Check for updates

Original Article

autism

Autism
2017, Vol. 21(3) 301–309
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1362361316636760
journals.sagepub.com/home/aut
SAGE

Psychometric analysis of the Systematic Observation of Red Flags for autism spectrum disorder in toddlers

Deanna Dow[1], Whitney Guthrie[1], Sheri T Stronach[2] and Amy M Wetherby[1]

Abstract

The purpose of this study was to examine the utility of the Systematic Observation of Red Flags as an observational level-two screening measure to detect risk for autism spectrum disorder in toddlers when used with a video-recorded administration of the Communication and Symbolic Behavior Scales. Psychometric properties of the Systematic Observation of Red Flags were examined in a sample of 247 toddlers of 16- to 24 months old: 130 with autism spectrum disorder, 61 with developmental delays, and 56 typically developing. Individual items were examined for performance to create an algorithm with improved sensitivity and specificity, yielding a total Composite score and Domain scores for Social Communication and Restricted Repetitive Behaviors. Codes indicating clear symptom presence were collapsed to yield a count of the number of Red Flags for the overall scale and each symptom domain. Results indicated significant group differences with large effects for the Composite, both Domain scores, and Red Flags score, and good discrimination (area under the curve = 0.84–0.87) between autism spectrum disorder and nonspectrum groups for the Composite, Social Communication Domain, and Social Communication Red Flags score. The Systematic Observation of Red Flags provides an observational screening measure for 16- to 24-month-olds with good discrimination, sensitivity, and specificity. A cutoff of 20 on the Composite is recommended to optimally detect autism spectrum disorder risk.

Keywords

autism spectrum disorders, screening, social cognition and social behavior, repetitive behaviors and interests

The American Academy of Pediatrics recommends that all children be screened for autism spectrum disorder (ASD) at 18 and 24 months (Johnson and Myers, 2007), as research indicates that reliable diagnoses can be made at these early ages (Chawarska et al., 2007; Guthrie et al., 2013) and early intervention can maximize child outcomes (Dawson et al., 2010; Wetherby et al., 2014). Despite research findings that parents often express concern by 18 to 24 months (Wetherby et al., 2008), the median age of diagnosis in the United States is approximately 4½ years (Baio, 2014). Challenges with early detection include lack of adequate ASD-specific screening tools, as well as time and cost restrictions that impede primary care providers from adequately screening to determine whether referral for a full assessment is needed. Although level-two screening provides an important opportunity to detect children early when potential concerns exist without adding the burden of a full diagnostic evaluation, this type of ASD-specific screening is not widely used in primary care.

Commonly used parent-report measures such as the Modified Checklist for Autism in Toddlers (M-CHAT; Robins et al., 2001) and Infant Toddler Checklist (ITC; Wetherby and Prizant, 2002) offer a practical and time efficient method for level-one screening, though additional follow-up interviews and evaluations are required to reduce false positives (Kleinman et al., 2008) and to differentiate between ASD and other developmental delays (Wetherby et al., 2008). The M-CHAT also may not successfully detect higher-functioning children with ASD, as samples of children identified with this tool have significant developmental delays on average (Robins et al., 2014).

Interactive level-two screening tools, such as the Screening Tool for Autism in Toddlers and Young Children

[1]Florida State University, USA
[2]University of Minnesota—Twin Cities, USA

Corresponding author:
Deanna Dow, Autism Institute , Mail Code 7814, Florida State University, Tallahassee, FL 32306, USA.
Email: deanna.dow@med.fsu.edu

Table 1. Participant demographics.

Characteristics, *M* (*SD*)	Diagnostic group		
	ASD	DD	TD
N	130	61	56
Age in months, *M* (*SD*)	20.75 (2.02)	20.82 (1.78)	20.82 (1.47)
Gender, *n* (%)			
Male	110 (84.6)	41 (67.2)	31 (55.4)
Female	20 (15.4)	20 (32.8)	25 (44.6)
Race, *n* (%)			
White	91 (70.0)	37 (60.7)	40 (71.4)
Black	22 (16.9)	17 (27.9)	7 (12.5)
Asian	3 (2.3)	0 (0.0)	1 (1.8)
Biracial	14 (10.8)	7 (11.5)	8 (14.3)
Ethnicity, *n* (%)			
Hispanic	21 (16.2)	6 (9.8)	0 (0.0)
Maternal education in years, *M* (*SD*)	14.87 (2.48)	14.26 (2.36)	16.33 (2.76)

SD: standard deviation; ASD: autism spectrum disorder; DD: developmentally delayed; TD: typically developing.

(STAT; Stone et al., 2000, 2008), offer an additional method of screening that allows direct observation of subtle impairments that may not be readily recognized by parents (Rutter, 2006). However, the STAT is designed for children older than the recommended screening age (i.e. 24–36 months) and requires administration by a trained professional familiar with ASD, greatly reducing its feasibility for implementation in community settings. A recent study (Gabrielsen et al., 2015) also examined the use of an observational rating tool during two 10-min samples of the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 1999). Experts missed 39% of children with ASD when asked whether they would refer the child for an evaluation after each 10-min sample, indicating that this rating tool is not effective in detecting risk for ASD in toddlers in a brief observation during a semi-structured interaction. Additional brief observational screening measures have been developed, including the Autism Detection in Early Childhood (ADEC; Dix et al., 2015) and the Rapid Interactive Test for Autism in Toddlers (RITA-T; Choueiri and Wagner, 2015), though further research with larger samples is needed to evaluate their potential utility. Despite the potential benefit of using level-two detection methods to determine early risk for ASD, adequate measures are not currently available for toddlers within the recommended screening age.

The Systematic Observation of Red Flags (SORF) is an observational measure that was designed for children who have received the Communication and Symbolic Behavior Scales (CSBS; Wetherby and Prizant, 2002). The CSBS behavior sample is a standardized, norm-referenced instrument designed to measure early social communication (SC) skills as a follow-up to the ITC and is not an ASD-specific measure; therefore, it cannot be used independently to determine ASD risk. However, the CSBS provides an ideal context to observe behaviors of ASD, as it offers a brief systematic sample with structured and unstructured activities to observe SC skills. Because it is often included as part of early communication evaluations, the SORF may be more feasible than administering diagnostic assessments that require extensive training and reliability requirements. The SORF was designed as a screening measure to identify children who should be referred for a diagnostic evaluation by a professional with expertise in ASD. The purpose of this study was (1) to examine diagnostic group differences and item-level performance to create an algorithm with the best SORF items, and (2) to examine sensitivity, specificity, and appropriate cutoff scores for the SORF as a level-two screener for ASD in 16–24 month toddlers when used with the CSBS and coded by individuals who are not experts in ASD.

Methods

Participants

This study included 247 toddlers evaluated at Florida State University through the FIRST WORDS® Project, a longitudinal, prospective study investigating early detection of ASD and other communication disorders. Parents completed the ITC in primary care settings between 9 and 24 months of age and were referred for a communication evaluation if their child scored in the bottom 10th percentile or if parents reported that they currently had concerns about their child's development. A small proportion of children were also referred to the FIRST WORDS Project directly because of parental or professional concern. These parents completed the ITC prior to or during their initial appointment. Children were included in this study if they had a SORF and diagnostic assessment completed between 16 and 24 months. Participant characteristics are presented in Table 1. Informed consent was obtained from all parents and the study was conducted in accordance with the Florida State University Institutional Review Board.

SORF coding system and procedures

The SORF is a coding system designed to detect 22 Red Flags (RF) for ASD in toddlers based on current diagnostic criteria (*Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; DSM-5)) providing an autism screening observation measure with 11 items from each domain—SC and Restricted Repetitive Behaviors (RRB). A previous version of the SORF included 29 items derived from diagnostic criteria in the fourth text revision of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; DSM-IV) and research on young children with ASD (Wetherby et al., 2004). Items were revised or removed based on initial research findings, and new items were added to capture additional behaviors and align with DSM-5 criteria. For this study, coding was completed by undergraduate research assistants blind to diagnostic classification while observing a video-recording of the CSBS behavior sample, which lasts approximately 20 min. Behaviors are coded on the SORF using a 0–3 graded response system, with 0 indicating an absence of relevant concern and 3 indicating the greatest level of severity or concern.

Two different types of scores were generated. Scores on the best performing items were included in a total Composite score and SC and RRB Domain score algorithms. Information on how these items were identified is presented later in this section. Scores indicating clear symptom presence (i.e. a score of 2 or 3) were collapsed to yield a count of the number of RF. All items were included in the total RF symptom count, and all 11 SC and 11 RRB items were included in separate domain RF symptom counts. The Composite and Domain scores provide continuous measures of severity of ASD behaviors, while the number of RF provides diagnostically useful information about the presence or absence of clinically significant behaviors.

Undergraduate research assistants were trained on the SORF coding and reached reliability when they completed 35 training videos with generalizability (g) intraclass correlation coefficients of at least 0.60. Approximately 15% of the SORFs included in this sample were scored by two coders to determine inter-rater reliability. Results indicated a g coefficient of 0.86 for the total including all SORF items, 0.84 for the SC Domain items, and 0.76 for the RRB Domain items. All coders reached reliability.

Diagnostic procedures and measures

Diagnostic procedures. A best estimate clinical diagnosis of ASD ($n=130$), developmentally delayed (DD; $n=61$), or typically developing (TD; $n=56$) was made between 18 and 24 months based on the ADOS Toddler Module (ADOS-T; Lord et al., 2012), Mullen Scales of Early Learning (MSEL; Mullen, 1995), Vineland Adaptive Behavior Scale–Second Edition (VABS-II; Sparrow et al., 1984), a video-recorded home observation, and a parent-report questionnaire, the Early Screening for Autism and Communication Disorders (ESAC). SORF coding was not used in diagnostic decision-making. Information regarding pregnancy complications, recurring medical problems, family history of learning and developmental problems, and parental concern about the child's development was also available for diagnostician review. Children were diagnosed with ASD if they demonstrated clinically significant impairment consistent with DSM-5 criteria in both symptom domains across multiple contexts. Children were classified as DD if ASD was ruled out and MSEL scores were in the delayed range (i.e. T score <38 on any subscale or 1.25 *SD* below the mean). The DD group comprised of a majority (59.1%) of children with receptive and/or expressive language delay, as well as children with nonverbal or global delays. Children were classified as TD if ASD was ruled out and DD was not present (i.e. T score $\geqslant 38$ on every MSEL subscale). A decision was made to defer diagnosis ($n=29$) if some ASD symptomatology was present, but there was insufficient evidence to meet diagnostic criteria (e.g. lack of symptoms in one of the two domains, symptoms present but judged to be not clinically significant, symptoms only present in one setting). These children were excluded from the present analyses because diagnosis was used as a predictor or outcome. Diagnostic evaluation characteristics are provided in Table 2.

Autism symptoms. The ADOS-T is a semi-structured, standardized assessment of communication, social interaction, and restricted and repetitive behaviors for toddlers to measure symptoms of ASD. The measure provides symptom domain scores (i.e. Social Affect, Restricted and Repetitive Behaviors) and a total score. Diagnostic cutoffs for the total score yield three categories reflecting degree of concern for ASD: "little-to-no concern," "mild-to-moderate concern," and "moderate-to-severe concern." Calibrated severity scores (Esler et al., 2015) were used to estimate ASD symptom severity across age range and language level.

Adaptive behavior. The VABS-II is a caregiver interview that measures adaptive functioning using a standard score, the Adaptive Behavior Composite, which combines four domain scores: Communication, Daily Living, Social, and Motor Skills.

Developmental level. The MSEL is a standardized assessment of cognitive functioning administered directly to the child. Standard scores for the Visual Reception, Fine and Gross Motor skills, and Expressive and Receptive Language subscales were obtained. Because of the substantial proportion of children who received a T-score of 20 (i.e. the floor) on one or more subtests (27.5%), developmental quotients (DQs) were calculated. DQs were calculated by first dividing the age equivalent for each subscale by the

Table 2. Descriptive statistics for outcome measures.

Characteristics, *M* (*SD*)	Diagnostic group		
	ASD	DD	TD
N	130	61	56
Age at ADOS	20.99 (2.14)	20.82 (1.78)	20.81 (1.45)
ADOS SA CSS score	7.25 (1.91)	3.10 (1.63)	2.39 (1.47)
ADOS RRB CSS score	6.43 (1.89)	3.30 (2.16)	3.39 (2.34)
ADOS Total CSS score	7.07 (1.79)	2.80 (1.38)	2.23 (1.29)
Age at MSEL	20.61 (2.29)	20.47 (1.68)	20.96 (2.14)
MSEL Gross Motor T	48.60 (9.99)	48.15 (10.03)	47.57 (10.35)
MSEL Fine Motor T	43.12 (10.78)	43.18 (10.58)	53.91 (7.67)
MSEL Visual Reception T	40.40 (11.73)	44.48 (12.72)	56.75 (10.17)
MSEL Receptive Language T	30.55 (12.28)	38.18 (13.96)	58.73 (10.00)
MSEL Expressive Language T	28.55 (9.47)	32.18 (10.25)	49.96 (8.17)
MSEL Nonverbal DQ	96.18 (16.83)	100.47 (15.05)	118.18 (12.51)
MSEL Verbal DQ	66.06 (22.07)	79.54 (20.58)	116.97 (17.32)
Age at VABS-II	20.36 (2.44)	20.48 (1.84)	20.55 (1.39)
VABS-II ABC	83.40 (8.26)	88.02 (8.61)	93.40 (6.56)

SD: standard deviation; ASD: autism spectrum disorder; DD: developmentally delayed; TD: typically developing; ADOS: Autism Diagnostic Observation Schedule; SA: Social Affect score; CSS: Calibrated Severity Score; RRB: Restricted Repetitive Behavior score; MSEL: Mullen Scales of Early Learning; T: T Score; DQ: developmental quotient; VABS-II: Vineland Adaptive Behavior Scale–Second Edition; ABC: Adaptive Behavior Composite.

child's chronological age and multiplying by 100; nonverbal DQ was then derived by averaging the visual reception and fine motor skills subscales, and verbal DQ was derived by averaging the expressive and receptive language subscales. Analysis of variance (ANOVA) results revealed that the ASD and DD groups did not differ significantly on nonverbal DQ ($F = 39.88$, $p = 0.14$), indicating that differences between these diagnostic groups cannot be accounted for by cognitive level alone.

ASD symptoms across contexts. A video-recorded home observation provided an opportunity for diagnosticians to examine the presence and severity of ASD symptoms in a naturalistic setting when determining diagnostic classification. Parents were instructed to interact with their child during a variety of everyday activities (e.g. snack, play, family chores, caregiving, book sharing) for 1 h. Diagnosticians also reviewed an ASD-specific questionnaire, the ESAC, to gain insight into parent report of ASD symptoms. Preliminary results provide strong support for the validity of the ESAC as an autism-specific parent-report screener (Wetherby et al., 2015).

Statistical analysis

Item-level analysis. Individual items were first examined using one-way ANOVA models to determine whether specific behaviors differentiated children with ASD from nonspectrum groups (DD and TD). A Welch correction was used to account for lack of homogeneous variances, and Dunnett's C post hoc testing corrected for Type I error when evaluating pairwise differences among means. Effect sizes were calculated using Cohen's *d* to reflect the size of diagnostic group differences ($\leqslant 0.2$ = small, 0.5 = medium, 0.8 = large). Children with DD and TD were then combined to form the nonspectrum group ($n = 117$) used for receiver operating characteristic (ROC) curve analyses, which provided information regarding each item's ability to discriminate the ASD and nonspectrum groups. Area under the curve (AUC) was examined to determine the strength of discrimination between groups for each item, with AUC values ranging from 0.5 denoting that no discrimination exists to 1.0 denoting perfect discrimination (Swets, 1988). All items with medium to large effect sizes and AUC values of at least 0.60 were included in the algorithm.

Summary score analysis. The Composite, RF, SC and RRB Domain scores, and SC and RRB RF scores were examined using one-way ANOVA models, generating *F*-statistics and pairwise group difference comparisons with a Welch correction for lack of homogeneous variances and Dunnett's C post hoc testing to correct for Type I error. Effect sizes were calculated using Cohen's *d*.

ROC curve analyses generated sensitivity, specificity, and optimal cutoff scores for each SORF score. The ROC curve plots the "true positive" rate (i.e. the proportion of children with ASD who are correctly identified), or sensitivity, against the "false positive" rate (i.e. the proportion of children who are identified incorrectly as having ASD) across the full range of cutoff scores. Specificity indicates the proportion of children without ASD who are correctly identified as not at risk ("true negatives"). The optimal cutoff score was determined by prioritizing sensitivity while

Table 3. Group differences of SORF summed scores and items.

	ASD ($n = 130$)	DD ($n = 61$)	TD ($n = 56$)		Pairwise group differences			
					ASD-DD		ASD-TD	
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*F*(2, 244)	*d*	*p*	*d*	*p*
Composite score	25.79 (8.18)	15.61 (7.36)	10.02 (6.76)	92.85***	1.3	0.000	2.1	0.001
Number of Red Flags	10.44 (3.84)	5.97 (3.31)	3.61 (3.12)	87.04***	1.2	0.000	2.0	0.001
SC Domain score	19.52 (6.26)	12.62 (6.10)	7.13 (5.44)	95.31***	1.1	0.000	2.1	0.000
SC number of Red Flags	7.27 (2.88)	4.33 (2.83)	2.11 (2.16)	91.81***	1.0	0.000	2.0	0.000
RRB Domain score	6.26 (3.42)	2.98 (2.52)	2.78 (2.69)	38.10***	1.1	0.000	1.1	0.000
RRB number of Red Flags	3.17 (1.88)	1.64 (1.42)	1.50 (1.66)	26.84***	0.9	0.000	0.9	0.000
1. Warm, joyful expressions	1.75 (1.04)	1.20 (1.09)	0.96 (0.99)	13.78***	0.5	0.001	0.8	0.000
2. Reduced facial expressions	1.57 (0.81)	1.23 (0.96)	0.88 (0.81)	14.79***	0.4	0.020	0.9	0.000
3. Sharing interests	1.80 (1.06)	1.23 (0.96)	0.63 (0.91)	29.83***	0.6	0.001	1.2	0.000
4. Response to name	1.97 (0.88)	1.08 (0.92)	0.52 (0.83)	62.07***	1.0	0.000	1.7	0.000
5. Eye gaze directed to faces	1.79 (0.75)	1.07 (0.96)	0.68 (0.77)	45.87***	0.8	0.000	1.5	0.000
6. Showing and pointing	2.05 (1.01)	1.30 (1.02)	0.61 (0.87)	50.23***	0.7	0.000	1.5	0.000
7. Using hand as a tool	0.21 (0.67)	0.00 (0.00)	0.07 (0.42)	3.64*	0.3	0.022	0.3	0.194
8. Directed consonant sounds	2.32 (0.83)	2.02 (0.79)	1.09 (0.90)	37.94***	0.4	0.042	1.4	0.000
9. Nonverbal communication	2.28 (0.95)	1.46 (0.96)	0.71 (0.97)	54.64***	0.9	0.000	1.6	0.000
10. Interest in objects over people	1.78 (0.88)	0.87 (0.85)	0.46 (0.63)	69.57***	1.1	0.000	1.7	0.000
11. Reciprocal social play	2.11 (0.10)	1.18 (0.98)	0.59 (0.89)	56.35***	1.3	0.000	2.4	0.000
12. Repetitive use of objects	1.22 (1.18)	0.41 (0.74)	0.60 (0.89)	17.20***	0.8	0.000	0.6	0.000
13. Repetitive body movement	1.16 (1.13)	0.75 (0.98)	0.86 (1.12)	3.61*	0.4	0.033	0.3	0.154
14. Repetitive speech	1.38 (1.16)	0.93 (1.05)	0.84 (1.11)	6.07**	0.4	0.020	0.5	0.005
15. Ritualized behavior	0.55 (0.87)	0.21 (0.58)	0.23 (0.66)	6.08**	0.5	0.009	0.4	0.018
16. Distress over change	0.75 (1.00)	0.28 (0.64)	0.23 (0.66)	10.90***	0.6	0.001	0.6	0.000
17. Excessive interest	0.80 (0.99)	0.38 (0.73)	0.37 (0.68)	7.58**	0.5	0.004	0.5	0.005
18. Clutches objects	0.57 (0.96)	0.56 (0.87)	0.48 (0.81)	0.19	0.0	0.995	0.1	0.786
19. Sticky attention	0.87 (1.09)	0.52 (0.79)	0.36 (0.75)	7.16**	0.4	0.040	0.6	0.002
20. Fixation on object parts	0.19 (0.56)	0.10 (0.44)	0.04 (0.27)	3.33*	0.2	0.359	0.3	0.078
21. Sensory aversion	0.34 (0.72)	0.54 (0.79)	0.30 (0.74)	1.94	−0.3	0.148	0.1	0.944
22. Sensory interest	0.68 (0.98)	0.25 (0.65)	0.13 (0.43)	14.73***	0.5	0.001	0.7	0.000

SORF: Systematic Observation of Red Flags; ASD: autism spectrum disorder; DD: developmentally delayed; TD: typically developing; *SD*: standard deviation; SC: Social Communication; RRB: Restricted Repetitive Behaviors.
Dunnett's C post hoc comparisons were used to correct for Type I error.
F-values are Welch corrected when necessary for violation of homogeneity of variance as assessed by Levene's test.
Cohen's *d*: ⩽0.2 = small, 0.5 = medium, 0.8 = large.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

maintaining an adequate level of specificity to maximize accurate detection of children with ASD while minimizing identification of children without ASD. Positive predictive value (PPV) and negative predictive value (NPV) were also calculated using optimal cutoffs for each summary score. PPV is the probability that children with a positive screen were diagnosed with ASD, while NPV is the probability that children with a negative screen were not diagnosed with ASD.

Results

Item-level analysis

Item-level analysis reported in Table 3 revealed significant group differences for 19 of 22 items (11 SC and 8 RRB), with the ASD group showing higher scores than at least one nonspectrum group. A total of 17 of these items discriminated between ASD and both diagnostic groups. Differences were observed between ASD and DD (but not ASD and TD) groups for using hand as tool and repetitive body movements. Clutching objects, fixation on object parts, and sensory aversion did not significantly differ between any of the groups. A total of 17 items (10 SC and 7 RRB) demonstrated medium to large effect sizes (i.e. $d \geqslant 0.5$) between ASD and at least one nonspectrum group, and 12 items (8 SC and 4 RRB) demonstrated medium to large effects sizes between ASD and both nonspectrum groups. Item-level performance was also examined using ROC curve analyses (see Table 4 for results). In all, 18 of the 22 items had statistically significant AUC

Table 4. Item-level ROC curve analysis: ASD ($n = 130$) versus nonspectrum (TD/DD; $n = 117$).

	AUC (*SE*)	95% CI
1. Warm, joyful expressions	0.67*** (0.03)	0.60–0.74
2. Reduced facial expressions	0.66*** (0.04)	0.59–0.73
3. Sharing interests	0.72*** (0.03)	0.65–0.78
4. Response to name	0.80*** (0.03)	0.74–0.86
5. Eye gaze directed to faces	0.77*** (0.03)	0.71–0.83
6. Showing and pointing	0.77*** (0.03)	0.71–0.83
7. Using hand as a tool	0.55 (0.04)	0.47–0.62
8. Directed consonant sounds	0.72*** (0.03)	0.66–0.78
9. Nonverbal communication	0.79*** (0.03)	0.74–0.85
10. Interest in objects over people	0.81*** (0.03)	0.76–0.87
11. Reciprocal social play	0.79*** (0.03)	0.74–0.85
12. Repetitive use of objects	0.66*** (0.04)	0.59–0.73
13. Repetitive body movement	0.59* (0.04)	0.51–0.66
14. Repetitive speech	0.62** (0.04)	0.55–0.69
15. Ritualized behavior	0.60** (0.04)	0.53–0.67
16. Distress over change	0.63*** (0.04)	0.56–0.70
17. Excessive interest	0.61** (0.04)	0.54–0.68
18. Clutches objects	0.50 (0.04)	0.42–0.57
19. Sticky attention	0.60** (0.04)	0.53–0.67
20. Fixation on object parts	0.54 (0.04)	0.47–0.61
21. Sensory aversion	0.46 (0.04)	0.39–0.54
22. Sensory interest	0.63*** (0.04)	0.56–0.70

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.
ROC: receiver operating characteristics; ASD: autism spectrum disorder; TD: typically developing; DD: developmentally delayed; AUC: area under the curve; *SE*: standard error; CI: confidence interval.

values. The remaining 4 items demonstrated nonsignificant discrimination between groups with AUC values near chance (i.e. 0.50): use of hand as tool, clutching objects, fixation on object parts, and sensory aversion, indicating that these items do not significantly predict diagnostic risk in 16- to 24-month-old toddlers in this clinical setting. A total of 17 items demonstrated individual AUC values of at least 0.60.

The 17 items with medium to large effect sizes and AUC values of at least 0.60 were included in an algorithm used to compute the Composite and SC and RRB Domain scores, providing continuous measures of ASD severity. Because the intended purpose of the RF scores is to provide diagnostically useful information about clinically significant behaviors, all 22 items were included in these scales. Including all items in the RF scales did not decrease sensitivity compared to including only algorithm items.

Summary score analyses

ANOVA results (see Table 3) revealed significant differences between the ASD and nonspectrum groups for the Composite, RF, SC Domain, SC RF, RRB Domain, and RRB RF scores with large effect sizes. Larger mean group differences were observed for the SC scores compared to RRB, even for Domain scores that included only optimally

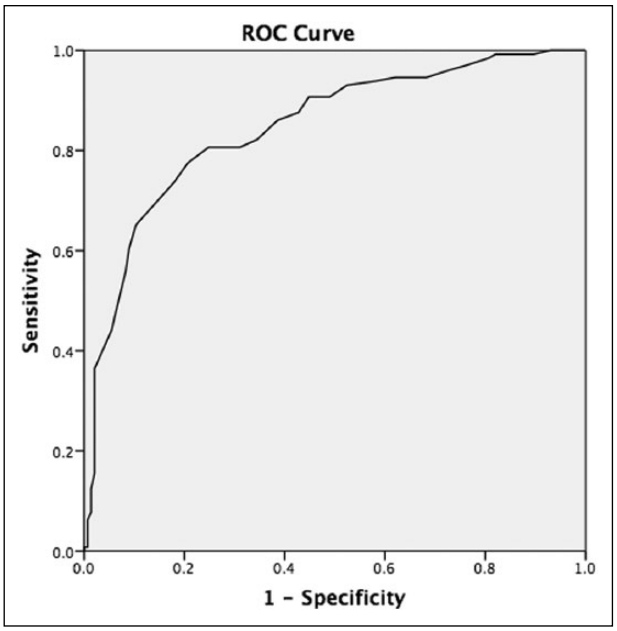

Figure 1. ROC curve for Composite score.

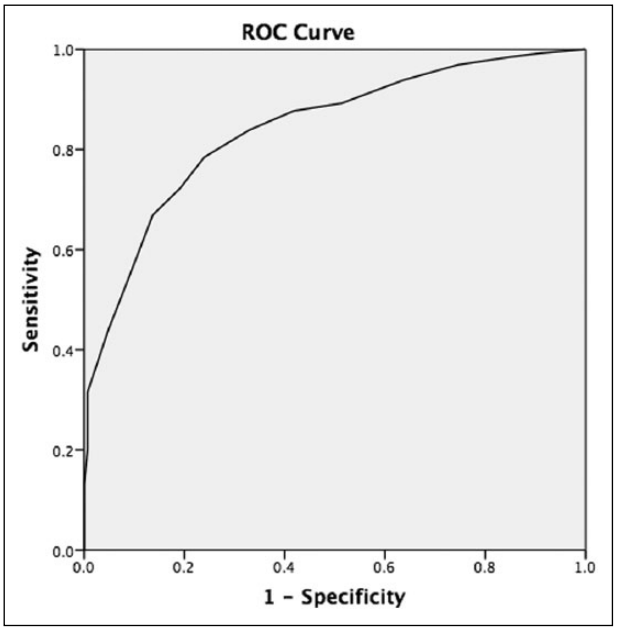

Figure 2. ROC curve for number of Red Flags (RF).

performing algorithm items. ROC curve analyses revealed that the Composite and RF scores provided good discrimination (see Figures 1 and 2; Table 5) between the ASD and nonspectrum groups, with AUC values of 0.87 and 0.86, respectively. The optimal cutoff for the Composite score was 20, with a sensitivity of 0.80, specificity of 0.78, PPV of 0.81, and NPV of 0.78. The optimal cutoff for the RF score was 8, with a sensitivity of 0.79, specificity of 0.75, PPV of 0.78, and NPV of 0.76. The SC Domain score demonstrated good discrimination (AUC = 0.85), yielding an optimal cutoff of 14 with a sensitivity of 0.80, specificity

Table 5. ROC curve analysis: ASD ($n = 130$) versus nonspectrum (TD/DD; $n = 117$).

	AUC (*SE*)	95% CI	Sensitivity	Specificity	Cutoff	PPV	NPV
Composite score	0.87*** (0.02)	0.82–0.91	0.80	0.78	20	0.81	0.78
Number of Red Flags	0.86*** (0.02)	0.81–0.90	0.79	0.75	8	0.78	0.76
SC Domain Score	0.85*** (0.02)	0.80–0.90	0.80	0.72	14	0.76	0.76
SC Red Flags	0.84*** (0.03)	0.79–0.89	0.79	0.68	5	0.73	0.74
RRB Domain Score	0.79*** (0.03)	0.73–0.84	0.79	0.66	4	0.72	0.73
RRB Red Flags	0.75*** (0.03)	0.69–0.81	0.79	0.52	2	0.65	0.69

***$p < 0.001$.
ROC: receiver operating characteristics; ASD: autism spectrum disorder; TD: typically developing; DD: developmentally delayed; AUC: area under the curve; *SE*: standard error; CI: confidence interval; PPV: positive predictive value; NPV: negative predictive value; SC: Social Communication; RRB: Restricted Repetitive Behaviors.

of 0.72, PPV of 0.76, and NPV of 0.76. The SC RF score demonstrated similar discrimination (AUC = 0.84) with an optimal cutoff of 5, which yields a sensitivity of 0.79, specificity of 0.68, PPV of 0.73, and NPV of 0.74. The RRB subscale demonstrated weaker discrimination (RRB Domain AUC = 0.79; RRB RF AUC = 0.75). The RRB Domain had an optimal cutoff of 4, sensitivity of 0.79, specificity of 0.66, PPV of 0.72, and NPV of 0.73, while the RRB RF had an optimal cutoff of 2, sensitivity of 0.79, specificity of 0.52, PPV of 0.65, and NPV of 0.69.

Discussion

Our findings support the utility of the SORF as an observational, level-two screening tool for ASD when used with the CSBS behavior sample. Using a Composite cutoff of 20 is recommended for optimal performance. The Composite and Domain scores provide continuous measures of ASD behaviors to quantify severity of current symptoms and may be useful to measure symptom change over time. In contrast, the RF scores may be most beneficial for clinicians who are interested in the presence and number of clinically significant symptoms that fit diagnostic criteria for ASD. One advantage of the SORF over other screeners is that it can be utilized in two meaningful ways, with the same tool serving as both a screener based on an optimal cutoff and a measure of symptom severity.

The SORF was developed to provide a practical alternative to other available screening measures in response to speculation that better tools are needed to justify routine ASD screening at 18–24 months (Al-Qabandi et al., 2011). While the M-CHAT offers an accessible, brief questionnaire, its rate of ASD detection is lower than expected based on population prevalence estimates, even when follow-up is completed (i.e. 67 cases per 10,000; Robins et al., 2014). In addition to at-risk children not presenting for diagnostic evaluations, this could also indicate that the rate of false negatives is higher than expected. Children who screened positive on the SORF in this sample also have higher developmental scores (based on the MSEL and VABS-II) than a sample identified by the M-CHAT (Robins et al., 2014), demonstrating a potential strength of the SORF in detecting high-functioning toddlers from a primary care sample. Children with ASD who were accurately detected by the M-CHAT had average MSEL scores over two *SD* below the mean for all subscales, whereas our ASD sample was within one *SD* of the mean for all subscales except language (receptive and expressive). The M-CHAT's lower-functioning ASD sample may reflect bias in sample recruitment and may inflate estimates of the measure's sensitivity. More research is needed for other observational screening tools reported in the literature (i.e. STAT, ADEC, and RITA-T) with large community-ascertained samples to determine effectiveness as level-two screeners in primary care settings. Furthermore, though the observational screening tool studied by Gabrielsen et al. (2015) intended to provide a brief 10-min observational measure similar to the SORF, the screener did not adequately predict risk for ASD. This finding suggests that a longer period of observation and/or observation of more behaviors may be needed in an effective screening measure. The SORF provides an important alternative to available screening methods, demonstrating efficacy in a sample ascertained from a primary care population with a larger percentage of high-functioning children with ASD than detected by other existing measures.

Item-level analysis

Children with ASD demonstrated more severe autism symptoms compared to other diagnostic groups (DD and TD) on most SORF items. A total of 19 of 22 items distinguished ASD from at least one other group, with 17 items selected for the algorithm due to optimal performance across analyses. SORF items are specific, easily observable behaviors that reflect manifestations of DSM-5 diagnostic criteria in toddlers. Results indicating that the large number of SORF items that discriminated between ASD and DD/TD suggest that a wide range of specific behaviors may be useful in accurately detecting risk for ASD, consistent with the heterogeneity of the clinical presentation of the disorder.

The SC Domain demonstrated similar performance to the Composite and RF scores when examined separately. This finding is consistent with evidence that prospective questions about SC deficits are effective when screening children under 2 for ASD (Charman et al., 1997; Wetherby et al., 2008). The RRB scale did not discriminate between groups as well when examined separately. While it is clear that RRBs are present in young toddlers (Elison et al., 2014; Guthrie et al., 2013; Morgan et al., 2008), this evidence is consistent with other studies that have shown that these symptoms do not discriminate children with ASD as effectively as SC symptoms when used in screening measures (Berument et al., 1999; Rowberry et al., 2015).

Results also indicated that some ASD-specific behaviors might not be effective items when using an observational screening tool for toddlers as a measure of ASD severity. The five items that were not included in the Composite and Domain score algorithm due to reduced performance across analyses were using hand as tool, repetitive body movement, clutching objects, fixation on object parts, and sensory aversion. Because these behaviors may indicate concern for ASD and their presence is clinically relevant to a child's symptomology based on DSM-5 criteria, these items were not removed from the measure and were included in the RF scores, which are intended for clinicians to provide diagnostically useful information about symptoms relevant to ASD.

Limitations and future directions

A possible advantage of the SORF over other observational screeners, such as the STAT, is that it is not coded based on specific, required activities and may be able to be applied in observational contexts other than the CSBS. Given that a structured observation such as the CSBS may not be feasible in many settings, future research will examine whether community professionals can utilize the SORF in a more unstructured, accessible context such as a home observation. Future studies will also expand on the utility of the SORF at younger ages (e.g. 12 months) to improve early detection. Use of the SORF in combination with a parent-report measure will be explored specifically to determine how it functions as a diagnostic measure, or if it can help to triage children who may have clear ASD from those needing careful diagnostic assessment to confirm or rule out ASD. While these results suggest that the SORF is a promising new observational screening tool, further research is needed to replicate these findings with an independent sample and to study methods to increase time- and cost-efficiency in order to improve screening options for detecting ASD risk in young toddlers.

Declaration of conflicting interests

Ms Guthrie receives royalties from use of the Autism Diagnostic Observation Schedule Toddler Module (ADOS-T). Dr Wetherby receives royalties from use of the Communication and Symbolic Behavior Scales (CSBS), but not from this study.

Funding

This research was supported in part by the Eunice Kennedy Shriver National Institute of Child Health & Human Development grants RO1HD078410 and RO1HD065272, the National Institute on Deafness and Other Communication Disorders grant R01DC007462, and the Centers for Disease Control and Prevention Cooperative Agreement U01DD000304.

References

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

Baio J (2014) Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network. *Morbidity and Mortality Weekly Report: Surveillance Summaries* 63(2): 1–21.

Berument SK, Rutter M, Lord C, et al. (1999) Autism screening questionnaire: diagnostic validity. *The British Journal of Psychiatry* 175(5): 444–451.

Charman T, Swettenham J, Baron-Cohen S, et al. (1997) Infants with autism: an investigation of empathy, pretend play, joint attention, and imitation. *Developmental Psychology* 33(5): 781–789.

Chawarska K, Paul R, Klin A, et al. (2007) Parental recognition of developmental problems in toddlers with autism spectrum disorders. *Journal of Autism and Developmental Disorders* 37(1): 62–72.

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

Dawson G, Rogers S, Munson J, et al. (2010) Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics* 125(1): e17–e23.

Dix L, Fallows R and 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.

Elison JT, Wolff JJ, Reznick JS, et al. (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 AN, Bal VH, Guthrie W, et al. (2015) The autism diagnostic observation schedule, toddler module: standardized severity scores. *Journal of Autism and Developmental Disorders* 45(9): 2704–2720.

Gabrielsen TP, Farley M, Speer L, et al. (2015) Identifying autism in a brief observation. *Pediatrics* 135(2): e330–e338.

Guthrie W, Swineford LB, Nottke C, et al. (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.

Johnson CP and Myers SM (2007) Identification and evaluation of children with autism spectrum disorders. *Pediatrics* 120(5): 1183–1215.

Kleinman JM, Robins DL, Ventola PE, et al. (2008) The modified checklist for autism in toddlers: a follow-up study

investigating the early detection of autism spectrum disorders. *Journal of Autism and Developmental Disorders* 38(5): 827–839.

Lord C, Luyster R, Gotham K, et al. (2012) *Autism Diagnostic Observation Schedule—Toddler Module Manual*. Los Angeles, CA: Western Psychological Services.

Lord C, Rutter M, DiLavore P, et al. (1999) *Autism Diagnostic Observation Schedule—Generic*. Los Angeles, CA: Western Psychological Services.

Morgan L, Wetherby AM and Barber A (2008) Repetitive and stereotyped movements in children with autism spectrum disorders late in the second year of life. *Journal of Child Psychology and Psychiatry* 49(8): 826–837.

Mullen EM (1995) *Mullen Scales of Early Learning*. Circle Pines, MN: American Guidance Service.

Robins DL, Casagrande K, Barton M, et al. (2014) Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). *Pediatrics* 133(1): 37–45.

Robins DL, Fein D, Barton ML, et al. (2001) The Modified Checklist for Autism in Toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. *Journal of Autism and Developmental Disorders* 31(2): 131–144.

Rowberry J, Macari S, Chen G, et al. (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.

Rutter M (2006) Autism: its recognition, early diagnosis, and service implications. *Journal of Developmental and Behavioral Pediatrics* 27(2): S54–S58.

Sparrow SS, Cicchetti DV, Balla DA (2005) *Vineland Adaptive Behavior Scales (Vineland II), Survey Interview Form/Caregiver Rating Form*. Livonia, MN: Pearson Assessments.

Stone WL, Coonrod EE and Ousley OY (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 WL, McMahon CR and Henderson LM (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.

Swets JA (1988) Measuring the accuracy of diagnostic systems. *Science* 240(4857): 1285–1293.

Wetherby AM and Prizant BM (2002) *Communication and Symbolic Behavior Scales Developmental Profile*. First Normed ed. Baltimore, MD: Paul H. Brookes Publishing Co.

Wetherby AM, Brosnan-Maddox S, Peace V, et al. (2008) Validation of the Infant-Toddler Checklist as a broadband screener for autism spectrum disorders from 9 to 24 months of age. *Autism* 12(5): 487–511.

Wetherby AM, Guthrie W, Petkova E, et al. (2015) Broadband and autism-specific screening using the Early Screening for Autism and Communication Disorders (ESAC). In: *International meeting for autism research*, Salt Lake City, UT, 13–16 May.

Wetherby AM, Guthrie W, Wood J, et al. (2014) Parent-implemented social intervention for toddlers with autism: an RCT. *Pediatrics* 134(6): 1084–1093.

Wetherby AM, Woods J, Allen L, et al. (2004) Early indicators of autism spectrum disorders in the second year of life. *Journal of Autism and Developmental Disorders* 34(5): 473–493.