Received: 8 May 2021 | Accepted: 16 November 2021

DOI: 10.1002/aur.2643

RESEARCH ARTICLE

# What are we optimizing for in autism screening? Examination of algorithmic changes in the M-CHAT

Synnve Schjølberg[1] | Frederick Shic[2] | Fred R. Volkmar[3,4] | Anders Nordahl-Hansen[5] | Nina Stenberg[6] | Tonje Torske[7] | Kenneth Larsen[8] | Katherine Riley[2] | Denis G. Sukhodolsky[3] | James F. Leckman[3] | Katarzyna Chawarska[3] | Roald A. Øien[3,8]

[1]Mental Health, Norwegian Institute of Public Health, Oslo, Norway

[2]Seattle Children's Research Institute, University of Washington School of Medicine, Seattle, Washington

[3]Yale University, New Haven, Connecticut

[4]Southern Connecticut University, New Haven, Connecticut

[5]Education, Østfold University College, Halden, Norway

[6]Oslo University Hospital, Oslo, Norway

[7]Vestre Viken Hospital, Drammen, Norway

[8]Education, UiT—The Arctic University of Norway, Tromso, Norway

**Correspondence**
Roald A. Øien, Psychology, UiT—The Arctic University of Norway, PB 6070, Tromso 9037, Norway.
Email: roald.a.oien@uit.no

**Funding information**
Research Council of Norway, Grant/Award Number: 196452; National Institute of Environmental Health Sciences (NIH/NIEHS), Grant/Award Number: NO-ES-75558; Norwegian Ministry of Health and Care Services, the Norwegian Ministry of Education and Research, the Research Council of Norway/FUGE, Grant/Award Number: 151918; National Institute of Neurological Disorders and Stroke (NIH/NINDS), Grant/Award Number: NS47537

**Abstract**

The present study objectives were to examine the performance of the new M-CHAT-R algorithm to the original M-CHAT algorithm. The main purpose was to examine if the algorithmic changes increase identification of children later diagnosed with ASD, and to examine if there is a trade-off when changing algorithms. We included 54,463 screened cases from the Norwegian Mother and Child Cohort Study. Children were screened using the 23 items of the M-CHAT at 18 months. Further, the performance of the M-CHAT-R algorithm was compared to the M-CHAT algorithm on the 23-items. In total, 337 individuals were later diagnosed with ASD. Using M-CHAT-R algorithm decreased the number of correctly identified ASD children by 12 compared to M-CHAT, with no children with ASD screening negative on the M-CHAT criteria subsequently screening positive utilizing the M-CHAT-R algorithm. A nonparametric McNemar's test determined a statistically significant difference in identifying ASD utilizing the M-CHAT-R algorithm. The present study examined the application of 20-item MCHAT-R scoring criterion to the 23-item MCHAT. We found that this resulted in decreased sensitivity and increased specificity for identifying children with ASD, which is a trade-off that needs further investigation in terms of cost-effectiveness. However, further research is needed to optimize screening for ASD in the early developmental period to increase identification of false negatives.

KEYWORDS

children, early detection, psychometrics

## INTRODUCTION

Early identification of children on a developmental path to autism spectrum disorder (ASD) is vital for providing early, tailored intervention. However, early identification

[Correction added on 1st December 2021, after first online publication: The third affiliation has been updated]

Synnve Schjølberg and Frederick Shic shared first authorship.

is challenging due to the heterogenous nature of ASD in terms of symptom patterns and the onset time of symptom patterns (Chawarska et al., 2007; Ozonoff et al., 2010; Zwaigenbaum et al., 2015). While for some children, symptoms are evident during infancy and early in development, for others, symptoms are difficult to detect until social expectations exceed social abilities (Ozonoff et al., 2015). Thus, screening instruments for children in the early developmental period might not pick up children that have more subtle ASD symptom expression, but rather those with more severe disabilities regardless of ASD diagnosis (Øien et al., 2018; Stenberg et al., 2021).

One of the first systematic and widely used early developmental screening instruments for ASD was the Checklist for Autism in Toddlers (CHAT; Baird et al., 2000; Baron-Cohen et al., 1992, 2000). The Modified Checklist for Autism in Toddlers (M-CHAT) was subsequently derived from the CHAT by Robins and colleagues in 2001 (Robins et al., 2001), broadening the symptom list to capture a larger proportion of the children with ASD. Since then, the M-CHAT and its derivative instruments have become some of the most widely used early screening instruments for ASD in young children, contributing to the early identification of children with ASD across the globe (Stewart & Lee, 2017). However, recent studies have proposed that the M-CHAT, like the CHAT, struggles with a high number of false negatives and false positives, showing clear and grave nonoptimal performance (Baird et al., 2011; Carbone et al., 2020; Guthrie et al., 2019; Øien et al., 2018, 2019; Stenberg et al., 2014, 2021).

The high number of false positives using the M-CHAT and its derivatives (Guthrie et al., 2019; Øien et al., 2018; Stenberg et al., 2014, 2021) add to the discussion regarding the utility and cost-effectiveness of universal screening (Baird et al., 2011; Guthrie et al., 2019; Hickey et al., 2021; McPheeters et al., 2016; Øien et al., 2018; Siu et al., 2016; Stenberg et al., 2014, 2021; Surén et al., 2019; Yuen, Carter, et al., 2018; Yuen, Penner, et al., 2018). A high rate of false positives may lead to unnecessary anxiety for some parents. However, one benefit of positive results when screening for ASD is the potential for identifying other disabilities and difficulties that also require specialized health services. Research is not clear on how adapting new criteria and/or new cut-off scores improve both the sensitivity and specificity of screening instruments or if there is a trade-off between rates of false positives and false negatives. As it has been debated that current lack of evidence for universal screening for ASD obtaining this knowledge is crucial for understanding how attempts at optimization may impact and influence the "true costs" of autism screening.

A 20-item revision of the M-CHAT that primarily focused on reducing the rate of false positives was published in 2014 (Robins et al., 2014): the Modified Checklist for Autism in Toddlers Revised (M-CHAT-R). The revision removed three items that were poor predictors of ASD and retained 20 items (with rewording and new exemplification for 12 of the 20 retained items and reordering of items). To help resolve potential ambiguity in item interpretation in the original M-CHAT, descriptive examples of each question were added in the M-CHAT-R. In addition, risk score calculation algorithms were modified in the revision. In addition, during the transition from the M-CHAT to the M-CHAT-R, the standard follow-up (Robins et al., 2001) was more rigorously operationalized as part of the standard operating procedure for screening administration (yielding the M-CHAT-R/F, i.e., the M-CHAT-R questionnaire with follow-up interview). The purpose of the follow-up interview, in both the case of the original M-CHAT and the M-CHAT-R was to provide additional diagnostic accuracy when children scored in an "intermediate range" of risk on the questionnaire portion. In this work, we do not consider the follow-up interview, which is often irregularly administrated in practice (Wallis et al., 2020).

However, more research is needed to assess the impact and tradeoffs of methodological optimizations in ASD screening in the general population and in children of different ages to answer the question, "what are we optimizing for?" This includes understanding factors related to different aspects of assessment: *false positives* with exploring symptom overlap to other neurodevelopmental disorders and *false negatives* with identifying broader symptom patterns than those currently considered as core ASD symptoms. A limitation of much of prior research on M-CHAT-related screening instruments, such as the original descriptive validation paper on the M-CHAT-R/F (Robins et al., 2014), is that they do not conduct prospective follow-up of all children, and as such tend to focus only on false positives while neglecting false negatives.

There are currently no studies that have simultaneously administered both the M-CHAT and the M-CHAT-R instruments with or without follow-up. For this reason, direct comparisons between the two measures are currently impossible. However, it is still possible to examine changes in screening performance due to algorithmic changes made in the transition from the M-CHAT to the M-CHAT-R.

This study aimed to evaluate the potential optimization of the original M-CHAT's efficacy in identifying ASD using the original recommended M-CHAT cut-off criteria as compared to a 20-item M-CHAT (M-CHAT$_{20}$) that was created from the original 23-item M-CHAT so as to replicate as closely as possible those changes incorporated into the M-CHAT-R. The 20 items of the M-CHAT$_{20}$ were the same as those retained in the M-CHAT-R, and the cut-off criteria applied for ASD risk-status were the same as those recommended by the M-CHAT-R. Specifically, this study examines trade-offs in rates of false positives and false negatives between the original M-CHAT and the M-CHAT$_{20}$.

# METHODS

## Participants

The present study utilizes data collected in the Norwegian Mother, Father, and Child Cohort Study (MoBa; Magnus et al., 2016). The Autism Birth Cohort (ABC) Study is a sub-study in the MoBa which aims to identify all ASD cases within MoBa (Skjærven et al., 2006; Stoltenberg et al., 2010; Surén et al., 2019). MoBa is a national prospective general population pregnancy cohort that includes 114,552 children born between 1999 and 2009 (Magnus et al., 2016). Parents who agreed to participate in MoBa and the ABC study signed an informed consent form in each study. The study was approved by the Regional Committee for Medical and Health Research Ethics South East. MoBa data version 9 was used. In the present study, the child's status as ASD or non-ASD was determined based by the discharge diagnosis listed in the National Patient Registry (NPR) or by diagnostic conclusion in the ABC study (at approximately 42 months). Diagnoses from the NPR are obtained prospectively, and are provided by specialized health services in Norway in clinics that conduct ASD-specific assessments utilizing gold standard instruments such as the ADOS and the ADI-R, together with other instruments including measures of cognitive and adaptive ability. The youngest children that participated in the MoBa study turn 12 years of age in 2021.

## Study sample

This study uses data collected prospectively in the MoBa study and the ABC study. The primary focus is on the early developmental period of children whose parents received and returned the 18-month questionnaire, relating ASD-relevant characteristics to a later diagnosis of ASD from the NPR linkage or by ABC discharge diagnosis.

The complete M-CHAT was included as one section (translated and back-translated, and items listed in the correct order) in the MoBa 18-month questionnaire from March 2005 through January 2011. Children whose parents returned the 18-month questionnaire and completed all 23 items from the M-CHAT (Robins et al., 2001) were included in the final sample ($N = 54{,}436$). Of the final sample, 332 children were later identified with an ASD diagnosis through the NPR or in the ABC clinic (mean age 42 months).

## Measures

### Original M-CHAT (2001) scoring

The original M-CHAT (Robins et al., 2001; for clarity, referred to here as the M-CHAT$_{23}$) is a 23-item, yes–no parent completed checklist developed for children 16–30 months. It was designed for completion by the childcare providers in the waiting room of well-baby clinics. A positive screening status depends on failing either (1) two or more of the six-critical discriminative items (i.e., the Crit6 criterion) and/or (2) three or more of the 23 items (Tot23 criterion). When the M-CHAT$_{23}$ is used as a screening measure, it is recommended to do a follow-up phone interview of screen positives to reduce false positives. These follow-up phone interviews were not conducted in the MoBa study due to its prohibitive cost.

### M-CHAT-R (2014) scoring and M-CHAT$_{20}$ adaptation

With the introduction of the M-CHAT-R by Diana Robins and colleagues, and subsequent validation of the instrument (Robins et al., 2014), 20 out of 23 items from the M-CHAT$_{23}$ constituted the revised version, as the revision found three items to perform below par.

The present study explores the optimization of a screening checklist by testing the efficacy of excluding the three least predictive items from the M-CHAT$_{23}$ (Robins et al., 2001) and by changing cut-off criteria in identifying children at risk for ASD. For clarity, we call this adapted measure the M-CHAT$_{20}$. It is important to note, for clarification, that the sample in this study was only administered the M-CHAT$_{23}$ at 18-months of age, and not the M-CHAT-R/F. Only the cut-offs and algorithm of the M-CHAT-R/F were applied to the original M-CHAT$_{23}$, similar to procedures employed by Guthrie et al. (2019), to generate M-CHAT$_{20}$ scores.

For this 20-item M-CHAT$_{20}$, we used the cut-off criteria developed for the M-CHAT-R (Robins et al., 2014) as the cut-off criteria for screen positives were changed in the revision: a total score of item failures across the 20-items (Tot20) of 0–2 is regarded as low-risk (no actions necessary), a score of 3–7 is considered as medium-risk (needs further follow-up to ascertain more information on the "at-risk" responses), and a score of 8–20 is regarded as high-risk (skip follow-up sequence and directly refer the child for a developmental and diagnostic assessment to determine if the child has ASD). In the present study, a score of 3 and above is regarded as "at-risk." It is important to note that children who failed two or fewer items would not have received a follow-up on either algorithm—even though some of these children would go on to receive a diagnosis of ASD. Also important to note is that, in the present study, neither children screening medium-risk nor at-risk received a follow-up interview as was implemented in the "F" portion of the M-CHAT-R/F.

### Statistical analyses

To examine if the M-CHAT$_{20}$ reduced false positives and false negatives compared to the original M-CHAT$_{23}$, 2 × 2 crosstab tables for each outcome group (ASD or non-ASD) comparing M-CHAT$_{20}$ versus M-CHAT$_{23}$ criteria screening results were assembled (Table 1). These tables were used to (1) calculate sensitivity (SE), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV; Table 2), and (2) identify significant differences in identification between criteria using McNemar nonparametric tests.

# RESULTS

In total 54,463 individuals included responded on the questionnaire at 18-months of age and returned to the Norwegian Institute at a mean age of 19-months of age (M = 19.02, *SD* = 1.21), 337 individuals were diagnosed with ASD later in childhood.

Value (NPV) with 95% confidence intervals for M-CHAT$_{23}$ and M-CHAT$_{20}$ algorithm and algorithm components.

## Performance of the M-CHAT$_{23}$ cut-off criteria

For children with an eventual outcome of ASD, 2 × 2 tables revealed that the M-CHAT$_{23}$ criteria (Crit6 or Tot23 cutoffs) correctly classified 105 (of 337) children (TP: true positives), and incorrectly classified 232 children as not ASD (FN: false negatives). For children with an eventual outcome of non-ASD, M-CHAT$_{23}$ *criteria* incorrectly classified 4048 (of 54,126) children as ASD though they did not develop ASD (FP: false positives), and correctly identified 50,078 children as non-ASD. This yielded a sensitivity of 31.16% (CI$^{95\%}$ 26.25%–36.40%), specificity of 92.52% (CI$^{95\%}$ 92.30%–92.74%), PPV of 4.43% (CI$^{95\%}$ 3.79%–5.16%), and NPV of 99.18% (CI$^{95\%}$ 99.12%–99.24%).

**TABLE 1** Comparison of original M-CHAT 23-item screener versus adapted M-CHAT 20-item analogue of M-CHAT-R by ASD/non-ASD prediction performance

| **Children with an ASD Outcome** | | | |
|---|---|---|---|
| **Criterion** | **M-CHAT$_{20}$-** | **M-CHAT$_{20}$+** | **Row N** |
| M-CHAT$_{23}$− | 232 | 0 | FN$_{23}$:232 |
| M-CHAT$_{23+}$ | 12 | 93 | TP$_{23}$:105 |
| Column N | FN$_{20}$:244 | TP$_{20}$:93 | N$_{ASD}$:337 |
| **Children with a non-ASD Outcome** | | | |
| **Criterion** | **M-CHAT$_{20}$-** | **M-CHAT$_{20}$+** | **Row N** |
| M-CHAT$_{23}$- | 50078 | 0 | TN$_{23}$:50078 |
| M-CHAT$_{23+}$ | 1291 | 2757 | FP$_{23}$:4048 |
| Column N | TN$_{20}$:51369 | FP$_{20}$:2757 | N$_{non-ASD}$:54126 |

*Note*: M-CHAT$_{23}$+/−: screen positive/screen negative on original M-CHAT criteria (Failed Crit6 *or* Tot23); M-CHAT$_{20}$+/−: screen positive/screen negative on 20-item reduced M-CHAT in line with scoring changes made in the development of M-CHAT-R component of M-CHAT-R/F (Failed Tot20); TN$_{xx}$ = true negative (correctly identified as non-ASD); TP$_{xx}$ = true positive (correctly identified as ASD); FN$_{xx}$ = false negative (incorrectly identified as non-ASD when actually ASD); FP$_{xx}$ = false positive (incorrectly identified as ASD when actually non-ASD); xx = corresponding to 20/23 in M-CHAT$_{20}$ or M-CHAT$_{23}$.

## Performance of the M-CHAT$_{20}$ cut-off criteria

For children with an eventual outcome of ASD, 2 × 2 tables revealed that the M-CHAT$_{20}$ criterion (Tot20 cut-off) led to 93 (of 337) TPs, 244 FNs, 2757 FPs, and 51,369 TNs. This yielded a sensitivity of 27.60% (CI$^{95\%}$ 22.89%–32.70%), specificity of 94.91% (CI$^{95\%}$ 94.72%–95.09%), PPV of 5.68% (CI$^{95\%}$ 4.81%–6.71%), and NPV of 99.16% (CI$^{95\%}$ 99.10%–99.21%).

## Comparison of the M-CHAT$_{23}$ original algorithm compared to M-CHAT$_{20}$

### ASD group

In total, 337 individuals were later diagnosed with ASD. Using M-CHAT$_{20}$ criteria decreased the number of correctly identified ASD children by 12 compared to M-CHAT$_{23}$, with no children with ASD screening negative on the M-CHAT$_{23}$ criteria subsequently screening positive on the M-CHAT$_{20}$ criterion. A nonparametric McNemar's test determined a statistically significant difference in identifying ASD by M-CHAT$_{23}$ compared to the M-CHAT$_{20}$ ($p$ = 0.0015), suggesting use of the M-CHAT$_{20}$ criterion increased false negatives.

### Non-ASD group

In total, 54,126 individuals did not later receive a diagnosis of ASD. Using M-CHAT$_{20}$ criterion decreased the number of false positive children by 1291 compared to M-CHAT$_{23}$, with no non-ASD children scoring below cut-off on the M-CHAT$_{23}$ criteria subsequently scoring above cut-off on the M-CHAT$_{20}$ criterion. A nonparametric McNemar's test determined a statistically

**TABLE 2** Sensitivity and specificity

| | **Sensitivity** | **Specificity** | **PPV** | **NPV** |
|---|---|---|---|---|
| M-CHAT$_{23}$ (Crit6 or Tot23) | 31.16% [26.25%,36.40%] | 92.52% [92.30%,92.74%] | 4.43% [3.79%,5.16%] | 99.18% [99.12%,99.24%] |
| M-CHAT$_{20}$ (Tot20) | 27.60% [22.89%,32.70%] | 94.91% [94.72%,95.09%] | 5.68% [4.81%,6.71%] | 99.16% [99.10%,99.21%] |

significant difference in non-ASD identification by M-CHAT$_{23}$ compared to M-CHAT$_{20}$ ($p < 0.0001$), suggesting use of the M-CHAT$_{20}$ criterion decreased false positives.

## DISCUSSION

The present study evaluated the impact of scoring algorithm changes from the M-CHAT$_{23}$ to the M-CHAT-R based on application of M-CHAT$_{23}$ and M-CHAT-R scoring criteria to a large sample of individuals with long-term developmental follow-up for ASD. Findings indicated that moving to M-CHAT-R scoring criteria (i.e., the M-CHAT$_{20}$ version) decreased false positives by 2.4% (1291/54126 children with no ASD diagnosis) at the cost of 3.6% increased false negatives (12/337 children with ASD). This tradeoff was in line with those findings observed from the validation study of the M-CHAT-R as compared to the original M-CHAT$_{23}$ (Robins et al., 2014).

These relatively small changes should be considered in the context of the performance of the M-CHAT$_{23}$, with or without M-CHAT-R scoring algorithms applied. As seen in Guthrie et al., 2019, Stenberg et al., 2014, and Øien et al., 2018, high numbers of false negatives and false positives were present, with most children with ASD (at least 68.8%, 232/337) screening negative and relatively few children without ASD (at most 7.5%, 4048/54126) screening positive. While screening positive in non-ASD children may have triggered undue alarm in families, for false negatives, these children with ultimate diagnoses of ASD would have not received any follow-up based on M-CHAT$_{20}$ criterion for children scoring between 0 and 2. This is in line with previous studies that have reported that most children with a later diagnosis of ASD who were screened at 18-months of age in prospective general population cohorts are not identified at 18-months (Guthrie et al., 2019; Øien et al., 2018; Stenberg et al., 2014; Yuen, Carter, et al., 2018).

Some of the M-CHAT-R's enhanced efficacy in identifying children at risk can be achieved by selecting the most efficient items from a checklist and deleting those with poor performance. To reiterate, our findings suggest that the M-CHAT$_{20}$ increases the false-negative rate while reducing the false positive rate, that is, improving specificity, but reducing sensitivity. However, neither the M-CHAT$_{23}$ nor M-CHAT$_{20}$ performs adequately in identifying children with a prospective diagnosis of ASD (high number of false negatives), as revealed in previous studies (Guthrie et al., 2019; Øien et al., 2018; Stenberg et al., 2014; Yuen, Carter, et al., 2018; Yuen, Penner, et al., 2018). The original CHAT showed similar difficulties in identifying ASD in a general population (Baird et al., 2001). It is important to note that the suboptimal performance and systematic identification of the more severe children (Stenberg et al., 2021) are not exclusive to the M-CHAT(-R) at 18 and 24 months, but seem present utilizing other screening instruments such as the social communication questionnaire (SCQ) at 36 months (Surén et al., 2019).

As shown in Stenberg et al. (2021), many children deemed as "at-risk" for ASD at 18-months were later diagnosed with other developmental disabilities, indicating that a change in criterion might reduce the identification of other developmental disabilities (Stenberg et al., 2021). Thus, a trade-off of increasing the specificity and decreasing the sensitivity might ultimately lead to fewer children being identified who go on to develop ASD as well as missing out on children with other developmental disabilities with valid needs of early identification. Reducing the sensitivity might definitely increase the age of diagnosis and access to early intervention. However, the authors want to acknowledge that there are advantages of using screening instruments in primary care to familiarize themselves with symptom patterns, and the instruments serve a purpose in that it identifies some children at 18 months of age. It might increase the knowledge in pediatric and well-visit clinics on early signs and symptoms of ASD. It is also established that these instruments work well when there is a parental concern. Screening instruments can help specify the difficulties that children have at a given time in their developmental course. In particular, efforts to identify more false negatives should be of most pressing concern in the research field on early identification. In this context, it is crucial to systematically assess behavioral differences in children at well-baby clinics using different developmental instruments, and, additionally, to use caution when interpreting both positive and a negative screens, because of the high number of false negatives. Due to the fact that symptoms might not be evident at 18 months, we might ask ourselves if we are asking the wrong questions or using inappropriate measures. In addition to developmental surveillance, sensitivity to parental concern, and using sound clinical judgment, it might be necessary to revisit constructs of ASD at different timepoints to improve early identification of the disorder. Thinking about screening instruments as "one measure to rule them all" may be utopian as various instruments serve different roles in identifying children with ASD.

When considering updates to screening measures, it may be critical to ask what is being changed and how does that change the weight of the diagnostic process. As recent studies have found (Guthrie et al., 2019; Øien et al., 2018; Stenberg et al., 2014, 2021; Sturner, Howard, Bergmann, Morrel, et al., 2017; Sturner, Howard, Bergmann, Stewart, & Afarian, 2017) screening for ASD is not as straightforward as would be implied by the original instrument publications and associated validation studies. In particular, replication of results from validations studies in longitudinal and prospective studies are necessary to understand mechanisms of symptom patterns and later outcomes for both false positives and false negatives. Reduction of false positives are important if the aim of the screening process is to identify only cases of ASD, however it might be debated that detecting other developmental disabilities are of equal importance.

## CONCLUSION

The results suggest that the performance of the original 23-item M-CHAT (labeled as the M-$\text{CHAT}_{23}$ in this work), after removing three items and using M-CHAT-R scoring criteria (labeled as the M-$\text{CHAT}_{20}$ in this work), has less sensitivity than the original M-$\text{CHAT}_{23}$ using all items with its original scoring method. Similar to expectations generated by original M-CHAT-R/F validation studies, however, we also found increased specificity for identifying children with ASD when using M-CHAT-R scoring on the M-$\text{CHAT}_{23}$. Still, a more extensive investigation into different pathways to diagnosis is needed to tailor more dynamic instruments to identify sets of markers for children who concurrent screening instruments miss at 18 months of age. One option is to compare two algorithms in epidemiological-type samples to study the trade-offs, while changing the number of items, and to use complementary analyses, such as moving cut-off points. The main advantage of doing this on epidemiological-type samples is the possibility to study the trade-offs and to optimize the performance.

To conclude; it is important that clinicians exhibit caution when interpreting the status of a screening, both in terms of a positive or negative result. In terms of interpreting a positive result, caution is of great importance as the PPV is universally suboptimal. In terms of a negative result, caution needs to be exhibited as more than 2/3 of children with a later diagnosis of ASD will not be identified by any extant or prior criteria or cut-off—not even being flagged as moderate-enough risk to receive a follow-up interview. These limitations may result from multiple factors. The M-CHAT R/F algorithm may improve the false positive issue; however, the false negative issue is still to be tackled, as it persists utilizing the new algorithm. This could be a result of symptoms not being evident or prototypical at 18 months of age, and thus it might be that these individuals would not meet the criteria for an ASD diagnosis at this age utilizing gold standard instruments either. Indicating that they might not have ASD at an immediate evaluation but meet the criteria for ASD later in the developmental period. As highlighted in Øien et al., 2018 (utilizing the MoBa), children screening negative while receiving a diagnosis later had atypicalities in development that did not compare to true negatives even if they had similar screening status at 18 months of age, which might indicate subtler developmental issues. This highlights the need for developmental surveillance, as it may also be the case that we are asking the wrong questions or performing the wrong tests. Thus, indicating that children should be followed up in terms of development at different timepoints regardless of screening status early in the developmental period. There is a clear need for continued improvement in this domain—but it is similarly essential to consider what is actually being changed and how those changes impact the weight of the diagnostic process.

## Limitations

As the MoBa did not include the M-CHAT-R item wording and item sequence, there is, of course, a limitation. More specifically, it is not possible to know if the rewording or resequencing of the items and the additional examples that are added to each item in the M-CHAT-R would affect the results in a positive direction. As noted, very few, if any, studies have abilities to conduct such analyses on the same set of children with ASD. The items still preserve the same phenomenology, even without the exemplification, and it seems like most children with a future diagnosis of ASD would still score below the cut-off for follow-up or "at-risk" status on the revised version compared to the original version of the instrument. Furthermore, we did not conduct the follow-up of individuals screening positive, and thus the reduction of false positives are solely based on the algorithmic change of the M-$\text{CHAT}_{23}$. We neither had full access to the MoBa study nor outcomes associated with developmental disabilities other than ASD, so providing information on how many of the false positives that went on to receive other diagnoses with the current dataset are not possible. However, we have previously reported information from the Autism Birth Cohort (ABC) study (Stenberg et al., 2021), and we have added the information from the sub-study ABC ($N = 1033$) to the Appendix of this article to highlight the large number of false positives that received other diagnoses at assessment at 42 months. This provides additional clarity on outcomes likely associated with the false positive group. Future studies aims at utilizing the NPR to show how many of the MoBa children that received other diagnoses however this data was not available to the authors at this time.

ACKNOWLEDGMENTS

The authors would like to acknowledge and thank all participating families and their children.

## ETHICS STATEMENT

The Autism Birth Cohort Study is funded by the National Institute of Neurological Disorders and Stroke (NIH/NINDS), Bethesda, MD, USA (Grant No. NS47537). The Norwegian Mother and Child Cohort Study is funded by the Norwegian Ministry of Health and Care Services, the Norwegian Ministry of Education and Research, the Research Council of Norway/FUGE (Grant No. 151918), the National Institute of Neurological Disorders and Stroke

(NIH/NINDS), Bethesda, MD, USA (Grant No. NS47537), and the National Institute of Environmental Health Sciences (NIH/NIEHS), Research Triangle Park, NC, USA (Contract No. NO-ES-75558). The present study was supported by a grant from the Research Council of Norway (Grant No. 196452) The MoBa study was approved by the Regional Committee for Medical and Health Research Ethics South East. The authors report no conflict of interest.

## AUTHOR CONTRIBUTIONS

Synnve Schjølberg, Frederick Shic, and Roald A. Øien conceptualized and designed the study, drafted the initial manuscript, and reviewed and revised the manuscript. Fred R. Volkmar, Tonje Torske, Kenneth Larsen, Nina Stenberg, James F. Leckman, Denis G. Sukhodolsky, Anders Nordahl-Hansen, Katherine Riley, and Katarzyna Chawarska critically reviewed and revised the manuscript for important intellectual content. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

## ORCID

*Frederick Shic* https://orcid.org/0000-0002-9040-1259
*Anders Nordahl-Hansen* https://orcid.org/0000-0002-6411-3122
*Denis G. Sukhodolsky* https://orcid.org/0000-0002-5401-792X
*Katarzyna Chawarska* https://orcid.org/0000-0002-1445-6346
*Roald A. Øien* https://orcid.org/0000-0003-3698-2184

## REFERENCES

Baird, G., Charman, T., Baron-Cohen, S., Cox, A., Swettenham, J., Wheelwright, S., & Drew, A. (2000). A screening instrument for autism at 18 months of age: A 6-year follow-up study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(6), 694–702. https://doi.org/10.1097/00004583-200006000-00007

Baird, G., Charman, T., Cox, A., Baron-Cohen, S., Swettenham, J., Wheelwright, S., & Drew, A. (2001). Screening and surveillance for autism and pervasive developmental disorders. *Archives of Disease in Childhood*, *84*(6), 468–475. https://doi.org/10.1136/adc.84.6.468

Baird, G., Douglas, H. R., & Murphy, M. S. (2011). Recognising and diagnosing autism in children and young people: summary of NICE guidance. *BMJ*, *343*(1), d6360. https://doi.org/10.1136/bmj.d6360

Baron-Cohen, S., Allen, J., & Gillberg, C. (1992). Can autism be detected at 18 months?: The needle, the haystack, and the CHAT. *British Journal of Psychiatry*, *161*(6), 839–843. https://doi.org/10.1192/bjp.161.6.839

Baron-Cohen, S., Wheelwright, S., Cox, A., Baird, G., Charman, T., Swettenham, J., Drew, A., & Doehring, P. (2000). Early identification of autism by the CHecklist for Autism in Toddlers (CHAT). *Journal of the Royal Society of Medicine*, *93*(10), 521–525. https://doi.org/10.1177/014107680009301007

Carbone, P. S., Campbell, K., Wilkes, J., Stoddard, G. J., Huynh, K., Young, P. C., & Gabrielsen, T. P. (2020). Primary care autism screening and later autism diagnosis. *Pediatrics*, *146*(2), e20192314. https://doi.org/10.1542/peds.2019-2314

Chawarska, K., Klin, A., Paul, R., & Volkmar, F. (2007). Autism spectrum disorder in the second year: Stability and change in syndrome expression. *Journal of Child Psychology and Psychiatry*, *48*(2), 128–138. https://doi.org/10.1111/j.1469-7610.2006.01685.x

Guthrie, W., Wallis, K., Bennett, A., Brooks, E., Dudley, J., Gerdes, M., Pandey, J., Levy, S. E., Schultz, R. T., & Miller, J. S. (2019). Accuracy of autism screening in a large pediatric network. *Pediatrics*, *144*(4), e20183963. https://doi.org/10.1542/peds.2018-3963

Hickey, E., Sheldrick, R. C., Kuhn, J., & Broder-Fingert, S. (2021). A commentary on interpreting the United States preventive services task force autism screening recommendation statement. *Autism*, *25*(2), 588–592. https://doi.org/10.1177/1362361320957463

Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A. K., Handal, M., Haugen, M., Høiseth, G., Knudsen, G. P., Paltiel, L., Schreuder, P., Tambs, K., Vold, L., & Stoltenberg, C. (2016). Cohort profile update: The Norwegian Mother and Child Cohort Study (MoBa). *International Journal of Epidemiology*, *45*(2), 382–388. https://doi.org/10.1093/ije/dyw029

McPheeters, M. L., Weitlauf, A., Vehorn, A., Taylor, C., Sathe, N. A., Krishnaswami, S., Fonnesbeck, C., & Warren, Z. E. (2016). Screening for autism spectrum disorder in young children. https://www.ncbi.nlm.nih.gov/books/NBK349703/

Øien, R. A., Cicchetti, D. V., Nordahl-Hansen, A., & Schjølberg, S. (2019). A commentary to "toddler screening for autism spectrum disorder: A meta-analysis of diagnostic accuracy". *Journal of Autism and Developmental Disorders*, *1–2*, 3440–3441. https://doi.org/10.1007/s10803-019-04226-3

Øien, R. A., Schjølberg, S., Volkmar, F. R., Shic, F., Cicchetti, D. V., Nordahl-Hansen, A., Stenberg, N., Hornig, M., Havdahl, A., Øyen, A. S., Ventola, P., Susser, E. S., Eisemann, M. R., & Chawarska, K. (2018). Clinical features of children with autism who passed 18-month screening. *Pediatrics*, *141*(6), e20173596. https://doi.org/10.1542/peds.2017-3596

Ozonoff, S., Iosif, A. M., Baguio, F., Cook, I. C., Hill, M. M., Hutman, T., Rogers, S. J., Rozga, A., Sangha, S., Sigman, M., Steinfeld, M. B., & Young, G. S. (2010). A prospective study of the emergence of early behavioral signs of autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*(3), 256–266.e2. https://doi.org/10.1016/j.jaac.2009.11.009

Ozonoff, S., Young, G. S., Landa, R. J., Brian, J., Bryson, S., Charman, T., Chawarska, K., Macari, S. L., Messinger, D., Stone, W. L., Zwaigenbaum, L., & Iosif, A. M. (2015). Diagnostic stability in young children at risk for autism spectrum disorder: a baby siblings research consortium study. *Journal of Child Psychology and Psychiatry*, *56*(9), 988–998. https://doi.org/10.1111/jcpp.12421

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*(1), 37–45. https://doi.org/10.1542/peds.2013-1813

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. https://doi.org/10.1023/a:1010738829569

Siu, B.-D., Grossman, D. C., Baumann, L. C., Davidson, K. W., Ebell, M., García, F. A. R., Gillman, M., Herzstein, J., Kemper, A. R., Krist, A. H., Kurth, A. E., Owens, D. K., Phillips, W. R., Phipps, M. G., & Pignone, M. P. (2016). Screening for autism spectrum disorder in young children: US Preventive Services Task Force recommendation statement. *JAMA*, *315*(7), 691–696. https://doi.org/10.1001/jama.2016.0018

Skjærven, R., Stoltenberg, C., & The Moba Study Group. (2006). Cohort profile: the Norwegian mother and child cohort study (MoBa). http://ije.oxfordjournals.org/content/35/5/1146.short

Stenberg, N., Bresnahan, M., Gunnes, N., Hirtz, D., Hornig, M., Lie, K. K., Lipkin, W. I., Lord, C., Magnus, P., Kjennerud, T. R., Schjølberg, S., Surén, P., Susser, E., Svendsen, B. K., Tetzchner, S., Øyen, A. S., & Stoltenberg, C. (2014). Identifying children with autism spectrum disorder at 18 months in a general population sample. *Paediatric and Perinatal Epidemiology*, *28*(3), 255–262. https://doi.org/10.1111/ppe.12114

Stenberg, N., Schjølberg, S., Shic, F., Volkmar, F. R., Øyen, A.-S., Bresnahan, M., Svendsen, B. K., Tetzchner, S. V., Thronæs, N. T., Macari, S., Cicchetti, D. V., Chawarska, K., Surén, P., & Øien, R. A. (2021). Functional outcomes of children identified early in the developmental period as at risk for ASD. *Journal of Autism and Developmental Disorders*, *51*, 922–932.

Stewart, L. A., & Lee, L.-C. (2017). Screening for autism spectrum disorder in low-and middle-income countries: A systematic review. *Autism*, *21*(5), 527–539.

Stoltenberg, C., Schjølberg, S., Bresnahan, M., Hornig, M., Hirtz, D., Dahl, C., Lie, K. K., Reichborn-Kjennerud, T., Schreuder, P., Alsaker, E., Øyen, A.-S., Magnus, P., Surén, P., Susser, E., & Lipkin, W. I. (2010). The autism birth cohort: A paradigm for gene–environment–timing research. *Molecular Psychiatry*, *15*(7), 676–680. https://doi.org/10.1038/mp.2009.143

Sturner, R., Howard, B., Bergmann, P., Morrel, T., Landa, R., Walton, K., & Marks, D. (2017). Accurate autism screening at the 18-month well-child visit requires different strategies than at 24 months. *Journal of Autism and Developmental Disorders*, *47*(10), 3296–3310. https://doi.org/10.1007/s10803-017-3231-0

Sturner, R., Howard, B., Bergmann, P., Stewart, L., & Afarian, T. E. (2017). Comparison of autism screening in younger and older toddlers. *Journal of Autism and Developmental Disorders*, *47*(10), 3180–3188. https://doi.org/10.1007/s10803-017-3230-1

Surén, P., Saasen-Havdahl, A., Bresnahan, M., Hirtz, D., Hornig, M., Lord, C., Kjennerud, T. R., Schjølberg, S., Øyen, A. S., Magnus, P., Susser, E., Lipkin, W. I., & Stoltenberg, C. (2019). Sensitivity and specificity of early screening for autism. *BJPsych Open*, *5*(3), e41. https://doi.org/10.1192/bjo.2019.34

Wallis, K. E., Guthrie, W., Bennett, A. E., Gerdes, M., Levy, S. E., Mandell, D. S., & Miller, J. S. (2020). Adherence to screening and referral guidelines for autism spectrum disorder in toddlers in pediatric primary care. *PLoS One*, *15*(5), e0232335. https://doi.org/10.1371/journal.pone.0232335

Yuen, T., Carter, M. T., Szatmari, P., & Ungar, W. J. (2018). Cost-effectiveness of universal or high-risk screening compared to surveillance monitoring in autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *48*(9), 2968–2979.

Yuen, T., Penner, M., Carter, M. T., Szatmari, P., & Ungar, W. J. (2018). Assessing the accuracy of the Modified Checklist for Autism in Toddlers: A systematic review and meta-analysis. *Developmental Medicine & Child Neurology*, *60*(11), 1093–1100. https://doi.org/10.1111/dmcn.13964

Zwaigenbaum, L., Bauman, M. L., Stone, W. L., Yirmiya, N., Estes, A., Hansen, R. L., McPartland, J. C., Natowicz, M. R., Choueiri, R., Fein, D., Kasari, C., Pierce, K., Buie, T., Carter, A., Davis, P. A., Granpeesheh, D., Mailloux, Z., Newschaffer, C., Robins, D., … Wetherby, A. (2015). Early identification of autism spectrum disorder: Recommendations for practice and research. *Pediatrics*, *136*, S10–S40. https://doi.org/10.1542/peds.2014-3667c

**How to cite this article:** Schjølberg, S., Shic, F., Volkmar, F. R., Nordahl-Hansen, A., Stenberg, N., Torske, T., Larsen, K., Riley, K., Sukhodolsky, D. G., Leckman, J. F., Chawarska, K., & Øien, R. A. (2021). What are we optimizing for in autism screening? Examination of algorithmic changes in the M-CHAT. *Autism Research*, 1–9. https://doi.org/10.1002/aur.2643

# APPENDIX A: AUTISM BIRTH COHORT STUDY—OVERVIEW OF DIAGNOSTIC OUTCOME VERSUS SCREENING STATUS

| | M-CHAT23 | | M-CHAT20 | |
|---|---|---|---|---|
| | ScreenNeg | ScreenPos | ScreenNeg | ScreenPos |
| Assessed, no Dx, or subthreshold | 195 | 12 | 200 | 8 |
| Autistic disorder | 23 | 17 | 26 | 13 |
| Profound disability with autism | 0 | 1 | 0 | 1 |
| PDD NOS | 22 | 5 | 23 | 4 |
| Asperger syndrome | 7 | 0 | 7 | 0 |
| Childhood disintegrative disorder | 1 | 0 | 1 | 0 |
| Intellectual disability | 1 | 20 | 2 | 17 |
| Language disorder | 101 | 36 | 108 | 29 |
| Other psychiatric/neurodevelopmental disorder (NDD) | 30 | 8 | 31 | 7 |
| Subthreshold autistic disorder | 0 | 1 | 0 | 1 |
| Subthreshold PDD NOS | 16 | 14 | 16 | 14 |
| Subthreshold Asperger syndrome | 2 | 0 | 2 | 0 |
| Subthreshold Language disorder | 20 | 0 | 20 | 0 |
| Subthreshold other psychiatric/NDD | 21 | 8 | 21 | 8 |
| Rett syndrome | 0 | 1 | | |