J Autism Dev Disord (2008) 38:827–839
DOI 10.1007/s10803-007-0450-9

ORIGINAL PAPER

The Modified Checklist for Autism in Toddlers: A Follow-up Study Investigating the Early Detection of Autism Spectrum Disorders

Jamie M. Kleinman · Diana L. Robins · Pamela E. Ventola · Juhi Pandey · Hilary C. Boorstein · Emma L. Esser · Leandra B. Wilson · Michael A. Rosenthal · Saasha Sutera · Alyssa D. Verbalis · Marianne Barton · Sarah Hodgson · James Green · Thyde Dumont-Mathieu · Fred Volkmar · Katarzyna Chawarska · Ami Klin · Deborah Fein

Published online: 20 September 2007

Abstract Autism spectrum disorders (ASD) often go undetected in toddlers. The Modified Checklist for Autism in Toddlers (M-CHAT) was used to screen 3,793 children aged 16–30 months from low- and high-risk sources; screen positive cases were diagnostically evaluated. Rescreening was performed on 1,416 children aged 42–54 months. Time1 Positive Predictive Value (PPV) was .36 for the initial screening and .74 for the screening plus follow-up telephone interview; values were similar for Time2 PPV. When separating referral sources, PPV was low for the low-risk sample but acceptable with the follow-up telephone interview. Children with ASD from the low-risk and high-risk samples were highly similar. Results indicate that the M-CHAT continues to be a promising instrument for the early detection of ASD.

Keywords Autism · Early identification · Pediatric screening

J. M. Kleinman (✉) · J. Pandey · H. C. Boorstein · E. L. Esser · L. B. Wilson · M. A. Rosenthal · S. Sutera · A. D. Verbalis · M. Barton · S. Hodgson · J. Green · T. Dumont-Mathieu · D. Fein
Department of Psychology, University of Connecticut, 406 Babbidge Rd., Storrs, CT 06269-1020, USA
e-mail: jamie_kleinman@yahoo.com

D. L. Robins
Georgia State University, Atlanta, GA, USA

P. E. Ventola · F. Volkmar · K. Chawarska · A. Klin
Yale University School of Medicine, New Haven, CT, USA

Introduction

Autism spectrum disorders (ASD) constitute a group of severe disorders of development, disrupting social relationships, communication, play, academic skills, and usually leading to life-long disability. ASD affects up to 60 children in 10,000 (Baird et al. 2000; Bertrand et al. 2001; Chakrabati and Fombonne 2001; Charman 2002; Fombonne 2003; Fombonne et al. 2006) or even more (Baird et al. 2006).

Autism can be difficult to detect in very young children, who are often referred for evaluation later than would be optimal. The average age at which parents first report concerns is generally reported to be around 17–18 months (and most recent data has first parent concerns at an average of 14–15 months, with a significant number below age 11 months, see Chawarska et al. 2007) but most children are not diagnosed until age 4 or even later, especially urban, low socio-economic status children (see review by Gray et al. 2006). Clear evidence exists, however, that early detection and subsequent early intervention can lead to substantially better prognosis, including improved language, social relationships, and adaptive functioning, as well as fewer maladaptive behaviors, which increases the chance of successful inclusion in public education (Eaves and Ho 2004; Harris and Handleman 2000). To facilitate early diagnosis, standardized early screening, in addition to ongoing developmental surveillance, is essential (AAP 2006).

Autism-Specific Screening

Developmental surveillance is a continuous process undertaken by pediatric providers whereby professional observations are incorporated into decision-making about a child's developmental needs (Glascoe and Dworkin 1995; Glascoe 2005). In the context of ongoing pediatric surveillance, however, the use of standardized screening instruments can increase the accuracy of detection of developmental disorders; the use of more informal, non-validated strategies leads to an unacceptably low sensitivity of 20–30% (Earls and Hay 2006; Sand et al. 2005).

An important question is whether to conduct autism-specific screening for the entire population (low-risk screening), or only after broad developmental surveillance or screening with general instruments such as the Parents Evaluation of Developmental Status (PEDS; Glascoe 2001), which could be considered high-risk screening. The American Academy of Pediatrics recently published guidelines (AAP 2006) endorsing autism-specific screening for *all* children at 18 months, but did not endorse any particular screeners, making the development of these instruments more pressing. It appears that most American pediatricians (82%) routinely screen for general developmental delays (although more than half used inadequately validated procedures), but only 8% reported screening for ASD (dosReis et al. 2006). Several autism-specific screening tools appropriate for young children have been published. However, further research on most instruments is still under way. Recent reviews can be found in Dumont-Mathieu and Fein (2005), Mawle and Griffiths (2006), Gray and colleagues (2006), and Robins and Dumont-Mathieu (2006).

Optimal Age for Autism Screening

One barrier to early screening, and an important factor in selecting the optimal age for screening, is doubt about the validity of early diagnosis. A growing body of literature indicates that a diagnosis of ASD is stable over time even when the diagnosis is made at age 2 (Charman et al. 2005; Cox et al. 1999; Eaves and Ho 2004; Gillberg et al. 1996; Kleinman et al. in press; Lord 1995; Lord et al. 2006; Moore and Goodson 2003; Stone et al. 1999; Sutera et al. 2007). While less is known about the stability of diagnosis made under the age of 2, the desire to detect autism as early as possible is driving ongoing research efforts to screen children before 2 years, and even in the first year of life.

Another issue in selecting the best age for autism screening is the age range for onset of the disorder. It is desirable to screen as early as can be done reliably, in order to maximize intervention opportunities. However, screening that is conducted too early may not be able to distinguish ASD from other forms of developmental delay, or even from typical development. Population screening at 14 months has been shown to yield many false positives (Dietz et al. 2006; Swinkels et al. 2006), suggesting that universal screening should occur after that age, possibly at an 18 month well child visit. Furthermore, approximately 30% of children with autism show a period of normal development followed by plateau or regression (Tuchman and Rapin 1997; Chawarska et al. 2007), and screening too early might miss some of these later onset children. A further factor to consider in determining the optimal screening age is the reluctance of many parents to participate in further screening or diagnostic evaluation when children are as young as 14–15 months (Dietz et al. 2006). Given the data on when an ASD diagnosis can reliably be made, the timing of most regressive cases, and the likelihood of parent (and physician) cooperation, a screening age range of 16–30 months was selected for the M-CHAT.

Current Instrument: M-CHAT

The M-CHAT (Robins et al. 1999; Robins et al. 2001) is a 23-item yes/no parent report checklist (see http://www2.gsu.edu/~wwwpsy/faculty/robins.htm or www.firstsigns.org for free download). It is an adaptation of the CHAT (Baron-Cohen et al. 1992, 1996) designed for the American healthcare system, eliminating the observation section and expanding parent report items. The format and the first nine items are from the CHAT, with the authors' permission, thus embedding the parent-report section of the CHAT. The M-CHAT does not require physician's observation of the child, although physicians may "flag" an M-CHAT when they suspect possible ASD, regardless of checklist responses. The format is simple, the reading level is approximately 6th grade, and no parent or physician training is required.

Eaves et al. (2006) examined the performance of the M-CHAT with a group of 84 children aged 24–48 months (mean age 37 months) referred for possible autism to a specialty clinic, of whom 64% were then diagnosed with ASD. The majority of the remaining children had more than one diagnosis, including intellectual delay and language disorder. Sensitivity was good: for the 2/6 critical item score sensitivity was 77% and for the 3/23 item score it was 92%. However, specificity was low (43% and 27% for the two scores). Some differences in sampling and procedure between Robins et al. (2001) and Eaves et al. (2006) may be partly responsible for the differences reported: (a) the Eaves et al. sample were severely affected, with a mean CARS score of 29 for the affected plus unaffected sample together, and 64% were found to have ASD, (b) Eaves et al. used as a cutoff 3 failed items out of the 19 autism-related items, whereas Robins et al. (2001) used 3 failed items out of any of the 23 items; (c) Robins et al. used a telephone follow-up to reduce false positives, and (d) the Robins et al. sample was aged 16–30 months whereas the Eaves, Wingert, and Ho sample was aged 24–48 months. In addition, Eaves et al. report the PPV of the M-CHAT for their sample to be .63–.68. The formula they use for calculating PPV (correct screening cases divided by total sample) is different from the formula used by Robins et al. and used here (true positives divided by all screen positive cases). When one uses that formula on their data, PPV is .70 and .69 for the two scoring methods, which is comparable to or better than the PPV reported by Robins et al. (2001), which was .36 before and .68 after the telephone interview for the whole screener, and .64 before and .79 after the telephone interview for the critical items.

Eaves and colleagues (2006) point out that sensitivity, which is good especially for the total score, may be more important than specificity, especially in initial, low-risk screening, but suggest that the M-CHAT may have insufficient specificity for identifying autism in high-risk samples already suspected of autism. However, Fine et al. (2005) used the M-CHAT successfully to screen children with 22q11.2 deletions for autism. Ventola et al. (2007) examined a group of children with ASD and other developmental disorders who had all screened positive on the M-CHAT. Eleven of the 23 items differentiated the ASD and non-ASD groups; after controlling for language level, four items remained different between the groups, all relating to joint attention.

The M-CHAT (and the CHAT) were translated into Chinese, and tested on a sample of 212 children with mental ages 18–24 months, about half of whom were diagnosed with ASD (Wong et al., 2004). The 7 most discriminating items were largely overlapping with, but not identical to, the 6 critical items on the M-CHAT identified by Robins et al. (2001). Using a cut-off of failing 2 of these 7 items produced a sensitivity of .93 and a specificity of .77, whereas failing any 6 of the 23 items produced a sensitivity of .84 and specificity of .85. The procedure recommended by the authors was initial screening with the parent report (M-CHAT) items followed by clinician observation for children screening positive, using the CHAT observation items. These results are very promising but it should be pointed out that this was not a low-risk screening. Mawle and Griffiths (2006) review available data and suggest that the M-CHAT has promising sensitivity for population screening but that additional follow-up data are needed on the initial sample.

Development of the M-CHAT and initial results on 1,293 children is described by Robins et al. (2001). Subsequent data by other groups (Wong et al. 2004; Eaves et al. 2006) provide support for the utility of the M-CHAT, but key results are still lacking. These include: replication of the initial results (including PPV and internal consistency) with a new sample, direct comparison of the high-risk versus low-risk children detected by the M-CHAT, and follow-up of the children to an age when diagnosis is more certain. The present paper reports additional data on the M-CHAT, to address the following questions:

Study 1: Replication Study. We report M-CHAT data on 3793 *new cases* in order to replicate: (a) positive predictive power from the checklist alone, and from the checklist plus the telephone interview, and (b) internal consistency reliability. The Robins et al. (2001) paper did not have a sufficient sample of children with ASD from high- versus low-risk sources to examine any differences between them. The present paper (c) examines key variables for the low-risk vs. high-risk sample to determine if the children with ASD detected by the M-CHAT from the general population differ from those from a high-risk referral sample; analyses examine differences on these developmental and diagnostic variables for low-risk ASD, low-risk non-ASD, high-risk ASD, and high-risk non-ASD. We also report positive predictive value (PPV) for the low and high-risk groups separately. Finally, (d) several procedures were used to identify children with ASD who might have been missed by the M-CHAT; we report on the characteristics and diagnoses of these potential misses.

Study 2: Follow-up study. Follow-up data have been obtained from 1416 children screened with the M-CHAT at 16–30 months and followed up around the age of 4 years, with two aims: (a) The present paper relates the child's M-CHAT scores at Time 1 (16–30 months) to a final diagnosis at Time 2 (4 years old), to estimate PPV based on a later diagnosis. (b) We also describe efforts to ascertain missed cases at age 4, to estimate the number of missed cases at age 2.

Study 1: Replication

Methods

Participants

Participants were a total of 3,793 new cases, drawn from one of two sources. The low-risk sample ($n = 3{,}309$) consisted of children screened during well-child care visits at their pediatrician's office. The high-risk sample ($n = 484$) was screened during intake with an early intervention service provider or because of referral from a developmental pediatrician or psychologist.

An additional 73 children were screened but are not included in the final sample of 3,793, because: (a) they had already received a diagnosis of an ASD or other disorder prior to screening, ($n = 2$), (b) they were older than 30 months or younger than 16 months when their caregiver filled out the screener ($n = 36$), (c) they had severe physical impairments that prevented the use of standardized evaluation instruments (e.g., blind, deaf, unable to sit independently; $n = 4$), or (d) they did not complete all of the components of the study they qualified for (telephone interview, evaluation), due to refusal or missing contact information ($n = 31$).

Table 1 presents sex distribution, age, and mean number of items failed after the M-CHAT and after the telephone interview for the low-risk, high-risk, and combined samples.

Materials

The *Modified Checklist for Autism in Toddlers* (M-CHAT; Robins et al. 2001) is a 23-item yes/no parent report screener for ASD. Screening positive (hereafter called "failing") on the screener is defined as failing any three items, or any two of six critical items (items 2, 7, 9, 13, 14, 15). Failed items are reviewed in a follow-up telephone interview; if the child still fails the screening, the family is offered a free developmental and diagnostic evaluation.

Diagnostic instruments include the *Autism Diagnostic Interview-Revised* (ADI-R; Lord et al. 1994), *Autism Diagnostic Observation Schedule* (ADOS; Lord et al. 1999), and the *Childhood Autism Rating Scale* (CARS; Schopler et al. 1988). This study used the ADI-R Toddler version, obtained from the instrument's author. This is an experimental edition of the widely used parent interview containing additional questions relating to early childhood behaviors and eliminating questions designed for use with older children, for example those relating to certain peer interactions. The algorithm items are the same as on the published ADI-R. The ADI-R and CARS classify children with autism or non-autism, and the ADOS also includes a classification for autism spectrum.

Clinical judgment by experienced clinicians is considered to be the "gold standard" for autism diagnosis (Volkmar et al. 2005). In the current study, the clinicians used the DSM-IV criteria for Autistic Disorder (APA 1994) for their clinical judgment to diagnose Autistic Disorder, PDD-NOS or as not on the autism spectrum. Children not on the autism spectrum could be given a diagnosis of language disorder, global developmental delay, or other condition. "No diagnosis" was given if the child was showing apparently typical development.

Other measures included the *Mullen Scales of Early Learning* (Mullen 1995), a standardized assessment of cognitive functioning and the *Vineland Adaptive Behavior Scales* (VABS; Sparrow et al. 1984), a widely used parent interview assessing adaptive functioning in the areas of communication, daily living, socialization, and motor skills.

Table 1 Characteristics of Study 1 replication sample

Mean (SD) Range	Low-risk n = 3309	High-risk n = 484	Total n = 3793	*F*
Male	1,649	354	2,003	N/A
Female	1,621	122	1,743	N/A
Sex not reported	39	8	47	N/A
Age in months at screening	20.53 (3.06)	24.25 (3.60)	21.01 (3.37)	576.56**
	16.00–30.85	16.23–30.62	16.00–30.85	
M-CHAT total items failed	.85 (1.35)	4.21 (4.98)	1.28 (2.44)	990.53**
	0–19	0–19	0–19	
M-CHAT critical items failed	.10 (.45)	1.44 (1.95)	.27 (.92)	1168.24**
	0–6	0–6	0–6	
Age in months at telephone interview*	22.71 (3.83)	25.63 (3.85)	24.09 (4.05)	53.88**
	17.51–36.01	16.62–34.07	16.62–36.01	
M-CHAT total items failed for screen positive cases*	4.50 (2.91)	8.84 (4.44)	6.74 (4.35)	130.21**
	2–19	2–19	2–19	
M-CHAT critical items failed for screen positive cases*	1.04 (1.33)	3.19 (1.83)	2.15 (1.93)	177.42**
	2–6	0–6	0–6	
Telephone interview total items failed*	1.51 (3.11)	5.26 (4.92)	3.41 (4.53)	82.12**
	0–19	0–18	0–19	
Telephone interview critical items failed*	.52 (1.21)	1.99 (1.95)	1.26 (1.79)	80.82**
	0–6	0–6	0–6	

* n for telephone interview is 189 low-risk, 196 high-risk, and 385 total

** $p < .01$

Procedure

A flow chart of how participants moved through Studies 1 and 2 is shown in Fig. 1.

Screening at Time 1. For the low-risk sample, physicians in Connecticut, southern Massachusetts, and Rhode Island were recruited to participate in the study. Physicians were recruited through mass mailings, direct contact between an investigator and a pediatrician, and through contacts by the Hezekiah Beardsley Connecticut chapter of the American Academy of Pediatrics. Physicians' offices ($n = 162$) participated in data collection by giving the M-CHAT to caregivers accompanying children to a well child visit between the ages of 16 and 30 months.

Although the M-CHAT was designed primarily to be used as a screening test with an unselected population, a high-risk group was added in order to obtain a larger sample of children diagnosed with an ASD. None of the children in the high-risk sample had received a diagnosis or had received more than minimal services (1–2 h per week) for a short time. The State of Connecticut's early intervention program, Birth-to-Three, invited offices statewide to participate, and several early intervention sites in Massachusetts also participated. Sixty-two early intervention offices participated in data collection. Several psychologists and developmental pediatricians also participated in the screening process, referring children for whom, as with the early intervention providers, there were general developmental concerns but no diagnosis had yet been made.

M-CHAT screening checklists for both groups were returned to the investigators for scoring. The telephone interview was administered to a caregiver of any child who failed the screening. The interview reviewed all failed items, following a script that asked for specific examples of behaviors and offered multiple examples against which to judge whether the child passed or failed the item. If the child failed the M-CHAT interview, the family was invited to bring the child for a free developmental/diagnostic evaluation.

In order to detect children with possible ASD who did not fail the M-CHAT, the screeners had a box labeled "for office use only" that provided the health care professional with the opportunity to indicate concern about possible ASD. When screeners were so flagged, the child was invited for an evaluation, regardless of the M-CHAT score. Any such flagged child who received an ASD diagnosis would be considered a missed case.

Evaluations at Time 1. A total of 203 children received a diagnostic evaluation (see Table 2) if they (a) failed the M-CHAT and follow-up telephone interview ($n = 185$), or (b) passed the M-CHAT but were flagged by the health care provider ($n = 6$) or failed another screening instrument under development (Yale Screener) ($n = 12$). Ethnicity data were available for 79% of the sample: 91% were Caucasian, 3% were African American, 2% were Asian, and 4% were Hispanic. The families had a mean maternal education of 13.2 years, a mean paternal education of 14.7 years, and an average household income of $40,000–$60,000.

Diagnostic evaluations took place at the child's home ($n = 1$), at the early intervention office ($n = 11$), at the University of Connecticut Psychological Services Clinic ($n = 174$), and at the Yale Child Study Center ($n = 17$). A

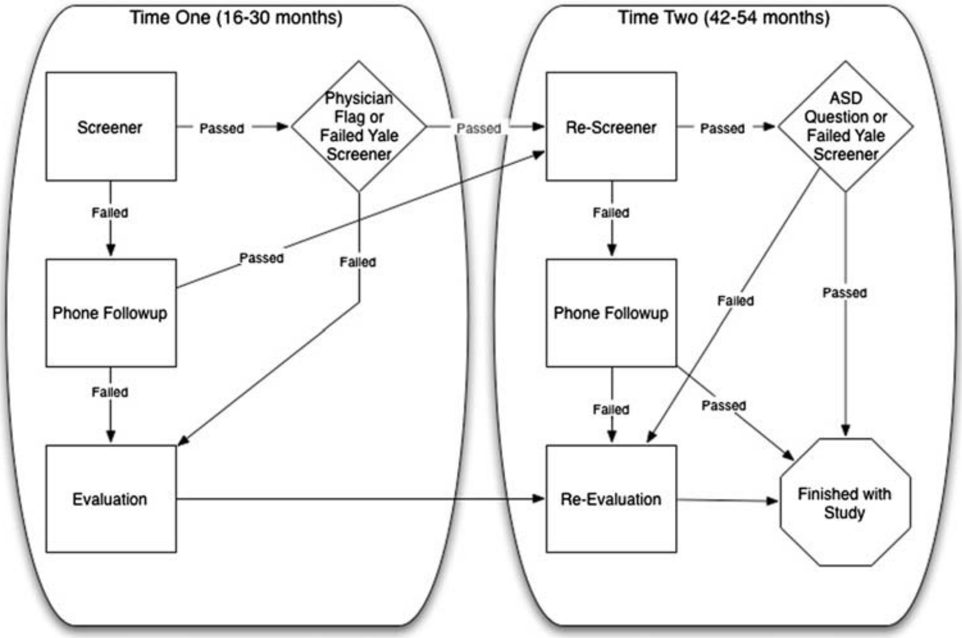

Fig. 1 Figure represents the process by which children moved through the study from Time 1 to Time 2

Table 2 Time 1 evaluated sample of children with failed M-CHAT ($n = 185$)

Mean (SD) Range	Low-risk non-ASD $n = 11$	High-risk non-ASD $n = 37$	Low-risk ASD $n = 20$	High-risk ASD $n = 117$	*F*
Demographic variables					
Age in months at evaluation	24.79 (3.45)	27.28 (4.00)	27.19 (4.84)	26.65 (4.58)	.952
	22.47–27.10	25.89–28.68	24.79–29.60	25.80–27.51	
Male	9	30	16	97	N/A
Female	2	7	4	20	N/A
Evaluation measures					
ADI-R total	13.73[a,b,c] (8.20)	21.73[e] (9.93)	25.00 (3.90)	27.90 (7.53)	12.46**
ADOS AB score	6.55[b,c] (4.40)	6.10[d,e] (3.85)	15.50 (4.12)	15.80 (4.28)	46.89**
CARS total	22.23[b,c] (2.75)	24.77[d,e] (4.60)	31.72 (5.55)	32.60 (4.69)	32.50**
DSM-IV # symptoms	1.27[a,b,c] (1.27)	3.06[d,e] (2.27)	5.61 (1.54)	6.18 (1.88)	37.88**
Mullen VR[1]	38.60[b,c] (14.92)	33.03 (12.42)	27.41 (11.27)	27.84 (9.22)	4.48**
Mullen FM	34.00 (10.59)	33.50 (15.12)	27.71 (9.81)	28.23 (10.21)	2.25
Mullen RL	32.00[b,c] (13.74)	28.22[e] (11.64)	22.53 (5.72)	21.83 (5.05)	8.76**
Mullen EL	32.70 (13.12)	31.14 (11.73)	23.19 (4.17)	24.02 (6.49)	8.20
Mullen ELC	86.67[a,b,c] (27.77)	69.75[d,e] (20.65)	56.31 (10.03)	58.01 (9.88)	14.80**
VABS[2] communication	74.36[b,c] (8.39)	71.48[d,e] (10.14)	64.33 (5.93)	65.22 (6.32)	10.40**
VABS daily living	74.18 (9.57)	70.88 (7.64)	68.83 (6.41)	68.49 (8.42)	2.15
VABS socialization	78.36[b,c] (10.81)	74.88[d,e] (9.15)	67.11 (7.14)	68.01 (7.94)	10.02**
VABS motor	77.45 (13.83)	82.78 (12.77)	82.00 (11.27)	82.56 (11.72)	.625
VABS ABC	70.70 (10.34)	76.28[e] (36.62)	65.33 (5.59)	65.36 (6.36)	3.33*

[a] Group difference between Low-Risk non-ASD and High-Risk non-ASD

[b] Group difference between Low-Risk non-ASD and Low-Risk ASD

[c] Group difference between Low-Risk non-ASD and High-Risk ASD

[d] Group difference between High-Risk non-ASD and Low-Risk ASD

[e] Group difference between High-Risk non-ASD and High-Risk ASD

* $p < .05$

** $p < .01$

[1] Mullen Scales of Early Learning; VR = Visual Reception, FM = Fine Motor, RL = Receptive Language, EL = Expressive Language, ELC = Early Learning Composite

[2] Vineland Adaptive Behavior Scales; ABC = Adaptive Behavior Composite

team of investigators performed the evaluations, including one licensed clinical psychologist or developmental pediatrician specializing in autism, one graduate student, and one research assistant videotaping the session. One of the team members collected history information and completed the caregiver interviews while the other member of the team evaluated the child. Blind assessment was not considered possible as all children presenting for an evaluation either had failed the M-CHAT, been flagged by their health care provider with possible autism concerns, or been flagged by another screener under development.

In addition to the instruments listed above, a full history of the child was also taken during an interview with the parents. Caregivers were provided with verbal feedback on the day of the evaluation and received a written report several weeks later, with treatment recommendations.

The screening and evaluation procedures were approved by the University of Connecticut and Yale University School of Medicine Institutional Review Boards.

Results

Replication of Positive Predictive Value (PPV) of Screening

The positive predictive value (PPV) for the initial screening was calculated as the proportion of children failing the M-CHAT who were diagnosed with an ASD. For the entire sample, 385 children failed the initial screening, and 137 of these children ultimately received a diagnosis of ASD, yielding a PPV of .36, with a 95% confidence interval of

.31 to .40. This compares to a value of .36 obtained by Robins et al. (2001) on the original, non-overlapping sample of 1,293 children. When examining results for the M-CHAT and telephone interview combined, 185 children failed both the screener and telephone follow-up and 137 of these were diagnosed with an ASD, yielding a PPV of 0.74, with a 95% confidence interval of .68 to .80. This compares to a value of .68 as obtained by Robins et al. (2001).

Replication of Internal Consistency

Cronbach's alpha was used to determine the internal consistency for the 23-item M-CHAT as well as the subset of six critical items for the current sample. Internal reliability was adequate for both the entire screener and for the six critical items, *alphas* = .85 and .84, respectively. This is consistent with the finding of Robins et al. (2001) who found the internal consistency for the entire screener and for the six critical items to be *alphas* of .85 and .83, respectively.

Comparison of Low-versus High-Risk Samples

General characteristics. Table 1 lists the mean, standard deviation, and range for age of the samples and M-CHAT scores, with and without the telephone interview. The high-risk sample is significantly older both at the time of their initial screening and telephone follow-up than the low-risk sample. As would be expected, they also fail significantly more total and critical items both on the screener and telephone follow-up than the low-risk group; this remains true when only those cases that qualified for the follow-up telephone interview were included (screen positive cases). This suggests that, as might be expected, the overall developmental risk was lower in the unselected group, whether considering the group as a whole, or only screen positive cases. Note that for both groups, the mean number of total and critical items failed is higher for the telephone interview than for the initial screening because only children who fail the screener go on to receive the telephone interview. However, the telephone interview scores are lower than the M-CHAT scores for the screen positive cases; since parents were only questioned about failed items, scores could only improve at the telephone interview.

Clinical characteristics. For the 31 children from the low-risk sample who were evaluated based on a failed M-CHAT and follow-up telephone interview, the diagnostic breakdown was as follows: Autistic Disorder *n* = 11; PDD-NOS, *n* = 9; language delay, *n* = 4; global developmental delay, *n* = 4; other diagnosis, *n* = 2; no diagnosis, *n* = 1. For the 154 children from the high-risk sample who were evaluated based on a failed M-CHAT and follow-up telephone interview, the diagnostic breakdown was as follows: Autistic Disorder *n* = 77; PDD-NOS, *n* = 40; language delay, *n* = 17; global developmental delay, *n* = 13; other, *n* = 6; no diagnosis, *n* = 1. Chi-square analyses indicated no group differences among the low- and high-risk groups on diagnosis, sex, ethnicity, or socioeconomic status.

Table 2 lists the means and standard deviations of the evaluation measures for children who were evaluated based on a failed M-CHAT and telephone interview. Children with a diagnosis of Autistic Disorder or PDD-NOS are included in the ASD group, and other diagnoses are considered non-ASD. Data are presented separately for low-risk ASD, low-risk non-ASD, high-risk ASD, and high-risk non-ASD, so that we could determine whether ASD cases detected from the general population differ from ASD cases from a referred sample. Analysis of variance (ANOVA) revealed differences among the 4 groups (high vs. low risk, ASD vs. non-ASD) on every variable except the Mullen fine motor and expressive language scores and the VABS daily living and motor scaled scores. Post hoc Tukey's test showed that *the low- and high-risk ASD groups did not differ from each other on any measure*. The two non-ASD groups differed significantly from each other on the ADI-R total score, DSM-IV number of symptoms endorsed, and Mullen ELC, with the high-risk group more affected or impaired on all measures. On all measures with the exception of the VABS motor scaled score, the ASD groups had more impaired functioning and more severe autistic symptomatology than the non-ASD groups.

Positive Predictive Value for Low- and High-Risk Samples. For the low-risk sample, 189 children failed the initial screening, of whom 20 ultimately were diagnosed with an ASD, yielding a PPV of .11, with a 95% confidence interval of .06 to .15. For the high-risk sample, 196 children failed the initial screening, of whom 117 ultimately were diagnosed with an ASD, yielding a PPV of .60, with a 95% confidence interval of .53 to .67. The PPV of the combined sample was .36, with a 95% confidence interval of .31 to .40. When examining results for the M-CHAT and telephone interview combined, in the low-risk sample, 31 children failed both the screener and telephone follow-up, of whom 20 were diagnosed with an ASD, yielding a PPV of .65, with a 95% confidence interval of .48 to .81. In the high-risk sample, 154 children failed both the screener and telephone follow-up, of whom 117 were diagnosed with an ASD, yielding a PPV of 0.76, with a 95% confidence interval of .69 to .83. The PPV for screener and follow-up for the combined sample was .74, with a 95% confidence interval of .68 to .80.

Characteristics of Children Representing Possible Misses

Of the 18 children evaluated because of referral from a pediatrician ($n = 6$) or failing another autism screener under development ($n = 12$), one missed case was identified. The remainder were diagnosed with language delay ($n = 10$), global developmental delay ($n = 3$), other diagnosis ($n = 1$), and no diagnosis ($n = 3$).

Study 2: Follow-up

Methods

Participants

To maximize the pool of participants available for analysis on the Time 2 testing, all participants eligible for re-screening or re-evaluation were included, including the new cases reported on in Study 1, and the earlier participants from the Robins et al. (2001) sample. Of the 2,469 children who were eligible for re-screening (see below for eligibility), data have been collected from 1,416 (57%; Table 3), of whom 1,160 were originally screened from low-risk sites and 256 from high-risk sites. Of the 1,053 children who were eligible for re-screening but on whom re-screening data were not collected, 1,043 did not return a re-screener after two attempts to contact them and 10 refused a telephone interview. Of the 161 children who were eligible for re-evaluation at Time 2, 120 (75% participation) have been re-evaluated (Table 4). Of those not evaluated, 20 (12%) refused to participate and 21 (12%) could not be reached or moved away. An additional 11 children who had not been evaluated at Time 1 were evaluated at Time 2 (see below).

Materials

The same measures were used in this study as in Study 1, except that for children who were re-evaluated after 60 months old ($n = 16$), the Differential Abilities Scale (DAS; Elliot 1990) replaced the Mullen. In addition, the Short Version of the ADI-R replaced the Toddler Version. The algorithm items used to assign diagnoses remained the same for all versions of the ADI-R.

The M-CHAT was used to re-screen children who had passed the screen or telephone follow-up at Time 1. This re-screening form had an additional question that asked if the child had ever been referred for possible ASD or another developmental disorder or diagnosed with such (following the procedure of Baird et al. 2000 to identify missed cases).

Table 3 Demographic information on Study 2 follow-up sample

	Low-risk n = 1,160	High-risk n = 256	Total n = 1,416
Male	606	194	800
Female	542	60	602
Sex not reported	12	2	14
Age in months at follow-up screening	59.12 (9.03) 47.97–88.28	54.89 (5.09) 48.03–64.56	58.32 (8.66) 47.97–88.28

Procedure

Screening at Time 2. Children who did not receive an evaluation at Time 1 because they passed the initial screening or the follow-up telephone interview were mailed a second M-CHAT approximately 2 years after the initial screening. If they did not return this mailing, an additional mailing was sent out within six months of the first mailing. All children who failed the re-screener received a follow-up telephone call to confirm their responses, as at Time 1.

Evaluation at Time 2. Children received a diagnostic evaluation at Time 2 if they (a) had received a diagnostic evaluation at Time 1 due to failed M-CHAT and telephone follow-up, or referral by the pediatrician or other screener under development, regardless of their diagnosis ($n = 120$), (b) failed the re-screener M-CHAT and telephone interview ($n = 8$), or (c) passed the re-screener but indicated they had been referred for possible ASD or developmental delay in the time between screenings ($n = 3$). Thus, the total number of children evaluated at Time 2 to date is 131. Diagnostic evaluations took place at the child's home ($n = 4$), at the early intervention office ($n = 1$), at the University of Connecticut Psychological Services Clinic ($n = 125$), and at the Yale Child Study Center ($n = 1$). At Time 2, children with a diagnosis of Autistic Disorder, PDD-NOS, or Asperger's Disorder were classified as ASD. Table 4 lists the demographic information and evaluation data for the ASD and non-ASD children. The children diagnosed with ASD did not differ in age or sex from the children not diagnosed with ASD, but the ASD group were more impaired on all developmental and diagnostic severity measures, except for DAS measures, probably because of low sample size.

Differential attrition was examined by comparing all variables for the Time 1 group who were re-evaluated at Time 2 ($n = 120$) to the Time 1 group who did not return for re-evaluation at Time 2 ($n = 41$). There were no significant differences (or trends) for age, sex, or any developmental or clinical variable. In examining attrition

Table 4 Time 2 evaluated sample

Mean (SD)	Non-ASD $n = 51$	ASD $n = 80$	F
Demographic variables			
Age in months at evaluation	55.87 (8.01)	52.17 (8.01)	N/A
	50.34–65.40	46.12–64.00	
Male	42	71	N/A
Female	9	9	N/A
Evaluation measures			
ADI-R total	10.67 (9.93)	19.84 (14.12)	12.92**
ADOS AB score	3.30 (3.67)	11.28 (5.70)	34.85**
CARS total	20.50 (4.49)	33.17 (5.39)	131.23**
DSM-IV # symptoms	1.52 (1.40)	6.78 (2.12)	134.97**
Mullen VR[a] ($n = 115$)	48.13 (17.13)	27.26 (13.57)	28.69**
Mullen FM	38.35 (15.74)	26.51 (10.08)	13.32**
Mullen RL	41.12 (14.45)	24.58 (7.15)	35.81**
Mullen EL	40.08 (12.05)	24.78 (7.78)	36.12**
Mullen ELC	87.55 (25.92)	58.28 (14.48)	31.89**
DAS NV[b] ($n = 16$)	89.33 (19.37)	79.38 (21.88)	3.17
DAS V	82.89 (15.24)	70.75 (18.86)	1.61
DAS GCA	85.22 (30.45)	74.00 (20.23)	.101
VABS[c] communication	85.88 (18.53)	63.00 (15.75)	48.50**
VABS daily living	77.90 (17.07)	57.75 (8.70)	47.21**
VABS socialization	83.11 (14.35)	62.40 (10.41)	66.19**
VABS motor	82.72 (19.52)	68.81 (15.85)	16.36**
VABS ABC	78.71 (16.68)	58.78 (10.66)	43.82**

** $p < .01$

[a] Mullen Scales of Early Learning; VR = Visual Reception, FM = Fine Motor, RL = Receptive Language, EL = Expressive Language, ELC = Early Learning Composite

[b] Differential Abilities Scale; V = Verbal, NV = Nonverbal, GCA = Global Composite of Abilities

[c] Vineland Adaptive Behavior Scales; ABC = Adaptive Behavior Composite

by diagnosis, 94 of the 120 Time 1 children with ASD returned (78%), whereas 26 of the 41 non-ASD children returned (63%), (Fisher's exact test ns). Ethnicity data were available for 135 of the 161 children eligible for re-evaluation (84%); 104 of 119 Caucasian children (87%) returned for re-evaluation, whereas 12 of 16 non-white children (75%) returned for re-evaluation (Fisher's exact test ns).

Results

For the 131 children who were evaluated at Time 2, the diagnostic breakdown is as follows: Autistic Disorder $n = 60$; PDD-NOS, $n = 18$; Asperger's Disorder, $n = 2$; language delay, $n = 13$; global developmental delay, $n = 12$; other, $n = 14$; no diagnosis, $n = 12$.

Positive Predictive Value (PPV) at Time 2

Based on the screening results at Time 1 and the evaluation outcome at Time 2 as described above, 76 of 201 children who failed the initial screening (without telephone interview) were diagnosed with ASD, for a PPV of .38, with a 95% confidence interval of .31 to .45. For the combined screening plus telephone interview, 73 of 124 children who failed the screening were diagnosed with an ASD, for a PPV of 0.59, with a 95% confidence interval of .50 to .68.

Time 2 Misses

A *miss* was defined as a child who passed the screener or telephone interview at Time 1, *and* was diagnosed with ASD at Time 2. Children representing possible missed cases who were old enough for Time 2 evaluations were ascertained from the following sources: (a) children flagged by their health care provider at Time 1 ($n = 3$), (b) children who failed another screener under development at Time 1 ($n = 1$), (c) children who failed the M-CHAT at Time 2, but not Time 1 ($n = 8$), and (d) children who passed the M-CHAT but were referred for possible ASD or developmental disorder by Time 2 ($n = 3$), for a total of 15 possible missed cases. All of these children were evaluated: 7 were diagnosed with an ASD, 1 with global developmental delay, 3 with language delay, and 4 were given no diagnosis.

The 7 missed cases of ASD were compared on Time 2 variables to the 73 children with ASD who were not missed, that is, who were detected at Time 1 with the M-CHAT. There were no significant differences on age, sex, ADI-R, ADOS, CARS, number of DSM-IV symptoms, Vineland Socialization or Vineland Motor scores. The missed cases were significantly higher functioning on Vineland Communication (mean standard score 74 vs. 63), $F(1, 79) = 6.03$, $p < .05$, Vineland Daily Living (71 vs. 58), $F\ (1,\ 79) = 20.09$, $p < .01$, Vineland Adaptive Behavior Composite (68 vs. 59) $F\ (1, 79) = 7.85$, $p < .01$, and DAS Verbal score (79 vs. 71), $F\ (1,\ 79) = 8.72$, $p < .05$.

Discussion

The purpose of the current study was to continue validating the M-CHAT as a screener for autism in young children. For the total sample, figures for PPV are very close to the original PPV's reported in Robins et al. (2001). Positive predictive value (PPV) for screening and diagnosis at Time 1 (16–30 months) was .36 ± .05 for the M-CHAT alone and .74 ± .06 for the M-CHAT plus telephone interview, indicating that the telephone follow-up is a critical step in eliminating false positives and improving the PPV. This was especially true of the low-risk, general population sample, where PPV of the M-CHAT alone was only .11 ± .05, but jumped to .65 ± .17 when including the telephone interview. Thus, PPV for the M-CHAT alone is unacceptably low for the low-risk sample, but increases to an acceptable level with the telephone follow-up. This suggests that pediatric practices screening low-risk children should have someone available to review answers, either on site or on the telephone, to avoid unnecessary referrals and parent concern. The telephone follow-up, while still improving the PPV, is less crucial for the high-risk sample, for whom PPV was .60 ± .07 for the M-CHAT alone, and .76 ± .07 for the M-CHAT plus telephone interview. Furthermore, the threshold for failing the screener was set low to avoid as many misses as possible, at the expense of positive predictive power; practitioners should be aware of this fact and might consider an intermediate type of evaluation, such as the Screening Tool for Autism in 2-Year-olds (Stone et al. 2004), or their own clinical assessment, before referring children with a borderline score for a full, specialized work-up.

The PPV for screen at Time 1 predicting to diagnosis at Time 2 (age 4) was .38 ± .07 for the screening alone, and .59 ± .09 for the screening plus telephone follow-up. Thus, the PPV of the initial screener is about the same for concurrent and predictive diagnoses, but the PPV of the screening plus telephone follow-up is lower for diagnosis 2 years later, but still in the acceptable range. In addition, most of the children who screened positive were diagnosed both at Time 1 and Time 2 with a developmental delay or disorder of some kind, and needed intervention referrals. At Time 2, however, 12 children did not meet criteria for any diagnosis; whether these children were falsely diagnosed at Time 1, or had improved because of maturation or early intervention, is not possible to determine. Kleinman et al. (in press) examined the question of diagnostic stability in this sample, and Sutera et al. (2007) investigated Time 1 characteristics of children who apparently moved off the spectrum by Time 2. They found that the children who moved off the spectrum were very similar at Time 1 to those who remained on the spectrum, with significantly better motor skills and a trend toward higher IQ and higher daily living skills, but similar on all other clinical and demographic variables. Thus, it seems likely that the children with excellent outcomes would have been hard to distinguish at Time 1 from the children who stayed on the autism spectrum.

It should be noted that the "telephone interview" does not need to be done on the telephone; it can be done on site in the physician's office following scoring of the initial screener, to clarify responses, or on the telephone following the visit in which a parent completed the M-CHAT. The telephone interview is available from the authors at the University of Connecticut. It should also be noted that if a child fails a large number of items on the initial screening (8 or more), a telephone follow-up may not be necessary, since the child is highly likely to still screen positive and a more detailed evaluation can be immediately recommended.

The M-CHAT was found to be internally consistent, as was the subset of the six most discriminating items (critical items) as defined in Robins et al. (2001). Internal consistency figures on the current sample were very similar to those reported by Robins et al. in the original, non-overlapping sample. Note, however, that there is no evidence that the six critical items would retain this reliability if they were disembedded from the rest of the screener.

The low-risk, general population sample of children evaluated was large enough in the current dataset to compare to the high-risk sample. M-CHAT total and critical scores were higher for the high-risk group; this difference remained significant when only the screen positive children were compared. This may explain the large improvement in PPV in the low-risk sample when the telephone interview is added: when scores are at the low end of the screen positive range (e.g., total of 3 or 4), it is more likely that the follow-up interview will change a child's classification to screen negative. In contrast, when the total score is higher, as it tended to be in the high-risk group, more items would need to be passed during the interview in order to change a child's status to screen negative; therefore, change in screen status is less likely. Hence, the difference between

PPV before and after telephone interview was small for the high-risk group, but large for the low-risk group.

Among those evaluated, children not diagnosed with ASD were generally more impaired if they came from the high-risk sample, which would be expected since they had already been referred for early intervention services. On the other hand, high- and low-risk groups with ASD were not found to be different on any demographic or clinical measure. Thus, although the M-CHAT operates differently in the high- and low-risk samples (different PPV's, different non-ASD children screening positive), it identified similar children with ASD from the two sources. It was interesting to note that the four groups (high- and low-risk ASD, high- and low-risk non-ASD) did not differ on expressive language, suggesting that expressive language appears to be a nonspecific symptom of developmental delay at this age, and is thus a weak discriminator among such disorders.

Eighteen children were identified as possible missed cases at Time 1 and evaluated, based on failing an additional screener under development or their health care provider indicating concern. Only one of these received a diagnosis of ASD. At Time 2, 15 children were identified as having possibly been missed at Time 1 screening, based on failing the M-CHAT at Time 2, health care provider concern at Time 1, or a referral for possible ASD by Time 2. Seven of these children were confirmed with ASD upon evaluation. Thus, of the total of 80 children diagnosed with ASD at Time 2, 7 had been missed by the M-CHAT at Time 1 (9%). Since it is not possible to determine how many children were missed at either time point without evaluating all the screen-negative children, which was beyond the resources of the study, this value of 91% of children detected could be regarded as an upper bound of sensitivity. This is consistent with the existing sensitivity figures for the M-CHAT, which estimated it as .77–.92 (Eaves et al. 2006) and .84–.93 (Wong et al. 2004), (but with high-risk samples only).

There are a number of limitations to the current study. The most significant problem lies in the potential for missed cases. There is currently no feasible method to detect all possible missed cases. Participation is voluntary, and the American healthcare system does not have the surveillance procedures in place to identify all children from the study who may receive a later ASD diagnosis. In addition, the M-CHAT is designed to detect children with Autistic Disorder or PDD-NOS who would be detectable around age 2, and would therefore be likely to miss some children with Asperger's Disorder, or those with high functioning autism who might not be detectable until later in childhood. In fact, the missed cases we did detect were higher functioning than the detected ASD cases on several developmental variables. Some caregivers may avoid filling out the Time 2 screener, which was the major way of finding missed cases, because of concerns about their child. Caregivers may also fill out the screener but under-report symptoms. At the Time 1 screening, this problem is potentially avoided by having the health care provider flag screeners for whom they suspect ASD. Despite this added procedure, the number of missed cases is not possible to determine with any certainty. An epidemiological study in which a large population is screened at age 2, and then tracked into the school years, to an age at which ASD is likely to have been detected in all children, would be the best way to determine the true sensitivity, specificity, and NPV of the screener, but this large effort was beyond the scope of our resources.

Parents may be less likely to come in for an evaluation at Time 2 than at Time 1 especially if their child has already been diagnosed and is receiving services with which they are satisfied. This has been a problem not only for ascertaining the missed cases, but also in re-evaluating the children who were seen at Time 1 for stability of diagnosis. A future direction of study is to include strategies to increase participation at Time 2, especially for those children who qualify for evaluations; another avenue is developing means of gathering data from those who do not wish to come for an in-person evaluation at Time 2, such as review of their medical and educational records, and obtaining data from parents and teachers/service providers.

In addition, the demographics of the screened sample indicate a sample skewed toward the upper SES range. Although there were minority and low-income families in the current sample, efforts are being made to increase the representation of these families by increasing the screening at large urban clinics. Although differential attrition was not found in the current study, either by Time 1 diagnosis or ethnicity, this will have to be reassessed with a larger sample.

In calculating the Positive Predictive Value from screening at Time 1 to diagnosis at Time 2, a false positive was any child who failed the MCHAT at Time 1 and was not diagnosed with an ASD at Time 2. It is important to note that some children identified as false positives in this way may in fact have been accurately identified as having an ASD at Time 1, but did not retain the ASD diagnosis after 2 years of intensive intervention; in the current study, treatment successes cannot be differentiated from children misdiagnosed at Time 1.

An additional limitation was that blind assessment was not possible. Although we attempted to keep the clinician who tested the child blind at Time 2 to diagnosis at Time 1, the parents frequently discussed the child's prior evaluation or intervention services with this person, so this attempt was not very successful. There was no attempt to keep the clinician who interviewed the parent blind to prior diagnosis, since discussion of intervention was considered clinically crucial. In addition, since the M-CHAT results often were discussed at the evaluation, the number of failed items was apparent. Diagnosis was done as objectively as possible; it depended on ADOS, ADI-R, CARS, and clinical judgment, and additional analyses (Ventola et al. 2006) indicated that clinical judgment had very good agreement with ADOS and CARS scores. Nevertheless, bias in arriving at a diagnosis based on knowledge of screening results cannot be excluded.

The current study was not designed to directly compare different methods of screening. Although the American Academy of Pediatrics recently (AAP 2006) recommended that all children be screened specifically for autism at 18 months, this recommendation was made on theoretical rather than empirical grounds. It is possible that an effective broad band screening and surveillance would detect children with ASD with adequate sensitivity. It would be useful to directly compare these two screening approaches empirically.

Overall, the results of this study suggest that the M-CHAT can be useful in detecting ASD in children 16–30 months. The scoring criteria established with the first set of participants (Robins et al. 2001) continue to be valid. Furthermore, the M-CHAT has been successfully incorporated into pediatric visits with minimal disruption. The biggest shortcoming appearing to date is the low PPV of the screener in unselected populations, before the telephone interview. Unless the number of failed items is high to begin with, the use of the telephone interview to review failed items, either in person or on the telephone, is crucial to avoid unnecessary referrals. Clinical decisions can be made in different settings. Pediatricians dealing with low-risk, unselected samples might choose to use a higher cutoff for failing; this would, of course, improve the PPV, but would lead to an increase in missed cases. Based on current data, the optimal course seems to be to use the current cut-offs, but for low-risk samples, to always use the follow-up interview to confirm responses, and perhaps to institute a second level screen or more in-depth parent interview and behavior observation for borderline cases, before a full-scale autism evaluation is instituted.

Acknowledgment We gratefully acknowledge the assistance of participating pediatricians and staff in their offices, as well as providers in the Connecticut and Massachusetts Birth-to-Three Early Intervention systems. We would like to thank Jillian Wood, Executive Director of the Hezekiah Beardsley Chapter of the American Academy of Pediatrics, who promoted the study state-wide, as well our dedicated undergraduate research assistants, and the children and families who participated. We would also like to extend sincere thanks to members of the Early Detection Advisory Board, especially Ho-Wen Hsu, MD, and Mark Greenstein, MD, for their wise advice and support. This study is supported by NIH grant R01 HD039961 and Maternal and Child Health Bureau grant R40 MC00270 and a prior grant from the National Alliance for Autism Research, and training grants from NIMH and the Dept. of Education.

References

American Academy of Pediatrics, Council on Children with Disabilities, Section on Developmental Behavioral Pediatrics, Bright Futures Steering Committee, Medical Home Initiatives for Children With Special Needs Project Advisory Committee (2006). Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. *Pediatrics, 118*, 405–420.

American Psychiatric Association. (1994). *Diagnostic and statistics manual of mental disorders* (4th ed., DSM-IV). Washington, DC: Author.

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

Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., & Charman, T. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: The Special Needs and Autism Project (SNAP). *Lancet, 368*, 210–215.

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*, 839–843.

Baron-Cohen, S., Cox, A., Baird, G., Swettenham, J., Nightengale, N., Morgan, K., Drew, A., & Charman, T. (1996). Psychological markers in the detection of autism in infancy in a large population. *British Journal of Psychiatry, 168*, 158–163.

Bertrand, J., Mars, A., Boyle, C., Bove, F., Yeargin-Allsopp, M., & Decoufle, P. (2001). Prevalence of autism in a United States population: The Brick Township, New Jersey, investigation. *Pediatrics, 108*, 1155–1161.

Chakrabarti, S., & Fombonne, E. (2001). Pervasive developmental disorders in preschool children. *Journal of the American Medical Association, 285*(24), 3093–3099.

Charman, T. (2002). The prevalence of autism spectrum disorders—Recent evidence and future challenges. *European Child & Adolescent Psychiatry, 11*(6), 249–256.

Charman, T., Taylor, E., Drew, A., Cockerill, H., Brown, J., & Baird, G. (2005). Outcome at 7 years of children diagnosed with autism at age 2: Predictive validity of assessments conducted at 2 and 3 years of age and pattern of symptom change over time. *Journal of Child Psychology and Psychiatry, 46*, 500–513.

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*, 128–138.

Cox, A., Klein, K., Charman, T., Baird, G., Baron-Cohen, S., Swettenham, J., Drew, A., & Wheelwright, S. (1999). Autism spectrum disorders at 20 and 42 months of age: Stability of clinical and ADI-R diagnosis. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 40*, 719–732.

Dietz, C., Swinkels, S., van Daalen, E., van Engeland, H., & Buitelaar, J. K. (2006). Screening for as*Journal of Autism and Developmental Disorders, 36*, 713–722.

dosReis, S., Weiner, C. L., Johnson, L., & Newschaffer, C. J. (2006). Autism spectrum disorder screening and management practices among general pediatric providers. *Journal of Developmental and Behavioral Pediatrics, 27*, S88–S94.

Dumont-Mathieu, T., & Fein, D. (2005). Screening for autism in young children: The Modified Checklist for Autism in Toddlers (M-CHAT) and other measures. *Mental Retardation and Developmental Disabilities Research Reviews, 11*, 253–262.

Earls, M., & Hay, S. (2006). Setting the stage for success: Implementation of developmental and behavioral screening and surveillance in primary care practice-The North Carolina assuring better child health and development 9ABCD) project. *Pediatrics, 118*, 183–188.

Eaves, L., & Ho, H. (2004). Brief report: stability and change in cognitive and behavioral characteristics of autism through childhood. *Journal of Autism and Developmental Disorders, 26*, 557–569.

Eaves, L., Wingert, H., Ho, & Helena H. (2006). Screening for autism: Agreement with diagnosis. *Autism, 10*, 229–242.

Elliot C. D. (1990) *Differential abilities scale*. San Antonio TX: Harcourt Brace.

Fine, S., Weissman, A., Gerdes, M., Pinto-Martin, J., Zackai, E., McDonald-McGinn, D., & Emanuel, B. (2005). Autism spectrum disorders and symptoms in children with molecularly confirmed 22q11.2 deletion syndrome. *Journal of Autism and Developmental Disorders, 35*, 461–470.

Fombonne, E. (2003). Epidemiological surveys of autism and other pervasive developmental disorders: An update. *Journal of Autism and Developmental Disorders, 33*, 365–382.

Fombonne, E., Zakarian, R., Bennett, A., Meng, L., & McLean-Heywood, D. (2006). Pervasive developmental disorders in Montreal, Quebec, Canada: Prevalence and links with immunizations. *Pediatrics*, *118*, 139–150.

Gillberg, C., Nordin, V., & Ehlers, S. (1996). Early detection of autism. Diagnostic instruments for clinicians. *European Child and Adolescent Psychiatry, 5*, 67–74.

Glascoe, F. (2005). Screening for developmental and behavioral problems. *Mental Retardation and Developmental Disabilities Research Reviews, 11*(3), 173–179.

Glascoe F. (2001). *Parents' evaluation of developmental status*. Nashville, TN: Ellsworth and Vandermeer Press, LLC.

Glascoe, F., & Dworkin, P. (1995). The role of parents in the detection of developmental and behavioral problems. *Pediatrics, 95*, 829–836.

Gray, K., Tonge, B., & Brereton, A. (2006). Screening for autism in infants, children, and adolescents. *International Review of Research in Mental Retardation,* 32, 197–227.

Harris, S., & Handleman, J. (2000). Age and IQ at intake as predictors of placement for young children with autism: A four to 6-year follow-up. *Journal of Autism and Developmental Disorders, 30*, 137–142.

Kleinman, J. Ventola, P., Pandey, J., Verbalis, A., Barton, M., Hodgson, S., Green, J., Dumont-Mathieu, T., Robins, D., & Fein, D. (in press) Diagnostic stability and prediction of outcome in very young children with autism spectrum disorders. *Journal of Autism and Developmental Disorders.*.

Lord, C. (1995). Follow-up of 2-year-olds referred for possible autism. *Journal of Child and Adolescent Psychiatry, 36*, 1365–1382.

Lord, C., Risi, S., DiLavore, P., 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. (1999). *Autism diagnostic observation schedule-WPS edition*. Los Angles, California: Western Psychological Services.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*, 659–685.

Mawle, E., & Griffiths, P. (2006) Screening for autism in pre-school children in primary care: Systematic review of English Language tools. *International Journal of Nursing Studies, 43*, 623–636.

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.

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

Robins, D., & Dumont-Mathieu, T. (2006). Early screening for autism spectrum disorders: Update on the modified checklist for autism in toddlers and other measures. *Journal of Developmental & Behavioral Pediatrics, 27*(Supplement 2), S111–S119.

Robins, D. L., Fein, D., & Barton, M. L. (1999). *The modified checklist for autism in toddlers (M-CHAT)*. Storrs, CT: Self-published.

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*, 131–144.

Sand, N., Silverstein, M., Glascoe, F. P., Gupta, V. B., Tonniges, T. P., & O'Connor, K. G. (2005). Pediatricians' reported practices regarding developmental screening: Do guidelines work? Do they help? *Pediatrics, 116*, 174–179.

Schopler, E., Reichler, R. J., & Renner, B. R. (1988). *The childhood autism rating scale*. Los Angeles: Western Psychological Services.

Sparrow S., Balla, D., & Cicchetti, D. (1984). *The vineland adaptive behavior scales*. Circles Pines, MN: American Guidance Services.

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*, 691–701.

Stone, W., Lee, E., Ashford, L., Brissie, J., Hepburn, S., Coonrod, E., & Weiss, B. (1999). Can autism be diagnosed accurately in children under 3 years? *Journal of Child Psychology and Psychiatry and Allied Disciplines, 40*, 219–226.

Sutera, S., Sutera, S., Pandey, J., Esser, E., Rosenthal, M., Wilson, L., Barton, M., Green, J., & Fein, D. (2007). Predictors of optimal outcome in toddlers diagnosed with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*, 98–107.

Swinkels, S. H. N., Dietz, C., van Daalen, E., Kerkhof, H. G. M., van Engeland, H., & Buitelaar, J. K. (2006). Screening for autistic spectrum in children aged 14–15 months. *I: The development of the Early Screening of Autistic Traits Questionnaire (ESAT). Journal of Autism and Developmental Disorders, 36*(6), 723–732.

Tuchman, & Rapin, I. (1997). Regression in pervasive developmental disorders: seizures and epileptiform electroencephalogram correlates. *Pediatrics, 99*, 560–566.

Ventola, P., Kleinman, B., Pandey, J., Barton, M., Allen, S., Green, J., Robins, D., & Fein, D. (2006) Agreement among four diagnostic instruments for autism spectrum disorders in toddlers. *Journal of Autism and Developmental Disorders, 36*, 839–847.

Ventola, P., Kleinman, J., Pandey, P., Wilson, L., Esser, E., Boorstein, H., Dumont-Mathieu, T., Marshia, G., Barton, M., Hodgson, S., Green, J., Volkmar, F., Chawarska, K., Babitz, T., & Fein, D. (2007). Differentiating between Autism Spectrum Disorders and other developmental disabilities in children who failed a screening instrument for ASD. *Journal of Autism and Developmental Disorders, 37*, 4325–436.

Volkmar, F., Chawarska, K., & Klin, A. (2005). Autism in infancy and early childhood. *Annual Review of Psychology, 56*, 315–336.

Wong, V., Hui, L., Lee, W., Leung, L., Ho, P., Lau, W., Fung, C., & Chung, B. (2004). A modified screening tool for autism (Checklist for Autism in Toddlers [CHAT-23] for Chinese children. *Pediatrics, 114*, e166–e176.