Separate Scoring Algorithms Optimize the Screening Properties of the Screening Tool for Autism in Toddlers for Different Screening Priorities

Shana Attar

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2021

Committee:

Wendy Stone

Kevin King

Program Authorized to Offer Degree:

Department of Psychology

University of Washington

Abstract

Separate Scoring Algorithms Optimize the Screening Properties
of the Screening Tool for Autism in Toddlers for Different Screening Priorities

Shana Attar

Chair of the Supervisory Committee:

Wendy Stone

Department of Psychology

Detecting autism in young children allows for timely access to specialized early intervention services. The Screening Tool for Autism in Toddlers (STAT) is a validated stage-2 Autism Spectrum Disorders (ASD) screening measure that takes 20 minutes to administer and comprises 12 play-based items that are scored according to specific criteria. An expanded version (STAT-E) includes the examiner's subjective ratings of children's social engagement and atypical behaviors. This study examines the screening properties of the STAT-E using the original STAT scoring algorithm and the extent to which an algorithm that includes the subjective ratings of social engagement and atypical behaviors improves the screening properties of the STAT-E relative to the original STAT scoring algorithm. Two-hundred and thirty-eight (238) families of children between 24 and 35 months old participated. The STAT-E was administered by assessors with limited experience who were trained using a scalable web-based platform and children received a comprehensive evaluation from a separate team of ASD research or clinical experts who were blind to the STAT-E results. Logistic regression, ROC curves, and classification

matrices and metrics (Youden's J and F1 score) were used to determine the screening properties of the STAT-E using the original STAT scoring algorithm and the extent to which an algorithm that included the subjective ratings of social engagement and atypical behaviors improved the screening properties of the STAT-E relative to the original STAT scoring algorithm. The concurrent validity of the STAT-E using the original STAT scoring algorithm in this sample was fair (sensitivity = .67, specificity = .66). Inclusion of the examiner ratings of social engagement and atypical behaviors on the STAT-E improved positive risk classification appreciably (F1 score = .80-.85 versus .74), while the specificity declined (specificity = .62). Results suggest that the STAT-E using the original STAT scoring algorithm optimizes specificity, while the STAT-E scoring algorithm with two new ratings optimizes the positive risk classification. Using multiple scoring algorithms on the STAT may provide improved scoring accuracy for diverse contexts and children. A fast and scalable web-based tutorial may be a pathway for increasing the number of community providers who can administer the STAT and contribute toward increased rates of autism screening.

Introduction

Access to specialized early intervention (EI) for autism spectrum disorder (ASD) is critical for improving long-term outcomes for children with ASD (Zwaigenbaum et al., 2015) and depends on timely detection of the disorder. However, while ASD can be diagnosed reliably by 24 months (Chawarska et al., 2009; Lord et al., 2006; Turner et al., 2006; Zwaigenbaum et al., 2015), a formal ASD diagnosis is often not conferred until preschool or school-age (Oswald et al., 2017; Sheldrick et al., 2017), and the median diagnostic age is 51 months (Maenner et al., 2020). Yet in the United States, a formal ASD diagnosis is often required to access insurance-covered ASD services and/or state-funded ASD-specific EI (Eisenhower et al., 2021).

The American Academy of Pediatrics (AAP) guideline to screen all children for ASD at 18- and 24-month well-child visits (Maenner et al., 2020) has contributed to increased use of ASD screening overall and a focus on optimizing stage-1 ASD screening tools for use by providers in diverse settings through the use of digital screening tools (Brooks et al., 2016; Campbell et al., 2017; Harrington et al., 2013; Steinman et al., 2021; Sturner et al., 2016). As such, recent ASD screening estimates range from 51%-93% for 18-month well-child visits (Mazurek et al., 2021; Monteiro et al., 2019) and 41%-82% for 24-month well-child visits (Mazurek et al., 2021; Monteiro et al., 2019). Stage-2 ASD screening tools have received less attention, although use of stage-2 screening tools in combination with stage-1 screening tools confers several advantages (Khowaja et al., 2018).

In the two-stage ASD screening model, an intermediary stage-2 ASD screener is performed between the first stage-1 screen and the comprehensive specialist diagnosis. Two-stage screening models are alluring because they can decrease the number of children considered potentially at-risk for ASD and increase the certainty of ASD in the children who continue to

screen positive for ASD. Two-tiered screening systems can therefore both reduce the number of children who receive a costlier and more time-intensive ASD diagnostic evaluation and allow children who have not received a diagnosis by ASD specialists to have "presumptive eligibility" for ASD specific EI (Rotholz et al., 2017).

The Screening Tool for Autism in Toddlers (STAT; Stone et al., 2000, 2004) is a validated stage-2 ASD screener that has been used across several settings in a sequential manner to increase sensitivity over broad screeners and direct at-risk children for ASD specific services. For example, primary care providers (PCPs) in Tennessee were trained to administer the STAT within their offices to children who screen positive on stage-1 screeners, which resulted in an 85% increase in the number of children diagnosed with ASD by PCPs (Swanson et al., 2014). A statewide program in South Carolina used the STAT to follow up on positive M-CHAT results to determine "presumptive eligibility" for early intervention prior to a formal diagnostic evaluation. This sequential process resulted in five times as many children referred for early intervention, only 2.5% of whom later failed to receive an ASD diagnosis after evaluation (Rotholz et al., 2017). A more recent two-tiered screening study conducted in EI contexts in Massachusetts found that 83.7% of children who screened positive on the STAT after having received a stage-1 screener (BITSEA: Brief Infant-Toddler Social and Emotional Assessment, Briggs-Gowan, 2004 or POSI: Parents Observations of Social Interaction, Smith et al., 2013) were diagnosed with ASD at a comprehensive evaluation (Eisenhower et al., 2021). In all studies, these two stages of screening prior to evaluation were embedded within EI agencies or PCP settings and implemented autonomously by the agencies, suggesting that community settings can adopt a tiered screening approach that incorporates the STAT.

The STAT has been shown to have strong screening properties when administered by ASD specialists in clinical research settings, with a sensitivity of .92 and specificity of .85 (Stone et al., 2004). Its screening properties and potential utility for use by providers with less ASD experience was examined through a Vanguard Study of the National Children's Study[1] (NCS) (Newschaffer et al., 2017). One aim of the Vanguard Study was to develop a time- and resource-efficient ASD assessment protocol for use with children who screen positive on the Modified Checklist for Autism in Toddlers – Revised with Follow-Up (M-CHAT-R/F; Robins et al., 2014). Criteria for putative assessments were that they required minimal training, took 20 minutes or less to administer, and could be used by practitioners with little-to-no expertise in ASD. An expanded version of the STAT (i.e., STAT-E) was developed and evaluated for possible inclusion in this protocol; for this version, subjective ratings of social engagement (SE) and atypical behavior (AB) were added to each of the 12 items on the STAT to examine their potential utility for enhancing its screening properties with this unique sample of providers. Additionally, a scalable, self-paced web-based module was developed to train evaluators on STAT-E administration and scoring. Although the main NCS project was cancelled, data from the Vanguard Study enabled a comparison of the screening properties of the STAT-E using the original scoring algorithm and those of the STAT-E using the expanded scoring algorithms. Additionally, the stipulation that the ASD assessment protocol require minimal assessor training and resources suggests that the results may extend beyond the goals of the Vanguard Study and translate to community settings.

[1] The NCS was initiated to investigate the effects of environmental exposures and gene-environment interactions on pregnancy outcomes, child health and development, and precursors of adult disease, using a nationally representative longitudinal cohort study of 100,000 children from before birth through age 21. The Vanguard Study was to precede the Main Study, and was designed to test study procedures that were being considered for inclusion in the Main Study. The Vanguard Study was launched in January 2009 and ultimately collected data on approximately 6,000 families. In December 2014, the National Institutes of Health (NIH) decided against launching the Main Study and stopped data collection for the Vanguard Study (National Children's Study Archive, 2016).

The purpose of the current study was to expand upon the findings from the ASD Vanguard Study. In the ASD Vanguard Study, the STAT-E was reported to have a sensitivity of .63 and specificity of .70; however, both two-year-old and three-year-old children were included in the sample data and the SE and AB ratings were not included in the scoring algorithm. Our study goals were: (1) to assess the concurrent validity of the STAT-E using the original STAT scoring algorithm for two-year-old children when administered by evaluators with minimal ASD experience who were trained using a scalable web-based module; and (2) to examine whether adding the SE and AB ratings improves the concurrent validity for two-year-old children in this sample above and beyond the original STAT scoring algorithm. Specifically, this study directly compared the performance of the STAT-E using the original STAT scoring algorithm and original threshold (2.0 or higher), with performance using the STAT-E scoring algorithm (i.e., incorporating subjective examiner ratings of SE and AB) in addition to the original threshold. Our aim was to improve the incremental validity of the STAT-E by increasing its positive risk classification, as we prioritized capturing children who exhibit risk for ASD over excluding children who do not. We focused on two-year-old children because: (a) two-year-olds may require scoring weights commensurate with their development; and (b) we prioritized improving screening tools for younger toddlers and thereby contribute to timelier ASD diagnosis. We predicted that the SE and AB ratings, in combination with the original STAT total score, would improve accurate identification of two-year-old children at risk for ASD above and beyond the STAT total score alone.

Method

Participants

The sample used in this study is a subset of the larger Vanguard Study (see Newschaffer et al., 2017). Our sample comprised 238 children aged 24-35 months who received the STAT along with an independent neurodevelopmental evaluation at either an ASD or general neurodevelopmental disorder clinic affiliated with one of eight study sites. Children both with and without prior suspicion of ASD were recruited in order to develop a sample similar to those who might screen positive on a level one screener such as the M-CHAT. Study enrollment began in March 2013 and concluded in January 2015. The study protocol was IRB approved at all participating sites, the coordinating center (Drexel University) and the data center (Battelle Memorial Institute). Because the project was funded as a National Children's Study [Panel on the Design of the National Children's Study et al., 2014] Formative Research Project, US OMB approval was also obtained.

Ten two-year-old children were excluded from data analysis due to invalid STAT-E administrations, defined *a priori* as the child refusing participation in four or more of the 12 items. The majority of the two-year-old sample (76%) was male, and the mean age of children was 30.88 months (SD=3.51 months). Seventy-three percent (174 children) received a diagnosis of ASD. Additional child demographic information is provided in Table 1.

Procedure

Children received the STAT-E either before or after their neurodevelopmental evaluation, which was randomly determined at the site-level. STAT-E assessors were independent from the neurodevelopmental evaluation team and were blinded to the results of the neurodevelopmental evaluation. By design, STAT-E assessors did not have extensive clinical experience working with children with ASD; fifty percent of STAT-E assessors reported having six months or less of

ASD experience. Additional demographic information for STAT-E assessors is provided in Table 2. All STAT-E assessors completed web-based training modules on the study protocol and the STAT-E administration and scoring procedures prior to beginning data collection. The web-based STAT tutorial used text and video clips to train assessors on administering and scoring the 12 original STAT items and the two additional SE and AB ratings, and took four hours to complete. The neurodevelopmental evaluation was completed independently by an expert research or clinical team at the same institution and included a cognitive assessment as well as a best-estimate clinical ASD diagnosis from an ADOS-reliable clinician.

To assess the quality of scoring and administration fidelity for the STAT-E, assessors at six of the eight sites video-recorded a small sample of their administrations and sent them for review by the STAT-E development team at University of Washington. A total of 39 videos, of which 34 were of sufficient quality for review, were received from 6 sites (University of Washington and Vanderbilt University sites did not submit tapes because these sites had certified STAT trainers available to support staff and assure quality and fidelity of administration). To evaluate STAT-E scoring reliability for each study staff, UW reviewers scored each of the 12 STAT-E items from the videotape and then compared their scores to those of the assessors. Overall percent agreement across the 12 items for each assessor was at least 80%.

Study Measures

The Screening Tool for Autism in Toddlers and Young Children [STAT]. The STAT (Stone et al., 2000, 2004) is an interactive screening measure that takes 20 minutes to administer and comprises 12 play-based items that assess different aspects of social communication: play, requesting, directing attention, and imitation. Items are scored as pass-fail according to specific

criteria and summed to obtain a total score. Children who receive a score of 2 or higher on the STAT are considered at risk for ASD and were coded as ASD risk. The STAT has a sensitivity of 0.92 and specificity of 0.85 (Stone et al., 2004).

The Screening Tool for Autism in Toddlers and Young Children – Expanded [STAT-E]. Consistent with the original version, the STAT-E generates a response-to-press total score that indicates risk for ASD when the total score is 2 or above. Of particular relevance to the current study is the addition of examiners' qualitative ratings of children's social engagement (SE) and atypical behaviors (AB) observed during each of the 12 items. Rating systems for these new scores were designed to be simple enough to be scored live without disrupting the screening process. Scoring conventions for the new rating formats were developed based on pilot-testing in the senior author's lab.

Social Engagement Rating.

A subjective rating of social engagement was added to capture the *quality* of a child's interaction rather than their ability to simply pass the task (e.g., roll a ball back and forth with an adult). This addition was designed to capture children: (a) who may have learned to perform the activities of the STAT but do so with less regard to the interpersonal nature of the situation; and/or (b) who may fail tasks due to difficulty with transitions or temperament but nonetheless engage socially. Social engagement was scored using a three-point scale, with a score of one indicating that a child "regards the examiner as a social partner most of the time," a score of two indicating that a child "regards the examiner as a social partner some of the time," and a score of three indicating that the child "shows very little regard for the examiner as a social partner." A mean social engagement score (summary social engagement) was computed across all 12 items

of the STAT-E for use in the candidate scoring algorithms and ranged from one to three, with higher scores indicating less social engagement.

Atypical Behavior Ratings. Four atypical behavior (AB) ratings were added to the STAT-E in order to capture a core diagnostic domain of ASD that was not previously measured on the STAT. Four categories that map onto DSM criteria and might be observed in a brief play-based session were included: atypical language features, repetitive actions on objects, repetitive body actions or posturing, and sensory-seeking behavior with objects. Each atypical behavior category was scored as "present" (i.e., one or more behaviors observed across all trials of the item) or "not present," to ease the scoring burden during a live administration for a construct that has shown to be difficult to score (Myers et al., 2018). Scores thus ranged from zero to four for each of the 12 STAT-E items. The mean atypical behavior score (summary atypical behavior) was computed across all 12 items of the STAT-E for use in the candidate scoring algorithms and ranged from zero to four, with higher scores indicating more atypical behaviors.

ASD Classification. Best-estimate clinical diagnosis was coded as ASD or No ASD and serves as the outcome variable. The diagnosis was made by expert research or clinical teams on the basis of a neurodevelopmental evaluation which included either the General or Second Edition of the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2013), structured parent interviews and developmental testing.

Analytic Approach

All analyses were conducted using the R software package [R Core Team, 2017]. Confusion matrices, which include screening properties such as sensitivity and specificity, were used to assess the screening properties of the STAT in the full sample. Next, SE and AB descriptive statistics were examined. In order to ascertain whether our candidate scoring models would generalize to an independent sample, we randomly split our data such that 60% was used in the training sample and 40% was used in the testing sample. There were no significant differences between the training sample and testing sample on demographic variables ($p>.20$ for all). Three logistic regression-based candidate scoring algorithms using the new ratings (SE, AB, or SE & AB) were trained on the training sample and a selected candidate scoring algorithm was tested on the testing sample. Akaike information criterion (AIC) model selection was used to distinguish between the logistic regression models. AIC estimates the quality of each model relative to the other models by measuring prediction error and penalizing additional parameters; lower AIC numbers indicate better model fit and parsimony of predictors (Kuha, 2004).

The screening properties for the candidate scoring algorithms in our training data were examined using sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The screening properties of the STAT-E using the original STAT scoring algorithm in the training sample was computed for comparison. Youden's J was computed to determine the utility of a candidate scoring algorithm when sensitivity and specificity are considered equally. The F1 score was computed to measure positive risk classification without accounting for negative risk. Youden's J and the F1 score have a value between 0 and 1 with higher numbers indicating greater classification accuracy.

Results

STAT-E using the original STAT scoring algorithm in the Full Sample

A pass or fail score using the STAT scoring algorithm yielded a STAT sensitivity of 0.67 and specificity of 0.66.

Descriptives of Social Engagement (SE) in the Full Sample

The summary SE score had a mean of 2.32 (SD=0.6) and a negative skew (-0.6), indicating that there were more endorsements of three ("shows very little regard for the examiner as a social partner") relative to two and one. Of all two-year-old children, 44% had a summary SE score between 2.0 and 2.9 and 44% had a summary SE score of 3. The summary SE score also showed a significant and strong, positive linear association with the STAT total score ($r=.74$, $p<.001$) and a significant and weak-moderate positive linear association with ASD diagnosis ($r=.36$, $p<.001$). Multi-collinearity between the summary SE score and the STAT-E total score was assessed due to their strong correlation; however, the variance inflation factor (VIF) was low (STAT-E VIF: 1.7; SE VIF: 1.62), indicating minimal concern for multi-collinearity.

Descriptives of Atypical Behavior (AB) in the Full Sample

Overall, the AB items were endorsed infrequently. The summary AB score had a mean of .37 (SD=0.49) and a positive skew (1.81), indicating that there were more endorsements of lower (less atypical) AB scores (i.e., 0 and 1) compared to higher AB scores (i.e., 2-4). Of all two-year-old children, 70% had a summary AB score between 0 and .9 and 27% had a summary AB score between 1 and 1.9. The summary AB score also had a weak-moderate positive linear association with the STAT total score ($r=.32$, $p>.05$) and a weak positive linear association with ASD diagnosis ($r=.23$, $p>.05$).

Development of New Algorithms Exploring SE and AB Ratings in Training Sample

Three candidate scoring algorithms were trained using logistic regression with the clinical best estimate (ASD or no ASD) as the outcome variable. The STAT risk classification was a predictor variable in each model. We also trained a scoring model with only the STAT for comparison. Our candidate scoring models tested (1) the STAT-E with SE; (2) the STAT-E with AB; and (3) the STAT-E with SE and AB. AIC model selection indicated that the best fit model included the STAT-E and the SE rating and not the AB rating. The model with the STAT-E and SE carried 32% of the cumulative weight and had the lowest AIC score, suggesting that the STAT-E and SE was the best model. Table 4 includes the logistic regression output with AIC scores.

Screening properties for the candidate scoring algorithms for the STAT-E are presented in Table 3. All of the candidate scoring algorithms improved upon the sensitivity of the STAT-E using the original STAT scoring algorithm. The STAT-E that included SE and AB yielded a sensitivity of 0.77 and specificity of 0.62. In our training sample, none of the candidate scoring methods achieved the same or better specificity than the STAT-E using the STAT scoring algorithm. ROC curves of the four scoring algorithms are included in Figure 1 and illustrate that the STAT-E using the STAT scoring algorithm performs better when the false positive rate is prioritized and the models with SE perform better when the true positive rate is prioritized. Overall, the candidate scoring algorithm with SE and AB had the highest Youden's J (0.39) and the highest F1 score (0.80) in comparison to the other candidate scoring algorithms and the STAT-E using the STAT scoring algorithm. Screening properties of all the scoring models are listed in Table 3.

Validation of Final Candidate Algorithm in Testing Sample

We assessed the candidate scoring algorithm with both additional ratings in our testing data, as this algorithm had the highest Youden's J and F1 score. Screening properties of the candidate scoring algorithm with SE and AB ratings in our testing data are listed in Table 3. The F1 score increased from .80 in our training data to .85 in our testing data.

Discussion

The first goal of this study was to examine the screening properties of the STAT-E with the original STAT scoring algorithm, as administered to two-year-old children by NCS assessors with minimal ASD experience who were trained using a self-paced, scalable, and cost-effective web-based training module. In this Vanguard Study sample, the STAT-E with the original STAT scoring algorithm had an appreciably lower sensitivity and specificity compared to the screening properties of the STAT from the original validation study (Stone et al., 2004). However, the screening properties of the STAT-E with the original STAT scoring algorithm in this sample were similar to the previously reported screening properties for two-year-old and three-year-old children (Newschaffer et al., 2016). Unfortunately, in this study, it was not possible to determine whether the weaker screening properties were due to the online training format or the decreased level of assessor experience, or a combination of both factors. Nonetheless, these weaker screening properties may indicate that more experience with ASD and/or more rigorous training on the STAT may lead to better concurrent validity.

Per the developer, the ideal way to obtain STAT training is to attend an in-person (or virtual) STAT training workshop and receive formal certification after submitting two videotapes that demonstrate administration and scoring fidelity, which enables assessors to solidify their ASD knowledge and practice STAT administration and scoring in the presence of an ASD expert

and certified STAT trainer. Nonetheless, when this more rigorous STAT training is not feasible, singular use of the web-based training module provides reasonable screening properties and the positive risk classification can be optimized by inclusion of the two experimental SE and AB ratings. This study therefore demonstrates that it is possible to train individuals of varying backgrounds to use the STAT with a briefer, predominantly online learning module. In this way, a less rigorous training model may be an acceptable path for scaling the use of the STAT more broadly to increase the number of providers, such as day-care providers and preschool teachers, who can contribute toward early identification of ASD.

The second goal of this study was to explore whether the SE and AB ratings optimize the screening properties of the STAT-E, with a focus on improving positive risk accuracy. Our results revealed that inclusion of SE and AB ratings did increase the positive risk accuracy of the STAT-E. However, the decision regarding which STAT scoring algorithm to use may depend on the screening context. For instance, when preventing unnecessary referrals to an already overburdened service delivery system is prioritized, such as within many primary care settings (Mazurek et al., 2020), the original STAT scoring method may be preferable to providers due to its higher specificity. However, when positive risk classification is prioritized, the candidate scoring algorithm with the two additional ratings confers a 6%-11% improvement in positive risk classification. One setting in which positive risk classification may be prioritized is in early intervention settings, where children are already identified with a disability and the STAT can highlight the need for ASD-specialized strategies and/or help determine whether or not to refer for additional diagnostic assessment (Rotholz et al., 2017).

Precedents for using several scoring algorithms within one measure have been established for ASD tools generally and the STAT in particular. For example, the ADOS-2

includes separate scoring algorithms for children with different language levels (Lord et al., 2013). Further, Roberts et al., 2019 suggested that using a two-threshold logistic regression method on the STAT has potential psychometric advantages over a single threshold and categorical scoring: a higher threshold can be used to maximize specificity and a low threshold can be used to maximize sensitivity.

While assessing for atypical behavior seemed conceptually important for providing coverage for both ASD diagnostic domains, our atypical behavior rating was endorsed infrequently and did not contribute to increased positive risk classification as we hypothesized. A possible explanation is that atypical behavior may be difficult for less experienced assessors to code, which has been supported in other studies (Myers et al., 2018). For example, Myers et al. used crowdsourcing to recruit novice observers to score social communication and "unusual behavior" and found that while social communication items correlated highly with the ratings of experts, the unusual behavior items did not. Our atypical behavior rating, however, did contribute to specificity, perhaps indicating that an absence of atypical behavior is more suggestive of *decreased* ASD risk rather than the presence of AB suggesting *increased* risk. This may be supported by the finding that children without ASD may also occasionally engage in atypical behavior, and that children with ASD may not show atypical behavior during a 20-minute duration. Our finding that atypical behavior does not contribute to the sensitivity of the STAT-E is consistent with other stage-2 screening tools which prioritize measuring social communication over atypical behavior (Dow et al., 2020; Nah et al., 2019). For example, the Brief Autism Detection in Early Childhood (BADEC; Nah et al., 2019) includes no atypical behavior items amongst the five critical items that are required for a child to screen at risk for ASD.

This study has several limitations that warrant discussion and provide directions for future research. First, combining our goal of assessing the screening properties of the STAT when administered by assessors with less ASD knowledge and familiarity with piloting two experimental ratings impedes our ability to generalize the concurrent validity of the two experimental ratings when used by assessors more experienced with ASD and the STAT. Second, our predominantly White and male sample hampers our ability to assess algorithmic parity across several important identity dimensions, such as race, ethnicity, and sex. A future study should assess the efficacy of SE and AB ratings when administered to a diverse sample in community settings by STAT assessors who have completed the recommended training.

To our knowledge, this study is unique in assessing the screening properties of a stage-2 ASD screening tool for two-year-old children when administered by assessors who are trained using a self-paced, cost-effective and scalable web-based training module. While additional research is needed, our results suggest that including two experimental ratings optimizes the positive risk classification of the STAT-E when used by novice assessors with two-year-old children in this Vanguard Study sample, and that a less rigorous training model may be an acceptable means for increasing the number of providers using the STAT and thereby contribute toward early identification of ASD. Additionally, this study supports previous research suggesting that providing only one scoring algorithm for an ASD screening tool may not work equally well across all contexts or for all children. We propose an alternative method of developing multiple scoring algorithms for use with the same screening measure to identify risk in diverse contexts and for diverse samples of children.

References

Briggs-Gowan, M. J. (2004). The Brief Infant-Toddler Social and Emotional Assessment: Screening for Social-Emotional Problems and Delays in Competence. *Journal of Pediatric Psychology*, *29*(2), 143–155. https://doi.org/10.1093/jpepsy/jsh017

Brooks, B. A., Haynes, K., Smith, J., McFadden, T., & Robins, D. L. (2016). Implementation of Web-Based Autism Screening in an Urban Clinic. *Clinical Pediatrics*, *55*(10), 927–934. https://doi.org/10.1177/0009922815616887

Campbell, K., Carpenter, K. L. H., Espinosa, S., Hashemi, J., Qiu, Q., Tepper, M., Calderbank, R., Sapiro, G., Egger, H. L., Baker, J. P., & Dawson, G. (2017). Use of a Digital Modified Checklist for Autism in Toddlers − Revised with Follow-up to Improve Quality of Screening for Autism. *The Journal of Pediatrics*, *183*, 133-139.e1. https://doi.org/10.1016/j.jpeds.2017.01.021

Chawarska, K., Klin, A., Paul, R., Macari, S., & Volkmar, F. (2009). A prospective study of toddlers with ASD: Short-term diagnostic and cognitive outcomes. *Journal of Child Psychology and Psychiatry*, *50*(10), 1235–1245. https://doi.org/10.1111/j.1469-7610.2009.02101.x

Dow, D., Day, T. N., Kutta, T. J., Nottke, C., & Wetherby, A. M. (2020). Screening for autism spectrum disorder in a naturalistic home setting using the systematic observation of red flags (SORF) at 18–24 months. *Autism Research*, *13*(1), 122–133. https://doi.org/10.1002/aur.2226

Eisenhower, A., Martinez Pedraza, F., Sheldrick, R. C., Frenette, E., Hoch, N., Brunt, S., & Carter, A. S. (2021). Multi-stage Screening in Early Intervention: A Critical Strategy for Improving ASD Identification and Addressing Disparities. *Journal of Autism and Developmental Disorders*, *51*(3), 868–883. https://doi.org/10.1007/s10803-020-04429-z

Harrington, J. W., Bai, R., & Perkins, A. M. (2013). Screening Children for Autism in an Urban Clinic Using an Electronic M-CHAT. *Clinical Pediatrics*, *52*(1), 35–41. https://doi.org/10.1177/0009922812463957

Hyman, S. L., Levy, S. E., & Myers, S. M. (2020). Identification, Evaluation, and Management of Children With Autism Spectrum Disorder. *Pediatrics*, *145*(1). https://doi.org/10.1542/peds.2019-3447

Khowaja, M., Robins, D. L., & Adamson, L. B. (2018). Utilizing two-tiered screening for early detection of autism spectrum disorder. *Autism*, *22*(7), 881–890. https://doi.org/10.1177/1362361317712649

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research*, *33*(2), 188–229. https://doi.org/10.1177/0049124103262065

Lord, C. (2013). Test Review: Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part II): Toddler Module. *Journal of Psychoeducational Assessment*, *32*, 88–92. https://doi.org/10.1177/0734282913490916

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*(6), 694–701. https://doi.org/10.1001/archpsyc.63.6.694

Maenner, M. J., Shaw, K. A., Baio, J., Washington, A., Patrick, M., DiRienzo, M., Christensen, D. L., Wiggins, L. D., Pettygrove, S., Andrews, J. G., Lopez, M., Hudson, A., Baroud, T., Schwenk, Y., White, T., Rosenberg, C. R., Lee, L.-C., Harrington, R. A., Huston, M., … Dietz, P. M. (2020). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United

States, 2016. *MMWR Surveillance Summaries*, *69*(4), 1–12. https://doi.org/10.15585/mmwr.ss6904a1

Mazurek, M. O., Harkins, C., Menezes, M., Chan, J., Parker, R. A., Kuhlthau, K., & Sohl, K. (2020). Primary Care Providers′ Perceived Barriers and Needs for Support in Caring for Children with Autism. *The Journal of Pediatrics*. https://doi.org/10.1016/j.jpeds.2020.01.014

Mazurek, M. O., Kuhlthau, K., Parker, R. A., Chan, J., & Sohl, K. (2021). Autism and General Developmental Screening Practices Among Primary Care Providers. *Journal of Developmental & Behavioral Pediatrics*, *42*(5), 355–362. https://doi.org/10.1097/DBP.0000000000000909

Monteiro, S. A., Dempsey, J., Berry, L. N., Voigt, R. G., & Goin-Kochel, R. P. (2019). Screening and Referral Practices for Autism Spectrum Disorder in Primary Pediatric Care. *Pediatrics*, *144*(4), e20183326. https://doi.org/10.1542/peds.2018-3326

Myers, E., Stone, W. L., Bernier, R., Lendvay, T., Comstock, B., & Cowan, C. (2018). The diagnosis conundrum: Comparison of crowdsourced and expert assessments of toddlers with high and low risk of autism spectrum disorder. *Autism Research*, *11*(12), 1629–1634. https://doi.org/10.1002/aur.2030

Nah, Y.-H., Young, R. L., & Brewer, N. (2019). Development of a brief version of the Autism Detection in Early Childhood. *Autism*, *23*(2), 494–502. https://doi.org/10.1177/1362361318757563

Newschaffer, C. J., Schriver, E., Berrigan, L., Landa, R., Stone, W. L., Bishop, S., Burkom, D., Golden, A., Ibanez, L., Kuo, A., Lakes, K. D., Messinger, D. S., Paterson, S., & Warren, Z. E. (2017). Development and validation of a streamlined autism case confirmation

approach for use in epidemiologic risk factor research in prospective cohorts: Streamlined ASD Case Confirmation. *Autism Research*, *10*(3), 485–501. https://doi.org/10.1002/aur.1659

Oswald, D. P., Haworth, S. M., Mackenzie, B. K., & Willis, J. H. (2017). Parental Report of the Diagnostic Process and Outcome: ASD Compared With Other Developmental Disabilities. *Focus on Autism and Other Developmental Disabilities*, *32*(2), 152–160. https://doi.org/10.1177/1088357615587500

Roberts, M. Y., Stern, Y., Hampton, L. H., Grauzer, J. M., Miller, A., Levin, A., Kornfeld, B., Davis, M. M., Kaat, A., & Estabrook, R. (2019). Beyond pass-fail: Examining the potential utility of two thresholds in the autism screening process: Roberts et al./Two thresholds in autism screening. *Autism Research*, *12*(1), 112–122. https://doi.org/10.1002/aur.2045

Robins, D. L., Casagrande, K., Barton, M., Chen, C.-M. A., & Fein, D. (2014). *Validation of the Modified Checklist for Autism in Toddlers, Revised With Follow-up (M-CHAT-R/F)*. *133*(1), 11.

Rotholz, D. A., Kinsman, A. M., Lacy, K. K., & Charles, J. (2017). Improving Early Identification and Intervention for Children at Risk for Autism Spectrum Disorder. *Pediatrics*, *139*(2), e20161061. https://doi.org/10.1542/peds.2016-1061

Sheldrick, R. C., Maye, M. P., & Carter, A. S. (2017). Age at First Identification of Autism Spectrum Disorder: An Analysis of Two US Surveys. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(4), 313–320. https://doi.org/10.1016/j.jaac.2017.01.012

Smith, N. J., Sheldrick, R. C., & Perrin, E. C. (2013). An Abbreviated Screening Instrument for Autism Spectrum Disorders. *Infant Mental Health Journal*, *34*(2), 149–155. https://doi.org/10.1002/imhj.21356

Steinman, K. J., Stone, W. L., Ibañez, L. V., & Attar, S. M. (2021). Reducing Barriers to Autism Screening in Community Primary Care: A Pragmatic Trial Using Web-Based Screening. *Academic Pediatrics*. https://doi.org/10.1016/j.acap.2021.04.017

Stone, W. L., Coonrod, E. E., & Ousley, O. Y. (2000). *Brief Report: Screening Tool for Autism in Two-Year-Olds (STAT): Development and Preliminary Data*. 6.

Stone, W. L., Coonrod, E. E., Turner, L. M., & Pozdol, S. L. (2004). Psychometric Properties of the STAT for Early Autism Screening. *Journal of Autism and Developmental Disorders*, *34*(6), 691–701. https://doi.org/10.1007/s10803-004-5289-8

Sturner, R., Howard, B., Bergmann, P., Morrel, T., Andon, L., Marks, D., Rao, P., & Landa, R. (2016). Autism Screening With Online Decision Support by Primary Care Pediatricians Aided by M-CHAT/F. *PEDIATRICS*, *138*(3), e20153036–e20153036. https://doi.org/10.1542/peds.2015-3036

Swanson, A. R., Warren, Z. E., Stone, W. L., Vehorn, A. C., Dohrmann, E., & Humberd, Q. (2014). The diagnosis of autism in community pediatric settings: Does advanced training facilitate practice change? *Autism*, *18*(5), 555–561. https://doi.org/10.1177/1362361313481507

Turner, L. M., Stone, W. L., Pozdol, S. L., & Coonrod, E. E. (2006). Follow-up of children with autism spectrum disorders from age 2 to age 9. *Autism*, *10*(3), 243–265. https://doi.org/10.1177/1362361306063296

Zwaigenbaum, L., Bauman, M. L., Choueiri, R., Fein, D., Kasari, C., Pierce, K., Stone, W. L., Yirmiya, N., Estes, A., Hansen, R. L., McPartland, J. C., Natowicz, M. R., Buie, T., Carter, A., Davis, P. A., Granpeesheh, D., Mailloux, Z., Newschaffer, C., Robins, D., ... Wetherby, A. (2015). Early Identification and Interventions for Autism Spectrum Disorder: Executive Summary. *Pediatrics*, *136*(Supplement 1), S1−S9. https://doi.org/10.1542/peds.2014-3667B

Tables

Table 1. Child Demographics for Full Sample and Split into Training and Testing Samples

	Sample (n=238)		Training (n=142)		Testing (n=96)	
	n	%	n	%	n	%
Sex						
Male	180	76%	103	73%	77	80%
Female	56	24%	38	27%	18	19%
Missing	2	1%	1	1%	1	1%
Age						
<30 months	80	34%	49	35%	31	32%
>= 30 months	158	66%	93	65%	65	68%
Verbal Ability						
Nonverbal	132	55%	82	58%	50	52%
Verbal	106	45%	60	42%	46	48%
Mullen ELC						0%
>70	79	33%	44	31%	35	36%
<= 70	100	42%	59	42%	41	43%
Missing	59	25%	39	27%	20	21%
ASD Diagnosis						
ASD	174	73%	103	73%	71	74%
Not ASD	64	27%	39	27%	25	26%
Ethnicity						
Hispanic or Latino	65	27%	36	25%	29	30%
Not Hispanic or Latino	171	72%	105	74%	66	69%
Missing	2	1%	1	1%	1	1%
Race						
Indigenous American	3	1%	1	1%	2	2%
Asian	22	9%	13	9%	9	9%
Black/African American	20	8%	13	9%	7	7%
Pacific Islander	1	0%	0	0%	1	1%
White	162	68%	100	70%	62	65%
More than one race	9	4%	5	4%	4	4%
Don't Know/Refused	19	8%	9	6%	10	10%
Missing	2	1%	1	1%	1	1%
Maternal Education Level						
<Bachelor's	118	50%	68	48%	50	52%
Bachelor's	69	29%	30	21%	26	27%
>Bachelor's	48	20%	43	30%	18	19%
Missing	1	0%	0	0%	1	1%
Family Income						
<$29, 999	58	24%	36	25%	22	23%
30,000−49,000	36	15%	19	13%	17	18%
50,000−74,999	38	16%	26	18%	12	13%
75,000−99,999	32	13%	17	12%	15	16%
$100,000+	64	27%	39	27%	25	26%
Don't Know/Refused	8	3%	4	3%	4	4%
Missing	2	1%	1	1%	1	1%

Notes:
† There were no significant differences between the training and testing sample on any demographic variable, (p>.20 for all)

Table 2. STAT Assessor Demographics for Full Sample and Split into Training and Testing Samples

Table 2: STAT-E Assessor Demographics

	Sample (n=238)		Training (n=142)		Testing (n=96)	
	n	%	n	%	n	%
Education Level						
<Bachelor's	86	36%	44	31%	42	44%
Bachelor's	72	30%	42	30%	30	31%
>Bachelor's	79	33%	55	39%	24	25%
Missing	1	0%	1	1%	0	0%
Field of Study						
Psychology	148	62%	89	63%	59	61%
Other	89	37%	52	37%	37	39%
Missing	1	0%	1	1%	0	0%
Years ASD Experience						
0-0.5	119	50%	67	47%	52	54%
1-2	74	31%	46	32%	28	29%
3-6	19	8%	13	9%	6	6%
6+	16	7%	10	7%	6	6%
Missing	10	4%	6	4%	4	4%
ASD Familiarity						
High	69	29%	38	27%	31	32%
Average	135	57%	82	58%	53	55%
Low	19	8%	12	8%	7	7%
Missing	15	6%	10	7%	5	5%

Notes:

† There were no significant differences between the training and testing sample on any demographic variable, ($p>.20$ for all)

Table 3. Logistic Regression Models for Scoring Algorithms in Training Data

	Dependent variable:			
	ASD Classification			
	STAT-E	SE	AB	SE + AB
STAT-E	1.475***	0.969*	1.284***	0.810
	(0.405)	(0.515)	(0.427)	(0.535)
Social Engagement		0.638		0.605
		(0.409)		(0.411)
Atypical Behavior			0.641	0.582
			(0.548)	(0.546)
Constant	0.260	-0.884	0.155	-0.918
	(0.256)	(0.773)	(0.271)	(0.775)
Observations	142	142	142	142
Log Likelihood	-76.274	-75.061	-75.506	-74.424
Akaike Inf. Crit.	156.549	156.122	157.012	156.848

Notes:

† Best-estimate clinical diagnosis of ASD was used as the outcome variable.

†† Each column includes the model results for a candidate scoring algorithm. The first column ("STAT-E") includes only the STAT-E risk outcome as a predictor variable. The second column ("SE") includes the STAT-E risk outcome and SE rating. The third column ("AB") includes the STAT-E risk outcome and AB rating. The fourth column ("SE + AB") includes the STAT-E risk outcome and the SE and AB rating.
* p<.1
*** p <.01

Table 4. Screening Properties of the STAT-E by Scoring Algorithm

Prediction Algorithm	Sensitivity	Specificity	NPV	PPV	Youden's J	F1
STAT-E	0.660	0.690	0.440	0.850	0.350	0.740
STAT-E + SE	0.770	0.510	0.470	0.800	0.280	0.790
STAT-E + AB	0.680	0.670	0.440	0.840	0.350	0.750
STAT-E + AB + SE	0.770	0.620	0.500	0.840	0.390	0.800
Testing Data	0.860	0.520	0.570	0.840	0.380	0.850

Notes.

† Best-estimate clinical diagnosis of ASD was used as the outcome variable.

Figure 1. ROC of Four STAT-E Scoring Algorithms.