*Original Article*

autism

Autism
2021, Vol. 25(2) 516–528

DOI: 10.1177/1362361320967529
journals.sagepub.com/home/aut


# Autism spectrum disorder screening in preschools

**Angel Hoe-chi Au[1], Kathy Kar-man Shum[2], Yongtian Cheng[3], Hannah Man-yan Tse[2], Rose Mui-fong Wong[4], Johnson Li[3] and Terry Kit-fong Au[2]**

## Abstract

Can non-clinicians spot preschoolers likely to have autism spectrum disorder by observing their everyday peer interaction? We set out to develop a screening tool that capitalizes on peer interaction as a naturalistic "stress test" to identify children more likely than their peers to have autism spectrum disorder. A total of 304 3- to 4-year-olds were observed at school with an 84-item preliminary checklist; data-driven item reduction yielded a 13-item Classroom Observation Scale. The Classroom Observation Scale scores correlated significantly with Autism Diagnostic Observation Schedule–2 scores. To validate the scale, another 322 2- to 4-year-olds were screened using the Classroom Observation Scale. The screen-positive children and randomly selected typically developing peers were assessed for autism spectrum disorder 1.5 years later. The Classroom Observation Scale as used by teachers and researchers near preschool onset predicted autism spectrum disorder diagnoses 1.5 years later (odds ratios = 14.6 and 6.7, respectively). This user-friendly 13-item Classroom Observation Scale enables teachers and healthcare workers with little or no clinical training to identify, with reliable and valid results, preschoolers more likely than their peers to have autism spectrum disorder.

## Lay abstract

With professional training and regular opportunities to observe children interacting with their peers, preschool teachers are in a good position to notice children's autism spectrum disorder symptomatology. Yet even when a preschool teacher suspects that a child may have autism spectrum disorder, fear of false alarm may hold the teacher back from alerting the parents, let alone suggesting them to consider clinical assessment for the child. A valid and convenient screening tool can help preschool teachers make more informed and hence more confident judgment. We set out to develop a screening tool that capitalizes on peer interaction as a naturalistic "stress test" to identify children more likely than their peers to have autism spectrum disorder. A total of 304 3- to 4-year-olds were observed at school with an 84-item preliminary checklist; data-driven item reduction yielded a 13-item Classroom Observation Scale. The Classroom Observation Scale scores correlated significantly with Autism Diagnostic Observation Schedule–2 scores. To validate the scale, another 322 2- to 4-year-olds were screened using the Classroom Observation Scale. The screen-positive children and randomly selected typically developing peers were assessed for autism spectrum disorder 1.5 years later. The Classroom Observation Scale as used by teachers and researchers near preschool onset predicted autism spectrum disorder diagnoses 1.5 years later. This user-friendly 13-item Classroom Observation Scale enables teachers and healthcare workers with little or no clinical training to identify, with reliable and valid results, preschoolers more likely than their peers to have autism spectrum disorder.

## Keywords

autism spectrum disorder, early identification, peer interaction, preschool, screening

[1]Autism Partnership Hong Kong, China
[2]University of Hong Kong, China
[3]University of Manitoba, Canada
[4]WAY Psychological Services, China

**Corresponding authors:**
Kathy Kar-man Shum, Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong, China.
Email: kkmshum@hku.hk

Terry Kit-fong Au, Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong, China.
Email: terryau@hku.hk

About 1 in 59 children has autism spectrum disorder (ASD), as estimated by the US Centers for Disease Control and Prevention (CDC) in 2019. The prevalence estimates for preschoolers are generally lower, for example, about 1 in 125 in the United States (Soke et al., 2017) or 1 in 132 in China (Wang et al., 2011). With early appropriate interventions, children with ASD—especially less severe ASD—stand a better chance of living more independently, having friends, and being in a steady relationship (Fein et al., 2013; Orinstein, Suh, et al., 2015; Orinstein, Tyson, et al., 2015; Roux et al., 2013). Such interventions are available for young children (e.g. Chang et al., 2016; Kasari et al., 2008; Reichow et al., 2012; Schreibman et al., 2015; Warren et al., 2011), but all too often children in need do not get them. In that case, better developmental outcomes tend to be elusive, even for those with relatively high IQs and intact verbal and nonverbal skills (Billstedt et al., 2005; Cederlund et al., 2007; Szatmari et al., 2003).

Many children miss out because they are not diagnosed in time, if at all. Children with ASD vary considerably in the severity of their deficits in social interaction, social communication, and social imagination (American Psychiatric Association, 2013). While severe cases can be diagnosed by age 2 or 3 years (Lord et al., 2006; Moore & Goodson, 2003), milder cases often go undiagnosed until age 6 or 7 years (CDC, 2012), and some are never diagnosed at all. Here in Hong Kong, about 10% of the cases of childhood autism based on the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10, F84.0; World Health Organization, 2004) and 17% of the cases of other ASD conditions (ICD-10, F84.1 and F84.5) are not referred for assessment until after the first grade, when social demands in the classroom and on the playground finally make the children's social impairments more obvious and problematic (Department of Health, Hong Kong, 2007).

Better tools are needed for early identification of children with ASD—especially less severe cases—for timely clinical assessment. Offering children with ASD effective treatment by age 3 years, for instance, instead of the usual age 6 or 7 years, can launch them on a better lifelong trajectory, giving them a crucial head start on understanding the social world and developing healthy social bonds. Parents are often the first to notice when something does not seem right. Reportedly, this first happens with children on the milder end of the autism spectrum around age 20 months (McConachie et al., 2005). Why then are they not diagnosed until so much older? For one thing, young children often relate better to supportive adults—like their parents—than to peers. Hence, especially in single-child families, children may not show obvious symptoms of ASD at home. With family size trending downward, recognizing signs of ASD in their young children is a major challenge for more and more parents (De Giacomo & Fombonne, 1998; Zwaigenbaum et al., 2005). It would be helpful to supplement parent reports (based on existing instruments such as Autism Behavior Checklist, Volkmar et al., 1988; Developmental Behavior Checklist—Early Screen, Gray & Tonge, 2005; Modified Checklist for Autism in Toddlers, Revised with Follow-up (M-CHAT-R/F), Robins et al., 2014; Pervasive Developmental Disorder Screening Test-II, Siegel, 2004) with other sources of information.

With professional training and regular opportunities to observe children interacting with their peers, preschool teachers are in a good position to notice children's ASD symptomatology (Duvekot et al., 2015). Yet even when a preschool teacher suspects that a child may have ASD, fear of false alarm may hold the teacher back from alerting the parents, let alone suggesting them to consider clinical assessment for the child.

A valid and convenient screening tool can help preschool teachers make more informed and hence more confident judgment. However, while there are many screening tools to help identify older children with less severe ASD in community settings (e.g. Asperger Syndrome Diagnostic Scale, Myles et al., 2001; Autism Spectrum Screening Questionnaire, Ehlers et al., 1999; Childhood Asperger Syndrome Test, Scott et al., 2002; Social and Communication Disorders Checklist, Skuse et al., 2005; Social Communication Questionnaire, Berument et al., 1999), there are far fewer tools to use with preschool children below age 4 years. For instance, the M-CHAT-R/F and the Rapid Interactive Screening Test for Autism in Toddlers (RITA-T) are screening tools widely used for children up to 30 and 36 months old, respectively (Choueiri & Wagner, 2015; Robins et al., 2001, 2014; Siu et al., 2016), but no preschool version is available to capitalize on preschool teachers' opportunities to see children in peer interaction regularly.

There is, however, the Diagnostic and Statistical Manual of Mental Disorders Autism Spectrum Problems Scale (DSM-ASD Scale; Achenbach, 2014) from the Child Behavior Checklist for Ages 1½–5 (CBCL/1½–5) and the Caregiver-Teacher Report Form (C-TRF; Achenbach & Rescorla, 2000) for ASD screening in preschool population. The 12 items on this scale can be grouped into a 7-item social communication/interaction (SCI) subscale and a 5-item restricted interests, repetitive behaviors (RRB) subscale (Rescorla, Ghassabian, et al., 2019). Rescorla, Given, et al. (2019) compared the item scores on the DSM-ASD Scale across preschool population samples from different countries and found lower similarity in mean item ratings between international samples for the C-TRF than the CBCL/1½–5. This might be due to greater variations in early childhood settings in schools (e.g. in the teacher–student ratio, classroom structure, and preschool program) than in families across societies (Rescorla, Given, et al., 2019). Such variations in schools may likely affect the relationships between preschool teachers and the children and hence the teachers' ability to observe and notice certain behaviors included in the checklist. For

instance, in preschool settings wherein the teacher–student ratio is less favorable, teachers will probably have less time to interact with and observe each child in class. As such, they may be less likely to pick up some of the behaviors described on the DSM-ASD Scale of the C-TRF, such as those that involve often fleeting social responding (e.g. "Seems unresponsive to affection") and those that require greater familiarity with the child (e.g. "Disturbed by any change in routine," "Can't stand having things out of place"). Moreover, while previous studies have provided evidence on the diagnostic accuracy of the CBCL/1½–5 as a screening instrument for ASD (Levy at al., 2019; Rescorla et al., 2015), such information is lacking for the C-TRF DSM-ASD Scale.

We therefore set out to develop a new ASD screening tool for use by teachers and other observers with minimal clinical training, who may not have known the child for very long or have had a lot of time to observe the child in his or her naturalistic settings. This new observation scale is based on the idea of a natural "stress test." According to the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; DSM-V; American Psychiatric Association, 2013), "social impairment" as a key diagnostic criterion for ASD includes "deficits in developing, maintaining, and understanding relationships" and "deficits in social-emotional reciprocity." The ICD-10 (World Health Organization, 2004) has listed "failure to develop (in a manner appropriate to mental age, and despite ample opportunities) peer relationships that involve a mutual sharing of interests, activities and emotions" as a manifestation of social impairment in childhood autism (pp. 147–149). Peer play elicits more stress (e.g. as indicated by sustained elevated cortisol) for children with ASD than for those without (Corbett et al., 2010). Coding of video-recorded peer interaction in semi-structured situations (sometimes including a child confederate) revealed that children with ASD engaged in more self-play and less cooperative play (e.g. Corbett et al., 2014). Video analysis of spontaneous free talk revealed robust differences between high-functioning preschoolers with ASD and children without ASD, especially when the conversation partners were not their friends (Bauminger-Zviely et al., 2014). As such, can children more likely to have ASD be identified using real-time observation of peer interaction in regular preschool classrooms—without requiring any child confederate, multiple video-recorders, and labor-intensive behavioral coding of videos?

For young preschoolers with less severe ASD, peer interaction without adult scaffolding and instructions can make their ASD-related deficits more apparent. When resources are sufficient, clinical psychologists sometimes make preschool visits if a clinical assessment suggests a case in a milder range of the autism spectrum. It can be telling to observe how a child interacts with peers—or does not—during free play. Indeed, a prior study has shown that preschool observation of children's free-field behavior in group activities and free play based on a protocol derived from the Autism Diagnostic Observation Schedule (ADOS) yielded similar information as that obtained at ADOS assessment performed by clinicians in a clinic (Westman Andersson et al., 2013). Yet preschool observation is not typically considered the most cost-effective use of a diagnostician's time, and hence, it is rarely done locally or in most other countries. Even where school observations are more common (e.g. in the United Kingdom), this practice can still be improved—that is, less experienced clinicians could benefit from having a valid and simple classroom observation scale.

Such a screening tool can be used by an assistant therapist or preschool teacher to gather valid results of peer interaction as a naturally occurring "stress test." We intended the classroom observation scale (1) to assist clinical diagnosis of milder cases of ASD in lieu of clinicians making preschool visits; (2) to identify children early on (e.g. first year in preschool) who are more likely to have ASD than their peers, so that these children can be kept under closer watch (i.e. surveillance); and (3) to help preschool teachers make better informed and more confident decisions about whether to discuss with parents of children whom they suspect perhaps to have ASD.

## Method

### Participants

Ethical approval for this research was granted by the Human Research Ethics Committee of the authors' university. Written parental consent was obtained prior to data collection. There were two phases to this study. The Classroom Observation Scale (COS) was developed in phase 1, which involved 304 children (age 3;0 to 4;11, $M = 3;11$, $SD = 6$ months, 162 boys and 142 girls) recruited from four ethnically diverse English-speaking international preschools serving families from middle to middle-upper socioeconomic backgrounds in Hong Kong. Parents of 185 children (98 boys and 87 girls) of the total sample of 304 gave further consent for their child to participate in an ASD assessment based on the ADOS-2 (Lord et al., 2012).

We validated the COS in phase 2 of this study. There was no overlap in the participants between phases 1 and 2. We received parental consent for 322 children (age 2;10 to 4;5, $M = 3;4$, $SD = 4$ months, with 161 boys and 161 girls) from five English-speaking international preschools—(1) to be observed in school and (2) to participate, if selected, in ADOS-2 assessment 1.5 years later. Their parents and preschool teachers ($n = 30$) also participated by providing information about the children, having first given written consent as well.

### Procedure

*Phase 1: development of the 13-item COS.* An item pool was generated from several sources: (1) an extensive

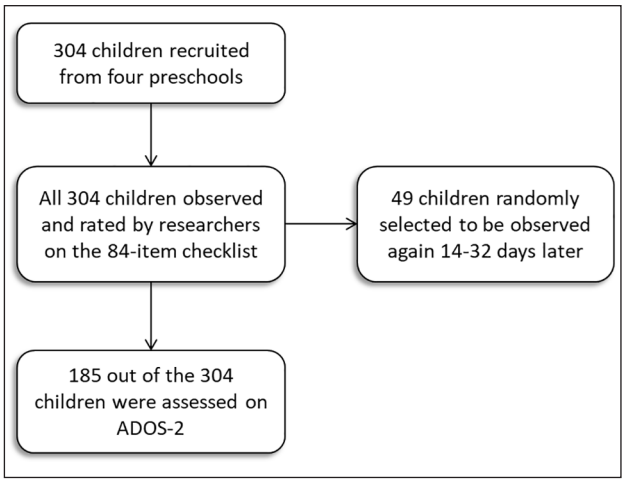


**Figure 1.** A flowchart indicating the number of children at each stage of data collection in phase 1 of the study.

review of research on peer interaction, (2) existing instruments for screening and assessing children with ASD, (3) a review of books with retrospective accounts from parents of children with ASD, and (4) interviews with experienced diagnosticians for ASD. The draft checklist consisted of over 100 items. Two local clinicians who specialized in ASD and had considerable experience of school observation suggested item revision in light of the preschool classroom context. The revised draft checklist was then sent to two clinical psychologists specialized in ASD in the United States and a developmental psychologist in Hong Kong for expert review. A preliminary checklist thus created comprised a set of 84 symptomatic/healthy behaviors and a 3-point rating system (1 = occurring rarely/most of the time, 2 = occurring less/more frequently than average, 3 = occurring at a similar rate as average peer).

An experienced clinical psychologist trained six research assistants (who had taken university-level psychology courses) to use the 84-item rating system in a special education classroom for high-functioning preschoolers with ASD. Both the clinical psychologist and the research assistants (one or two assistants at a time) observed the children simultaneously on site, and each research assistant's ratings were compared item-by-item against the psychologist's ratings at the end of the observation. After about 9 h of training, each rater achieved an item-by-item agreement greater than 75% with the psychologist.

The six raters then observed the children in the four preschools—1 school day per child, and four to five children per school day. To establish interrater reliability, a portion of the cases ($n = 96$) were seen by two raters at the same time, with different combinations among the six raters. Each observation interval focused on a target child and lasted a minute, and each child on average was the focus of around 30 1-min observations. The order in which children were observed was randomized using computer-generated random number strings. To facilitate observation, the 84 items were grouped by school routines: structured learning time, social time (e.g. free play), and transition time (e.g. clean-up). All 304 children were observed in all these contexts. During structured teaching times, observers sat or stood at the side or back of the classrooms where they could clearly see the children's behaviors (e.g. fiddling objects, doing repetitive behaviors, talking to peers, looking at teachers). During free play, observers stayed near the target children so they could hear the children's conversations and observe their interactions with peers without interrupting. The observers took notes and later gave each target child a score on each of the 84 items at the end of each day. Forty-nine children were randomly selected to be observed again 14–32 days later by the same raters who had carried out the original observation to assess test–retest reliability of the instrument.

Data-driven item reduction of our 84-item preliminary list yielded a much shorter 13-item COS. Then, as a first validity check, we examined whether COS scores were related to ASD symptomatology, as indicated by scores based on a widely used assessment tool, namely the ADOS-2 (Lord et al., 2012). ADOS-2 was administered by a clinical psychologist formally trained and qualified to use it for both research and clinical purposes and kept blind to the children's COS scores. Of the 304 children observed in phase 1, 185 of them—whose parents granted further consent—underwent the ADOS-2 assessment (Figure 1).

*Phase 2: validation of the COS.* We further evaluated how well observers with little or no clinical training (i.e. research assistants and preschool teachers) could use the 13-item COS to identify preschoolers under age 4.5 years more likely than their peers to have ASD. Parents were invited to participate about 2 months after their children had started preschool. Interested parents returned a signed consent form to the school.

The same clinical psychologist from phase 1 trained two new research assistants (with university-level psychology coursework but no prior clinical training) to use the 13-item COS, reaching good interrater reliability after about 6 h of training using the same method and criteria. The two research assistants then observed each child participant on 2 school days no more than 19 days apart ($M = 4.7$ days; $SD = 3.8$ days), with four to seven children per school day in random order for each round of 1-min observations. Each target child was observed for about 30 1-min intervals in total. The observers took notes and later gave each target child a COS score on each of the 13 items at the end of each day. All 322 children recruited in phase 2 were observed and rated by both research assistants on COS.

A teacher in each classroom was asked to use COS and Social Responsiveness Scale–2 (SRS-2 teacher-report;

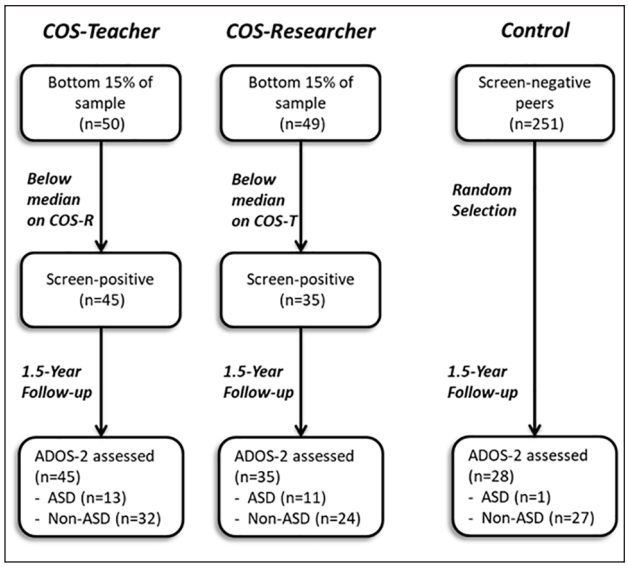


**Figure 2.** The two screening approaches to identifying ASD in phase 2 of the study ($N = 322$). The two approaches together identified 54 of the 322 children as more likely to have ASD, noting considerable overlap of screen-positives between the two approaches. The screen-negative peers ($n = 251$) did not meet the bottom 15% cutoff on either COS-Teacher or COS-Researcher. Teachers completed both the COS and SRS near preschool onset, and ADOS-2 assessments were conducted at the 1.5-year follow-up.

Constantino, 2012) to rate the children. The SRS-2 was used as a measure of convergent validity for the COS. Altogether, 30 teachers from the five preschools did so; they were all briefed beforehand on the scoring of the checklist items for about 30–45 min by a clinical psychologist on our research team.

Children of interest were identified based on the COS teacher-report (COS-Teacher) and researcher-report (COS-Researcher). Between the 15th and 85th percentile (i.e. within about one standard deviation of the mean) is typically considered within the normal range for clinical measures (e.g. IQ scores; Sattler, 2008), so we used the bottom 15% as a cutoff for COS-Teacher and COS-Researcher. This cutoff seemed like a reasonable first approximation for bootstrapping our way to find an evidence-based cutoff for the COS. We adopted two approaches to identify young children more likely than their peers to have ASD near preschool onset (Figure 2):

1. *Bottom 15% on COS-Teacher and below the median on COS-Researcher* ($n = 45$)—the second criterion helped reduce false positives (e.g. in case COS-Teacher data happened to be collected on atypical days for a child);
2. *Bottom 15% on COS-Researcher and below the median on COS-Teacher* ($n = 35$).

We did not give ASD assessment to all 322 children in this community sample for two obvious reasons: (1) ethical concerns of clinically assessing a large number of children without clinical referrals (further discussed in a later section) and (2) financial costs. Instead, we used these two approaches and identified 54 of 322 children as more likely to have ASD, noting considerable overlap of screen-positives between the two approaches.

In the second semester of the children's second preschool year—generally about 1.5 years after the COS data collection—these 54 screen-positive children were mixed with 28 randomly selected screen-negative peers (i.e. typically-developing control) for ASD assessment using ADOS-2. Hence, a total of 82 children were assessed on ADOS-2. The clinical assessments were done by the same clinical psychologist as in phase 1, who was trained and qualified for using ADOS-2 for research and clinical purposes and did not know the children's COS screen-positive versus control status or their scores on other measures.

## Instruments

*COS*: The 13 items selected in phase 1 were used, but the 3-point scale was expanded to a 5-point scale for finer-grained ratings (1 = very rarely or never; 2/3/4/5 = less often than/about as often as/more often than/much more often than most students, respectively; Appendix 1).

*ADOS-2*: This is a semi-structured, standardized tool for autistic disorder and ASD (Lord et al., 2012; Oosterling et al., 2010). It provides opportunities for children to engage in communication, social interaction, and play (or imaginative use of materials). The 185 children in phase 1 and 82 children in phase 2 were all assessed on module 2 of ADOS-2, as they all spoke in multiword utterances. The COS scores would be checked against ADOS-2 raw scores, which measure ASD symptomatology.

*SRS-2* (*Preschool Form*): This 65-item teacher-rating scale assesses a child's social awareness, social information processing, capacity for reciprocal social communication, social anxiety/avoidance, and characteristic autistic preoccupation/traits (Constantino, 2012). Cronbach's alpha for this sample was 0.96.

## Statistical analysis

Data analyses were performed using SPSS-25. Raw scores were used for all analyses unless specified otherwise. To yield the 13-item COS from the list of 84 items used in phase 1, variance in item scores and the collinearity among items were checked. The latter was examined by computing Spearman's correlation coefficients between items. Psychometric properties of the COS were calculated based on data collected in both phase 1 and phase 2. Cronbach's alphas were reported for internal reliability. Intraclass correlation coefficients (ICCs) were reported for interrater and test–retest reliabilities. Cross-informant agreement between teachers' and researchers' ratings on COS was assessed by calculating the Pearson correlation coefficient between the

two measures. Convergent validity was assessed based on the correlations of the COS with ADOS-2 in phase 1 and with SRS-2 in phase 2 of this study.

In the validation phase, ADOS-2 was conducted around 1.5 years after the classroom observation. One-way analysis of variance (ANOVA) compared the mean scores on COS-Researcher and COS-Teacher between the non-ASD and ASD groups classified based on the ADOS-2 assessment. To examine the predictive validity of the two screening approaches (Figure 2) in detecting ASD versus non-ASD, Pearson chi-square tests were conducted for each approach to test for significant relations between the categorization based on screening and subsequent diagnoses of ASD. Cramer's *V* was reported on the strength of association between the two nominal variables (i.e. classification status and ASD diagnosis). Furthermore, decision statistics including sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR−), and odds ratio (OR) were calculated for each screening approach. Likelihood ratios were used as they are independent of prevalence (Spitalnic, 2004; Wales, 2003). Note that the formula for LR+ = sensitivity / (1 − specificity), and LR− = (1 − sensitivity) / specificity. The OR is the ratio between LR+ and LR−, and can be calculated by the formula OR = LR+ / LR− = (sensitivity × specificity) / ((1 − sensitivity) × (1 − specificity)). We used OR to quantify the strength of association between screen-positive status and ASD diagnosis. Test of significance for OR with a *p* value less than 0.05 indicates that the OR value is significantly greater than 1, and the null hypothesis (i.e. no association between screen-positive status and ASD diagnosis) can thus be rejected (Szumilas, 2010). In general, a higher OR represents better accuracy in prediction.

The receiver operating characteristic (ROC) analyses were done separately for the COS-Teacher and COS-Researcher to further evaluate their predictive validity. Here a larger area under the ROC curve (AUC) would suggest a higher screening accuracy for COS-Teacher and COS-Researcher in classifying ASD cases versus non-cases (Fawcett, 2006), with an area of 1 representing perfect classification and an area of 0.5 as random results. Cutoff criterion at a fixed level of sensitivity (i.e. 0.8 and 0.9), and the corresponding specificity and OR were also reported. These analyses speak to whether a specific cutoff criterion on either COS-Teacher or COS-Researcher might be informative in clinical practice—specifically, how well it could classify "cases" versus "non-cases" (Grund & Sabin, 2010). ROC analyses help describe the sensitivity and specificity of a cutoff criterion.

## Results

### Item reduction

Drawn from existing tools, many items had been developed using clinical samples, resulting in items with low variance in scores in community samples (<5% of children getting "1 = occurring rarely/most of the time"). Consequently, 62 of the 84 items were removed due to low variance in the community sample. Spearman's correlation coefficients were computed for the remaining 22 items, and 9 of them were further excluded due to high collinearity with other items (Spearman's $\rho > 0.8$), indicating substantial overlap between them. Of the 13 items retained for the COS (Appendix 1), 10 focused on challenges in peer interaction (e.g. "Directs facial expressions to peers"), 2 on restricted and repetitive behaviors (e.g. "Engages in repetitive behaviors or unusual mannerisms"), and 1 on self-regulation challenge (e.g. "Sits down or stays seated during structured teaching times"). Spearman correlations among the 13 items for the COS are presented in Appendix 2. The 10 items on peer interaction (COS items 1–10) were significantly correlated with each other with Spearman's ρs ranging from 0.23 to 0.76 ($p$s < 0.001), but less strongly correlated with the 2 items on restricted and repetitive behaviors (items 12 and 13; Spearman's ρs ranging from 0.15 to 0.34, $p$s < 0.01), and even less so with the item on self-regulation (item 11; Spearman's ρs ranging from 0.07 to 0.24).

### Internal reliability

Cronbach's alpha for COS for the phase 1 sample was 0.91, and internal reliabilities for COS-Researcher and COS-Teacher in phase 2 were 0.88 and 0.89, respectively.

### Interrater reliability

ICC estimates were calculated based on a mean-rating ($k = 2$), absolute-agreement, two-way random-effects model (Koo & Li, 2016). The ICC estimates for the 13-item COS between researchers were 0.94 in phase 1 and 0.98 in phase 2. High interrater reliability between the independent observers suggests that the target behaviors are easy to capture by observation.

Cross-informant agreement between teachers' and researchers' ratings was assessed by calculating the Pearson correlation coefficient between COS-Teacher and COS-Researcher (Gresham et al., 2010). Results indicated significant correlation between the teachers' and researchers' ratings on COS ($r = 0.55$, $p < 0.001$), suggesting reasonable agreement between different kinds of observers with little or no clinical training (i.e. research assistants vs preschool teachers).

### Test–retest reliability

Test–retest reliability of COS in phase 1 was calculated based on a random selection of 49 children from the total sample of 304 children, observed again 14–32 days later. The ICC estimate between the observations was 0.73, based on a mean-rating ($k = 2$), absolute-agreement, two-way mixed-effects model (Koo & Li, 2016). All 322 children recruited in phase 2 were observed on 2 school days

**Table 1.** Mean scores (standard deviations) on the COS-Teacher and COS-Researcher for the non-ASD and ASD groups among the 82 clinically assessed children in phase 2 of the study.

| Measures | Non-ASD ($n = 68$) | ASD ($n = 14$) | $F$ value | $p$ value | Cohen's $d$ |
|---|---|---|---|---|---|
| COS-Teacher | 32.6 (6.5) | 27.1 (6.4) | 7.73 | 0.007 | 0.84 |
| COS-Researcher | 35.7 (5.8) | 29.3 (4.9) | 15.31 | <0.001 | 1.21 |

COS: Classroom Observation Scale; ASD: autism spectrum disorder.

(on average 4.7 days apart), and the ICC estimate of COS-Researcher between the 2 days was 0.73.

## Content validity

This refers to the extent to which an instrument measures the targeted construct (Anastasia, 1988). Content validity of the COS was high because the final 13 items were distilled from the preliminary 84 items drawn from prior research and modified by input from experienced clinicians and autism experts. The COS spans stereotypical behaviors during structured learning times and less structured social play times, making it appropriate for the preschool setting and population.

## Convergent validity

Convergent validity refers to the extent to which measures of theoretically related constructs are correlated. We assessed convergent validity based on the correlations of the COS with ADOS-2 in phase 1 and with SRS-2 in phase 2 of this study.

*Correlation with ADOS-2 in phase 1.* The Pearson correlation between the COS total scores and ADOS-2 raw scores showed significant association between the two measures ($r = -0.37$, $p < 0.001$). Sequential multiple regression revealed that the COS total scores significantly predicted ADOS-2 raw scores in the second step ($\beta = -0.37$, $t = -5.60$, $p < 0.001$), after controlling for gender and age in the first step of the multiple regression. The COS total scores explained a significant proportion of variance in ADOS-2 scores: $\Delta R^2 = 0.13$; $F(1, 206) = 31.3$, $p < 0.001$. In short, observers who had little clinical training could use the 13-item COS to identify 3- to 4-year-olds with ASD symptomatology.

*Correlation with SRS-2 in phase 2.* Correlations between the 13-item COS and 65-item SRS-2 were significant (COS-Teacher: Pearson $r(157) = -0.67$, $p < 0.001$; COS-Researcher: Pearson $r(160) = -0.50$, $p < 0.001$).

## Predictive validity

In phase 2, 82 preschoolers (age 4;3 to 5;7, $M = 4;8$, $SD = 4$ months, with 55 boys and 27 girls) received ASD assessments on average around 1.5 years after the classroom observation (the lag ranging from 13 to 18 months, $M = 16.8$ months, $SD = 1$ month), and the clinical assessment was primarily based on the ADOS-2 classification scheme. Among the 82 children assessed, 14 (10 boys and 4 girls; age 4;4 to 5;5, $M = 4;9$, $SD = 4.3$ months) met diagnostic criteria for ASD (with ADOS-2 Comparison Scores at or above 3), while 68 children did not meet diagnostic criteria.

*Contrasting non-ASD and ASD on COS.* Table 1 shows the mean scores and standard deviations on the screening scales (COS-Teacher and COS-Researcher) for the clinically assessed children classified into non-ASD and ASD. Univariate ANOVAs revealed group differences on COS-Teacher scores ($F(1, 67) = 7.73$, $p = 0.007$, Cohen's $d = 0.84$) and COS-Researcher scores ($F(1,80) = 15.31$, $p < 0.001$, Cohen's $d = 1.21$) obtained about 1.5 years prior to the ASD assessment. Specifically, the non-ASD group had significantly higher scores (i.e. better peer interaction) than the ASD group on both COS-Teacher and COS-Researcher.

*Predicting ASD near the end of year 2 in preschool.* Pearson's chi-square tests (Table 2) showed that the categorization based on the two screening approaches (Figure 2) significantly predicted the classification of ASD versus non-ASD cases (COS-Teacher: $\chi^2 = 9.83$, $p = 0.002$; COS-Researcher: $\chi^2 = 8.89$, $p = 0.003$). Using the bottom 15% on the COS-Teacher as the cutoff produced a higher OR (OR = 14.63, 95% confidence interval (CI) = (1.81–118.12), $p = 0.01$) than using the bottom 15% on the COS-Researcher (OR = 6.72, 95% CI = (1.71–26.46), $p = 0.01$). Note that the null effect value should be 1. Importantly, the lower limits of both CIs here were larger than 1, thereby indicating that both COS-Teacher and COS-Researcher can significantly predict ASD diagnosis.

ROC analyses were further conducted separately on the COS-Teacher and COS-Researcher to examine the predictive validity of these scales in discriminating ASD cases in our sample (Table 3; Figure 3). The AUC represents a single-value index of discriminative ability across the full range of cutoffs. Note that an AUC within the range of 0.7–0.9 denotes moderate accuracy, while an AUC above 0.9 indicates high test accuracy.

Both COS-Teacher and COS-Researcher showed moderate accuracy in differentiating ASD from non-ASD cases with AUCs of 0.76 and 0.80, respectively ($ps \leqslant 0.001$). Cutoff scores and specificities were estimated at the sensitivity levels of 0.8 and 0.9, and reported in Table 3 along with LR+, LR−, and OR. Relatively large OR values were

**Table 2.** Predictive validity of the two screening approaches in identifying preschoolers with ASD in phase 2 of the study.

| | Sensitivity (95% CI) | Specificity (95% CI) | LR+ (95% CI) | LR− (95% CI) | OR (95% CI) | z statistic for OR (*p* value) | $\chi^2$ (*p* value) | Cramer's *V* (*p* value) |
|---|---|---|---|---|---|---|---|---|
| Approach 1[a] | 0.93 (0.66–1.00) | 0.53 (0.40–0.65) | 1.97 (1.48–2.64) | 0.14 (0.02–0.90) | 14.63 (1.81–118.12) | 2.52 (0.01) | 9.83 (0.002) | 0.35 (0.002) |
| Approach 2[b] | 0.79 (0.49–0.95) | 0.65 (0.52–0.76) | 2.23 (1.46–3.40) | 0.33 (0.12–0.92) | 6.72 (1.71–26.46) | 2.73 (0.01) | 8.89 (0.003) | 0.33 (0.003) |

COS: Classroom Observation Scale; ASD: autism spectrum disorder; CI: confidence interval; OR: odds ratio; LR+: positive likelihood ratio; LR−: negative likelihood ratio.
[a]Below the 15th percentile on the COS-Teacher and below the median on the COS-Researcher. [b]Below the 15th percentile on the COS-Researcher and below the median on the COS-Teacher.

observed at the 0.9 sensitivity level for either the COS-Teacher (OR = 10.49) or COS-Researcher (OR = 10.87). Indeed, both the COS-Teacher and COS-Researcher helped identify children more likely than their peers to have ASD by predicting ASD diagnosis later well above the chance level.

These results indicated that (1) COS proved useful for identifying preschool children under age 4.5 years more likely than their peers to have ASD diagnosable about 1.5 years down the road and (2) COS proved to be useful across different types of potential users with little or no clinical training.

## Discussion

Our new screening tool for identifying—during the first semester of preschool—children more likely than their peers to have ASD is based on a very simple idea. While severe cases of ASD may be noticed by parents and preschool teachers and readily diagnosed by clinicians early on, milder cases often go undiagnosed until the first or second grade. We use peer interaction without adults hovering around as a naturally occurring "stress test" to identify children who have difficulty navigating the social world they share with their peers—difficulty that may foretell long-term social impairments.

This new screening tool works well for young children: all the children in the validation phase ($n = 322$) were under age 4.5 years, with a mean age of 3;4. Unlike most existing screening tools that relied on clinical samples for validation, the COS was developed and validated using a community sample, thereby boosting its applicability in community settings. Note that ASD cases in mainstream preschools tend to be less severe, which are more difficult to diagnose in young children. Our results provided evidence to support that the COS can help spot young preschoolers more likely to have ASD, so teachers and parents can keep a close watch for clearer symptomatology before seeking clinical assessment.

The COS developed in our study was easy to use for observers with little or no clinical training. The teachers in preschools were able to use the COS with reliable and valid results to help identify preschoolers under age 4.5 years more likely than their peers to have ASD, after receiving a 30- to 45-min group briefing at their preschools by a member of our research team. The eight research assistants who acted as observers in this study had taken university-level psychology courses and had received only a few hours of training from a clinical psychologist. Yet, they could use the COS to help identify children more likely than their peers to have ASD without knowing the children beforehand. Moreover, the COS has good psychometric properties in terms of reliability (internal consistency, interrater reliability, and test–retest reliability) and validity (convergent and crucially—for

**Table 3.** Receiver operating characteristic (ROC) analyses with areas under the curve (AUC), validity indexes, and cutoff scores for the COS-Teacher and COS-Researcher in predicting ASD cases in phase 2 of the study.

| | AUC (95% CI) | *z* (*p* value) | Estimated at fixed sensitivity | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Sensitivity | Specificity | LR+ | LR− | OR | Cutoff criterion |
| COS-Teacher | 0.76 (0.64–0.85) | 3.49 (<0.001) | 0.80 | 0.67 | 2.39 | 0.30 | 7.96 | ⩽28.7 |
| | | | 0.90 | 0.54 | 1.95 | 0.19 | 10.49 | ⩽31.2 |
| COS-Researcher | 0.80 (0.70–0.88) | 4.78 (<0.001) | 0.80 | 0.56 | 1.81 | 0.36 | 5.04 | ⩽34.6 |
| | | | 0.90 | 0.55 | 1.99 | 0.18 | 10.87 | ⩽34.9 |

COS: Classroom Observation Scale; ASD: autism spectrum disorder; CI: confidence interval; OR: odds ratio; LR+: positive likelihood ratio; LR−: negative likelihood ratio.

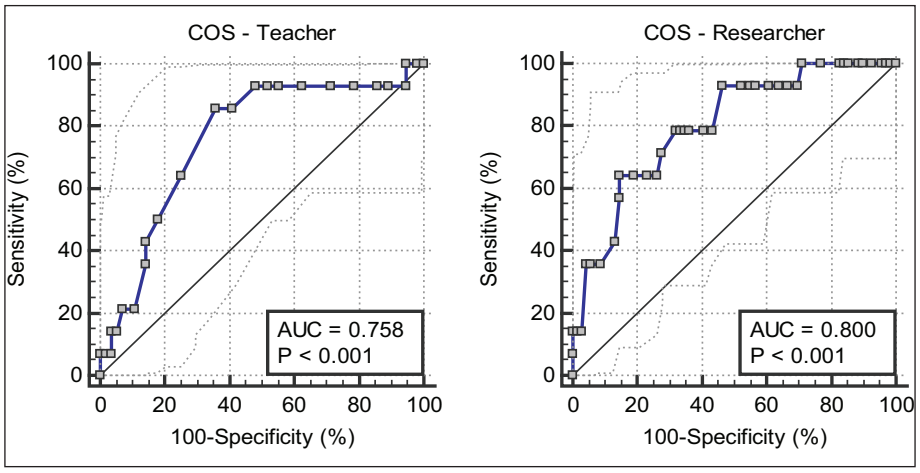


**Figure 3.** Receiver operating characteristic (ROC) curves for COS-Teacher (left) and COS-Researcher (right) in predicting the diagnosis of ASD based on ADOS-2. Screening accuracy was measured by the area under the ROC curve (AUC).

screening purposes—predictive validity for meeting ASD diagnosis prospectively) based on data collected from both types of informants, making it a potentially robust screening tool.

The results provided support to the ecological validity of the COS for use by preschool teachers, as well as assistant therapists who major in psychology at the undergraduate level but have not received extensive clinical training on ASD. Importantly, both types of data collection methods are plausible in real-life preschool settings. In cases for which ASD is suspected, preschool teachers can rate the child on his or her peer interaction based on the COS, while at the same time, an independent observer who may not be familiar with the child can rate the child's social behaviors on the COS prior to a formal clinical assessment by a diagnostician. As seen from our results, although the cross-informant agreement was statistically significant, the medium-level correlation between the two kinds of observers suggested that the teachers and independent observers might likely pick up different aspects of children's behaviors in their observations. Moreover, the observation by independent observers for each child was conducted during just 2 school days. This method is thus affordable in terms of manpower, time, and cost.

Compared to the 12-item DSM-ASD Scale from the CBCL/1½–5 and C-TRF which consists of 7 items on SCI and 5 items on RRB (Rescorla, Ghassabian, et al., 2019), the 13-item COS derived from data-driven item reduction consists of only 2 items on RRB, while the majority of the items are on social interaction. In their international comparisons of the DSM-ASD Scale scores on the C-TRF, Rescorla, Given, et al. (2019) noted greater societal differences for the RRB than the SCI subscale, revealing less consistency in teachers' ratings on RRB behaviors across societies. Rescorla, Given, et al. (2019) further speculated that the societal differences might be due to the varying level of sensitivity of teachers to RRB problems, and the degree to which group settings allow preschoolers to engage in these behaviors. In this study, 8 items on RRB were originally included in the 84-item preliminary checklist. Nonetheless, except for the two items eventually retained in the COS, the rest of the RRB items were removed due to low variance in scores in the community sample. This may suggest that RRB behaviors may not be readily picked up by observers in preschool settings in Hong Kong.

Furthermore, it is noteworthy that there is minimal overlap of items between the COS and the DSM-ASD Scale of the C-TRF. In particular, the SCI items of the

C-TRF DSM-ASD Scale focus more on social responding behaviors, that is, whether the child responds to others' initiation of interaction (e.g. "Doesn't answer when people talk to him or her," "Seems unresponsive to affection," "Avoids looking other in the eye"). By contrast, the COS items refer more to the social initiation behaviors of preschoolers (e.g. "Initiates to point out things in the environment to other children or adults," "Initiates conversation with other children," "Shows empathy for the feelings of peers and tries to make them feel better," "Initiates the sharing of toys or food with other children"). Perhaps social initiation behaviors, in contrast to more fleeting social responding behaviors, might be more easily spotted by observers who are not familiar with the child (such as the research assistants in this study), and likewise by teachers who may not have a lot of time to interact with and observe the child. As such, the COS may prove versatile in its utility as a screening tool for ASD in preschool populations, along with other existing ASD screening instruments (e.g. CBCL/1½–5 and C-TRF).

### *Limitations and implications for future research*

We are mindful that the COS cutoff scores (i.e. bottom 15% of our full sample) can only be used by preschool teachers and clinicians and their assistants as references for the time being. Moreover, the wide CIs obtained for the ORs of the two screening approaches in predicting ASD diagnoses indicated uncertainty in the estimates, and thus, the results here should be interpreted with caution. To further validate the results and to establish normative cutoffs for this new screening tool, future studies will need to use larger random and representative norming samples.

To address the ethical concerns of giving clinical assessments (e.g. ADOS-2) to too many children without clinical referrals (and being pragmatic about financial constraints), we decided to give ASD assessments to all the screen-positive children ($n=54$) and only a random sample of the screen-negative children ($n=28$) in the ratio of about 2:1. Indeed, we had reservations about administering clinical assessments using standardized testing instruments to a relatively large number of children with no particular clinical concerns, in view of the possibility of item leakage and a breach of test security. While we did not clinically assess all 322 children and could not know the number of true positive and true negative cases for the full sample, we did our best with our data to estimate sensitivity and specificity based on the clinically assessed sub-sample and estimated the OR for the two screening approaches (Glas et al., 2003).

In our study, we found that at least 14 children in our sample of 322 (4.3%) were diagnosed of ASD. This rate was higher than the CDC (2018) estimate of 1.7% or published preschool estimates around 0.8% (Soke et al., 2017; Wang et al., 2011), perhaps due to (1) random sampling error of our modest sample size, (2) differential parental consent to the study (i.e. parents who were concerned about their child's development might have been more likely to give consent), and/or (3) under-estimation of ASD cases in the medical record–based surveillance system studies for large populations that contributed to the published prevalence estimates (for a similar point, see CDC, 2012).

The current research design, while pragmatic, could not tell how many of the screen-negative children who were not clinically assessed might turn out to meet diagnosis for ASD. Fortunately, our screening worked quite well: (1) the screen-positive children were much more likely to meet ASD diagnostic criteria subsequently than would be expected by chance and (2) by contrast, only one child in the control group (i.e. randomly selected screen-negative children) turned out to meet ASD diagnostic criteria. To be more confident that the screen-negative children in general are truly without ASD, future studies should consider involving larger random samples of screen-negative children in the clinical assessments.

We waited about 1.5 years in the validation phase to do clinical assessment for ASD because we expected that ASD cases found in community settings such as mainstream preschools would tend to be less severe and hence be more difficult to diagnose in the first year of preschool. Given the positive findings reported here, future evaluation of the COS as a preschool screening tool can consider screening near preschool onset and waiting 1 year or even less to clinically assess the screen-positive children.

More research is also needed to find out how the COS can be more easily and more effectively used. For example, a good user manual with helpful answers for frequently asked questions may suffice to replace in-person briefings for preschool teachers; evidence-based guidelines can inform teachers on how to observe the screen-positive children more closely and effectively, perhaps by using the COS more than once to track the children for a few months to help inform whether clinical assessment is called for. Such continuing surveillance can help reduce false positives based on only one-off screening when the children are quite young (e.g. near the onset of preschool). Also, it may be helpful to use the COS again on children initially screen-negative should concerns become evident. Further psychometric evaluation of the COS with teachers as the raters (e.g. test–retest reliability, interrater reliability, predictive validity for ASD diagnosis) will also provide valuable information on whether teachers as raters without being supplemented by researchers' ratings will suffice.

## Conclusion

This study aimed to develop a convenient COS that can help preschool teachers and observers with little or no formal clinical training to identify in mainstream preschools, with reliable and valid results, children under age 4.5 years more likely than their peers to have ASD. We are mindful that there are good alternate approaches for developing simple ASD screening tools for preschoolers under age

4 years. For example M-CHAT-R/F and the RITA-T (Choueiri & Wagner, 2015) could expand the target age range upward from toddlerhood to early childhood. Nonetheless, the present study constitutes a first step in developing an easy-to-use, reliable, and valid tool to help teachers and healthcare workers capitalize on peer interaction as a naturally occurring stress test to identify, during the first semester of preschool, children more likely than their peers to have ASD. With this COS joining forces with existing screening tools (e.g. CBCL/1½–5 and C-TRF), young children in community settings such as mainstream preschools should stand a better chance of early identification and getting effective intervention for ASD, launching them on a better lifelong trajectory in understanding the social world and developing healthy social bonds.

## Acknowledgements

We are grateful to the children, parents, preschools, and Autism Partnership Hong Kong for their support, and to a team of dedicated research assistants for data collection and coding.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Research Grants Council of Hong Kong (HKU 740213) and the Karen Lo Eugene Chuang Professorship in Diversity and Equity.

## ORCID iDs

Kathy Kar-man Shum https://orcid.org/0000-0003-4340-3160

Hannah Man-yan Tse https://orcid.org/0000-0002-5477-5691

## References

Achenbach, T. M. (2014). *DSM-oriented guide for the Achenbach System of Empirically Based Assessment (ASEBA)*. Research Center for Children, Youth, & Families, University of Vermont.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles* (Vol. 30). Research Center for Children, Youth, & Families, University of Vermont.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.

Anastasia, A. (1988). *Psychological testing* (6th ed.). Macmillan Publishing.

Bauminger-Zviely, N., Karin, E., Kimhi, Y., & Agam-Ben-Artzi, G. (2014). Spontaneous peer conversation in preschoolers with high-functioning autism spectrum disorder versus typical development. *Journal of Child Psychology and Psychiatry*, *55*, 363–373.

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

Billstedt, E., Gillberg, C., & Gillberg, C. (2005). Autism after adolescence: Population-based 13- to 22-year follow-up study of 120 individuals with autism diagnosed in childhood. *Journal of Autism and Developmental Disorders*, *35*, 351–360.

Cederlund, M., Hagberg, B., Billstedt, E., Gillberg, I. C., & Gillberg, C. (2007). Asperger Syndrome and Autism: A comparative longitudinal follow-up study more than 5 years after original diagnosis. *Journal of Autism and Developmental Disorders*, *38*, 72–85.

Centers for Disease Control and Prevention. (2012). Prevalence of autism spectrum disorders—autism and developmental disabilities monitoring network, 14 sites, United States, 2008. *Morbidity and Mortality Weekly Report*, *61*, 1–19.

Centers for Disease Control and Prevention. (2018). Press release. https://www.cdc.gov/ncbddd/autism/data.html

Chang, Y. C., Shire, S. Y., Shih, W., Gelfand, C., & Kasari, C. (2016). Preschool deployment of evidence-based social communication intervention: JASPER in the classroom. *Journal of Autism and Developmental Disorders*, *46*, 2211–2223.

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

Constantino, J. N. (2012). *Social responsiveness scale, second edition (SRS-2)*. Western Psychological Services.

Corbett, B. A., Newsom, C., Key, A. P., Qualls, L. R., & Edmiston, E. K. (2014). Examining the relationship between face processing and social interaction behavior in children with and without autism spectrum disorder. *Journal of Neurodevelopmental Disorders*, *6*, 35.

Corbett, B. A., Schupp, C. W., Simon, D., Ryan, N., & Mendoza, S. (2010). Elevated cortisol during play is associated with age and social engagement in children with autism. *Molecular Autism*, *1*, 13.

De Giacomo, A., & Fombonne, E. (1998). Parental recognition of developmental abnormalities in autism. *European Child & Adolescent Psychiatry*, *7*(3), 131–136.

Department of Health, Hong Kong. (2007). CAS Epidemiological data on Autistic Spectrum Disorders from 2003 to 2005. *Child Assessment Service Epidemiology and Research Bulletin*, *3*, 1–24.

Duvekot, J., van der Ende, J., Verhulst, F. C., & Greaves-Lord, K. (2015). The screening accuracy of the parent and teacher-reported Social Responsiveness Scale (SRS): Comparison with the 3Di and ADOS. *Journal of Autism and Developmental Disorders*, *45*, 1658–1672.

Ehlers, S., Gillberg, C., & Wing, L. (1999). A screening questionnaire for Asperger syndrome and other high-functioning autism spectrum disorders in school age children. *Journal of Autism and Developmental Disorders*, *29*, 129–141.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Fein, D., Barton, M., Eigsti, I. M., Kelley, E., Naigles, L., Schultz, R. T., . . . Troyb, E. (2013). Optimal outcome in individuals with a history of autism. *Journal of Child Psychology and Psychiatry*, *54*, 195–205.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*, *56*, 1129–1135.

Gray, K. M., & Tonge, B. J. (2005). Screening for autism in infants and preschool children with developmental delay. *Australian and New Zealand Journal of Psychiatry*, *39*, 378–386.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System—Rating Scales. *Psychological Assessment*, *22*(1), 157–166.

Grund, B., & Sabin, C. (2010). Analysis of biomarker data: Logs, odds ratios, and receiver operating characteristic curves. *Current Opinion in HIV and AIDS*, *5*(6), 473–479.

Kasari, C., Paparella, T., Freeman, S., & Jahromi, L. B. (2008). Language outcome in autism: Randomized comparison of joint attention and play interventions. *Journal of Consulting and Clinical Psychology*, *76*, 125–137.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Levy, S. E., Rescorla, L. A., Chittams, J. L., Kral, T. J., Moody, E. J., Pandey, J., . . . Rosenberg, C. R. (2019). ASD screening with the Child Behavior Checklist/1.5-5 in the study to explore early development. *Journal of Autism and Developmental Disorders*, *49*(6), 2348–2357.

Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of General Psychiatry*, *63*, 694–701.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule second edition (ADOS-2) manual (Part 1): Modules 1–4*. Western Psychological Services.

McConachie, H., Couteur, A. L., & Honey, E. (2005). Can a diagnosis of Asperger syndrome be made in very young children with suspected autism spectrum disorder? *Journal of Autism and Developmental Disorders*, *35*, 167–176.

Moore, V., & Goodson, S. (2003). How well does early diagnosis of autism stand the test of time? Follow-up study of children assessed for autism at age 2 and development of an early diagnostic service. *Autism*, *7*, 47–63.

Myles, B. S., Bock, S. J., & Simpson, R. L. (2001). *Asperger syndrome diagnostic scale*. Western Psychological Service.

Oosterling, I., Roos, S., Bildt, A., Rommelse, N., Jonge, M., Visser, J., Lappenschaar, M., . . . Buitelaar, J. (2010). Improved diagnostic validity of the ADOS Revised algorithms: A replication study in an independent sample. *Journal of Autism and Developmental Disorders*, *40*, 689–703.

Orinstein, A. J., Suh, J., Porter, K., De Yoe, K. A., Tyson, K. E., Troyb, E., . . . Fein, D. A. (2015). Social function and communication in optimal outcome children and adolescents with an autism history on structured test measures. *Journal of Autism and Developmental Disorders*, *45*, 2443–2463.

Orinstein, A. J., Tyson, K. E., Suh, J., Troyb, E., Helt, M., Rosenthal, M., . . . Schultz, R. T. (2015). Psychiatric symptoms in youth with a history of autism and optimal outcome. *Journal of Autism and Developmental Disorders*, *45*, 3703–3714.

Reichow, B., Barton, E. E., Boyd, B. A., & Hume, K. (2012). Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders (ASD). *Cochrane Database of Systematic Reviews*, *10*, CD009260.

Rescorla, L., Kim, Y. A., & Oh, K. J. (2015). Screening for ASD with the Korean CBCL/1½–5. *Journal of Autism and Developmental Disorders*, *45*(12), 4039–4050.

Rescorla, L. A., Ghassabian, A., Ivanova, M. Y., Jaddoe, V. W., Verhulst, F. C., & Tiemeier, H. (2019). Structure, longitudinal invariance, and stability of the child behavior checklist 1½–5's diagnostic and statistical manual of mental disorders–autism spectrum disorder scale: Findings from generation R (Rotterdam). *Autism*, *23*(1), 223–235.

Rescorla, L. A., Given, C., Glynn, S., Ivanova, M. Y., & Achenbach, T. M. (2019). International comparisons of autism spectrum disorder behaviors in preschoolers rated by parents and caregivers/teachers. *Autism*, *23*(8), 2043–2054.

Robins, D. L., Casagrande, K., Barton, M., Chen, C. M. A., Dumont-Mathieu, T., & Fein, D. (2014). Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). *Pediatrics*, *133*, 37–45.

Robins, D. L., Fein, D., Barton, M. L., & Green, J. A. (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.

Roux, A. M., Shattuck, P. T., Cooper, B. P., Anderson, K. A., Wagner, M., & Narendorf, S. C. (2013). Postsecondary employment experiences among young adults with an autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*, 931–939.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations*. JM Sattler.

Schreibman, L., Dawson, G., Stahmer, A. C., Landa, R., Rogers, S. J., McGee, G. G., . . . McNerney, E. (2015). Naturalistic developmental behavioral interventions: Empirically validated treatments for autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *45*, 2411–2428.

Scott, F. J., Baron-Cohen, S., Bolton, P., & Brayne, C. (2002). The CAST (Childhood Asperger Syndrome Test): Preliminary development of a UK screen for mainstream primary-school-age children. *Autism*, *6*, 9–31.

Siegel, B. (2004). *Pervasive Developmental Disorder Screening Test–II (PDDST-II)*. Harcourt.

Siu, A. L., Bibbins-Domingo, K., Grossman, D. C., Baumann, L. C., Davidson, K. W., Ebell, M., . . . Krist, A. H. (2016). Screening for autism spectrum disorder in young children: US Preventive Services Task Force recommendation statement. *Journal of the American Medical Association*, *315*, 691–696.

Skuse, D. H., Mandy, W. P., & Scourfield, J. (2005). Measuring autistic traits: Heritability, reliability and validity of the Social and Communication Disorders Checklist. *The British Journal of Psychiatry*, *187*(6), 568–572.

Soke, G. N., Maenner, M. J., Christensen, D., Kurzius-Spencer, M., & Schieve, L. A. (2017). Brief report: Estimated prevalence of a community diagnosis of autism spectrum disorder by age 4 years in children from selected areas in the United States in 2010: Evaluation of birth cohort effects. *Journal of Autism and Developmental Disorders*, *47*, 1917–1922.

Spitalnic, S. (2004). Test properties 2: Likelihood ratios, Bayes' formula, and receiver operating characteristic curves. *Hospital Physician*, *40*(10), 53–58.

Szatmari, P., Bryson, S., Boyle, M., Streiner, D., & Duku, E. (2003). Predictors of outcome among high functioning children with autism and Asperger syndrome. *Journal of Child Psychology and Psychiatry*, *44*, 520–528.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *19*(3), 227–229.

Volkmar, F. R., Cicchetti, D. V., Dykens, E., Sparrow, S. S., Leckman, J. F., & Cohen, D. J. (1988). An evaluation of the autism behavior checklist. *Journal of Autism and Developmental Disorders*, *18*(1), 81–97.

Wales, N. S. (2003). Moving beyond sensitivity and specificity: Using likelihood ratios to help interpret diagnostic tests. *Australian Prescriber*, *26*(5), 111–113.

Wang, X., Yang, W. H., Jin, Y., Jing, J., Huang, X., Li, X. H., Wei, W., . . . Fan, Y.-B. (2011). Prevalence of Autism Spectrum Disorders in preschool children of Guangzhou kindergartens. *Chinese Mental Health Journal*, *25*, 401–408.

Warren, Z., McPheeters, M. L., Sathe, N., Foss-Feig, J. H., Glasser, A., & Veenstra-VanderWeele, J. (2011). A systematic review of early intensive intervention for autism spectrum disorders. *Pediatrics-English Edition*, *127*(5), e1303.

Westman Andersson, G., Miniscalco, C., Johansson, U., & Gillberg, C. (2013). Autism in toddlers: Can observation in preschool yield the same information as autism assessment in a specialised clinic? *The Scientific World Journal*, *2013*, 384745.

World Health Organization. (2004). *ICD-10: International statistical classification of diseases and related health problems: Tenth revision* (2nd ed.).

Zwaigenbaum, L., Bryson, S., Rogers, T., Roberts, W., Brian, J., & Szatmari, P. (2005). Behavioral manifestations of autism in the first year of life. *International Journal of Developmental Neuroscience*, *23*(2–3), 143–152.

## Appendix 1

Items of the Classroom Observation Scale (COS).

| No. | Item description |
|---|---|
| 1. | Spends social time alone. |
| 2. | Initiates to point out things in the environment to other children or adults. |
| 3. | Initiates conversation with other children (at least 4 turns). |
| 4. | Directs facial expressions to peers. |
| 5. | Shows empathy for the feelings of peers and tries to make them feel better (e.g., stops annoying behaviors, reports about others to the teacher, comforts peers), instead of showing no reaction or an inappropriate reaction. |
| 6. | Plays pretend with other children. |
| 7. | Shows things to other children (e.g., toys or actions). |
| 8. | Initiates the sharing of toys or food with other children. |
| 9. | Copies or imitates the behaviors (e.g. action, language, and facial expression) of other children appropriately and timely. |
| 10. | Pays attention to other children's conversation or speech. |
| 11. | Sits down or stays seated during structured teaching times. |
| 12. | Fiddles with objects (e.g., spins, scratches, touches, or fumbles with them). |
| 13. | Engages in repetitive behaviors or unusual mannerisms (e.g., flicking fingers, flapping hands, walking on toes, jumping, grimacing, squirming, staring sideways). |

Reverse scoring for items 1, 12, and 13.
Rating: 1 = *very rarely or never*; 2 = *less often than most students*; 3 = *about as often as most students*; 4 = *more often than most students*; 5 = *much more often than most students*.

## Appendix 2

Spearman correlations among the 13 items on COS.

| Items | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | – | 0.50*** | 0.69*** | 0.64*** | 0.26*** | 0.76*** | 0.71*** | 0.57*** | 0.56*** | 0.54*** | 0.15* | 0.24*** | 0.17** |
| 2 | | – | 0.59*** | 0.53*** | 0.28*** | 0.49*** | 0.59*** | 0.48*** | 0.49*** | 0.48*** | 0.12* | 0.16** | 0.26*** |
| 3 | | | – | 0.66*** | 0.23*** | 0.64*** | 0.74*** | 0.58*** | 0.56*** | 0.58*** | 0.19** | 0.20*** | 0.22*** |
| 4 | | | | – | 0.31*** | 0.62*** | 0.69*** | 0.52*** | 0.61*** | 0.61*** | 0.07 | 0.17** | 0.17** |
| 5 | | | | | – | 0.24*** | 0.26*** | 0.42*** | 0.25*** | 0.30*** | 0.24*** | 0.26*** | 0.34*** |
| 6 | | | | | | – | 0.71*** | 0.54*** | 0.61*** | 0.55*** | 0.13* | 0.17** | 0.22*** |
| 7 | | | | | | | – | 0.62*** | 0.62*** | 0.64*** | 0.09 | 0.15** | 0.22*** |
| 8 | | | | | | | | – | 0.49*** | 0.46*** | 0.15** | 0.27*** | 0.22*** |
| 9 | | | | | | | | | – | 0.60*** | 0.22*** | 0.24*** | 0.28*** |
| 10 | | | | | | | | | | – | 0.17** | 0.29*** | 0.29*** |
| 11 | | | | | | | | | | | – | 0.50*** | 0.48*** |
| 12 | | | | | | | | | | | | – | 0.46*** |
| 13 | | | | | | | | | | | | | – |

COS: Classroom Observation Scale.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.