Screening for Autism Spectrum Disorder in a Naturalistic Home Setting Using the Systematic Observation of Red Flags (SORF) at 18–24 months

Deanna Dow, Taylor N. Day, Timothy J. Kutta, Charly Nottke, and Amy M. Wetherby

The purpose of this study was to examine the utility of the Systematic Observation of Red Flags (SORF; Dow et al., 2016) as a level 2 screener for autism spectrum disorder (ASD) in toddlers during a naturalistic video-recorded home observation. Psychometric properties of the SORF were examined in a sample of 228 toddlers—84 with ASD, 82 with developmental delay (DD), and 62 with typical development (TD). Trained undergraduate research assistants blind to diagnosis rated 22 red flags (RF) of ASD associated with DSM-5 diagnostic criteria using a 4-point scale. The following scores were computed: a total score summing all items, domain scores summing social communication and restricted, repetitive behavior items, and number of RF counting items with scores of 2 or 3 indicating clear symptom presence. The performance of the total, domain, and RF scores and individual items were examined. A composite score was formed with six items with the best psychometric performance: poor eye gaze directed to faces, limited showing and pointing, limited coordination of nonverbal communication, less interest in people than objects, repetitive use of objects, and excessive interest in particular objects, actions, or activities. The 6-item composite provides a brief measure with optimal performance, while the RF may be instrumental for clinicians who are interested in characterizing the range of observed symptoms. The SORF shows promise as a practical alternative to currently available screening methods for implementation by nonexperts with the potential to increase feasibility and reduce common obstacles to access to care. ***Autism Res*** *2020, 13: 122–133.* © 2019 The Authors. *Autism Research published by International Society for Autism Research* published by Wiley Periodicals, Inc.

Lay Summary: Research suggests that current autism spectrum disorder (ASD) screening tools are not accurate enough to use in routine screening. The Systematic Observation of Red Flags was developed as a practical option for children at high risk for ASD. It can be used with video-recorded samples of parent–child interactions in the home and by raters who are not experts in ASD. It shows promise in predicting ASD risk in toddlers to determine if a full diagnostic evaluation is necessary.

Keywords: early detection; early signs; psychometrics; red flags

Introduction

Although the American Academy of Pediatrics recommends universal screening for autism spectrum disorder (ASD) in primary care settings at 18 and 24 months [Johnson & Myers, 2007], some have expressed skepticism in routinizing this practice due to the lack of effective screening tools available [Al-Qabandi, Gorter, & Rosenbaum, 2011; Campos-Outcalt, 2011]. Furthermore, there remain substantial barriers to making referrals for children identified through ASD-specific screening in primary care settings for further diagnostic evaluation [Pierce et al., 2011; Bauer, Sturm, Carroll, & Downs, 2013]. Commonly reported challenges to screening and appropriate referral include time and resource constraints, difficulty obtaining responses on caregiver questionnaires, lack of provider concern for ASD, and failure to complete necessary follow-up with families that screen positive [Bauer et al., 2013; Daniels, Halladay, Shih, Elder, & Dawson, 2014]. Despite the age of first parent concern averaging around 19 months in toddlers with ASD [Rosenberg, Landa, Law, Stuart, & Law, 2011], first diagnosis in the United States remains at approximately 4.5 years old [Baio, Wiggins, Christensen, et al., 2018], with longer delays and lower diagnosis rates for children from non-white and socioeconomically disadvantaged families [Daniels & Mandell, 2013]. In spite of this, current evidence suggests the importance of beginning

From the Department of Psychology, Florida State University, Autism Institute, Tallahassee, Florida (D.D., T.N.D.); College of Medicine, Florida State University Autism Institute, Tallahassee, Florida (T.J.K., C.N.); Department of Clinical Sciences, College of Medicine, Florida State University, Autism Institute, Tallahassee, Florida (A.M.W.)

Received April 2, 2019; accepted for publication September 26, 2019

Address for correspondence and reprints: Deanna Dow, Department of Psychology, Florida State University, Autism Institute, 2312 Killearn Center Blvd, Building A, Tallahassee, FL 32309. E-mail: deannatracydow@gmail.com

Published online 23 October 2019 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/aur.2226

intervention in early childhood to reduce symptom severity and impairment, as earlier services may lead to more positive outcomes [Harris & Handleman, 2000; Granpeesheh, Dixon, Tarbox, Kaplan, & Wilke, 2009; Wetherby et al., 2018]. Methods that make the screening process more automated and require less time and initiative may improve barriers to implementation of screening [Bauer et al., 2013; Campbell et al., 2017].

ASD Screening and Diagnostic Evaluation

Screening tools for ASD detection fall into two levels, based on the method and target population (see Zwaigenbaum et al., 2015 for a review). Broadband and ASD-specific level 1 caregiver-report screeners are intended for all children in primary care settings, regardless of risk. Because of the need for widespread use, they should be brief, easy to complete, and should not require substantial staff time or expertise to score or interpret. Level 2 screeners are intended for children at increased risk for ASD based on a positive level 1 screening result and/or other risk factors. Level 2 screeners may utilize trained individuals to assess a child's behavior with standardized or systematic observational techniques. Though some ASD-specific knowledge is needed, level 2 screeners should require substantially less staff time and training compared to a full diagnostic evaluation. Alternatively, diagnostic evaluations using gold-standard tools require extensive training and resources to complete and consequently may have high out-of-pocket expenses, long waits for families to access, and may not be recommended unless a child exhibits ASD-specific concerns. Therefore, using a 2-tiered screening approach can be advantageous for children at increased risk before initiating a full diagnostic evaluation.

Tiered Screening Approach Benefits

Use of a level 1 broadband screener alone (i.e., the Infant Toddler Checklist (ITC; Wetherby & Prizant, 2002) without a second level screen has been shown to predict a range of developmental concerns (i.e., ASD, DD, learning disabilities) as early as 12 months [Pierce et al., 2011]. However, ASD was initially missed at a substantial rate in those with other language and developmental delays [Pierce et al., 2019], with 23.8% who were later diagnosed at 3–4 years old missed during their initial visit between 12 and 36 months. While there has been limited research attention on the potential for community providers to detect and diagnose ASD in toddlers with stability (as opposed to researchers and university-associated clinics), there is some evidence to support methods that streamline children from primary care into early intervention (EI) services without a full diagnostic evaluation, either through a 2-tiered screening process [Rotholz, Kinsman, Lacy, & Charles, 2017], pediatrician diagnosis (only in high certainty cases), or use of level 2 screening methods in a multidisciplinary team setting [Ahlers et al., 2019]. These studies support the value of implementing level 2 screening and flexible diagnostic processes to improve earlier access to EI services.

Level 2 Screening Measures

There are currently three level 2 ASD screening tools available that have been empirically tested in samples of young children in clinical settings: the Screening Tool for Autism in Two-Year-Olds (STAT; Stone, Coonrod, & Ousley, 2000; Stone, Coonrod, Turner, & Pozdol, 2004), Autism Detection in Early Childhood (ADEC; Hedley et al., 2010; Nah, Young, Brewer, & Berlingeri, 2014), and the Systematic Observation of Red Flags of Autism (SORF; Wetherby et al., 2004; Dow, Guthrie, Stronach, & Wetherby, 2016). Although other measures have been developed and tested [Bryson, McDermott, Rombough, Brian, & Zwaigenbaum, 2000; Dix, Fallows, & Murphy, 2015; Choueiri & Wagner, 2015], they have not yet been validated in large enough community samples to evaluate their utility. One study showed that even when experienced psychologists rated two 10-min segments of the ADOS as part of a short observational screening measure, they missed 39% of toddlers with ASD [Gabrielsen et al., 2015].

The STAT is an observational play-based screening measure that was developed to assess children 24–36 months of age. In a sample of children with considerable cognitive impairment (e.g., ASD group: mean mental age = 16.1 months, mean chronological age = 28.5 months), the STAT demonstrated good discrimination between children with ASD and other developmental delays (sensitivity = 1.00; specificity = 0.85; Stone et al., 2004). An exploratory study further examined use of the STAT in 71 children at high genetic risk who were under 24 months [Stone, McMahon, & Henderson, 2008], 19 of whom were diagnosed with ASD at their follow-up evaluation between 24 and 42 months old. Final results revealed good sensitivity (0.93) and specificity (0.83), though 21 12- to 13-month old children were removed from the sample due to a high rate of false positives, resulting in a sample of only 50 children. Further investigation in a large community sample is needed to determine the validity of the STAT for children younger than 24 months of age.

The ADEC is an interactive, behavior-based ASD screening instrument intended for use in children 12–36 months. It has been tested in a relatively small (N = 114) clinical sample with a mean age of 28.67 months [Hedley, Nevill, & Monroy-Moreno, 2015]; 48 children were diagnosed with ASD between 19 and 36 months. Similar to the STAT sample, children with ASD in this study had significant developmental delays (i.e., average nonverbal developmental scores over two *SDs* below the mean). The ADEC was

found to be psychometrically sound (sensitivity = 0.93, specificity = 0.64; cutoff = 11) in this population, though additional study is needed in a larger sample of children under 24 months to determine its validity.

The SORF was originally developed as an observational measure that was based upon DSM-IV diagnostic criteria [Wetherby et al., 2004]. Following preliminary testing and the release of DSM-5, items were modified according to updated diagnostic criteria and other relevant behaviors. The current version of the SORF includes 22 items and has been validated with the Communication and Symbolic Behavior Scales (CSBS; Wetherby & Prizant, 2002) Behavior sample in a large community-based sample ($N = 247$; Dow et al., 2016), with results indicating good discrimination, sensitivity, specificity, and positive and negative predictive values (PPV and NPV) between ASD and nonspectrum groups. The CSBS Behavior Sample is often utilized for children with suspected language or social communication delays in clinic settings, allowing for the SORF to be coded without requiring an additional evaluation. However, because many children with developmental concerns will not receive a clinical evaluation, utility in this context may be limited. The current study addresses the need to explore how and when the SORF may be used to broaden clinical utility and provide screening opportunities for those who would not otherwise be seen in a clinic.

Level 2 Screening Opportunities in the Home Context

The home context offers an ideal setting for administering screening measures, as it provides accessibility to services for families regardless of their participation in scheduled clinical evaluations. Observing a child's behavior in this naturalistic setting gives a view of the child during everyday activities in a familiar environment that would not otherwise be accessible to practitioners. It also could be pivotal to building consensus with families on the early signs of ASD. Child behavior can be assessed in the home through in-person visits and remotely through the use of video-recorded observations, maximizing opportunities for evaluation while reducing cost and burden for families.

In-person home visits are often used for screening and preventative services, both to determine eligibility for federally funded intervention programs through Part C of the Individuals with Disabilities Education Act and for families with increased risk factors from birth. Approximately 75% of home intervention programs serve children birth to age 3 [Sweet & Appelbaum, 2004], consistent with the recommended timing for screening and diagnosis of ASD.

Purpose of this Study

The purpose of this study was to examine the psychometric properties of the SORF when used in a naturalistic home context during everyday activities. We addressed two research aims: (1) to study group differences by examining item-level performance to create an algorithm with best-performing items and (2) to calculate sensitivity, specificity, PPV and NPV, and determine optimal cutoff scores in 18- to 24-month-old toddlers during a home observation.

Methods

Participants

Toddlers who were evaluated by the FIRST WORDS® Project at Florida State University were included in this study. Parents of all participants provided written informed consent and the present study were prospectively approved by the Florida State University Institutional Review Board. The FIRST WORDS® Project is a prospective, longitudinal study of early detection of communication disorders, including ASD. The ITC was completed by parents of children 9 through 24 months of age in primary care settings. Families were referred for an evaluation if their child received a score within the bottom 10th percentile on the questionnaire or if parents indicated concern about their child's development. A small percentage was referred directly from professionals or parents due to developmental concerns and/or concerns specific to ASD; these individuals completed the questionnaire at or before the first appointment. Our sample was randomly selected from children who had a home observation and diagnostic evaluation to assess for ASD between 18 and 24 months. We stratified by race to best approximate minority groups according to the demographic makeup in the region (see Table 1 for participant characteristics). There were 228 children (84 ASD, 82 DD, 62 TD) with a mean age of 20.45 months enrolled in the study following parental informed consent. The TD group consisted of children who were initially flagged for concern on the ITC, but whose developmental scores were within the expected range and for whom there was no concern for ASD. The TD group was smaller than the ASD and DD groups due to a more limited number of children who did not demonstrate developmental concerns after concern was indicated on initial screening. The prevalence of ASD in our sample was also higher than expected; this is likely due to families' increased likelihood to follow through with a clinical evaluation if they had specific concerns for ASD.

Diagnostic Procedures and Measures

Diagnostic procedures. Study participants received a diagnostic evaluation that included the ADOS–Toddler Module (ADOS-T; Lord, Luyster, Gotham, & Guthrie, 2012), Mullen Scales of Early Learning (MSEL; Mullen, 1995), Vineland Adaptive Behavior Scale–Second Edition

Table 1. Participant Demographics

Characteristic, *Mean* (*SD*)	Diagnostic group		
	ASD	DD	TD
N	84	82	62
Age in months—*Mean* (*SD*)	20.66 (1.81)	20.29 (1.56)	20.38 (1.46)
Sex—*n* (%)			
Male	72 (85.7)[a,b]	60 (73.2)[b,c]	34 (54.8)[a,c]
Female	12 (14.3)[a,b]	22 (26.8)[b,c]	28 (45.2)[a,c]
Race—*n* (%)			
White	59 (70.2)	57 (69.5)	46 (74.2)
Black	13 (15.5)	13 (15.9)	6 (9.7)
Asian	2 (2.4)	2 (2.4)	1 (1.6)
Biracial	9 (10.7)	10 (12.2)	7 (11.3)
Ethnicity—*n* (%)			
Hispanic	17 (20.2)[a,b]	7 (8.5)[b,c]	1 (1.6)[a,c]
Maternal education in years—*Mean* (*SD*)	14.81 (2.50)	14.34 (2.62)	16.17 (2.82)[a,c]

[a]Significant difference with DD group.
[b]Significant difference with TD group.
[c]Significant difference with ASD group.

(VABS-II; Sparrow, Balla, & Cicchetti, 1984), video-recorded home observation, and parent-report questionnaire, the Early Screening for Autism and Communication Disorders (ESAC; Wetherby et al., 2015). Children were diagnosed with ASD if their symptoms (based on behavioral observation and parent report) met DSM-5 criteria. Children were diagnosed with DD if ASD was ruled out and delays were found based on MSEL scores, and TD if both ASD and DD were ruled out. A conservative cutoff of 1.25 SDs below the mean (i.e., *T* score < 38) was used to determine delay, as in previous research on a similarly high-functioning community sample [Guthrie, Swineford, Nottke, & Wetherby, 2013]. The majority (53.7%) of children in the DD group had language delay (see Table 2 for diagnostic characteristics).

ASD diagnostic assessment. Autism symptoms were evaluated in the clinic using the ADOS–T, which provides symptom domain scores (i.e., Social Affect, Restricted, and Repetitive Behaviors) and a total score, with cutoffs reflecting "little-to-no concern," "mild-to-moderate concern," and "moderate-to-severe concern" for ASD. Calibrated severity scores were used to estimate ASD symptom severity, as they provide a consistent measure regardless of the child's age range and language level [Esler et al., 2015].

Home observation. In addition to the ADOS-T, parents were asked to complete a 1-hr home observation to allow diagnosticians to observe symptoms across contexts. A videographer was present to record the observation and was instructed not to interact with the child or give feedback about the child's behavior. Parents were given written and verbal instructions and asked to interact with their child during a variety of everyday activities. Examples were provided for playing with toys, playing with people, having mealtime, caregiving, completing family chores, and book sharing. Parents were encouraged to participate in activities that happen regularly in the home, and if possible, to spend 5–10 min on activities within each of the suggested categories for up to an hour. Though the home observation was also used to code the SORF, SORF data were not used to make diagnostic determinations.

Table 2. Descriptive Statistics for Diagnostic Outcome Measures

Characteristic, *Mean* (*SD*)	Diagnostic groups				Pairwise *P*-values	
	ASD	DD	TD	*F*-value	DD-ASD	TD-ASD
ADOS-T SA CSS	7.04 (2.14)	3.37 (1.78)	2.55 (1.49)	128.61***	0.000	0.000
ADOS-T RRB CSS	6.19 (2.29)	3.71 (2.27)	3.81 (2.30)	30.22***	0.000	0.000
ADOS-T Total CSS	6.89 (2.12)	3.04 (1.51)	2.42 (1.30)	155.47***	0.000	0.000
MSEL Gross Motor T	46.84 (9.88)	47.10 (10.52)	54.89 (8.26)	14.76***	1.000	0.000
MSEL Fine Motor T	43.64 (10.60)	44.43 (9.59)	53.61 (8.43)	22.23***	1.000	0.000
MSEL Visual Reception T	40.54 (11.21)	44.54 (11.47)	57.92 (10.65)	45.70***	0.065	0.000
MSEL Receptive Language T	31.44 (13.46)	38.20 (13.36)	57.94 (9.76)	83.19***	0.002	0.000
MSEL Expressive Language T	30.52 (10.61)	33.18 (9.13)	50.13 (8.82)	83.10***	0.229	0.000
MSEL ELC composite	75.81 (16.53)	81.34 (13.40)	110.03 (20.47)	105.04***	0.053	0.000
MSEL Nonverbal DQ	90.13 (15.45)	93.94 (12.75)	110.66 (13.08)	41.93***	0.244	0.000
MSEL Verbal DQ	64.45 (23.86)	74.92 (17.47)	108.03 (15.81)	91.24***	0.002	0.000
VABS-II Socialization	84.85 (8.72)	87.51 (7.87)	89.79 (7.83)	6.65[a]	0.110	0.001
VABS-II Daily Living Skills	86.65 (10.06)	89.79 (10.37)	94.47 (7.65)	11.86***	0.108	0.000
VABS-II Communication	81.86 (13.68)	88.72 (10.72)	100.63 (8.91)	47.87***	0.000	0.000
VABS-II Motor Skills	92.98 (9.28)	91.28 (10.05)	93.60 (8.49)	1.23	0.734	1.000
VABS-II ABC	84.02 (9.28)	87.09 (8.89)	93.23 (7.41)	20.38***	0.071	0.000

$^{**}P < 0.01$, $^{***}P < 0.001$.

DD, developmental delayed; TD, typically developing; ADOS-T, Autism Diagnostic Observation Schedule-Toddler Module; SA, Social Affect; CSS, calibrated severity score; MSEL, Mullen Scales of Early Learning; T, T Score; DQ, Developmental Quotient; VABS-II, Vineland Adaptive Behavior Scales, Second Edition; ABC, Adaptive Behavior Composite.

Parent report of symptoms. The ESAC was also used as a parent-report measure to assess ASD symptoms. Unpublished evidence supports the use of the 30-item ESAC as an ASD-specific level 1 screener [Wetherby et al., 2015], with sensitivity ranging from 0.78 to 0.86 and specificity ranging from 0.81 to 0.84 in a sample of 451 12- to 36-month-old children. Preliminary data from a replication study (Kutta et al., in preparation) on a new sample of 464 children (12–36 months old) further supports these findings (sensitivity: 0.69–0.76 for 12–17 months and 0.83–0.95 for 18–36 months; specificity: 0.75–0.86 for 12–36 months).

Developmental level. The MSEL [Mullen, 1995] was used as a standardized assessment of developmental level. It yields standard scores for overall development—Early Learning Composite—and *T*-scores for five subscales—Visual Reception, Expressive and Receptive Language, and Fine and Gross Motor skills. Developmental quotients (DQs) were also calculated, as recommended to provide a familiar IQ measure while avoiding possible floor effects [Munson et al., 2008], given the large proportion of the sample with at least one *T*-score of 20 (i.e., the floor). The DD and ASD groups demonstrated statistically similar nonverbal DQs, suggesting that cognitive level cannot be solely explaining differences between these groups.

Adaptive behavior. The VABS-II [Sparrow et al., 1984] is a caregiver interview that measures adaptive functioning *via* the Adaptive Behavior Composite and domain scores in Social, Communication, Daily Living, and Motor Skills. Results are reported as standard scores.

SORF Measure and Coding

The SORF is an observational coding system developed to identify red flags (RF) for ASD in toddlers. The SORF has been validated during the CSBS Behavior Sample in a clinical setting, with recommended cutoff summary scores and number of RF to indicate a child's risk for ASD based on behaviors exhibited [Dow et al., 2016]. The SORF has 22 items, with 11 items in each of the two DSM-5 symptom domains—Social Communication (SC) and Restricted Repetitive Behaviors (RRB). Behaviors are coded using a graded response system, with zero signifying no concern and three signifying clear, clinically significant severity or concern. Codes of 2 or 3 represent clinically significant symptoms and are included in the number of RF score. Scores from six items that demonstrated the best psychometric performance were added together to create a composite score. The entire 1-hr recording was considered when coding the SORF, and coders took an average of 10 min beyond the length of the video to complete scoring.

Coding Procedures and Inter-Rater Reliability

Undergraduate research assistants blind to diagnostic status completed training relevant to the core diagnostic features of ASD and early detection (i.e., 2.5 hr total) through the Autism Navigator for Primary Care, a web-based course on the early signs of ASD with video illustrations of toddlers with ASD. They also received training on the SORF coding system through individual practice with feedback and consensus coding. Coders who were previously reliable on the SORF coding using the CSBS Behavior Sample also achieved subsequent reliability for the home observation. Coder reliability was established with the completion of 15 training videos, with 80% reliability per video required for at least 12 videos (i.e., 80% of the videos). Additionally, multiple coders scored 15% of the video-recorded home observations in order to determine inter-rater reliability. Reliability was measured using intraclass correlation generalizability (*g*) coefficients to assess inter-rater agreement for measurement of continuous clinical outcomes [Cicchetti, 1994]. Results indicated excellent reliability, with *g* coefficients of 0.75 for both the Composite items and the number of RF items [Cicchetti & Sparrow, 1981]. Individual item *g* coefficients ranged from 0.52 to 0.94, with an average of 0.70.

Analyses

Item examination. One-way ANOVA models and Cohen's *d* effect sizes [Cohen, 1988] were used to determine whether individual items differentiated children with ASD from the DD and TD groups. Receiver operating characteristic (ROC) curve analyses were conducted to inform on individual item's ability to differentiate between children with and without ASD. Sensitivity (or the "true positive" rate) denotes the proportion of correctly-identified children with ASD. Specificity signifies the proportion of children correctly identified as not at risk who do not have ASD (i.e., "true negatives"). Area under the curve (AUC) shows the strength of discrimination between groups, ranging from 0.5 (i.e., no better than chance) to 1.0 (i.e., perfect discrimination; Swets, 1988).

Summary scores and analysis. The following summary scores were computed: a total score summing all 22 items, domain scores summing the SC and RRB items, and number of RF counting items with scores of 2 or 3 indicating clear symptom presence. The number of RF were counted for each domain and summed for a total number of RF. Analysis of variance (ANOVA) and ROC curve analyses were conducted for each item. Scores from items that demonstrated the best psychometric performance in discriminating groups were then added together to create a composite score.

Table 3. Diagnostic Group Differences of SORF Summed Scores and Items

	ASD (n = 84) *Mean* (*SD*)	DD (n = 82) *Mean* (*SD*)	TD (n = 62) *Mean* (*SD*)		Pairwise group differences			
					ASD-DD		ASD-TD	
				F (2,225)	d	P	d	P
Composite score	7.36 (3.60)	4.27 (2.80)	2.45 (2.28)	50.60***	0.96	0.00	1.63	0.00
Total score	20.23 (7.83)	14.71 (6.62)	10.32 (4.85)	40.07***	0.76	0.00	1.52	0.00
Number of RF	6.81 (3.10)	4.91 (2.41)	3.24 (1.78)	35.60***	0.68	0.00	1.41	0.00
SC domain score	16.06 (6.09)	12.12 (5.23)	8.98 (4.19)	32.34***	0.69	0.00	1.35	0.00
SC number of RF	5.61 (2.48)	4.13 (2.00)	2.94 (1.59)	29.58***	0.66	0.00	1.28	0.00
RRB domain score	4.17 (3.72)	2.59 (3.06)	1.34 (1.71)	15.85***	0.46	0.00	0.98	0.00
RRB number of RF	1.20 (1.30)	0.78 (1.01)	0.31 (0.59)	13.31***	0.36	0.03	0.88	0.00
1 Limited sharing warm, joyful expressions	1.89 (0.82)	1.55 (0.96)	1.55 (0.90)	3.94*	0.38	0.04	0.39	0.07
2. Flat affect/reduced facial expressions	0.35 (0.65)	0.18 (0.59)	0.18 (0.43)	2.18	0.27	0.21	0.31	0.25
3. Limited sharing interests, enjoyment	2.27 (1.09)	2.12 (1.13)	1.71 (1.29)	4.38*	0.14	1.00	0.47	0.01
4. Lack of response to name	1.33 (1.21)	0.83 (1.03)	0.71 (1.06)	6.90**	0.44	0.09	0.55	0.18
5. Poor eye gaze directed to faces	1.06 (0.88)	0.52 (0.69)	0.32 (0.65)	19.20***	0.68	0.00	0.96	0.00
6. Limited showing and pointing	2.49 (0.84)	1.79 (1.09)	1.26 (1.14)	26.67***	0.72	0.00	1.23	0.00
7. Using another person's hand as tool	0.19 (0.63)	0.06 (0.36)	0.06 (0.40)	1.83	0.25	0.27	0.25	0.37
8. Limited directed consonant sounds	1.48 (1.40)	1.26 (1.29)	0.23 (0.64)	21.17***	0.16	0.72	1.15	0.00
9. Limited coordination of nonverbal communication	1.19 (1.30)	0.57 (0.99)	0.16 (0.55)	18.56***	0.54	0.00	1.03	0.00
10. Less interest in people than objects	1.37 (0.94)	0.83 (0.87)	0.47 (0.67)	20.94***	0.60	0.00	1.10	0.00
11. Limited reciprocal social play	2.44 (0.84)	2.40 (0.87)	2.34 (0.85)	0.25	0.05	1.00	0.12	1.00
12. Repetitive use of objects	0.45 (0.77)	0.17 (0.54)	0.11 (0.41)	6.95**	0.42	0.01	0.55	0.00
13. Repetitive body movements	0.76 (1.06)	0.60 (0.86)	0.16 (0.45)	9.05***	0.17	0.65	0.74	0.00
14. Repetitive speech/intonation	1.15 (1.87)	0.63 (1.25)	0.35 (0.79)	6.08**	0.33	0.06	0.56	0.00
15. Ritualized patterns of behavior	0.04 (0.24)	0.00 (0.00)	0.02 (0.13)	1.02	0.24	0.47	0.10	1.00
16. Marked distress over change	0.44 (0.81)	0.49 (0.93)	0.24 (0.50)	1.85	−0.06	1.00	0.30	0.41
17. Excessive interest in particular objects, actions, or activities	0.80 (1.00)	0.38 (0.71)	0.13 (0.46)	13.81***	0.48	0.00	0.86	0.00
18. Clutches particular objects	0.18 (0.50)	0.15 (0.52)	0.13 (0.34)	0.21	0.06	1.00	0.12	1.00
19. Sticky attention to objects	0.10 (0.40)	0.04 (0.19)	0.10 (0.35)	0.88	0.19	0.73	0.00	1.00
20. Fixation on parts of objects	0.13 (0.49)	0.05 (0.31)	0.03 (0.18)	1.66	0.20	0.43	0.27	0.31
21. Adverse response to sensory stimuli	0.01 (0.11)	0.02 (0.22)	0.02 (0.13)	0.13	−0.06	1.00	−0.08	1.00
22. Unusual sensory exploration/interest	0.11 (0.41)	0.06 (0.29)	0.05 (0.28)	0.65	0.14	1.00	0.17	0.90

ASD, autism spectrum disorder; DD, developmental delay; TD, typically developing; SORF, Systematic Observation of Red Flags; SC, social communication; RRB, restricted, repetitive behaviors.

*P < 0.05, **P < 0.01, ***P < 0.001.

Dunnett's C post hoc comparisons were used to correct for Type I error.

Cohen's d: ≤0.20 = small, 0.50 = medium, 0.80 = large.

Next, ANOVA and ROC curve analyses were used to examine the composite, total, SC and RRB domain scores, number of RF, and number of SC and RRB RF. Sensitivity was prioritized when determining optimal cutoff scores, while ensuring maintenance of an adequate level of specificity. PPV and NPV were also calculated. A fivefold cross-validation logistic regression approach was conducted with all summary scores as predictors of diagnostic group (i.e., ASD vs. nonspectrum) to confirm that performance was consistent across samples.

Because the video-recorded home observation was used for both SORF coding and as a component of the diagnostic evaluation procedures, a supplemental analysis was completed using the ADOS-T concern classification to examine whether results were conflated in predicting diagnostic group. ROC curve analyses were conducted on all SORF summary scores to measure discrimination between children who fell in the moderate-to-severe level of concern on the ADOS-T (indicating a high likelihood of clinically significant ASD symptoms) vs. children who fell in either the little-to-no concern or mild-to-moderate concern categories. The video-recorded home observation was viewed as part of the diagnostic evaluation after the ADOS-T administration was completed, ensuring that ADOS scores were not biased from behaviors exhibited in the home.

Results

Item Level Analyses

Results of item-level analyses (see Table 3) showed significantly higher scores for the ASD group than at least one nonspectrum group on 12 items (8 SC and 4 RRB). Six of these demonstrated significant differences between ASD

Table 4. Item level ROC curve analysis based on diagnostic classification: ASD (*n* = 84) *versus* nonspectrum (TD/DD; *n* = 144)

	AUC (*SE*)	95% CI	AUC (*SE*)
1. Limited sharing warm, joyful expressions	0.60* (0.04)	0.52–0.67	0.60* (0.04)
2. Flat affect/reduced facial expressions	0.56 (0.04)	0.48–0.64	0.56 (0.04)
3. Limited sharing interests or enjoyment	0.58* (0.04)	0.50–0.66	0.58* (0.04)
4. Lack of response to name	0.63** (0.04)	0.55–0.70	0.63** (0.04)
5. Poor eye gaze directed to faces	0.70*** (0.04)	0.63–0.77	0.70*** (0.04)
6. Limited showing and pointing	0.73*** (0.03)	0.67–0.80	0.73*** (0.03)
7. Using another person's hand as tool	0.53 (0.04)	0.45–0.61	0.53 (0.04)
8. Limited directed consonant sounds	0.63** (0.04)	0.55–0.70	0.63** (0.04)
9. Limited coordination of nonverbal communication	0.66*** (0.04)	0.58–0.74	0.66*** (0.04)
10. More interest in people than objects	0.70*** (0.04)	0.63–0.77	0.70*** (0.04)
11. Limited reciprocal social play	0.52 (0.04)	0.45–0.60	0.52 (0.04)
12. Repetitive use of objects	0.61** (0.04)	0.53–0.69	0.61** (0.04)
13. Repetitive body movements	0.58* (0.04)	0.50–0.66	0.58* (0.04)
14. Repetitive speech/intonation	0.60** (0.04)	0.53–0.68	0.60** (0.04)
15. Ritualized patterns of behavior	0.51 (0.04)	0.43–0.59	0.51 (0.04)
16. Marked distress over change	0.52 (0.04)	0.44–0.60	0.52 (0.04)
17. Excessive interest in particular objects, actions, activities	0.64*** (0.04)	0.56–0.72	0.64*** (0.04)
18. Clutches particular objects	0.52 (0.04)	0.44–0.59	0.52 (0.04)
19. Sticky attention to objects	0.51 (0.04)	0.43–0.59	0.51 (0.04)
20. Fixation on parts of objects	0.53 (0.04)	0.45–0.61	0.53 (0.04)
21. Adverse response to sensory stimuli	0.50 (0.04)	0.42–0.58	0.50 (0.04)
22. Unusual sensory exploration/interest	0.52 (0.04)	0.44–0.59	0.52 (0.04)

*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

Table 5. Summary score ROC curve analysis based on diagnostic classification: ASD (*n* = 84) vs. nonspectrum (TD/DD; *n* = 144)

	AUC (*SE*)	95% CI	Sensitivity	Specificity	Cutoff	PPV	NPV
Composite score[a]	0.81*** (0.03)	0.75–0.86	0.77	0.72	5	0.62	0.84
Total score	0.77*** (0.03)	0.71–0.84	0.70	0.67	15	0.55	0.79
Number of RF[b]	0.75*** (0.03)	0.68–0.81	0.73	0.63	5	0.54	0.80
SC domain score[c]	0.74*** (0.03)	0.68–0.81	0.73	0.63	12	0.54	0.80
SC RF[d]	0.73*** (0.04)	0.66–0.80	0.64	0.74	5	0.59	0.78
RRB domain score[c]	0.68*** (0.04)	0.61–0.76	0.70	0.54	2	0.47	0.76
RRB RF[d]	0.64*** (0.04)	0.57–0.72	0.62	0.62	1	0.49	0.74

CI, confidence interval.

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

[a]Composite score = Sum of best performing items: (1) Poor eye gaze directed to faces, (2) Limited showing and pointing, (3) Limited coordination of nonverbal communication, (4) Less interest in people than objects, (5) Repetitive use of objects, (6) Excessive interest in particular objects, actions, and activities.

[b]Number of RF = Count of items with clinically significant severity (i.e., score of 2 or 3) across all 22 items.

[c]Domain scores = Sum of all 11 items within each symptom domain.

[d]Domain RF = Count of items with clinically significant severity (i.e., score of 2 or 3) in each symptom domain.

Table 6. Summary score ROC cure analysis based on ADOS concern range: Moderate-to-severe concern (n = 69) *versus* little-to-no/mild-to-moderate concern (n = 159)

	AUC (*SE*)	95% CI	Sensitivity	Specificity	Cutoff
Composite score[a]	0.85*** (0.03)	0.77–0.89	0.81	0.74	7
Total score	0.83*** (0.03)	0.77–0.89	0.74	0.74	16
Number of RF[b]	0.80*** (0.03)	0.74–0.87	0.81	0.64	5
SC domain score[c]	0.80*** (0.03)	0.74–0.87	0.77	0.70	13
SC RF[d]	0.79*** (0.04)	0.72–0.86	0.74	0.74	5
RRB domain score[c]	0.69*** (0.04)	0.61–0.76	0.73	0.53	2
RRB RF[d]	0.67*** (0.04)	0.59–0.75	0.65	0.61	1

CI, confidence interval.

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

[a]Composite score = sum of best performing items: (1) Poor eye gaze directed to faces, (2) Limited showing and pointing, (3) Limited coordination of nonverbal communication, (4) Less interest in people than objects, (5) Repetitive use of objects, (6) Excessive interest in particular objects, actions, & activities.

[b]Number of RF = count of items with clinically significant severity (i.e., score of 2 or 3) across all 22 items.

[c]Domain scores = sum of all 11 items within each symptom domain.

[d]Domain RF = count of items with clinically significant severity (i.e., score of 2 or 3) in each symptom domain.

Table 7. Five-fold cross-validation logistic regression analysis results based on diagnostic classification: ASD (*n* = 84) *versus* nonspectrum (TD/DD; *n* = 144)

	Average AUC	Average sensitivity	Average specificity	Correct classification	Average PPV	Average NPV
Composite score[a]	0.81	0.76	0.75	0.76	0.66	0.84
Total score	0.77	0.74	0.69	0.71	0.59	0.80
Number of RF[b]	0.74	0.66	0.76	0.72	0.63	0.79
SC domain score[c]	0.74	0.65	0.76	0.75	0.64	0.80
SC RF[d]	0.73	0.64	0.73	0.70	0.59	0.78
RRB domain score[c]	0.68	0.58	0.74	0.68	0.56	0.75
RRB RF[d]	0.64	0.56	0.67	0.62	0.49	0.73

CI, confidence interval.

[a]Composite score = sum of best performing items: (1) Poor eye gaze directed to faces, (2) Limited showing and pointing, (3) Limited coordination of nonverbal communication, (4) Less interest in people than objects, (5) Repetitive use of objects, (6) Excessive interest in particular objects, actions, & activities.

[b]Number of RF = count of items with clinically significant severity (i.e., score of 2 or 3) across all 22 items.

[c]Domain scores = sum of all 11 items within each symptom domain.

[d]Domain RF = count of items with clinically significant severity (i.e., score of 2 or 3) in each symptom domain.

and both nonspectrum groups: poor eye gaze directed to faces, limited use of showing and pointing gestures, limited coordination of nonverbal communication, less interest in people than objects, repetitive movements of objects, and excessive interest in particular objects, actions, or activities. Eleven items did not significantly differ between either ASD-DD or ASD-TD in the home context. Twelve of the 22 items had statistically significant AUC values (see Table 4); 10 of these exhibited individual AUC values of 0.60 or greater.

The overlapping six items with significant group differences between ASD and both nonspectrum diagnostic groups and AUC values of 0.60 or greater were used to derive an algorithm to compute a composite score, which provides the optimal continuous measure of ASD severity in the home context. All 22 items were included in the RF scores and tallied in order to provide diagnostically relevant information about the presence and number of clinically significant behaviors. Including all items in the number of RF resulted in higher sensitivity and specificity in comparison to including only the best-performing items.

Summary Score Analyses

As detailed in Table 3, significant differences were found between the ASD and nonspectrum groups for all summary scores (i.e., composite, total, RF, SC domain, SC RF, RRB domain, and RRB RF). The SC scores demonstrated larger overall mean group differences and effect sizes compared to RRBs. The Composite score resulted in good discrimination (Table 4) between the nonspectrum and ASD groups (AUC = 0.81). In order to balance sensitivity and specificity, the optimal cutoff for the Composite score was 5, with a sensitivity of 0.77, specificity of 0.72, PPV of 0.62, and NPV of 0.84. The Total score showed slightly lower discrimination (AUC = 0.79), with sensitivity of 0.70, sensitivity of 0.67, PPV of 0.55, and NPV of 0.79 at a cutoff of 15. The RF score showed fair discrimination (AUC = 0.75; Table 5), with an optimal cutoff of 5, sensitivity of 0.75, specificity of 0.63, PPV of 0.54, and NPV of 0.80. For improved sensitivity, alternate cutoffs with lower specificity resulted in the following psychometric properties: Composite: cutoff = 4; sensitivity = 0.88; specificity = 0.58; Total score: cutoff = 13; sensitivity = 0.85; specificity = 0.53; RF score: cutoff = 4; sensitivity = 0.86; specificity = 0.42. Supplementary ROC curve summary score analyses revealed that the summary scores discriminate similarly between ADOS concern classification groups, with slightly higher AUC values and improved sensitivity compared to diagnostic classification (Table 6). Age and developmental level were examined as moderators of group differences for the Total Score; developmental level significantly moderated diagnosis ($\beta = -9.83$, $P < 0.01$). Cross-validation revealed consistency in AUC, sensitivity, and specificity for summary scores (see Table 7 for results) and supported use of the composite as the best performing measure to predict risk for ASD on the SORF.

Discussion

Our findings support use of the SORF as an observational, level 2 screening tool for children 18–24 months of age during a naturalistic video-recorded observation in a home environment. The composite score, which is comprised of the best performing six items, is the optimal method for predicting ASD risk (as shown in the overall sample and across five cross-validation samples), with good discrimination, sensitivity, and specificity (i.e., indicated by an AUC value >0.80; sensitivity and specificity between 0.70 and 0.80; Glascoe, 1996) at a cutoff of 5. The RF score may be instrumental for diagnosticians because it highlights the presence and number of symptoms included in the diagnostic criteria for ASD, though it provides slightly lower sensitivity and specificity. Similarly, the total score also provides a comparable

measure of ASD risk with a breadth of symptoms from both diagnostic domains.

Alternate cutoffs should also be considered in order to prioritize sensitivity in certain clinical contexts. Lowering the composite and RF score cutoffs by 1 point and the total score cutoff by 2 points resulted in sensitivity above 0.85 for all three summary score measures. Using these lower cutoffs is advantageous in capturing children with ASD at a higher rate, especially in settings where children with other developmental delays may be detected and referred on to appropriate EI services. However, the feasibility of conducting diagnostic evaluations on a higher number of cases (due to the higher rates of false positives) in these clinical contexts must be considered to avoid misclassification of ASD.

The SORF cutoffs established for use in the home observation were different than the cutoffs established for the CSBS in the clinical setting [Dow et al., 2016]. Using the SORF during a home observation provides a practical, accessible, and feasible option for level 2 screening. While the STAT, ADEC, and SORF used in clinical settings provide appropriate options for families who are able to access timely clinical services when first concerns arise, these methods have practical limitations due to clinic waitlists, provider training requirements, and many families' reduced access to services, often leading to substantial delays in evaluation and diagnosis. One advantage to using the SORF during a home observation is that it can be rated by individuals without specialized education or certification in a field related to developmental disabilities. Moreover, our results indicated the SORF is efficacious in a naturalistic setting among a diverse sample ascertained from a primary care population. Though results based on this clinical sample do not directly support application through telehealth services, they indicate a positive first step toward implementation by nonexperts with the potential to increase feasibility and reduce common obstacles to access to care. Additionally, the profile on the SORF could provide important information in decision-making about intervention targets related to SC and RRB.

While the ITC alone has an estimated PPV of 0.75 for detecting developmental delays at 12 months [Pierce et al., 2011], many of these children are flagged for concern due to language or general developmental delays and would not need a specialized diagnostic evaluation for ASD. Adding the SORF as an intermediary step between the ITC and a full diagnostic evaluation would prevent longer clinic waitlists and unnecessary burden on families that would likely result from direct referrals for psychological assessment following a positive level 1 screen. The PPV for ASD using the SORF composite (i.e., 0.62 in the original sample and an average of 0.66 in the cross-validation samples) was similar to results found using the STAT between 14 and 23 months (i.e., PPV = 0.68; Stone et al., 2008). Given that the composite only requires six items to be coded and can be rated without an additional clinic assessment, the benefit of adding this intermediate screening step could greatly streamline the diagnostic system.

SORF Item Analyses

Children with ASD demonstrated higher severity on approximately half of the SORF items in the home context compared to DD and TD groups. Twelve items distinguished ASD from at least one other group; six of these were selected for the composite algorithm based on best performance across analyses. While all 22 items represent the heterogeneity of ASD symptoms, it is not surprising that the symptoms that are most prominent and easily detectable in a time-limited observational sample may vary across several contextual factors. The structure of the setting, the materials available, the prompts given by the clinician or caregiver, and whether behaviors are intentionally elicited or avoided may all have significant impacts on the utility of specific items. For example, intentionally calling a child's name to see whether they respond is necessary for a Response to Name item to predict risk, and a parent making an object in which the child has an excessive interest unavailable could mask RRB symptoms. Given these inherent contextual differences, certain items did not demonstrate adequate discrimination in a naturalistic home observation, despite their utility in a more structured clinical setting. Additionally, while the unstandardized nature of the context is a strength for community implementation, lack of standardization of materials and activities could inflate scores for families with lower socioeconomic backgrounds and/or knowledge of ASD symptomatology; therefore, careful interpretation of scores across demographic samples should be considered.

Many of the RRB items had low frequency of occurrence and subsequently demonstrated weak discrimination in the home, consistent with past research that suggests these behaviors are more difficult to detect in this context compared to the clinic [Stronach & Wetherby, 2014] and are generally less effective in screening than SC items [Berument, Rutter, Lord, Pickles, & Bailey, 1999; Rowberry et al., 2015]. Parents who are aware of their child's RRBs may also use strategies in a naturalistic home observation to reduce the time they engage in them or prevent them from occurring (e.g., through limiting or removing objects, distraction, or interruption). SC items perform better overall than RRB items in the home, likely because there are ample opportunities to observe the child's social abilities when parents are asked to engage their child in interaction. However, certain SC items are dependent on presented opportunity in a naturalistic environment, such as the child's name being called (for response to name) and the availability of items appropriate for reciprocal play (for limited sharing of reciprocal social play), making these items

less effective in discriminating between groups in the home setting.

Sample Characteristics

It is noteworthy and a strength of this study that the ASD sample detected by the SORF has high overall developmental scores, consistent with the most recent CDC study suggesting that almost half (i.e., 44%) of children with ASD have average or above average intellectual abilities [Baio et al., 2018]. These scores were significantly higher than the STAT and ADEC ASD samples [Stone et al., 2004; Hedley et al., 2015], likely due to differences in recruitment (i.e., the STAT and ADEC samples were clinically referred; the SORF sample was referred from primary care screening). Children with ASD in the STAT sample had an average mental age of approximately half their chronological age (i.e., mental age means = 16–17 months, chronological age means = 31–32-months), and the ADEC ASD sample had a nonverbal developmental level over two *SDs* below average (i.e., MSEL Nonverbal DQ = 65.11). In contrast, the average *T*-score for our ASD sample was within one *SD* of the mean for cognitive and motor subscales and within two *SDs* of the mean for language subscales. Given that developmental level moderated the prediction of SORF total by diagnosis, psychometric properties would likely improve if applied to a sample with lower cognitive abilities and greater symptom severity; relatedly, the lower functioning samples detected with the STAT and ADEC may have inflated estimates of sensitivity. However, specificity may be higher in our sample due to the nature of including a large percentage of children with high cognitive functioning.

SORF Prediction of ADOS Concern Classification

SORF performance as measured by summary scores was similar when used to predict ADOS-T concern classification instead of best-estimate diagnostic classification. This evidence provides support that significant findings are not solely a result of conflation due to the dual-purpose use of the home observation in both diagnostic procedures and SORF coding. In fact, discrimination was slightly improved when applied to the ADOS-T concern classification, demonstrating agreement between the SORF as a screening measure and the ADOS-T as a diagnostic tool. These findings further support the utility of the SORF as a valid measure of current ASD symptoms based on a home observation.

Limitations and Future Directions

Replication is necessary in young children with and without ASD, especially given the heterogeneity in ASD symptoms across children. Although the composite score comprised of six items demonstrated improved sensitivity and specificity in this sample and across cross-validation subsamples, caution should be taken in generalizing the use of a limited number of items to predict risk for a complex disorder that presents with various symptoms across individuals, without evidence from replication. In comparison, the total and RF scores provide measures to characterize the range and severity of ASD symptoms.

The current study utilized the entire 1-hr home observation collected as part of the diagnostic evaluation to code SORF items, which requires a significant amount of time and consequently impacts its feasibility for implementation. Consideration of a shorter observation sample will be critical in creating a more efficient screening method. However, previous studies have found that using too brief of an observation may not adequately detect risk, even when an observation from a gold standard evaluation such as the ADOS is used [Gabrielsen et al., 2015]. Therefore, studying the SORF in time intervals to determine the minimum length of observation required without sacrificing accuracy is planned as our next step of analysis in order to broaden its utility.

Additionally, although universal ASD screening is recommended at 18 and 24 months [Johnson & Myers, 2007], evidence suggests that symptoms emerge and can be detected in some children as early as 12 months of age [Zwaigenbaum et al., 2015; Elison et al., 2014]. Given the rapid brain development that occurs early in a child's life, beginning screening and intervention services earlier may further improve child outcomes. There are currently no level 2 screening methods validated for use at this young age. A future research aim is to develop a revised algorithm for the SORF that can be utilized in the naturalistic home observation at 12 months of age. The SORF's utility in combination with a parent-report questionnaire should also be examined to determine whether it can triage children who have clear ASD symptomatology from those who need a thorough diagnostic evaluation. In addition to in-person home visits, the use of telehealth technology should also be examined, as it may provide additional opportunities to assess risk in families when a home visit is not feasible or desired. Mobile technology (e.g., cameras, smart phones, tablet computers) is accessible to most families or can be without substantial cost, and video recordings can be taken by the family themselves or the child's behavior can be viewed *via* a live streaming video system. These methods may be especially convenient and desirable to families, as they allow participation in screening without invasive evaluations. Finally, further research is needed to replicate these findings among independent samples and study methods to confirm that the SORF can be applied in a time- and cost-efficient manner to a naturalistic home setting in an effort to improve ASD screening in toddlers.

Conflicts of Interest

A.W. is a co-author of the Communication and Symbolic Behavior Scales, which is published by Paul H. Brookes

Publishing Co, and receives royalties for its use, but not from this study. The remaining authors do not have any conflict of interest to declare.

References

Ahlers, K., Gabrielsen, T. P., Ellzey, A., Brady, A., Litchford, A., Fox, J., … Carbone, P. S. (2019). A pilot project using pediatricians as initial diagnosticians in multidisciplinary autism evaluations for young children. Journal of Developmental & Behavioral Pediatrics, 40(1), 1–11.

Al-Qabandi, M., Gorter, J. W., & Rosenbaum, P. (2011). Early autism detection: are we ready for routine screening? Pediatrics, 128(1), e211–e217.

Baio, J., Wiggins, L., Christensen, D. L., et al. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 Sites, United States, 2014. MMWR Surveillance Summary, 67(SS-6), 1–23.

Bauer, N. S., Sturm, L. A., Carroll, A. E., & Downs, S. M. (2013). Computer decision support to improve autism screening and care in community pediatric clinics. Infants & Young Children, 26(4), 306–317.

Berument, S. K., Rutter, M., Lord, C., Pickles, A., & Bailey, A. (1999). Autism screening questionnaire: diagnostic validity. The British Journal of Psychiatry, 175(5), 444–451.

Bryson, S. E., McDermott, C., Rombough, V., Brian, J., & Zwaigenbaum, L. (2000). The autism observation scale for infants. Unpublished Scale, Toronto, ON.

Campbell, K., Carpenter, K. L., Espinosa, S., Hashemi, J., Qiu, Q., Tepper, M., … Dawson, G. (2017). Use of a digital Modified Checklist for Autism in Toddlers–Revised with follow- up to improve quality of screening for autism. The Journal of Pediatrics, 183, 133–139.

Campos-Outcalt, D. (2011). Should all children be screened for autism spectrum disorders? No: screening is not ready for prime time. American Family Physician, 84(4), 377.

Choueiri, R., & Wagner, S. (2015). A new interactive screening test for autism spectrum disorders in toddlers. The Journal of Pediatrics, 167(2), 460–466.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. American Journal of Mental Deficiency, 86(2), 127–137.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Daniels, A. M., Halladay, A. K., Shih, A., Elder, L. M., & Dawson, G. (2014). Approaches to enhancing the early detection of autism spectrum disorders: a systematic review of the literature. Journal of the American Academy of Child & Adolescent Psychiatry, 53(2), 141–152.

Daniels, A. M., & Mandell, D. S. (2013). Explaining differences in age at autism spectrum disorder diagnosis: A critical review. Autism, 18(5), 583–597.

Dix, L., Fallows, R., & Murphy, G. (2015). Effectiveness of the ADEC as a level 2 screening test for young children with suspected autism spectrum disorders in a clinical setting. Journal of Intellectual and Developmental Disability, 40(2), 179–188.

Dow, D., Guthrie, W., Stronach, S. T., & Wetherby, A. M. (2016). Psychometric analysis of the Systematic Observation of Red Flags for autism spectrum disorder in toddlers. Autism, 21(3), 301–309.

Elison, J. T., Wolff, J. J., Reznick, J. S., Botteron, K. N., Estes, A. M., Gu, H., … Piven, J. (2014). Repetitive behavior in 12-month-olds later classified with autism spectrum disorder. Journal of the American Academy of Child & Adolescent Psychiatry, 53(11), 1216–1224.

Esler, A. N., Bal, V. H., Guthrie, W., Wetherby, A., Weismer, S. E., & Lord, C. (2015). The autism diagnostic observation schedule, toddler module: standardized severity scores. Journal of Autism and Developmental Disorders, 45(9), 2704–2720.

Gabrielsen, T. P., Farley, M., Speer, L., Villalobos, M., Baker, C. N., & Miller, J. (2015). Identifying Autism in a Brief Observation. Pediatrics, 135(2), e330–e338.

Glascoe, F. P. (1996). Developmental screening. In M. Wolraich (Ed.), Disorders of development and learning: A practical guide to assessment and management (pp. 89–128). St. Louis, MO: Mosby.

Granpeesheh, D., Dixon, D. R., Tarbox, J., Kaplan, A. M., & Wilke, A. E. (2009). The effects of age and treatment intensity on behavioral intervention outcomes for children with autism spectrum disorders. Research in Autism Spectrum Disorders, 3, 1014–1022.

Guthrie, W., Swineford, L. B., Nottke, C., & Wetherby, A. M. (2013). Early diagnosis of autism spectrum disorder: stability and change in clinical diagnosis and symptom presentation. Journal of Child Psychology and Psychiatry, 54(5), 582–590.

Harris, S. L., & Handleman, J. S. (2000). Age and IQ at intake as predictors of placement for young children with autism: A four- to six-year follow-up. Journal of Autism and Developmental Disorders, 30(2), 137–142.

Hedley, D., Nevill, R. E., & Monroy-Moreno, Y. (2015). Efficacy of the ADEC in identifying autism spectrum disorder in clinically referred toddlers in the US. Journal of Autism and Developmental Disorders, 45, 2337–2348.

Johnson, C. P., & Myers, S. M. (2007). Identification and evaluation of children with autism spectrum disorders. Pediatrics, 120(5), 1183–1215.

Lord, C., Luyster, R., Gotham, K., & Guthrie, W. (2012). Autism diagnostic observation schedule—Toddler module manual. Los Angeles, CA: Western Psychological Services.

Mullen, E. M. (1995). Mullen scales of early learning (AGS ed.). Circle Pines, MN: American Guidance Service.

Munson, J., Dawson, G., Sterling, L., Beauchaine, T., Zhou, A., Koehler, E., … Abbott, R. (2008). Evidence for latent classes of IQ in young children with autism spectrum disorder. American Journal on Mental Retardation, 113(6), 439–452.

Nah, Y. H., Young, R. L., Brewer, N., & Berlingeri, G. (2014). Autism Detection in Early Childhood (ADEC): Reliability and validity data for a level 2 screening tool for autistic disorder. Psychological Assessment, 26(1), 215.

Pierce, K., Carter, C., Weinfeld, M., Desmond, J., Hazin, R., Bjork, R., & Gallagher, N. (2011). Detecting, studying, and

treating autism early: the one-year well-baby check-up approach. The Journal of Pediatrics, 159(3), 458–465.

Pierce, K., Gazestani, V. H., Bacon, E., Barnes, C. C., Cha, D., Nalabolu, S., … Courchesne, E. (2019). Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. JAMA Pediatrics, 173(6), 578–587.

Rosenberg, R. E., Landa, R., Law, J. K., Stuart, E. A., & Law, P. A. (2011). Factors affecting age at initial autism spectrum disorder diagnosis in a national survey. Autism Research and Treatment, 2011, 874619–874619.

Rotholz, D. A., Kinsman, A. M., Lacy, K. K., & Charles, J. (2017). Improving early identification and intervention for children at risk for autism spectrum disorder. Pediatrics, 139(2), e20161061. https://doi.org/10.1542/peds.2016-1061

Rowberry, J., Macari, S., Chen, G., Campbell, D., Leventhal, J. M., Weitzman, C., & Chawarska, K. (2015). Screening for autism spectrum disorders in 12-month-old high-risk siblings by parental report. Journal of Autism and Developmental Disorders, 45(1), 221–229.

Sparrow, S. S., Balla, D., & Cicchetti, D. (1984). Vineland adaptive behavior scales (survey form). Circle Pines, MN: American Guidance Service.

Stone, W. L., Coonrod, E. E., & Ousley, O. Y. (2000). Brief report: screening tool for autism in two-year-olds (STAT): development and preliminary data. Journal of Autism and Developmental Disorders, 30(6), 607–612.

Stone, W. L., Coonrod, E. E., Turner, L. M., & Pozdol, S. L. (2004). Psychometric properties of the STAT for early autism screening. Journal of Autism and Developmental Disorders, 34(6), 691–701.

Stone, W. L., McMahon, C. R., & Henderson, L. M. (2008). Use of the Screening Tool for Autism in Two-Year-Olds (STAT) for children under 24 months: An exploratory study. Autism, 12 (5), 557–573.

Stronach, S., & Wetherby, A. M. (2014). Examining restricted and repetitive behaviors in young children with autism spectrum disorder during two observational contexts. Autism, 18 (2), 127–136.

Sweet, M. A., & Appelbaum, M. I. (2004). Is home visiting an effective strategy? A meta- analytic review of home visiting programs for families with young children. Child Development, 75(5), 1435–1456.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. Science, 240(4857), 1285–1293.

Wetherby, A., Woods, J., Allen, L., Cleary, J., Dickinson, H., & Lord, C. (2004). Early indicators of autism spectrum disorders in the second year of life. Journal of Autism and Developmental Disorders, 34, 473–493.

Wetherby, A.M., Guthrie, W., Petkova, E., Woods, J., Lord, C., Voccola, D., … Rozenblit, L. (2015). Broadband and autism-specific screening using the Early Screening for Autism and Communication Disorders (ESAC). Presented at the International meeting for autism research, Salt Lake City, UT, 13–16 May.

Wetherby, A. M., & Prizant, B. M. (2002). Communication and Symbolic Behavior Scales Developmental Profile-First (Normed ed.). Baltimore, MD: Paul H. Brookes.

Wetherby, A. M., Woods, J., Guthrie, W., Delehanty, A., Brown, J. A., Morgan, L., … Lord, C. (2018). Changing developmental trajectories of toddlers with autism spectrum disorder: Strategies for bridging research to community practice. Journal of Speech, Language, and Hearing Research, 61(11), 2615–2628.

Zwaigenbaum, L., Bauman, M. L., Fein, D., Pierce, K., Buie, T., Davis, P. A., … Wagner, S. (2015). Early screening of autism spectrum disorder: recommendations for practice and research. Pediatrics, 136(Supplement 1), S41–S59.