, N., Gamliel, I., Pilowsky, T., Feldman, R., Baron Cohen, S., & Sigman,

- M. (2006). The development of siblings of children with autism at 4 and 14 months: Social engagement, communication, and cognition. *Journal of Child Psychology and Psychiatry*, *47*, 511–523.
- Yirmiya, N., & Ozonoff, S. (2007). The very early autism phenotype. *Journal of Autism and Developmental Disorders*, *37*, 1–11.
- Zwaigenbaum, L., Bryson, S., Rogers, T., Roberts, W., Brian, J., & Szatmari, P. (2005). Behavioral manifestations of autism in the first year of life. *International Journal of Developmental Neuroscience*, *23*, 143–152.
- Zwaigenbaum, L., Thurm, A., Stone, W., Baranek, G., Bryson, S., Iverson, J., et al. (2007). Studying the emergence of autism spectrum disorders in high-risk infants: Methodological and practical issues. *Journal of Autism and Developmental Disorders*, *37*, 466–480.