Journal of Autism and Developmental Disorders, Vol. 34, No. 6, December 2004 (© 2004)

Psychometric Properties of the STAT for Early Autism Screening

Wendy L. Stone,[1,3] Elaine E. Coonrod,[2] Lauren M. Turner,[2] and Stacie L. Pozdol[2]

The STAT is an interactive screening measure for autism that assesses behaviors in the areas of play, communication, and imitation skills. In Study 1, signal detection procedures were employed to identify a cutoff score for the STAT using developmentally matched groups of 2-year-old children with autism and with nonspectrum disorders. The resulting cutoff yielded high sensitivity, specificity, and predictive values for the development sample as well as for an independent validation sample. Study 2 examined psychometric properties of the STAT and revealed acceptable levels of interrater agreement, test-retest reliability, and agreement between STAT risk category and ADOS-G classification. The STAT demonstrates strong psychometric properties and shows promising utility as a Level 2 screening measure for autism.

KEY WORDS: AUTISM; Screening; Early identification; Young children.

The past decade has seen a tremendous increase in research focused on the early manifestations of autism. It has been demonstrated that the diagnosis of autism can be made accurately in children as young as 2 years (Cox *et al*., 1999; Lord, 1995; Stone *et al*., 1999), and that the measurement of early social-communicative behaviors—such as imitation, play, and joint attention—is key to early identification (Baron-Cohen, Allen, & Gillberg, 1992; Stone, Coonrod, & Ousley, 2000). Unfortunately, there is a gap between research and clinical practice such that many children fail to receive definitive diagnoses of autism until the age of 4 or older (Howlin & Moore, 1997; Siegel, Pliner, Eschler, & Elliot, 1988).

Early identification of children with autism has come to be recognized as a critical aspect of their medical management and treatment, and its importance has been highlighted in recent practice guidelines issued by the American Academy of Pediatrics (American Academy of Pediatrics Committee on Children with Disabilities, 2001), the American Academy of Neurology (Filipek *et al*., 2000), and the National Academy of Sciences (Committee on Educational Interventions for Children with Autism, 2001). The impetus for early detection derives from several intervention studies demonstrating significant gains in language, social, and cognitive functioning for young children with autism participating in early intervention programs (Harris, Handleman, Gordon, Kristoff, & Fuentes, 1991; Lovaas, 1987; Rogers & Lewis, 1989; Strain, Hoyson, & Jamieson, 1985). Early intervention is thought to be critical for preventing a cascade of effects that result from early deficits and interfere with later functioning (Happe, 1994; Mundy & Crowson, 1997). Identification of autism at young ages would allow broader participation in these specialized early intervention services,

[1] Vanderbilt Children's Hospital.
[2] Vanderbilt University.
[3] Correspondence should be addressed to Wendy L. Stone, Vanderbilt Child Development Center, 426 Medical Center South, 2100 Pierce Avenue, Nashville, TN 37232-3573; USA; Tel.: 615-936-0280; e-mail: wendy.stone@vanderbilt.edu

0162-3257/04/1200-0691/0

which could potentially lead to improved outcomes for more children.

However, the early identification of autism is not without its challenges. Autism is a behaviorally based diagnosis, and there is often a lack of familiarity with its early behavioral manifestations among front-line professionals. Moreover, in many communities, diagnostic evaluations are provided in specialized multidisciplinary centers, which often have long waiting lists. Both situations can present obstacles to early diagnosis and can delay enrollment in early intervention. One approach to overcoming these obstacles and facilitating early identification is the development of effective early screening measures for autism. The availability of screening tools designed for community service providers could increase knowledge about the early behavioral features of autism as well as enable providers to direct children to specialized assessment and/or intervention services at young ages (Baird *et al.*, 2001).

Screening has been defined as "a brief assessment procedure designed to identify children who, because of the risk of a possible learning problem or handicapping condition, should proceed to a more intensive level of diagnostic assessment" (Meisels & Atkins-Burnett, 1994, p. 1). There are many dimensions on which screening measures may differ, one of which is the setting and population for which the tool was designed, i.e., the level of the screening (Siegel, 1998). Level 1 screening measures for autism are used to identify children at risk for autism from the general population. These screenings are designed for use in settings such as pediatric practices, where they are administered to all children—whether or not there are concerns about developmental problems—during their well-child visits. Level 2 screening for autism involves the identification of children at risk for autism from a population of children demonstrating a broad range of developmental concerns, such as global developmental delay or language impairment. These screenings are used in settings such as child find agencies, early intervention programs, or evaluation clinics serving children with a variety of developmental problems.

Another dimension on which screening measures can differ is the method by which information is gathered. Screening information can be gathered through parental report, through observations of the child, or through direct interactions with the child. Each method has its merits and drawbacks. Among the advantages of parental reports are the ease and speed of administration, while disadvantages are the potential sources of reporter bias (Glascoe, 2000; Stone, Hoffman, Lewis, & Ousley, 1994). For example, some parents may fail to recognize abnormal behaviors, while others may view developmentally appropriate behaviors as pathological. Interactive methods enable clinicians to directly observe subtle social and communicative deficits that parents might not recognize; however, context-specific or low frequency behaviors (e.g., peer interactions, motor stereotypies) may not be observed readily in clinical settings. In addition, interactive screenings may require more time—and sometimes training—to administer.

Existing screening tools for young children with autism have focused primarily on screening at the population level (i.e., Level 1) and on gathering information via parental report. A comparison of the early screening measures for autism is presented in Table I. The Checklist for Autism in Toddlers (CHAT; Baron-Cohen *et al.*, 1992) was the first published measure, and has been the subject of the most research. The CHAT is a Level 1 screener that was designed for use during 18 month well-child pediatric visits. It consists of parental report and interactive items tapping behaviors that include protodeclarative pointing, gaze monitoring, and pretend play. In a large population-based sample of 16,235 children, 12 children screened at 18 months failed the CHAT, 10 of whom (83%) obtained a subsequent diagnosis of autism at age 3½ (Baron-Cohen *et al.*, 1996). However, a 6-year follow-up study revealed that many children diagnosed with autism at age 7 had not been identified as at-risk at the 18 months screening, resulting in a sensitivity of

Table I. Comparison of Early Screening Measures for Autism

Feature	CHAT[a]	M-CHAT[b]	PDDST[c]	STAT[d]
Ages designed for	18 months	24 months	Under 6 years	24–35 months
Type of screening	Level 1	Level 1	Levels 1– 2	Level 2
Nature of items	Interview & Interactive	Questionnaire	Questionnaire	Interactive
# Interactive items	5	0	0	12

[a] Checklist for Autism in Toddlers.
[b] Modified Checklist for Autism in Toddlers.
[c] Pervasive Developmental Disorders Screening Test.
[d] Screening Tool for Autism in Two-year-olds.

.38 (Baird *et al.*, 2000). Initial efforts to examine the utility of the CHAT as a Level 2 screener have also been described (Scambler, Rogers, & Wehner, 2001).

The Modified Checklist for Autism in Toddlers (M-CHAT) (Robins, Fein, Barton, & Green, 2001) was originally developed as an adaptation of the CHAT and as a Level 1 screener. The M-CHAT is a parental questionnaire for 24-month-olds that consists of 23 items, 9 of which were taken directly from the CHAT. The authors report the estimated sensitivity and specificity to be .87 and .99, respectively; however, because the majority of the children in their nonautistic sample did not receive diagnostic evaluations, the actual screening properties of the M-CHAT are not yet known.

The Pervasive Developmental Disorders Screening Test (PDDST) (Siegel, 1996) is a parental questionnaire that was developed for use with children under 6 years old. Questions focus on children's early behaviors in areas such as nonverbal communication, temperament, sensory responses, play, attachment, and social interaction. A distinctive feature of the PDDST is that different versions are available for Levels 1 and 2 screening. Preliminary reports of psychometric properties revealed sensitivity levels of .85 and .69, and specificity levels of .71 and .63, for Levels 1 and 2, respectively (Siegel & Hayer, 1999). However, this measure has yet to be published.

The Screening Tool for Autism in Two-Year-Olds (STAT) (Stone *et al.*, 2000) is unique among the existing screeners in that it is the only Level 2 measure comprised of interactive items. The advantage of an interactive measure is its provision of a standard set of items or activities that afford direct observation of key behaviors. The STAT was designed for use with children from 24 through 35 months of age, and consists of 12 activities for observing children's early social-communicative behaviors. Items were selected for inclusion on the STAT based on their effectiveness in differentiating 2-year-old children with autism from developmentally matched children with nonautistic developmental disorders. Initial research with the STAT has revealed strong sensitivity, specificity, and predictive values (Stone *et al.*, 2000).

The purpose of the present study was to extend our investigation of the psychometric properties of the STAT in a larger sample. Our specific aims were to: (1) derive a scoring algorithm for the STAT using signal detection methods; (2) examine the reliability and validity of the STAT.

STUDY 1: DEVELOPMENT OF SCORING ALGORITHM

Method

Participants

Fifty-two children participated in this study, 26 with a clinical diagnosis of autism and 26 with developmental delay and/or language impairment (DD/LI). Children were recruited for participation between 1997 and 2000. The majority of children ($n = 41$) were recruited from a regional, university-based diagnostic evaluation center, and the others were referrals from a university-affiliated speech and hearing center ($n = 7$) or from a state network providing early identification and service coordination ($n = 4$). Eligibility requirements for participation included: (1) chronological age from 24 through 35 months (i.e., between 2 years, 0 months, 0 days and 2 years, 11 months, 29 days); (2) absence of an identified genetic or metabolic disorder; (3) absence of a severe sensory or motor impairment.

Children in the two groups were individually matched on chronological age and mental age, resulting in 26 matched pairs. The matched pairs were then randomly divided into two subsamples, each consisting of 13 pairs. One subsample was used to identify a cutoff score demonstrating the highest levels of sensitivity and specificity (i.e., development sample), and the other subsample was used to obtain independent validation of the identified cutoff (i.e., validation sample).

Demographic characteristics of the development and validation samples are presented in Table II. For the development sample, significant group differences were found for gender, $X^2(1, N = 26) = 3.9$, $p = .047$ and group differences for mental age approached significance, $t\ (24) = 1.9$, $p = .064$, while all other group comparisons were nonsignificant, $ps > .16$. For the validation sample, no significant diagnostic group differences were found for any of the demographic variables, all $ps > .38$.

Measures and Procedures

The (STAT) (Stone *et al.*, 2000) is an interactive measure that is administered within the context of play, and takes about 20 minutes to complete.

Table II. Demographic Characteristics for Signal Detection Sample

	Development sample		Validation sample	
	Autism (n = 13)	DD/LI (n = 13)	Autism (n = 13)	DD/LI (n = 13)
Chronological age (months)				
M (*SD*)	31.2 (3.8)	31.1 (3.5)	32.2 (3.5)	31.2 (4.1)
Range	24–35	24–35	25–35	26–35
Mental age (months)				
M (*SD*)	16.3 (3.7)	19.3 (4.0)	17.3 (7.1)	19.6 (6.1)
Range	11–25	11–25	11–39	14–38
Race (%)				
Caucasian	92	61	77	85
African-American	8	31	23	15
Other	0	8	0	0
Male (%)	77	39	85	69
Mothers with high school education or beyond (%)	92	85	85	85

The STAT consists of 12 items that were derived from three measures: the Play Assessment Scale (Fewell, 1991), the Prelinguistic Communication Assessment (Stone, Ousley, Yoder, Hogan, & Hepburn, 1997), and the Motor Imitation Scale (Stone, Ousley, & Littleford, 1997) (see Stone *et al.*, 2000 for additional information about the development of the STAT). Items assess behaviors in four social-communicative domains: Play, Requesting, Directing Attention, and Motor Imitation. A brief description of each item is presented in Table III.

Each item is scored as Pass, Fail, or Refuse according to specific criteria described in the STAT manual (Stone & Ousley, 1997). Because the four STAT domains contain different numbers of items, equal weighting for the domains is achieved by expressing domain scores as the proportion of failed items to total items. Thus, scores for the domains with two items (i.e., Play and Requesting) can be 0, .5, or 1, and scores for domains with four items (i.e., Directing Attention and Motor Imitation) can be 0, .25, .50, .75, or 1. The total STAT score is derived by summing the four domain scores, and can therefore range from 0 to 4, with higher scores representing greater impairment.

All data were collected during the course of the child's diagnostic evaluation, after informed consent was obtained from parents. Administration of the STAT took place during a break in the child's evaluation schedule, most often after the diagnostic evaluation had been completed and while the clinicians and parents were discussing the results. The diagnostic evaluations were usually conducted by a team of clinicians that included a licensed psychologist and a licensed speech-language pathologist. A social worker and/or developmental pediatrician also participated in some of the evaluations. All evaluations included standardized assessments of cognitive/developmental level using the Bayley Scales

Table III. Description of STAT Items

Domain	Item	Description
Play	Turn-taking	Examiner rolls a ball or toy car to the child to engage him/her in back-and-forth play.
	Doll play	Examiner presents the child with a doll or stuffed animal, along with furniture and eating utensils, to observe the use of functional play.
Requesting	Snack	Examiner presents the child with a clear, tightly sealed jar filled with desirable food treats.
	Bubbles	Examiner blows soap bubbles and then hands the tightly sealed jar to the child.
Directing Attention	Balloon	Examiner inflates a balloon and then lets go so that it flies across the room as it deflates.
	Puppet	Examiner places an animal puppet on his/her own hand when the child is not looking and then begins writing with it within the child's view.
	Bag of toys	Examiner presents an opaque bag containing interesting toys to the child and encourages him/her to look inside.
	Noisemaker	Examiner activates a noisemaker out of view of the child.
Motor Imitation	Rattle	Examiner shakes a rattle, then encourages the child to do the same.
	Car	Examiner rolls a small car back and forth across the table, then encourages the child to do the same.
	Drum hands	Examiner drums his/her hands on the table, then encourages the child to do the same.
	Hop dog	Examiner hops a small toy dog across the table, then encourages the child to do the same.

of Infant Development—Second Edition (Bayley, 1993), the Mullen Scales of Early Learning (Mullen, 1995), or the Battelle Developmental Inventory (Newborg, Stock, Wnek, Guidubaldi, & Scinicki, 1984). Some evaluations also included assessments of adaptive behavior and/or speech-language skills. Autism diagnoses were made by the team psychologist, and were based on criteria provided in DSM-IV (APA, 1994). All psychologists had over 15 years of experience in the assessment of young children.

The STAT was administered by examiners with college degrees who had received prior training on administration and scoring. The STAT examiners were independent from the diagnostic team and were blind to the results of the diagnostic evaluation. Likewise, the team clinicians were blind to the results of the STAT screening. These procedures ensured that the child's STAT score and the child's formal diagnosis were obtained independently.

Results

Signal detection methodology was used to identify the cutoff score that demonstrated optimal sensitivity and specificity for the development sample. Because the STAT was developed as a screening measure, greater weight was placed on sensitivity (i.e., correctly identifying all children at risk for autism) than specificity (i.e., correctly identifying all children *not* at risk for autism). The sensitivity and specificity associated with different cutoff scores for this sample are presented in Table IV. As the table reveals, the optimal cutoff for maximizing sensitivity while maintaining adequate specificity appeared to be between 1.75 and 2.13. A STAT score of 2 (or higher) was therefore selected as the cutoff for autism risk.

Table IV. Sensitivity and Specificity for Different STAT Cutoff Scores for Development Sample

Cutoff *	Sensitivity	Specificity
−50	1.00	.00
.75	1.00	.31
1.13	1.00	.54
1.38	1.00	.69
1.75	**1.00**	**.77**
2.13	**1.00**	**.85**
2.50	.92	.85
2.88	.85	.85
3.13	.77	1.00
3.38	.69	1.00
3.63	.54	1.00
3.89	.31	.00
5.00	.00	.00

Notes: * A score greater than or equal to the cutoff indicates autism risk.

Use of this cutoff score with the validation sample resulted in a sensitivity of .92 and specificity of .85, indicating a good hit rate for identifying the children who did and did not receive a clinical diagnosis of autism. The positive predictive value (i.e., proportion of children who screen at high risk for autism who actually have autism) was .86 and the negative predictive value (i.e., proportion of children who screen at low risk for autism who actually do not have autism) was .92 for the validation sample.

Across the development and validation samples, a total of 6 children were identified incorrectly by the STAT. Only one child with a clinical diagnosis of autism scored as low-risk on the STAT. This child had a chronological age of 35 months and a mental age of 39 months, and was thus older and functioning at a higher cognitive level than most of the children in the autistic sample. Five children in the DD/LI group were identified incorrectly by the STAT. There was no clear pattern that explained the overidentification, though the misidentified children tended to be somewhat younger and/or to have somewhat lower mental ages than the majority of the DD/LI sample.

STUDY 2: EXAMINATION OF PSYCHOMETRIC PROPERTIES

Method

Participants

Participants in this study were 104 children, 50 with a clinical diagnosis of autism, 15 with PDD-NOS, and 39 with DD/LI. Children were recruited between 1999 and 2001 from two primary sources: a university-affiliated speech and hearing center ($n = 47$) and a state network providing early identification and service coordination ($n = 32$). Additional referrals came from parents or local pediatricians ($n = 14$) or from a regional, university-based diagnostic evaluation center ($n = 11$). Children whose parents reported a wide range of developmental concerns were recruited (i.e., not just those suspected of having autism), in order to obtain a fairly representative sample of children referred for developmental evaluations.

Eligibility requirements for participation included: (1) chronological age from 24 through 35 months; (2) absence of an identified genetic or metabolic disorder; and (3) absence of a severe sensory or motor impairment. Twelve of the children in Study 2 (12%) had also participated in Study 1. Demographic characteristics for this sample are presented in Table V. As expected from unselected clinic-based samples (Stone *et al.*, 2000), there were significant group differences for mental age, F (2, 101) = 36.5 and expressive language age, F (2, 101) = 19.9, $ps = .00$. *Post hoc* comparisons revealed that the autistic group had lower mean mental ages and expressive language ages relative to the DD/LI and PDD-NOS groups, $ps = .00$. In addition, children with PDD-NOS had a lower mean mental age than children with DD/LI, $p = .03$. Group differences were also found for chronological age, F (2, 101) = 4.02, $p = .02$, with *post hoc* comparisons revealing a lower mean chronological age for the autistic group relative to the DD/LI group, $p = .03$. No significant group differences were found for race, gender, or maternal education. The lower mental ages and language ages in the autistic group appear to reflect the nature of clinic-based populations (Stone *et al.*, 2000), and these group differences were not expected to exert undue influence on the results of the test-retest reliability, interobserver reliability, or concurrent validity analyses that are the focus of this study.

Table V. Demographic Characteristics for Psychometric Sample

	Autism (n = 50)	DD/LI (n = 39)	PDD-NOS (n = 15)
Chronological age (months)			
M (SD)	28.5 (3.3)	30.4 (3.3)	30.0 (3.7)
Range	24–35	24–35	24–35
Mental age (months)			
M (SD)	16.1 (4.1)	24.2 (5.0)	20.5 (3.8)
Range	8–30	16–33	15–28
Expressive language age (months)			
M (SD)	11.2 (4.5)	18.3 (6.5)	16.7 (5.0)
Range	3–24	6–28	8–24
Race (%)			
Caucasian	80	70	93
African-American	16	26	0
Other	4	4	7
Male (%)	80	82	87
Mothers with high school education or beyond (%)	94	90	93

Concurrent validity of the STAT was evaluated using the entire sample of 104. A subset of 29 children (14 with autism, 2 with PDD-NOS, and 13 with DD/LI) was used to assess interobserver agreement. A subset of 21 children (9 with autism, 6 with PDD-NOS, and 6 with DD/LI) was used to assess test-retest reliability.

Measures and Procedures

Informed consent was obtained from parents prior to the administration of any measures. All children received an assessment battery that included the STAT, the Mullen Scales of Early Learning (Mullen, 1995), the Autism Diagnostic Observation Schedule-Generic (ADOS-G) (Lord *et al.*, 2000), the Childhood Autism Rating Scale (CARS) (Schopler, Reichler, & Renner, 1988), and the Sequenced Inventory of Communication Development-Revised (SICD-R) (Hedrick, Prather, & Tobin, 1984). Evaluations were conducted or supervised by a licensed psychologist and a licensed speech-language pathologist. As in Study 1, the STAT was administered by examiners with college degrees who had received prior training on administration and scoring. The STAT and ADOS-G were administered by different examiners who were blind to the results of the other's evaluation. Clinical psychologists determined children's diagnoses on the basis of DSM-IV (APA, 1994) or DSM-IV-TR criteria (APA, 2000). These clinical diagnoses were based on observations made throughout the evaluation, but were independent from information about STAT risk status. The cutoff score for the STAT derived in Study 1 was used in the present study.

Interobserver agreement was evaluated by having two examiners score the STAT independently as it was administered. Test–retest reliability was examined by asking parents to return with their children for a second STAT administration approximately two weeks following the first. Different examiners administered the STAT at each time point to reduce any potential familiarity bias. Concurrent validity was assessed by comparing STAT results with ADOS-G results.

Results

Interobserver agreement for STAT risk category (i.e., high risk for autism vs. low risk for

autism) was 1.00 using Cohen's kappa. Children in the test–retest reliability sample received their second STAT administration an average of 20 days following their first (range = 4–44 days). Test–retest reliability for STAT risk category, using Cohen's kappa, was .90. The one disagreement for risk category occurred for a child who had a clinical diagnosis of PDD-NOS. This child scored as low-risk on the STAT at Time 1 and high-risk at retest.

Concurrent validity of the STAT was assessed by comparing children's STAT risk category with their ADOS-G classification (see Table VI). Because the STAT was designed to screen specifically for autism, rather than for autism spectrum disorders, children classified as PDD-NOS on the ADOS-G were removed from initial analyses. The resulting sample thus consisted of 82 children. For this sample, Cohen's kappa for agreement between STAT risk category and ADOS-G classification was .95. Only two children were identified incorrectly by the STAT; both were identified as high-risk by the STAT but did not meet ADOS-G criteria for autism (i.e., false positives).

To ensure that these high levels of concurrent validity were not influenced by mental age differences between the autistic and nonspectrum samples, kappas were also calculated for a subsample of children who were matched on mental age. Twelve children with an ADOS-G classification of autism were individually matched to 12 children with an ADOS-G nonspectrum classification (mean CAs = 29.8 and 30.2, respectively; mean MAs = 21.4 and 21.6, respectively) to examine agreements between the STAT and the ADOS-G. Results revealed a kappa of .92, suggesting that the mental age differences between the two samples did not account for the classification agreements using these two measures.

Although children with PDD-NOS were not included in the derivation of the STAT scoring, it was of interest to determine how they performed on the STAT. Table IV reveals that of the 22 children classified as PDD-NOS on the ADOS-G, 14 (64%) were classified as low-risk and 8 (36%) as high-risk by the STAT. Descriptive statistics for STAT scores obtained by children with ADOS-G classifications of Autism, PDD-NOS, and Nonspectrum are presented in Table VII. Significant group differences for mean STAT scores were obtained, F (2,

Table VII. Descriptive Statistics for STAT Scores by ADOS-G Classification

	Unmatched sample		
	Autism (n = 52)	PDD-NOS (n = 22)	Nonspectrum (n = 30)
Mean (SD)	3.19 (.57)	1.86 (.76)	1.08 (.64)
Median	3.25	1.75	1.00
Range	2.00–4.00	.50–3.50	0–3.00
	Matched sample		
	Autism (n = 12)	PDD-NOS (n = 12)	Nonspectrum (n = 12)
Mean (SD)	2.75 (.44)	2.08 (.86)	1.10 (.61)
Median	2.88	2.00	1.13
Range	2.00–3.25	.50–3.50	0–2.25

Table VI. Concurrent Validity of the STAT with ADOS-G Classification and Clinical Diagnosis

	ADOS-G classification		
STAT Risk Category	Autism (n = 52)	Nonspectrum (n = 30)	PDD-NOS (n = 22)
High	52 (100%)	2 (7%)	8 (36%)
Low	0 (0%)	28 (93%)	14 (64%)
	Clinical diagnosis		
	Autism (n = 50)	DD/LI (n = 39)	PDD-NOS (n = 15)
High	50 (100%)	4 (10%)	8 (53%)
Low	0 (0%)	35 (90%)	7 (47%)

101) = 110.3, $p = .000$. Post hoc comparisons revealed significant differences between all three groups, $ps = .000$, with children in the autistic group demonstrating the highest scores, children in the PDD-NOS group demonstrating intermediate scores, and children in the nonspectrum group demonstrating the lowest scores.

Because there were mental age differences between these groups, this analysis was repeated using subgroups of children matched on mental age to determine whether the higher STAT scores in the autistic group were simply a reflection of their lower mental ages. Again significant group differences were obtained, F (2, 33) = 19.1, $p = .000$. Post hoc comparisons revealed significant differences between the autistic and nonspectrum groups, $p = .000$, and between the PDD-NOS and nonspectrum groups, $p = .004$. Group differences between the autistic and PDD-NOS groups did not attain statistical significance, $p = .059$.

Though not a primary focus of this study, the relation between STAT risk category and clinical diagnosis was also examined, and is presented in Table VI. This comparison is somewhat less stringent because clinical diagnoses were based on observations that included the STAT administration. When the children with a clinical diagnosis of PDD-NOS were removed from the sample, Cohen's kappa for agreement between STAT risk category and clinical diagnosis was .91. Four children were misidentified by the STAT; all four were in the DD/LI group but were identified as high-risk by the STAT. The 15 children with clinical diagnoses of PDD-NOS were split between the two STAT risk categories (see Table VI).

DISCUSSION

The results of this study suggest that the STAT demonstrates strong psychometric properties as a Level 2 screening measure for autism. The STAT has high sensitivity, specificity, and predictive value in identifying young children at risk for autism within a clinic-based sample. Moreover, STAT scores are reliable across examiners and across test–retest administrations. In contrast to most other screening measures that were developed for use with children at a single age, the STAT can be used effectively with children spanning a 12-month age range. This feature may increase its utility for clinical settings.

It is important to note that the STAT was designed as a screening measure specifically for autism, rather than for all autism spectrum disorders. In the development phase, children with PDD-NOS were excluded from the samples used to derive and validate the scoring algorithm. As a result, one might expect the STAT to be less accurate in identifying children at risk for a PDD-NOS diagnosis than for identifying those at risk for an autism diagnosis. This was, in fact, the case. Children with an ADOS-G or clinical diagnosis of PDD-NOS were not classified consistently into either risk category on the STAT. Thus, children with milder autism spectrum symptomatology may be less likely to be identified by this screening relative to those with autism.

The phenomenon of underidentification of children with milder symptoms is not unique to the STAT, as others have reported that children with milder variants of autism spectrum disorders are less likely to be identified by autism screening procedures (Filipek *et al.*, 1999). However, because the STAT was designed to differentiate children with autism from those with other developmental disorders, and not from those with typical development, even those children who screen low-risk on the STAT can still have significant language, cognitive, and/or social impairments that require intervention services. Research with the CHAT has indicated that identification of children with milder variants of autism spectrum disorders could be improved by using less stringent cutoff criteria (Baird *et al.*, 2000). Future work with the STAT might likewise focus on developing a scoring algorithm that is more sensitive to PDD-NOS without overidentifying children with nonspectrum disorders. However, the early identification of children with PDD-NOS may be further complicated by findings that PDD-NOS may be a less stable diagnosis for young children compared to a diagnosis of autism (Stone *et al.*, 1999).

The relation of mental age to performance on the STAT—and other early screening measures – is complex and warrants further investigation. Our research with children referred to a university-based diagnostic evaluation center has suggested that two-year-old children who receive diagnoses of autism, as a group, obtain lower cognitive scores than same-aged children with other (nongenetic) developmental disorders. It is not known whether this phenomenon is unique to the setting, whether it represents some type of clinician bias, or whether it

reflects the impact of autism symptomatology on early performance on cognitive measures. Regardless, the relative advantages and limitations of using naturally-occurring samples vs. mental-aged matched samples for developing screening measures require careful consideration. Because information about cognitive scores may not be available in screening settings, autism screening measures must have adequate psychometric properties for unmatched samples. At the same time, one would expect screening measures for autism to pick up on differences in key behavioral features between children, rather than differences in mental age alone. Matching groups on mental age is not an ideal solution to this dilemma, as it results in subsamples that are not representative of their respective populations (i.e., children with autism who have higher mean mental ages than in the naturally occurring sample and children with nonspectrum disorders who have lower mean mental ages than in the naturally occurring sample). Resolution of this issue awaits further research in different settings and with larger samples.

The interactive nature of the STAT has several advantages over parental report instruments. First, it provides a standard context for eliciting early social-communicative behaviors. Within this context, clinicians and service providers can observe children's play, imitation, and communication behaviors directly, rather than relying only on parental reports for this information. Second, direct interactions offer opportunities to observe more qualitative aspects of behavior (e.g., the extent to which eye contact and vocalizations are coordinated during interactions) that may be difficult to obtain through questionnaires. Third, because the STAT items tap important developmental skill areas (i.e., play, communication, and imitation), the qualitative information obtained during the screening can be used to inform intervention goals for individual children. A final advantage of interactive screening measures such as the STAT is the potential they hold for promoting community awareness and education about the early characteristics of autism. Increased awareness of the critical behaviors to look for in young children—and the types of activities that can be used to elicit those behaviors—may enable parents and community professionals to recognize and refer more at-risk children at younger ages.

However, there are some limitations to the interactive format of the STAT. For example, the STAT takes about 20 minutes to administer, thus incurring more professional time and expense relative to parental report measures. In addition, the STAT requires training in administration and scoring to ensure that it is used and interpreted in a manner consistent with its design. Because the key social-communicative behaviors that characterize young children with autism represent negative symptoms (i.e., the absence of expected behaviors), they can be more difficult to detect than positive symptoms, such as hand-flapping. STAT training workshops are used to help service providers and community professionals learn to recognize these more elusive behavioral deficits.

There are several directions for future research that would contribute to the further development of the STAT. Although the STAT was designed for use in community settings, its properties have been evaluated only in a university-based clinic setting to date. Moreover, in the present study the STAT was administered by examiners who received training on its administration and scoring under the direct supervision of a psychologist specializing in early diagnosis, which may have contributed to the strong psychometric findings. Additional research on the use of the STAT by community professionals and in community-based settings should be undertaken. Research examining the utility of different cutoff scores for different types of settings may also be useful. In addition, future work could evaluate the utility of the STAT in evaluating treatment effects and tracking developmental gains in social-communication skills in young children. The continuous total or domain scores (rather than the cutoff) might be used effectively toward this end. For instance, data from a recent study of children with autism suggest that play and imitation scores are more likely to increase from age 2 to age 3 than are directing attention scores (Turner, Pozdol, & Stone, 2002). Finally, preliminary work with the STAT has begun to examine its utility in identifying children with autism below the age of two years. We hope that this work will contribute to our understanding of the early features of autism and will enable more children to reap the benefits of early intervention.

ACKNOWLEDGMENTS

We are extremely grateful to the parents and children who generously donated their time to this project. We would like to acknowledge the valuable

contributions of many people to this research: Evon Lee and Linda Ashford for their help in recruiting families; Kathleen Mahn, Tanya Klepper, Sarah McGrath, and Tricia Lipani for their contributions to the assessment process; and Stephanie Milan and Irene Feurer for their assistance with statistical analyses. This research was supported in part by grants from the Department of Education (H324C990039), the National Institute of Mental Health (MH50620), and the National Institute of Child Health and Human Development (T3207226).

REFERENCES

American Academy of Pediatrics Committee on Children with Disabilities. (2001). The pediatrician's role in the diagnosis and management of autistic spectrum disorder in children. *Pediatrics*, *107*, 1221–1226.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington, D. C.: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (*4th ed., rev*.). Washington, D. C.: Author.

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 and Adolescent Psychiatry,* 39, 694–702.

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, 468–475.

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., Nightingale, N., Morgan, K. Drew, A., & Charman, T. (1996). Psychological markers in the detection of autism in a large population. *British Journal of Psychiatry*, 168, 158–163.

Bayley, N. (1993). *The Bayley Scales of Infant Development (2nd ed.)*. San Antonio, TX: Harcourt Brace.

Committee on Educational Interventions for Children with Autism. (2001). *Educating Children with Autism*. Washington, DC: National Academy Press.

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,* 40, 719–732.

Fewell, R. R. (1991). Play Assessment Scale (5th revision). Unpublished manuscript, University of Miami School of Medicine.

Filipek, P. A., Accardo, P. J., Ashwal, S., Baranek, G. T., Cook, E. H., Jr., Dawson, G., Gordon, B., Gravel, J. S., Johnson, C. P., Kallen, R. J., Levy, S. E., Minshew, N. J., Ozonoff, S., Prizant, B. M., Rapin, I., Rogers, S. J., Stone, W. L., Teplin, S. W., Tuchman, R. F., Volkmar, F. R. (2000). Practice parameter: Screening and diagnosis of autism: Report of the Quality Standards Subcommittee of the American Academy of Neurology and the Child Neurology Society. *Neurology*, 55, 468–479.

Filipek, P. A., Accardo, P. J., Baranek, G. T., Cook, E. H., Dawson, G., Gordon, B., Gravel, J. S., Johnson, C. P., Kallen, R. J., Levy, S. E., Minshew, N. J., Ozonoff, S., Prizant, B. M., Rapin, I., Rogers, S. J., Stone, W. L., Teplin, S., Tuchman, R. F., & Volkmar, F. R. (1999). The screening and diagnosis of autistic spectrum disorders. *Journal of Autism and Developmental Disorders*, 29, 439–484.

Glascoe, F. P. (2000). Evidence-based approach to developmental and behavioral surveillance using parents' concerns. *Child: Care, Health, and Development*, 26, 137–49.

Happe, F. (1994). Annotation: Current theories of autism: The "theory of mind" account and rival theories. *Journal of Child Psychology and Psychiatry*, 35, 215–229.

Harris, S. L., Handleman, J. S., Gordon, R., Kristoff, B., & Fuentes, F. (1991). Changes in cognitive and language functioning of preschool children with autism. *Journal of Autism and Developmental Disorders*, 21, 281–290.

Hedrick, D. L., Prather, E. M., & Tobin, A. R. (1984). *Sequenced Inventory of Communication Development (Rev. ed.)*. Seattle, WA: University of Washington Press.

Howlin, P., & Moore, A. (1997). Diagnosis in autism. *Autism*, 1, 135–162.

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

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., & Rutter, M. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30, 205–223.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology*, 55, 3–9.

Meisels, S. J., & Atkins-Burnett, S. (1994). *Developmental screening in early childhood: A guide*. Washington, DC: National Association for the Education of Young Children.

Mullen, E. M. (1995). *Mullen Scales of Early Learning: AGS Edition*. Circle Pines, MN: American Guidance Service.

Mundy, P., & Crowson, M. (1997). Joint attention and early social communication: Implications for research on intervention with autism. *Journal of Autism and Developmental Disorders*, 27, 653–676.

Newborg, J., Stock, J. R., Wnek, L., Guidubaldi, J., Scinicki, J. (1984). *Battelle Developmental Inventory*. Chicago: Riverside Publishing.

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.

Rogers, S. J., & Lewis, H. (1989). An effective day treatment model for young children with pervasive developmental disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, 28, 207–214.

Scambler, D., Rogers, S. J., & Wehner, E. A. (2001). Can the checklist for autism in toddlers differentiate young children with autism from those with developmental delays? *Journal of the American Academy of Child and Adolescent Psychiatry*, 40, 1457–1463.

Schopler, E., Reichler, R. J., & Renner, B. R. (1988). *The Childhood Autism Rating Scale*. Los Angeles: Western Psychological Services.

Siegel, B. (1996). Pervasive Developmental Disorders Screening Test. Unpublished manuscript, University of California at San Francisco.

Siegel, B. (1998, June). *Early screening and diagnosis in autistic spectrum disorders: The Pervasive Developmental Disorders Screening Test (PDDST)*. Paper presented at the NIH State

of the Science in Autism: Screening and Diagnosis Working Conference, Bethesda, MD.

Siegel, B., & Hayer, C. (1999, April). *Detection of autism in the 2nd and 3rd year: The Pervasive Developmental Disorders Screening Test (PDDST)*. Poster presented at the Biennial Meeting for the Society for Research in Child Development, Albuquerque, NM.

Siegel, B., Pliner, C., Eschler, J.,& Elliot, G. R. (1988). How children with autism are diagnosed: Difficulties in identification of children with multiple developmental delays. *Developmental and Behavioral Pediatrics*, 9, 199–204.

Stone, W.L., Coonrod, E.E., & Ousley, O.Y. (2000). Screening tool for autism in two-year-olds (STAT): Development and preliminary data. *Journal of Autism and Developmental Disorders*, 30, 607–612.

Stone, W. L., Hoffman, E. L., Lewis, S. E., & Ousley, O. Y. (1994). Early recognition of autism: Parental reports vs. clinical observation. Archives of Pediatric and *Adolescent Medicine,* 148, 174–179.

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

Stone, W. L. & Ousley, O. Y. (1997). STAT Manual: Screening tool for autism in two-year-olds. Unpublished manuscript, Vanderbilt University.

Stone, W. L., Ousley, O.Y., & Littleford, C. D. (1997). Motor imitation in young children with autism: What's the object? *Journal of Abnormal Psychology*, 25, 475–485.

Stone, W. L., Ousley, O. Y., Yoder, P. J., Hogan, K. L., & Hepburn, S. L. (1997). Nonverbal communication in two- and three-year-old children with autism. *Journal of Autism and Developmental Disabilities*, 27, 677–696.

Strain, P. S., Hoyson, M., & Jamieson, B. (1985). Normally developing preschoolers as intervention agents for autistic-like children: Effects on class deportment and social interaction. *Journal of the Division for Early Childhood, Spring*, 105–115.

Turner, L. M., Pozdol, S. L., & Stone, W. L. (2002, November). *Changes in social-communicative skills from age 2 to age 3 in children with autism*. Poster session presented at the International Meeting for Autism Research, Orlando, FL.