


Development of an interactive tool of early social responsiveness to track autism risk in infants and toddlers

REINA S FACTOR^{1,2}  | ROSA I ARRIAGA³ | MICHAEL J MORRIER^{4,5} | JENNIFER B MATHYS⁶ | MONICA DIRIENZO⁵ | CHANEL A MILLER³ | AUDREY M SOUTHERLAND³ | GREGORY D ABOWD³ | OPAL Y OUSLEY^{4,5}

1 Department of Psychology, Virginia Polytechnic Institute & State University, Blacksburg, VA; **2** Virginia Tech Center for Autism Research, Blacksburg, VA; **3** School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA; **4** Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA; **5** Emory Autism Center, Atlanta, GA; **6** TEACCH Autism Program, University of North Carolina at Chapel Hill, Charlotte, NC, USA.

Correspondence to Reina S Factor at UCLA Semel Institute for Neuroscience and Human Behavior, 760 Westwood Plaza, Los Angeles, CA 90095, USA. E-mail: rfactor@mednet.ucla.edu

PUBLICATION DATA

Accepted for publication 30th July 2021.
Published online 24th August 2021.

ABBREVIATIONS

ASD	Autism spectrum disorder
CSBS DP:	Communication and Symbolic
ITC	Behavior Scales Developmental Profile: Infant-Toddler Checklist
ESR	Early social responsiveness
M-CHAT-23	23-item Modified Checklist for Autism in Toddlers
PPV	Positive predictive value
ROC	Receiver operator curve

AIM To evaluate the psychometric properties of a 4-minute assessment designed to identify early autism spectrum disorder (ASD) status through evaluation of early social responsiveness (ESR).

METHOD This retrospective, preliminary study included children between 13 and 24 months (78 males, 79 females mean age 19.4mo, SD 3.1) from two independent data sets (an experimental/training sample [$n=120$] and a validation/test sample [$n=37$]). The ESR assessment examined social behaviors (e.g. eye contact, smiling, ease-of-social-engagement) across five common play activities (e.g. rolling a ball, looking at a book). Data analyses examined reliability and accuracy of the assessment in identifying ESR abilities and in discriminating children with and without ASD.

RESULTS Results indicated adequate internal consistency and test-retest reliability of the ESR assessment. Receiver operator curve analysis identified a cutoff score that discriminated infants with ASD-risk from peers in the training sample. This score yielded moderate sensitivity and high specificity for best-estimate ASD diagnosis in the validation sample.

INTERPRETATION Preliminary findings indicated that brief, systematic observation of ESR may assist in discriminating infants with and without ASD, providing concrete evidence to validate or supplement parents', pediatricians', or clinicians' concerns. Future studies could examine the utility of ESR 'growth curves'.

Autism spectrum disorder (ASD) is characterized by impaired reciprocal social interactions and the presence of restricted and repetitive behaviors.¹ For infants who are eventually diagnosed with ASD, parents often report non-specific developmental concerns that do not yet meet full diagnostic criteria for ASD, thus creating a dilemma for frontline clinicians needing to address parent concerns.^{2,3} Systematic observation of an infant's early social responsiveness (ESR) may provide a means for identifying subtle social delays, before the age when more obvious symptoms of ASD emerge.⁴⁻⁶ In this study, we examined individual differences in ESR and determined whether brief, systematic ESR observations can differentiate infants and toddlers with ASD, or at risk for ASD, from their peers.

ESR is marked by three critical developmental stages, each characterized by increased social awareness, responsiveness, and engagement.⁷⁻¹⁰ At 2 to 3 months, infants become more socially aware of others and begin to detect how their own actions are related to others' actions (e.g. observable behaviors include increased smiling and increased vocalizing in the presence of a social partner).¹¹ Around 6 months,

infants exert greater influence on social interactions using positive affect, gestures, and vocalizations to maintain social routines or repair disrupted social routines. Response to social signals also emerges (e.g. following another's gaze toward an object of interest, often called response to joint attention). They also display increased positive affect during face-to-face play, increased attention to facial expressions during a disruption in social play, and increased use of gestures and vocalizations to re-engage a social partner in play.¹¹⁻¹³ Between 9 and 12 months, infants consolidate their ability to initiate joint attention (e.g. pointing to an object of interest and shifting gaze between an object and a person), which allows them to initiate and regulate the pacing of social interactions.¹⁴ These capacities serve as the foundation for sharing experiences with others and represent a fundamental social development turning point.¹⁵⁻¹⁷

Infants eventually diagnosed with ASD demonstrate atypical patterns of ESR development, characterized by a delay in ESR or the emergence of ESR behaviors, followed by a decline.^{18,19} Both retrospective and prospective studies reveal these infants show reduced frequency of ESR,

poorly timed social interactions, and reductions in socially directed eye gaze, positive affect, and social-communicative bids; these differences are often evident by 12 months.^{18,20} Further, response to and initiation of joint attention signals are infrequent, and considered pathognomonic of young children with ASD.^{18–21} Therefore, systematic tracking of ESR may provide early and robust ASD indicators and could lead to earlier diagnosis and intervention.²²

We examined a 4-minute, interactive ESR assessment that provides a system for documenting real-time observations of ESR (e.g. eye contact, smiling, pointing, turn-taking, and ease of social engagement) across five simple play activities. To substantially reduce time burdens for the clinician and hopefully ultimately be used by pediatricians and parents, the assessment uses scripted instructions, and behaviors are scored concurrently with administration of the item; a total score is then derived.

We sought to determine whether (1) individual differences in ESR can be measured reliably during a brief, interactive assessment, (2) individual ESR differences are stable, (3) how the ESR compares with other measures of sociability or ASD behavior, (4) ESR can be used to differentiate infants and toddlers with ASD, or at risk for ASD, from their same-age peers, and (5) ESR can supplement or stand apart from parent report. Given the coronavirus pandemic, where many diagnostic services have been affected, a brief and accessible assessment is especially pertinent.

METHOD

Participants

This retrospective study included a total of 157 children ($n=78$ males [49.7%], $n=79$ females [50.3%]). The mean age was 19.4 months (SD 3.1; range 13–24mo) across two sites (Georgia Institute of Technology, Atlanta, GA, USA [hereafter Georgia Tech] and Emory Autism Center). The Georgia Tech sample represented a subset of children from a larger National Science Foundation study repository,²³ who constituted the experimental/training sample ($n=120$), whereas the Emory Autism Center sample represented a subset of children from a clinical research database, who constituted the validation/test sample ($n=37$) (Table 1). Children were selected from these databases on the basis of age (13–24mo) and combined into a single database for the present study. For both sites, parents were required to complete the study questionnaires in English. No other inclusionary or exclusionary criteria existed. The total sample represented the following self-reported demographics: African-American/Black ($n=33$, 21.0%), Asian ($n=4$, 2.5%), Hispanic/Latino ($n=2$, 1.3%), White ($n=95$, 60.5%), multiple ethnicities or other ($n=17$, 10.9%), or not reported ($n=6$, 3.8%). Maternal education included less than high school ($n=1$, 0.6%), completion of high school ($n=8$, 5.1%), some college or college degree ($n=89$, 56.7%), a master's degree or higher ($n=47$, 29.9%), or not reported ($n=12$, 7.6%). Data were collected between December 2008 and September 2014.

The experimental/training sample included 120 children (mean age 19.6mo; SD 3.0, range 15–24mo) who

What this paper adds

- Early social responsiveness (ESR) can be reliably measured via a brief observation.
- Brief observation of ESR in infants is stable across time.

completed assessments at the Child Study Lab at Georgia Tech; 40 (33.3%) children completed a follow-up appointment within approximately 4 months (mean 4.6mo; SD 1.8; range 2–8mo). Children were recruited via mailed advertisements, via flyers placed at local preschools/daycares, and from an online study portal. General recruitment efforts targeted a community sample of children, whereas targeted recruitment focused on children with known or suspected developmental delays.

The validation/test sample included 37 children (mean age 18.7mo; SD 3.3, range 13–24mo) who completed two research assessments at the Emory Autism Center in Atlanta, GA, USA, with follow-up assessments occurring approximately 5 months after the first visit (mean 5.3mo; SD 3.0; range 1–13mo). Children in this sample were recruited via advertisements at local daycare centers or from a clinical database of families seeking assistance. Parents identified themselves as either having or not having a specific ASD concern when entering the study. These children represent a subset of a larger clinical research database.

Institutional review boards at each site (i.e., the Emory University Institutional Review Board and the Georgia Tech Institutional Review Board) approved procedures and written informed consent from parents was obtained before study enrollment. Clinical referrals were provided as needed.

Table 1: Demographic information for the experimental ($n=120$) and validation ($n=37$) samples

Measure	Experimental sample ^a	Validation sample ^b
<i>n</i>	120	37
Age, mean (SD), mo	19.6 (3.0)	18.7 (3.31)
Male:female ratio	1.03:1 (61:59)	0.85:1 (17:20)
M-CHAT-23 _{STEP 1} ^c	15.8% ASD-risk ($n=19$)	29.7% ASD-risk ($n=11$)
M-CHAT-23 _{STEP 1+F/U} ^c	11.7% ASD-risk ($n=14$)	—
Best estimate ^d	—	29.7% ($n=11$)
ASD diagnosis	—	—
CSBS DP: ITC social composite ^e	19.2% concern ($n=23$)	40.5% concern ($n=15$)

^aThe experimental sample represented a community sample with targeted recruitment for children with developmental delay. ^bThe validation sample included children whose parents had specific concerns about autism as well as those with no concerns. ^c_{STEP 1} refers to the initial score derived from the parents' M-CHAT-23 questionnaire responses, whereas _{STEP 1+F/U} refers to the final risk determination after the follow-up interview, if required. ^dBest-estimate clinical judgement was based on all-information-available for the validation sample. ^eCSBS DP: ITC social composite yields a 'concern' designation for scores <10th centile on the basis of age. M-CHAT-23, 23-item Modified Checklist for Autism in Toddlers; ASD, autism spectrum disorder; CSBS DP: ITC, Communication and Symbolic Behavior Scales Developmental Profile: Infant-Toddler Checklist.

Measures

Participants completed a battery of assessments, which lasted approximately 45 minutes. These included the ESR assessment, an ASD screening questionnaire (the 23-item Modified Checklist for Autism in Toddlers [M-CHAT-23] questionnaire),^{24,25} a developmental questionnaire (the Communication and Symbolic Behavior Scales Developmental Profile: Infant-Toddler Checklist [CSBS DP: ITC]),²⁶ and a background questionnaire.

ESR assessment

Following questionnaire completion, the 4-minute ESR assessment proceeded with both the child and parent in the room. The child typically sat on the parent's lap at a table, although the assessment continued in other locations in the testing room as needed (Fig. S1, online supporting information). Trained clinicians or research assistants administered the ESR, following the same procedures across sites, which included scripted language for item administration. Training consisted of live observation, practice administrations with live feedback, and achieving 80% or greater co-coding reliability. The ESR assessment included five structured play activities with standardized verbal prompts and standardized pauses during play (e.g. saying 'hello', rolling a ball, looking at a book, a silly interaction [e.g. book on head], and tickling). The ESR assessment yielded 17 behavior codes that recorded the presence or absence of eye contact, smiling, pointing, and/or turn-taking (Appendix S1, online supporting information). The absence and presence of behaviors were coded as 1 point and 0 points, respectively. For each activity, an ease-of-social-engagement rating was also recorded for each of the five activities, using a 0- to 2-point rating scale. A rating of 0 indicated that the child was easy to engage during the play activity (the child was attuned to the examiner's actions, was readily available for interaction, and attended to the examiner with anticipation and expectancy, requiring minimal effort from the examiner). A rating of 1 indicated that the examiner experienced some difficulty engaging the child, and a rating of 2 indicated that the examiner had difficulty engaging the child. Before determining an overall score, the ease-of-social-engagement ratings were transformed into a dichotomized score (0 or 2), with a score of 1 transformed to a 2. Calculation of the total ESR score involved summing the 17 behavior ratings and the five ease-of-social-engagement ratings. Higher scores indicated poorer ESR (range 0–27). For the experimental sample, administration took 2.5 to 3 minutes, on average, and scoring was completed via video tape by two raters separately. For the validation sample, scoring occurred in real-time, with the combined administration and scoring procedures requiring a total of 3 to 4 minutes, on average.

M-CHAT-23/F

The M-CHAT-23/F is based on a 23-item questionnaire designed to assess for early symptoms of ASD and ASD-risk.^{24,25} It involves determining an initial score based on a

parent questionnaire (hereafter referred to as M-CHAT-23_{STEP 1}) and, if indicated, a follow-up (F/U) interview (M-CHAT-23_{STEP 1+F/U}). For step 1, a low-risk group can be identified on the basis of a score of 0 to 2. A low-risk score requires no action or referral and does not require the follow-up interview. A moderate risk score requires a follow-up interview, with a score of 2 or more after the interview indicating ASD-risk and requiring referral (D Robins, personal communication, 14 May 2021). A high-risk score indicates that an immediate referral is needed (the interview is not required). This two-step process improves accuracy of identifying ASD-risk, compared with the questionnaire only, and yields a higher positive predictive value (PPV); thus it is the preferred method for ASD-risk determination when relying on a single measure.^{24,25} As described below, the M-CHAT-23_{STEP 1+F/U} was used to determine ASD-risk status for the Georgia Tech site; however, for the Emory site, the M-CHAT-23_{STEP 1} was used to obtain information about symptoms of ASD, but ASD diagnostic status was determined on the basis of an all-information-available determination rather than the follow-up interview.

CSBS DP: ITC

The CSBS DP: ITC parent questionnaire solicits information about social communication, speech, and symbolic understanding and development; it yields three composite subscores and a total score.²⁶ The social composite has been used to quantify social communication delays in children with autism, and a score at or below the 10th centile is considered a potential cutoff score for deciding whether a child should be referred for an autism assessment; however, as a stand-alone instrument, it may not reliably discriminate ASD from other developmental or communication delays.^{26,27} In this study, the social composite score was used to describe the sample and in a regression analysis as a predictor of ASD status.

Determination of ASD status

Determination of ASD status differed across the samples, with the experimental sample focusing on ASD-risk determination at one point in time, and the validation sample focusing on best-estimate, all-information-available diagnosis determination across two time points.²⁸

Experimental sample

For the experimental sample ($n=120$), ASD-risk status was determined on the basis of the two-step M-CHAT-23_{STEP 1+F/U} screening procedure involving completion of the M-CHAT-23 questionnaire (step 1) and, as indicated, a follow-up interview. This procedure yielded the best discrimination of ASD-risk from non-ASD-risk.^{24,25,29} If the follow-up interview was required, item scores were modified on the basis of the interview and a new total score was recalculated.^{24,25,29} Nineteen children failed the M-CHAT-23_{STEP 1} initially, although five children passed the follow-up interview (indicating a 'negative' screen for

ASD). Ultimately, 14 children screened positive for ASD on the M-CHAT-23/F ($n=14$; final M-CHAT-23 score: mean 7.9; SD 4.8, range 2–16). Thus, 14 children constituted the ASD-risk group for the experimental sample and 106 children constituted the non-ASD group.

Validation sample

For the validation sample ($n=37$), an ASD best-estimate diagnosis was based on all-information-available from data gathered across two research visits.²⁸ The following information was obtained during each visit: parental report of developmental milestones, parental endorsement of ASD-related concerns on a background questionnaire, the M-CHAT-23 parent checklist (without follow-up interview), standardized behavioral questionnaires (e.g. CSBS DP: ITC), incidental behavioral observations made throughout the child's research visit, and behavioral observations during the ESR assessment without knowledge of the proposed cutoff score. Information from the Autism Diagnostic Observation Schedule^{30,31} was also considered, if available. Using a best-estimate, all-information-available diagnosis, children were identified as ASD versus non-ASD; these groups were used for data analyses. All research staff who conducted the research visits held master's or PhD degrees, and worked regularly within an ASD diagnostic clinic. The initial determination was made by the assessment clinician and final determinations were confirmed independently by two PhD-level staff (OYO and MJM) with ASD diagnostic expertise using all-information-available across the two visits, yielding 97.3% agreement. Final best-estimate ASD diagnosis discrepancies ($n=1$) were resolved by a consensus meeting with all research staff. For descriptive purposes only, designations were given within the non-ASD category (developmental delay [16.2% of total sample; $n=6$ on the basis of the CSBS DP: ITC scores, report of developmental milestones, the absence of ASD, and ultimately an all-information-available clinical judgement], or typically developing [54.1%; $n=20$]) and, for the ASD category, severity of symptoms were rated as moderate (10.8%; $n=4$) or severe (18.9%; $n=7$), based on all-information-available clinical judgement.

Data analysis plan

The data analysis plan included examination of the following measures within the experimental sample: means and standard deviations of the ESR total score, the test–retest correlation, internal consistency of the ESR items, the intraclass correlation, and interrater reliability. Concurrent validity was examined using correlation analysis among the total scores from the ESR and M-CHAT-23, and the CSBS DP: ITC social composite scaled score. Using the experimental/training sample, a receiver operator curve (ROC) analysis was performed to determine a potential cutoff score for identifying ASD-risk, which was tested subsequently in the validation/test sample.³² The ROC analysis helps determine the best cutoff score for a

continuous measure (ESR) to match a categorical outcome (ASD-risk vs non-ASD-risk). The curve is generated by plotting sensitivity against 1 minus specificity for each potential cutoff score. The cutoff is chosen by identifying the score that corresponds to the highest sum of the sensitivity plus specificity, thus choosing the cutoff score that maximizes both.³² An area under the curve is also generated to describe the curve. On the basis of the ROC result, multiple measures of accuracy were reported (e.g. sensitivity, specificity, PPV) for the experimental and validation samples; where appropriate, a prevalence of 1 in 54 was assumed for calculations (e.g. PPV).³³ Further, for the validation sample, hierarchical regression analyses examined the incremental validity of using the ESR measure in addition to parent report (M-CHAT-23_{STEP 1} and CSBS DP: ITC social composite) to predict best-estimate diagnosis. Dichotomized scores for each measure were examined in the regression analysis (low vs moderate to high risk for the M-CHAT-23_{STEP 1}, and concern vs no concern for the CSBS DP: ITC social composite).

Assuming $\alpha=0.05$ and power=0.8, power analyses indicated sample sizes for the ASD and non-ASD subgroups were adequate. For the experimental/training sample, the sample size required for the comparison of means was 63 (57 non-ASD and six ASD), assuming a difference of five points, SD 4, and a 10:1 ratio for non-ASD to ASD. For the ROC analysis, the required sample was 88 (80 non-ASD and eight ASD), assuming an area under the curve of 0.8, a null value of 0.5, and a ratio of 10:1 non-ASD to ASD. For moderate correlations ($r=0.5$), the required sample size was 29. Effect sizes from the experimental/training sample were used in power analyses relevant to the validation/test sample. The sample size required for a comparison of means was $n=35$ (27 non-ASD and eight ASD), assuming a difference of 6.5 (SD 4.6 for non-ASD and SD 5.7 for ASD) and a ratio of 3.4:1 non-ASD to ASD. The sample size required for the regression analysis was $n=22$, assuming two test predictors, and a multiple partial correlation of 0.63. SciStat (<https://www.scistat.com/index.php>; MedCalc Software Ltd, Ostend, Belgium) and SPSS 27.0 (IBM, Armonk, NY, USA) were used to conduct power analyses.

RESULTS

Examinations of means, standard deviations, and ranges revealed a wide distribution of ESR scores and significant subgroup mean differences within each sample. Analyses of the experimental sample ($n=120$) revealed a mean of 8.6 (SD 5.1; 95% confidence interval [CI] 7.6–9.5; range 0–24), whereas analysis of the validation sample revealed a mean of 9.0 (SD 5.6; 95% CI 7.2–10.9; range 1–20) (Figs 1 and S2, online supporting information). A comparison of subgroup means revealed significant differences between subgroups for each sample. For the experimental sample, the non-ASD group ($n=106$, mean 7.9, SD 4.6; 95% CI 7.0–8.8) had significantly lower scores than the ASD group ($n=14$, mean 13.7, SD 5.9; 95% CI 10.9–16.8), $t_{118}=-4.3$,

$p < 0.001$. Similarly, for the validation sample, the non-ASD group ($n = 26$, mean 6.35, SD 3.5; 95% CI 4.9–7.7) had significantly lower scores than the ASD group ($n = 11$, mean 15.4, SD 4.1; 95% CI 12.7–18.2), $t_{35} = -6.8$, $p < 0.001$.

Reliability analyses revealed good psychometric properties of the ESR assessment. High test–retest values for the ESR total score were found using Pearson correlation ($r = 0.70$, $p < 0.001$; $n = 40$) and an internal consistency analysis yielded a high Cronbach's alpha (22 items; $\alpha = 0.79$; $n = 120$) for the ESR item ratings (17 dichotomous behavioral ratings and five dichotomous ease-of-engagement ratings). Intraclass correlation was also calculated for the experimental sample. All administrations of the ESR assessment were completed by three assessors (MD, AMS, CAM), and at least two of three raters co-coded the administrations for 43 of 120 (35.8%) participants. Considering pairs of ratings from the three available raters, the intraclass correlation coefficient estimates and their 95% confident intervals were calculated using SPSS 27.0 on the basis of a mean-rating ($k = 2$), consistency, one-way random-effects model: intraclass correlation coefficient (1,2) = 0.79 (95% CI = 0.63–0.89). This estimate indicated good reliability for the ESR total score.

Interrater reliability was examined for each pair of three clinical research assistants across 20 co-coding sessions per pair, yielding a mean percentage agreement of 0.92 (SD 0.02, range 0.68–1.00), for the experimental sample. A similar percentage agreement was found for the validation sample for pairs of raters across 14 co-coded assessments (80.6% [SD 0.13]).

Correlation analyses revealed significant, moderate correlations of the ESR with the M-CHAT-23/F total score

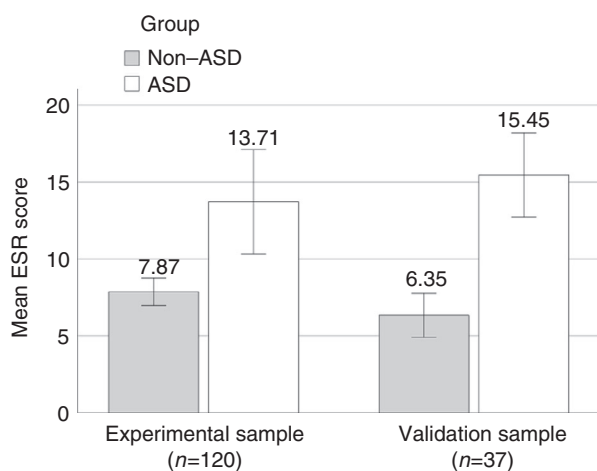


Figure 1: Means and 95% confidence intervals (CIs) for the early social responsiveness (ESR) total score for the experimental and validation samples. Groups represented include non-ASD and ASD. For the experimental sample, ASD-risk was determined; for the validation sample, best-estimate ASD diagnosis was determined. The mean ESR score is significantly different between the non-ASD and ASD groups ($p < 0.001$) for both the experimental and validation samples. Error bars depict the 95% CIs.

(Pearson correlation 0.41; 95% CI 0.20–0.60; $p < 0.001$) and the CSBS DP: ITC social composite scaled score (Pearson correlation -0.23 , 95% CI -0.39 to -0.06 ; $p < 0.01$), suggesting that each measure provides both shared and unique information (Table S1, online supporting information).

Using the experimental sample, we completed a series of ROC curve analyses that plotted sensitivity against 1 minus specificity, using the ESR total score as a predictor of ASD-risk status (Fig. 2). The ROC analyses included a comparison of ASD-risk versus non-ASD groups (area under the curve = 0.781 [95% CI = 0.643–0.919])³² and yielded an optimal cutoff score of greater than or equal to 12, which corresponded to the highest sum between the sensitivity and specificity scores.³² Using a cutoff score of greater than or equal to 12, the sensitivity and specificity are reported for both the experimental and validation groups (Table 2). In addition, assuming an ASD prevalence of 1 in 54,³³ the PPV, negative predictive value, and accuracy values are reported (Table 2); these values are also reported for each sample without consideration of the population prevalence, for comparison. Calculations revealed generally high specificity (range 83.2–93.2), negative predictive value (range 96.0–99.8), and accuracy values (range 82.5–92.3), and moderate to high sensitivity (range 76.9–90.9), but generally low PPV values (range 7.9–83.3).

To examine the performance of the ESR in conjunction with parental report, hierarchical regression analysis was completed with the validation sample. Results indicated that ESR significantly improves prediction of ASD best-

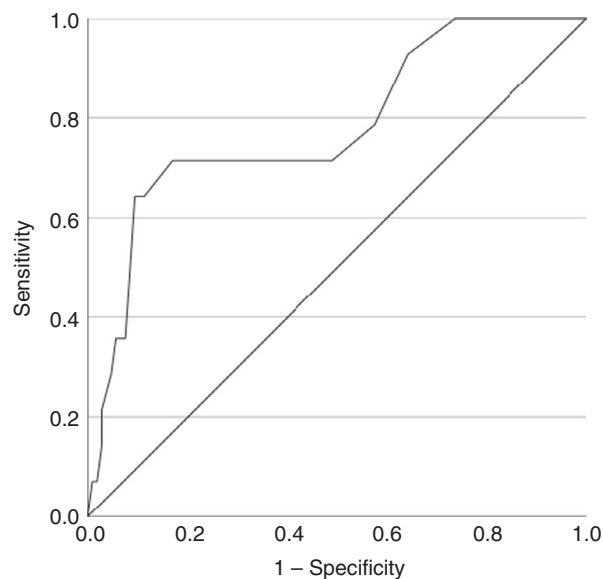


Figure 2: Receiver operator curve (ROC) analysis for the early social responsiveness (ESR) total score. The ROC analysis examined potential cutoff scores for the ESR assessment total score. Values for sensitivity versus 1 minus specificity were plotted to form the curve, yielding an area under the curve = 0.781. An optimal cutoff score of greater than or equal to 12 was identified by determining the largest sum of sensitivity plus specificity.

Table 2: Sensitivity and specificity table for experimental ($n=120$) and validation ($n=37$) samples

Measure	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	PPV (95% CI) ^a	NPV (95% CI)	Accuracy (95% CI)	Accuracy (95% CI) ^a
Experimental sample								
ESR (≥ 12)	76.9 (46.1–95.0)	83.2 (74.7–89.7)	35.7 (24.9–48.2)	96.7 (91.6–98.8)	(7.93) (4.9–12.6)	96.7 (91.6–98.8)	82.5 (74.5–88.3)	(83.1) (75.1–89.3)
Validation sample								
ESR (≥ 12)	90.9 (58.7–99.8)	92.3 (74.9–99.1)	83.3 (56.6–95.1)	96.0 (78.7–99.4)	(18.2) (5.49–46.1)	96.0 (78.7–99.4)	91.9 (78.1–98.3)	(92.3) (78.6–98.5)
M-CHAT-23 _{STEP 1}	81.8 (48.2–97.7)	92.3 (74.9–99.1)	81.8 (53.6–94.6)	92.3 (77.3–97.7)	(16.7) (4.9–43.9)	92.3 (77.3–97.7)	89.2 (74.6–97.0)	(92.1) (78.4–98.4)
ESR (≥ 12) and M-CHAT-23 _{STEP 1}	81.8 (48.2–97.7)	100.0 (86.8–100.0)	100.0 (n/a)	92.9 (78.8–97.9)	(100.0) (n/a)	92.9 (78.8–97.9)	94.6 (81.8–99.3)	(99.7) (89.9–100.0)
CSBS DP: ITC soc (≤ 10 th centile)	81.8 (48.2–97.7)	76.9 (56.4–91.0)	60.0 (41.3–76.1)	90.1 (73.7–97.3)	(6.3) (3.1–12.5)	90.1 (73.7–97.3)	78.4 (61.8–90.2)	(77.0) (60.3–89.2)
ESR (≥ 12) and CSBS DP: ITC soc (≤ 10 th centile)	81.8 (48.2–97.7)	96.2 (80.4–99.9)	90.0 (56.3–98.4)	92.6 (78.1–97.8)	(28.6) (5.4–73.7)	92.6 (78.1–97.8)	91.9 (78.1–98.3)	(95.9) (83.7–99.7)

^aBased on a prevalence of 1 in 54 (Maenner et al.³⁹). CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; ESR, early social responsiveness; M-CHAT-23_{STEP 1}, 23-item Modified Checklist for Autism in Toddlers initial score, which was dichotomized to indicate low vs moderate to high risk for this analysis; CSBS DP: ITC soc, Communication and Symbolic Behavior Scales Developmental Profile: Infant-Toddler Checklist social composite.

estimate diagnosis, compared with parent report only. A significant change in R^2 was found when ESR was added to separate regression models for the M-CHAT-23_{STEP 1} ($R^2_{\text{change}}=F_{1,33}=17.82, p<0.001$) and the CSBS DP: ITC social composite ($R^2_{\text{change}}=F_{1,33}=30.63, p<0.001$) (Table S2, online supporting information). Further analyses revealed that the combination of parent report and ESR measures resulted in higher classification accuracy of ASD best-estimate diagnosis. Notable improvements were observed in specificity and PPV for the combination of ESR and M-CHAT-23_{STEP 1} (specificity changed from 92.3 to 100.0, PPV changed from 81.8 to 100.0 and 16.7 to 100.0 when considering the sample only and the population prevalence, respectively). Improvements in specificity were found for the ESR and CSBS DP: ITC social composite combination (specificity changed from 76.9 to 96.2); changes in PPV were also observed (PPV changed from 60.0 to 90.0 and from 6.3 to 28.6 when considering the sample only and the population prevalence, respectively; Table 2).

DISCUSSION

The results of this study support the premise that ESR observations can provide critical information for determining ASD status for 13- to 24-month-old infants and toddlers. Our preliminary results suggest that brief, standardized observations of infants and toddlers during common play activities yield a reliable, stable metric for assessing ESR and may assist in identifying ASD status in infants and young children. Using an ROC-derived ESR cutoff score, our results suggest that a single score can discriminate children with and without a best-estimate diagnosis of ASD. The proposed cutoff score (≥ 12) yielded high specificity (range 83.2–93.2) within the validation sample, suggesting that a brief ESR assessment can reliably identify ‘true negatives,’ or children without ASD. Similarly, the high negative predictive value (range 96.0–99.8) revealed that when the ESR cutoff was not met, a child has a high probability of not having a best-estimate diagnosis of ASD. Thus, a brief ESR assessment could provide a critical source of information for professionals who seek to avoid over-identification of ASD.

Regarding sensitivity, or the ability to identify true positives, the proposed ESR cutoff scores yielded moderate to high sensitivity values (range 76.9–90.9). In contrast, the PPV was low when considering only the ESR measure; thus caution might be needed if using ESR without an additional indicator, such as parental report. Using hierarchical regression analysis, we were able to examine the utility of combining a parent report measure with ESR. Results showed ESR significantly improves prediction of ASD status compared with parent report only. The results suggest that the ESR may be particularly effective in conjunction with parent report using the M-CHAT-23 questionnaire initial score (step 1), as this combination increased the specificity and the PPV substantially (Table 2). Improvements in specificity and PPV were also found for the combination of the CSBS DP: ITC social composite and ESR,

although not to the same level. If findings are replicated, the ESR measure could complement other sources of information. Thus, these results support gathering information from multiple sources (e.g. observation, parent report) when identifying ASD status, which mirrors clinical best practices.

This work directly addresses the need for brief ASD assessments, outlined by the Centers for Disease Control and Prevention and the American Academy of Pediatrics.^{33,34} The American Academy of Pediatrics recommends pediatricians screen all infants and toddlers for ASD, initially at 18 months and again at 24 months, during well checkups.³⁴ Recent work suggested that ASD diagnoses can be stable before 18 months of age, which highlights the need to assess ESR early in life.³⁵ Taking a more conservative stance, the US Preventive Services Task Force advocates for ASD screening, but secondary to parent or clinician concerns, given the finding that population screeners can over-identify ASD-risk.³⁶ Despite these recommendations, historically only about 8% of pediatricians report routine screening, despite increasing pressure to do so.^{37–39}

Benefits of this ESR assessment include the assessment's brevity, standardized administration, use of commonly available materials (a ball and a book), and reliance on a single score that can be quickly calculated. Thus, the assessment might be easily incorporated into other screening or assessment protocols and provide both quantitative and qualitative information about social behavior that cannot be obtained from parent report questionnaires alone or provide a method for repeated screening during early development. The ESR was developed with clinicians, including pediatricians, in mind, as it requires no specific training other than practicing the standard prompts. The 17 behavior items scored during task administration require a determination of the presence or absence of discrete behaviors after a specific prompt, but also provide a global ease-of-social-interaction rating and a rich amount of information quickly.

Although not tested directly in this study, the ESR assessment could provide a means for shared observation of child behaviors of concern between a parent and clinician, possibly leading to a shared conversation about the level of concern about ASD and the need for monitoring, referral, and/or treatment. This mutual understanding is often lacking between parents and clinicians/pediatricians, which might be improved through shared observations, which also allows parents to bring up additional concerns.³⁸ Further, the use of standardized, observational methods for assessing ESR behavior may improve professionals' acute awareness of early ASD social differences and provide a means for distinguishing ASD symptoms from typical early variation in social development. Future studies focused on quality of healthcare delivery could examine how structured behavioral observations of ESR increase clinician/pediatrician confidence in decision-

making or impact parent's perceptions of service. Examining how repeated assessments of ESR at multiple well-baby visits impact ASD referrals and diagnosis will also add to the future findings. Recent questions about the utility of existing screeners that rely only on parental report leave room for interactive, observational screeners, such as the ESR assessment, as an additional source of information when examining early social behavior.³⁹ Further tests of the feasibility of using this assessment in medical settings and for pediatricians to employ and score is needed in future work.

Several limitations exist that could be addressed in future studies. First, relying on the M-CHAT-23/F two-step screening for the experimental/training sample allowed an estimate of outcome for a relatively large group of children, but did not provide the final definitive outcome. However, the cutoff score derived from using this method did translate well to a clinical sample (the validation/test sample), which mirrored a clinic setting where parents presented with or without a priori ASD concerns. Future studies require larger sample sizes and larger resources; to overcome these barriers, mobile or online data capture systems for both the screening and the best-estimate diagnostic outcomes might be used.⁴⁰ Second, the difference in training and testing group sizes is a limitation that could be overcome with careful infrastructure planning and resources needed for a prospective study. Third, comparison with more comprehensive ASD screening or assessment outcome measures, longitudinal follow-up, and investigation of interrater reliability within a pediatrician's office setting would provide additional validity indicators. Finally, future studies should incorporate the latest version of the M-CHAT series (M-CHAT-R/F comprising 20 items and a follow-up interview).

Although these preliminary results indicate the ESR assessment is a reliable measure to help identify ASD status, future studies should examine larger samples, including those known to be at genetic or other risk for ASD (e.g. Down syndrome, fragile X, 22q11.2 deletion syndrome, infants born preterm), as well as longitudinal analyses. Tracking development might yield an ESR 'growth curve' to increase the precision of the ESR assessment in identifying risk or a change in the developmental trajectory. In sum, we present preliminary evidence that the ESR assessment deserves further study as it might provide a rapid, effective method for early pediatric assessment for atypical social responsiveness and consideration of ASD-risk, leading to earlier awareness, intervention, and improved developmental outcomes.

ACKNOWLEDGEMENTS

We thank the families and children for their participation in the study. This study was funded in part by Emtech Bio and the National Science Foundation Expeditions in Computing Award 1029679. The authors have stated that they had no interests that might be perceived as posing conflict or bias.

DATA AVAILABILITY STATEMENT

Data are available upon request.

SUPPORTING INFORMATION

The following additional material may be found online:

Figure S1: Early social responsiveness (task administration).

Figure S2: Histogram of the early social responsiveness total score for the experimental group ($n=120$).

Table S1: Assessment results

Table S2: Hierarchical multiple regression for ASD diagnosis for the validation sample ($n=37$)

Appendix S1: Early social responsiveness assessment protocol.

REFERENCES

- American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5). Washington, DC: American Psychiatric Association, 2013.
- Chawarska K, Klin A, Paul R, Volkmar F. Autism spectrum disorder in the second year: stability and change in syndrome expression. *J Child Psychol Psychiatry* 2007; **48**: 128–38.
- Wiggins LD, Baio JO, Rice C. Examination of the time between first evaluation and first autism spectrum diagnosis in a population-based sample. *J Dev Behav Pediatr* 2006; **27**: S79–87.
- Bryson SE, Zwaigenbaum L, McDermott C, Romboough V, Brian J. The Autism Observation Scale for Infants: scale development and reliability data. *J Autism Dev Disord* 2008; **38**: 731–8.
- Ozonoff S, Iosif AM, Baguio F, et al. A prospective study of the emergence of early behavioral signs of autism. *J Am Acad Child Adolesc Psychiatry* 2010; **49**: 256–66.
- Zwaigenbaum L, Bryson S, Rogers T, Roberts W, Brian J, Szatmari P. Behavioral manifestations of autism in the first year of life. *Int J Dev Neurosci* 2005; **23**: 143–52.
- Aschersleben G, Hofer T, Jovanovic B. The link between infant attention to goal-directed action and later theory of mind abilities. *Dev Sci* 2008; **11**: 862–8.
- Adamson LB, Bakeman R, Deckner DF, Romski M. Joint engagement and the emergence of language in children with autism and down syndrome. *J Autism Dev Disord* 2009; **39**: 84–96.
- Rochat PR. Social contingency detection and infant development. *Bull Mem Clin* 2001; **65**: 347–60.
- Tomasello M. Joint attention as social cognition. In: Moore C, Dunham PJ, editors. Joint attention: its origins and role in development. New York: Lawrence Erlbaum Associates; 1995. p. 103–30.
- Millar WS. Smiling, vocal, and attentive behavior during social contingency learning in seven- and ten-month-old infants. *Merrill-Palmer Quart* 1988; **34**: 301–25.
- Phillips W, Baron-Cohen S, Rutter M. The role of eye contact in goal detection: evidence from normal infants and children with autism or mental handicap. *Dev Psychopathol* 1992; **4**: 375–83.
- Striano T, Vaish A. Seven- to 9-month-old infants use facial expressions to interpret others' actions. *Br J Dev Psychol* 2006; **24**: 753–60.
- Morgan B, Maybery M, Durkin K. Weak central coherence, poor joint attention, and low verbal ability: independent deficits in early autism. *Dev Psychol* 2003; **39**: 646–56.
- Mundy P, Jarrold W. Infant joint attention, neural networks and social cognition. *Neural Networks* 2010; **23**: 985–97.
- Mundy P, Sullivan L, Mastergeorge AM. A parallel and distributed-processing model of joint attention, social cognition and autism. *Autism Res* 2009; **2**: 2–1.
- Mundy P. Joint attention and social-emotional approach behavior in children with autism. *Dev Psychopathol* 1995; **7**: 63–82.
- Baranek GT. Autism during infancy: a retrospective video analysis of sensory-motor and social behaviors at 9–12 months of age. *J Autism Dev Disord* 1999; **29**: 213–24.
- Jones W, Klin A. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature* 2013; **504**: 427–31.
- Osterling J, Dawson G. Early recognition of children with autism: a study of first birthday home videotapes. *J Autism Dev Disord* 1994; **24**: 247–57.
- Dawson G, Meltzoff AN, Osterling J, Rinaldi J, Brown E. Children with autism fail to orient to naturally occurring social stimuli. *J Autism Dev Disord* 1998; **28**: 479–85.
- Lambert-Brown BL, McDonald NM, Mattson WI, Martin KB, Ibañez LV, Stone WL, Messinger DS. Positive emotional engagement and autism risk. *Dev Psychol* 2015; **51**: 848.
- Rehg J, Abowd G, Rozga A, et al. Decoding children's social behavior. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2013. Portland, OR: IEEE; 2013. p. 3414–21.
- Kleinman JM, Robins DL, Ventola PE, et al. The modified checklist for autism in toddlers: a follow-up study investigating the early detection of autism spectrum disorders. *J Autism Dev Disord* 2008; **38**: 827–39.
- Chlebowski C, Robins DL, Barton ML, Fein D. Large-scale use of the modified checklist for autism in low-risk toddlers. *Pediatrics* 2013; **131**: e1121–7.
- Wetherby AM, Brosnan-Maddox S, Peace V, Newton L. Validation of the Infant-Toddler Checklist as a broadband screener for autism spectrum disorders from 9 to 24 months of age. *Autism* 2008; **12**: 487–511.
- Veness C, Prior M, Bavin E, Eadie P, Cini E, Reilly S. Early indicators of autism spectrum disorders at 12 and 24 months of age: a prospective, longitudinal comparative study. *Autism* 2012; **16**: 163–77.
- Klin A, Lang J, Cicchetti DV, Volkmar FR. Interrater reliability of clinical diagnosis and DSM-IV criteria for autistic disorder: results of the DSM-IV autism field trial. *J Autism Dev Disord* 2000; **30**: 163–7.
- Robins DL. Scoring the M-CHAT [Internet]. Published online 2013. <https://mchatscreen.com/m-chat/scoring-2/>. Accessed 7 March 2021.
- Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop S. Autism diagnostic observation schedule. 2nd ed. Torrance, CA: Western Psychological Services; 2012.
- Luyster R, Gotham K, Guthrie W, et al. The Autism Diagnostic Observation Schedule—Toddler Module: a new module of a standardized diagnostic measure for autism spectrum disorders. *J Autism Dev Disord* 2009; **39**: 1305–20.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29–36.
- Maenner MJ, Shaw KA, Baio J, et al. Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2016. *MMWR Surveill Summ* 2020; **69**: 1–12.
- Myers SM, Johnson CP. Management of children with autism spectrum disorders. *Pediatrics* 2007; **120**: 1162–82.
- Pierce K, Gazestani VH, Bacon E, et al. Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. *JAMA Pediatr* 2019; **173**: 578–87.
- Siu AL, Bibbins-Domingo K, Grossman DC, et al. Screening for autism spectrum disorder in young children: US Preventive Services Task Force recommendation statement. *JAMA* 2016; **315**: 691–6.
- Dosreis S, Weiner CL, Johnson L, Newschaffer CJ. Autism spectrum disorder screening and management practices among general pediatric providers. *J Dev Behav Pediatr* 2006; **27**: S88–94.
- Bellesheim KR, Kizzee RL, Curran A, Sohl K. ECHO autism: integrating maintenance of certification with extension for community healthcare outcomes improves developmental screening. *J Dev Behav Pediatr* 2020; **41**: 420–7.
- Zwaigenbaum L, Maguire J. Autism screening: where do we go from here? *Pediatrics* 2019; **144**: e20190925.
- Berger NI, Wainer AL, Kuhn J, et al. Characterizing available tools for synchronous virtual assessment of toddlers with suspected autism spectrum disorder: a brief report. *J Autism Dev Disord* 2021; **19**: 1–2.

DESARROLLO DE UNA HERRAMIENTA INTERACTIVA DE RESPUESTA SOCIAL TEMPRANA PARA RASTREAR EL RIESGO DE AUTISMO EN BEBÉS Y NIÑOS PEQUEÑOS

OBJETIVO

Este estudio preliminar retrospectivo evaluó las propiedades psicométricas de una evaluación de 4 minutos diseñada para identificar el estado del trastorno del espectro autista (TEA) temprano a través de la evaluación de la capacidad de respuesta social temprana (RST).

MÉTODO

El estudio incluyó a niños de entre 13 y 24 meses (78 varones, 79 mujeres con una edad media de 19,4 meses, DE 3,1) de dos conjuntos de datos independientes (una muestra experimental / de entrenamiento [n = 120] y una muestra de validación / prueba [n = 37]). La evaluación de RST examinó los comportamientos sociales (por ejemplo, contacto visual, sonreír, facilidad de participación social) en cinco actividades de juego comunes (por ejemplo, hacer rodar una pelota, mirar un libro). Los análisis de datos examinaron la confiabilidad y precisión de la evaluación para identificar las habilidades de RST y para discriminar a los niños con y sin TEA.

RESULTADOS

Los resultados indicaron una adecuada consistencia interna y confiabilidad prueba-reprueba de la Evaluación de RST. El análisis de la curva del operador del receptor identificó una puntuación de corte que discriminaba bebés con riesgo de TEA de sus compañeros en la muestra de entrenamiento. Esta puntuación arrojó una moderada Sensibilidad y alta especificidad para el diagnóstico de TEA, mejor estimado en la muestra de validación.

INTERPRETACIÓN

Los hallazgos preliminares indicaron que la observación breve y sistemática de la RST puede ayudar a discriminar a los bebés con y sin TEA, proporcionando evidencia concreta para validar o complementar las preocupaciones de los padres, pediatras o médicos. Los estudios futuros podrían examinar la utilidad de las "curvas de crecimiento" de la RST.

DESENVOLVIMENTO DE UMA FERRAMENTA INTERATIVA DE RESPONSABILIDADE SOCIAL PRECOCE PARA RASTREAR O RISCO DE AUTISMO EM LACTENTES E CRIANÇAS PEQUENAS

OBJETIVO

Este estudo preliminar e retrospectivo avaliou as propriedades psicométricas de uma avaliação de 4 minutos projetada para identificar o status precoce do transtorno do espectro autista (TEA) por meio da avaliação da responsividade social precoce (RSP).

MÉTODO

O estudo incluiu crianças entre 13 e 24 meses (78 do sexo masculino, 79 do sexo feminino com idade média de 19,4 meses, DP 3,1) de dois conjuntos de dados independentes (uma amostra experimental/treinamento [n=120] e uma amostra de validação/teste [n=37]). A avaliação da RSP examinou comportamentos sociais (por exemplo, contato visual, sorriso, facilidade de envolvimento social) em cinco atividades lúdicas comuns (por exemplo, rolar uma bola, olhar um livro). A análise dos dados examinou a confiabilidade e a precisão da avaliação na identificação das habilidades de RSP e na discriminação de crianças com e sem TEA.

RESULTADOS

Os resultados indicaram consistência interna adequada e confiabilidade teste-reteste da avaliação da RSP. A análise da curva do operador receptor identificou uma pontuação de corte que discriminou bebês com risco de TEA de pares na amostra de treinamento. Esta pontuação rendeu moderada sensibilidade e alta especificidade para melhor estimar o diagnóstico de TEA na amostra de validação.

INTERPRETATION

Achados preliminares indicaram que a observação breve e sistemática da RSP pode auxiliar na discriminação de bebês com e sem TEA, fornecendo evidências concretas para validar ou complementar as preocupações dos pais, pediatras ou médicos. Estudos futuros podem examinar a utilidade de "curvas de crescimento" da RSP.