Published in final edited form as:
Autism Res. 2012 August ; 5(4): 267–276. doi:10.1002/aur.1238.

The ADOS Calibrated Severity Score: Relationship to Phenotypic Variables and Stability over Time

Stacy Shumway, **Cristan Farmer**, **Audrey Thurm**, **Lisa Joseph**, **David Black**, and **Christine Golden**
Pediatrics and Developmental Neuroscience Branch, National Institute of Mental Health, Maryland, USA

Abstract

LAY ABSTRACT—Measuring the severity of autism is a challenge for researchers and clinicians. Recently, Gotham et al., (2009) addressed this issue by creating calibrated severity scores (CSS) based on raw total scores of the Autism Diagnostic Observation Schedule (ADOS), a standardized measure commonly used in autism diagnosis. We tested the utility of the CSS by comparing its scores to raw scores from the ADOS in a sample of 368 children aged 2–12 years with autism, PDD-NOS, non-spectrum delay, or typical development. As expected, we found that the CSS were more uniformly distributed within diagnostic categories and across ADOS modules than were raw scores. In particular, CSS were useful in controlling for differences in verbal development. Follow-up evaluations showed good temporal stability of the CSS over 12–24 months in children with autism. The results of this study support the use of the CSS to measure the severity of the core symptoms of autism. Further research is needed to determine if the CSS can also be used to measure changes in symptom severity and serve as a tool for clinical research.

SCIENTIFIC ABSTRACT—Measurement of the severity of autism at a single time point, and over time, is a widespread challenge for researchers. Recently, Gotham et al., (2009) published a severity metric (calibrated severity scores; CSS) that takes into account age and language level and is based on raw total scores of the Autism Diagnostic Observation Schedule (ADOS), a standardized measure commonly used in autism diagnosis. The present study examined psychometric characteristics of the CSS compared to raw scores in an independent sample of 368 children aged 2–12 years with autism, PDD-NOS, non-spectrum delay, or typical development. Reflecting the intended calibration, the CSS were more uniformly distributed within clinical diagnostic category and across ADOS modules than were raw scores. Cross sectional analyses examining raw and severity scores and their relationships to participant characteristics revealed that verbal developmental level was a significant predictor of raw score, but accounted for significantly less variance in the CSS. Longitudinal analyses indicated overall stability of the CSS over 12–24 months in children with autism. Findings from this study support the use of the CSS as a more valid indicator of autism severity than the ADOS raw total score, and extend the literature by examining the stability over 12–24 months of the CSS in children with ASD.

Keywords

Autism Diagnostic Observation Schedule (ADOS); Autism Spectrum Disorders; Severity; Diagnosis

Correspondence concerning this article should be addressed to: Cristan Farmer, Pediatrics & Developmental Neuroscience Branch, National Institute of Mental Health, 10 Center Drive MSC 1C250, Bethesda, MD 20892, Phone (301) 435-3999, Fax (301) 402-8497, farmerca@mail.nih.gov.
Stacy Shumway is now in the Department of Communication Sciences and Disorders at the University of Utah, Salt Lake City, Utah, USA.

Few measures currently exist to assess the severity of the core symptoms of autism spectrum disorders [ASD; autistic disorder, pervasive developmental disorder-not otherwise specified (PDD-NOS), and Asperger's disorder]. These core symptoms (manifesting in varying degrees based on specific diagnosis) include impairments in social and communication domains, as well as restricted, repetitive, and stereotyped patterns of interest (APA, 2000). The measurement of ASD symptom severity is important for several reasons, including (a) the comparison of samples across studies, (b) tracking the course of the disorder over time, and (c) quantification of treatment effects. The need for a valid and reliable measure of ASD symptom severity is well-established. However, numerous factors have contributed to the dearth of such instruments. The inter- and intra-diagnostic heterogeneity of ASD makes the assessment of core symptom severity quite challenging (Lord et al., 2006). This heterogeneity is, in part, accounted for by differences in phenotypic characteristics, such as IQ, language level, and age, which influence the presentation of ASD and may mask the true severity of core ASD symptoms. Behaviors that are commonly associated with ASD (e.g., inattention, hyperactivity, irritability, aggression) also may confound ratings of core symptom severity. These associated behaviors can also influence ratings of ASD symptoms across time. Several existing clinician-rated measures do provide ratings of autism severity (e.g., Autism Behavior Checklist, Krug et al., 1980; Childhood Autism Rating Scale, Schopler et al., 1986; Gilliam Autism Rating Scale, Gilliam, 1995); however, data suggest that the scores generated by these instruments are not independent of phenotypic characteristics such as age, IQ, and language level (Gotham, Pickles, & Lord, 2009).

The current "gold standard" for ASD diagnosis is widely considered to consist of best estimate diagnosis based on administration of the (a) Autism Diagnostic Interview-Revised (Rutter, Le Couteur, & Lord, 2003), (b) Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000), and application of (c) the DSM-IV criteria (APA, 2000) (Risi et al., 2006). The ADOS is the only component of the gold standard based entirely on observation and interaction with the individual suspected of having ASD. The ADOS is a semi-structured observation of a child's behavior in response to a variety of social presses, with different modules tailored to the language level and age of the individual (e.g., Module 1 is for nonverbal and usually young children, while Module 4 is for adolescents and adults with fluent speech). These modules were created in order to control for the strong effects of expressive language level on behaviors such as reciprocal social interaction and play, which are coded to measure autism symptoms (Lord et al., 2000). Specifically the three core domains of ASD symptoms are assessed by all four ADOS modules: reciprocal social interaction, communication, and restricted and repetitive behaviors. The ADOS scoring algorithm aids in diagnosis and broadly classifies individuals into non-spectrum, PDD-NOS, and autism. The raw scores were not intended to be, nor are they suitable for, interpretation as a measure of autism severity (Gotham et al., 2009). For example, differences in the composition and number of items of the algorithms across the modules have limited the extent to which scores can be compared across modules, over time, and across individuals.

In 2007, Gotham et al. published revised ADOS algorithms (referred to in this paper as *raw scores*), which included the same number of items and similar content across modules. With the number and content of items in the revised algorithms similar across modules, the comparability of raw scores across modules was increased, facilitating use of the ADOS longitudinally to track severity of symptoms across individuals with varying language levels and within individuals when language level changes. It is especially important to be able to track change in autism severity across language level, given that children can show improvements in language, necessitating different module administration, but at the same time can show improvement, stability, or worsening of autism symptoms overall (Gotham et al., 2009).

The revised algorithms comprise items from two domains, Social Affect (SA) and Restricted and Repetitive Behaviors (RRB), and have shown increased specificity relative to the original algorithms (Gotham et al., 2007; Gotham et al,. 2008; Oosterling et al., 2010), which excluded RRB. Increased sensitivity between non-spectrum and ASD was also observed with the addition of the RRB domain; however, diagnostic sensitivity between autism and PDD-NOS was not altered by the addition of the RRB domain. In addition, participant characteristics, such as verbal IQ, were still found to be associated with raw scores within and across ADOS modules when using the revised algorithms (Gotham et al., 2007; Gotham et al., 2008), indicating the continued need for a severity metric of core autism symptoms.

Gotham et al. (2009) used the revised ADOS algorithms to develop a Calibrated Severity Score (CSS) to more accurately capture core autism symptom severity as measured on the ADOS. Data were culled from 1,807 ADOS administrations to 1,118 individuals with ASD between the ages of 2 and 16 years. Raw total scores (the sum of the SA and RRB domains) from the five revised ADOS algorithms were used to generate severity scores for: (a) Module 1 No Words, (b) Module 1 Some Words, (c) Module 2 younger than 5, (d) Module 2 older than 5, and (e) Module 3 (Module 4 was not included). The sample was divided amongst cells in an 18-cell module-by-age matrix. Within each of the 18 calibration cells, raw scores were standardized (calibrated) using a 10-point severity rating scale. CSS ratings 1–3 represented non-spectrum cases, 4–5 non-autism ASD, and 6–10 autism. The authors found that when compared with the raw score, the CSS was relatively independent of participant characteristics. Of primary importance was Verbal IQ, which was primarily responsible for the 43% of variance accounted for in the raw score versus 10% accounted for by an identical model predicting CSS.

To date, only one study has evaluated the psychometric characteristics of the CSS in an independent sample. de Bildt and colleagues (2011) examined 1,455 ADOS assessments from 1,248 Dutch children aged 2 to 16 years (*n* = 542 autism, *n* = 486 non-autism ASD, and *n* = 427 non-spectrum). The non-spectrum group was comprised of non-clinical cases as well as individuals with various psychiatric diagnoses. Importantly, some statistically significant differences were reported between the de Bildt and Gotham samples. Compared to Gotham et al. (2009), the de Bildt sample had a larger proportion of children with non-autism ASD, higher average verbal ability, and lower average RRB scores. The Gotham et al. (2009) CSS were replicated in the Dutch sample for Modules 1 and 3, but not for Module 2. Within Module 2 (in which 2 year olds were excluded from the sample), the CSS were more independent from cognitive functioning than the raw scores, but the distribution of individuals across calibration cells was not noticeably more uniform than the raw scores. The authors hypothesized that differences between the two study samples, including the larger proportion of children with non-autism ASD relative to the Gotham sample, likely contributed to differences in findings.

Given the limited replication of the original development paper for the CSS, it is essential that ADOS severity scores be generated from multiple sites, with differing research subject cohorts, in order to further assess the psychometric properties of the CSS across ADOS modules 1–3. Specifically, independence from language and IQ needs to be further considered, particularly in young children, who are often referred for an initial diagnostic evaluation due to language or other developmental delays. Finally, no study has yet to report on the stability of CSS over time. This is especially important to consider, as studies have started to utilize the CSS to quantify change in treatment trials (Dawson et al., 2010).

Current Study

Aims

ADOS CSS (Gotham et al., 2009) potentially allow for greater understanding of the manifestation of core autism symptom severity over time, independent of factors such as age, IQ, and language level. This type of data is crucial in the quest to understand and treat ASD. Thus, the goal of the current study was to evaluate the independence of the ADOS CSS from measures of related constructs in an independent sample of young children. Further, this study extends the existing literature by examining the stability of the CSS over 12-to-24 months in a subsample of young children with autism.

Hypotheses

We hypothesized that the ADOS CSS (a) will be distributed as expected in relation to clinical diagnosis; (b) will be less sensitive than raw scores to the effects of cognitive, language, and age variables; and (c) will be stable over a time period of one-to-two years, supporting the validity and reliability of the severity metric.

Methods

Participants

Participants included 368 children aged 2–12 years who screened for autism research in the Pediatrics and Developmental Neuroscience Branch of the National Institute of Mental Health. The sample comprised children with autism (AUT; *n* = 157, 85% male), PDD-NOS (*n* = 47, 87% male), non-spectrum delay (*n* = 95, 76% male) and typical development (TD; *n* = 69, 71% male). Children in the AUT, PDD-NOS, and non-spectrum delay groups were recruited based on referral for ASD or developmental delay; typically-developing children were recruited for lack of developmental concerns. The non-spectrum delay group was quite heterogeneous, and consisted of children with Intellectual Disability, language delays, and other specific delays/disorders (e.g., motor delays, psychiatric disorders). As described below, all children in the non-spectrum delay group completed a diagnostic evaluation, and ASD was ruled out for all children in this group. All diagnostic evaluations were conducted by an experienced, Ph.D.-level clinician who had met standard requirements for research reliability (80%) on diagnostic measures (i.e., ADI-R and ADOS). DSM-IV criteria, in conjunction with clinical judgment from all assessments and results obtained from the ADI-R and ADOS, were used in making diagnoses. As the revised ADOS algorithms became available during the course of the study, they were computed alongside the old algorithm.

Participant characteristics are summarized in Table 1. Average age ranged from 46 months (TD) to 55 months (AUT); the AUT group was significantly older than the non-spectrum delay and TD groups. The diagnostic groups differed significantly on cognitive, behavioral, and language measures; post-hoc tests revealed significantly lower scores in the AUT group relative to the other groups on most measures. ADOS scores are presented in Table 2. A subsample of children with AUT (*n* = 89; mean age 67 months) were re-assessed 12 to 24 months after the initial evaluation; these children did not differ significantly from the full AUT group on age, gender, or cognitive level.

Procedures

This research was approved by the Institutional Review Board of the Neuroscience Institutes at the National Institutes of Health. All participants completed a diagnostic evaluation conducted by a doctoral-level clinician that included assessment of autism symptoms, adaptive behavior, and cognitive functioning. To assess the core symptoms of autism, the ADI-R and the ADOS were administered; the Vineland Adaptive Behavior Scales, Second

Edition (Sparrow, Cichetti & Balla, 2005) was used as a measure of adaptive behavior; and either the Mullen Scales of Early Learning (Mullen, 1995) or the Differential Ability Scales, Second Edition (DAS-II; Elliott, 2007) was used to evaluate cognitive functioning. The Mullen was administered to the vast majority of participants (90% of the ASD group, 93% of the non-spectrum delay group, and 90% of the typical group received the Mullen). Developmental Quotients (DQ; age equivalents divided by chronological age multiplied by 100) were calculated in lieu of IQ, in order to provide a consistent metric between the two cognitive measures and to accommodate floor effects or children out of the age range for the developmentally appropriate test. The Mullen VDQ and NVDQ have been shown to have good convergent validity with the DAS-II (Bishop et al., 2011). Within the ASD group, the DQ was strongly correlated with the DAS-II standard scores (Verbal $r = 0.95$, $p < 0.001$; Nonverbal $r = 0.88$, $p < 0.001$). Measures of expressive and receptive language were also completed on a subset of children in the AUT group, according to ability level. These included the Expressive One Word Picture Vocabulary Test (EOWPVT; Brownell, 2000), the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997); and the MacArthur-Bates Communicative Developmental Inventories (MCDI; Fenson et al., 1993). Two language variables (number of words produced and number of words understood) were computed from one of two versions of the MCDI. Both words produced and words understood were calculated from the MCDI Words and Gestures form ($n = 85$). Words produced was also calculated for children who received the MCDI Words and Sentences (based on age and language level; $n = 36$). The MCDI Words and Sentences form does not measure the number of words understood, so this variable was not present for children who received this form. During the diagnostic evaluation, the PPVT was administered to children able to point or indicate a word when named ($n = 53$), and the EOWPVT was administered to children able to name pictures ($n = 44$). A subset of the AUT group ($n = 89$) was administered a second ADOS between 12 and 24 months [$M(SD) = 17.27(4.34)$ months] from the initial evaluation (Time 2). When possible (i.e., in 91% of cases), raters at Time 1 and Time 2 were different, and ADOS scores from previous evaluations were not consulted prior to Time 2.

Analyses

CSS were calculated using the guidelines published in Gotham et al. (2009), wherein the ADOS module and child's age are used to convert the total ADOS raw score (SA+RRB). We replicated the multiple linear regression models used in Gotham et al. (2009) to assess the relative contributions of demographic and cognitive variables (nonverbal and verbal DQ) to the two dependent variables: ADOS raw score and CSS. We also examined specific language variables to further delineate how specific aspects of language affect ADOS raw scores versus CSS. A method for comparing the effects of a set of independent variables on two different outcomes in one sample was used in which all variables are standardized before performing the first regression (Cohen, Cohen, West, & Aiken, 2003). The predicted values ($\hat{y}_{dv1}$) are subtracted from the observed values (y_{dv2}) of the second dependent variable. This difference ($y_{dv2} - \hat{y}_{dv1}$) is then regressed on the set of independent variables. The overall test of R^2 indicates whether the effect of the independent variables differs between the two dependent variables. Fisher's Z transformation for correlated correlation coefficients (Meng et al., 1992) was used to statistically compare the magnitude of simple correlations. Repeated-measures ANOVA was used to evaluate change in ADOS scores between initial evaluation and Time 2 (12–24 months post initial evaluation). Alpha was set at $p < 0.01$. SPSS Version 19 was used for all statistical calculations.

Results

Figures 1 and 2 show the distributions of ADOS raw scores and CSS (respectively) by diagnostic groups. As the ADOS is part of the diagnostic gold standard contributing to clinical diagnosis, it was expected that ADOS CSS classification and clinical diagnosis would be closely related.

An important impetus for the development of ADOS CSS was the need for comparison across modules in both cross-sectional and longitudinal data. Figures 3 and 4 illustrate the effect of ADOS module on raw scores and CSS, respectively. Within each diagnostic group, a clear pattern of lower raw score with higher module is observed. However, the CSS is relatively immune to this effect; scores are uniformly distributed across modules in the cross-sectional sample. These patterns are likely explained by the impact of subject characteristics such as age, verbal ability, and cognitive level on module selection. Thus, the effects of these variables were directly evaluated.

The effect of cognitive and demographic variables within the ASD group (AUT + PDD-NOS; $n = 179$ with complete data on all variables) was evaluated using multiple linear regression (see Table 3). Following the methods of Gotham et al. (2009), the Verbal Developmental Quotient (VDQ) and Nonverbal Developmental Quotient (NVDQ) were added in block one, followed by the remaining variables in block two. The full model accounted for 56% of the variance in ADOS raw score, driven primarily by a significant effect of VDQ ($\beta = -0.60$, $p < 0.001$). In contrast, the full model accounted for 18% of the variance in CSS. Although no variables reached statistical significance ($p < 0.01$), VDQ did approach significance in predicting CSS (VDQ $\beta = -0.27$, $p = 0.02$). Following methodology outlined by Cohen et al. (2003), the difference between models was tested. The difference in variance explained for the two scores was significant [$F(3, 175) = 24.33$, $p < 0.001$], driven by a stronger effect of VDQ in the raw score than in the CSS ($t = 4.62$, $p < 0.001$).

While VDQ is in some ways an index of language ability, it is not synonymous. Thus, in order to further delineate how specific aspects of language affect CSS versus ADOS raw scores, data from specific language measures (PPVT, EOWPVT, and MCDI words produced and words understood) were examined from a subset of the children with AUT. As with the overall AUT sample, in which 86% of children were administered ADOS module 1, the majority of children with specific language data were administered ADOS module 1 (EOWPVT = 68.2% of children; PPVT = 67.9%; MCDI words understood = 96.5%; MCDI words produced = 88.4%). The following analyses reflect only those children that completed each of the measures. Individually, each language variable significantly predicted ADOS raw scores: R^2 values were 0.18 (MCDI Words Understood), 0.32 (MCDI Words Produced), 0.38 (EOWPVT), and 0.44 (PPVT). For the CSS, MCDI Words Produced, PPVT, and EOWPVT were significant predictors (respective $R^2 =$ 0.07, 0.15, 0.21). MCDI Words Understood was not a significant predictor of CSS ($R^2 =0.01$). In all cases, the correlation between each language variable and the raw score was significantly ($p < 0.01$) larger than that between the language variable and the CSS. Confidence intervals (95%) for the differences were as follows: EOWPVT: (0.04, 0.38); PPVT: (0.18, 0.51); MCDI Words Understood: (0.19, 0.46); MCDI Words Produced: (0.23, 0.51).

As in Gotham et al. (2009), we examined the relationship between cognitive and demographic variables and ADOS scores within the combined ASD and non-spectrum delay group. Again, the full model accounted for more variance in the raw score (43%) than in the CSS (22%), in both cases accounted for by the effects of VDQ. However, the addition of the non-spectrum delay group weakened the relationship between the language variables and the

ADOS scores; this was due to non-significant and very weak ($|r| < 0.20$) correlations in the non-spectrum delay group.

A subset of 89 children from the AUT group was reassessed at an interval between 12 and 24 months [*M(SD)* = 17.27(4.34)]. As predicted, scores did not differ significantly between initial testing and Time 2 for the ADOS CSS [repeated-measures ANOVA; $F(1, 88) = 1.55$, $p = 0.22$] or for the raw scores [$F(1, 88) = 2.87$, $p = 0.09$] (Cohen's $d_{CSS} = 0.27$; $d_{raw} = 0.36$). For over half of the participants ($n = 57$, 64%), CSS at Time 2 were within one point of the CSS at the initial evaluation (see Figure 5). For nearly a quarter of participants ($n = 21$, 23.6%), CSS at Time 2 were improved (reduced) by more than one point, while 11 (12.4%) participants' CSS worsened (increased) by more than one point. A small proportion of the sample received a different module at Time 2 than at initial testing ($n = 20$, 22%), an expected result of the substantial period of time between administrations. Most of these children moved from Module 1 to Module 2 ($n = 16$). Within this group, the change in CSS at Time 2 was very similar to that in the whole sample: 62.6% of CSS were within 1 point, 18.8% worsened by more than 1 point, and 18.8% improved by more than 1 point. Change in CSS was not significantly predicted by Time 1 age, NVDQ, VDQ, Vineland Adaptive Behavior Composite, or any language variable. Further, change in CSS was not correlated with change between Time 1 and Time 2 on any of the aforementioned variables. Overall, both the CSS and raw scores appeared to be stable over a 12-to-24 month period for most individuals; though the change in these variables was not statistically significant, the effect was slightly larger for the raw scores than the severity score.

Discussion

The purpose of this study was to extend the small set of literature on the ADOS CSS by evaluating the CSS in an independent sample of young children, and to examine the stability of the CSS over 12–24 months in children with autism. The ADOS raw total score has been shown to be significantly influenced by individual characteristics, such as cognitive level (Gotham et al., 2009), and is therefore problematic as an indicator of autism symptom severity. Gotham et al. (2009) presented an alternative, the CSS, which is derived from the ADOS raw total score using age and language level. The CSS for Modules 1 and 3 have been validated in a Dutch sample (de Bildt et al., 2011), but no other data on the CSS are available. Results of the current study provide data indicating that the CSS is a more valid indicator of autism symptom severity than the ADOS raw total, and is stable over 12–24 months.

Upon initial inspection of the current data, the distribution of the CSS between diagnostic groups appears to be similar to the distribution of the raw scores, with relatively clear separation between diagnostic categories. This is consistent with our hypotheses; the CSS were not expected to discriminate between diagnostic groups any differently than the raw scores. However, an important effect becomes obvious when considering scores by module. Within all diagnostic groups, a trend for lower raw scores (i.e., less severe symptoms) in the more advanced ADOS modules was observed. This pattern was consistent with the observed distributions by calibration cell in both Gotham et al. (2009) and de Bildt et al. (2011). Characteristics such as age and language level have an obvious impact on the expression of autism symptoms, though it is *not* the case that an older child with fluent speech will necessarily have reduced symptom severity. It is here that the strength of the CSS is demonstrated: within diagnostic group, average severity scores are stable across module. This finding has the important implication of allowing inter-module comparison within cross-sectional data.

The stability of the CSS across modules is likely due to the fact that CSS were relatively impervious to the effect of verbal cognitive ability, which was responsible for a great deal of the variance in ADOS raw scores. Gotham et al. (2009) found that phenotypic and demographic variables (primarily verbal IQ) accounted for 43% of the variance in the raw score, while these variables only explained 10% of the variance in CSS. de Bildt et al. (2011) found a similar pattern of association. Our data supported these findings; we found levels of explained variance nearly identical to those reported by Gotham. Further, we determined that the difference in variance explained was statistically significant, with a more robust effect of VDQ upon the raw score than on the CSS. This suggests that the CSS captures autism symptom severity as measured by the ADOS more cleanly than does the raw score.

Given the significant influence of verbal cognitive ability (VDQ) on ADOS raw total scores, and its decreased relationship with the CSS found by Gotham et al. (2009), the current study further explored the relationship between language and ADOS scores (raw and CSS) to illustrate how specific aspects of language (related to vocabulary comprehension and production) may affect ADOS scores. As hypothesized, the specific language measures were more strongly associated with raw scores than with CSS, with the correlation between each language variable and the raw score being significantly larger than that between the language variable and the CSS. Taken together, these results mirror Gotham et al. (2009) in showing that ADOS raw scores were more sensitive to verbal ability than the CSS. Although the correlations between the individual language variables and the ADOS CSS decreased compared to the raw scores, most of the language variables examined in this study were still significant predictors of the CSS. As previously noted, while VDQ is an index of language ability, it is not synonymous with language ability, and results of this study show that some aspects of language ability are perhaps more entangled with the measurement of autism severity than others (Lord et al., 2000). This study examined language measures related to vocabulary comprehension and production; however, continued research is needed to examine other aspects of language ability in relation to measurement of autism severity.

In the original calibration study, Gotham et al. (2009) presented case summaries consisting of longitudinal data for four children with ASD to provide a preliminary view of patterns of change in scores over time (age range was approximately 3–13 years). Patterns indicated that CSS were relatively stable in two of the four children, but showed more variability over time for the other two children (one increasing and the other decreasing). Findings from the present study that included a group of children with autism showed relative stability in CSS over a 12-to-24 month period for the majority of children with autism. Although the severity of the individual symptom domains of ASD (e.g., social, language, and restricted/repetitive behavior) is known to vary over time (Charman et al., 2005), categorical diagnosis is thought to be relatively stable (e.g., Lord et al., 2006; Stone et al., 1999). Thus, a valid measure of *overall* severity would be expected to remain stable over a period of 12–24 months, as found in the current study. Although change in both the raw scores and CSS was non-significant, the effect size was slightly larger in the raw scores than in the CSS. This may suggest that the CSS are somewhat more stable than raw scores; however, without some index of "true" change in autism symptoms, it is impossible to interpret this observation. That is, there is currently no other reliable and valid measure of autism symptom severity with which to compare the ADOS CSS. Regardless, the apparent stability of the CSS over 12–24 months found in this study, in tandem with the relative independence from the effect of developmental level, has important implications for both clinical and research applications. Continued research is needed to parse out the effects of participant characteristics on the stability of CSS over time, including the effects of treatment. Similar to the studies conducted to evaluate diagnostic stability (e.g., Lord et al., 2006; Stone et al., 1999), treatment was not controlled in the current study, and was allowed to reflect what children were receiving in the community.

Limitations and Future Directions

A clear need exists to examine the recently developed ADOS CSS across multiple sites and samples. This study evaluated the CSS in a relatively large, independent sample. However, as the authors of the original study stated, the severity metric should be recalibrated as larger and more representative samples become available (Gotham et al. 2009). The young age as well as limited communication/language ability in some members of the present sample precluded complete data on some language measures (e.g., EOWPVT), which prevented generalization of some of the findings in this report to that segment of the population. Although well-characterized, the small sample sizes for specific modules, language levels and age should also be considered when generalizing the results of this study. This is the first study to report on the stability of the CSS over time; however, Time 2 data were collected on only a portion of the sample, and the time to follow up was not uniform, spanning 12 to 24 months. Future research following larger samples over a longer period of time, in addition to examining potential correlates of change in severity, will help elucidate the stability of the CSS and will be important in determining when and how severity scores change in individuals with ASD. In addition, further analysis of larger samples that compare ADOS severity with other measures of autism severity will be necessary for the convergent validity of the CSS to be established. This is especially important because the heterogeneity in autism symptoms, particularly in the language domain, may differentially impact measures often used in treatment studies that capture overall severity of impairment and/or functioning (e.g., Clinical Global Impressions [Guy, 1976], Children’s Global Assessment Scale [Schaffer et al. 1983], Ohio Autism Clinical Impressions Scale [see Hurt, Arnold, & Aman 2010]). Thus, future research should also specifically compare the CSS to these other types of measures cross-sectionally and over time.

Conclusion

The ADOS Calibrated Severity Score proposed by Gotham et al. (2009) has good psychometric properties; it discriminates well between diagnostic groups, is less influenced by verbal cognitive abilities than is the raw score, and is stable over a period of 12–24 months. The CSS is a good option for characterizing samples and exploring relationships between autism severity and external variables, and may be especially useful in longitudinal studies where comparison between modules is necessary. Although this measure of autism symptom severity appears to have good psychometric properties, the available literature does suggest that autism symptoms may be best measured by domain rather than in aggregate in order to best capture variation. Still, in practice, the ADOS raw score is frequently used as the sole measure of symptom severity. Thus, findings from this study add to the literature in suggesting that the use of the CSS would represent an improvement upon current practice in the field.

Acknowledgments

This research was supported by the Intramural Program of the National Institute of Mental Health (NIMH). The views expressed in this article do not necessarily represent the views of the NIMH, NIH, HHS, or the United Stated Government. The authors extend our gratitude for the children and their families who volunteered their time and efforts during this research.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4. Washington, DC: Author; 2000. Text Revision

Bishop S, Guthrie W, Coffing M, Lord C. Covergent validity of the Mullen Scales of Early Learning and the Differential Ability Scales in children with autism spectrum disorders. American Journal on Intellectual and Developmental Disabilities. 2011; 116:331–343. [PubMed: 21905802]

Brownell, R. Expressive One-Word Picture Vocabulary Test. Novato, CA: Academic Therapy Publications; 2000.

Charman T, Taylor E, Drew A, Cockerill H, Brown J, Baird G. 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. 2005; 46:500–513. [PubMed: 15845130]

Cohen, J.; Cohen, P.; West, S.; Aiken, L. Applied multiple regression/correlation analysis for the behavioral sciences. 3. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2003.

Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J, et al. Randomized, controlled trial of an intervention for toddlers with autism: The Early Start Denver Model. Pediatrics. 2009; 125:e17–e23. [PubMed: 19948568]

de Bildt A, Oosterling I, van Lang N, Sytema S, Minderaa R, van Engeland H, et al. Standardized ADOS scores: Measuring severity of autism spectrum disorders in a Dutch sample. Journal of Autism and Developmental Disorders. 2011; 41:311–319. [PubMed: 20617374]

Dunn, L.; Dunn, L. Peabody Picture Vocabulary Test. 3. Circle Pines, MN: American Guidance Service Publishing; 1997.

Elliott, C. Introductory and technical handbook. 2. San Antonio, TX: The Psychological Corporation; 2007. Differential Ability Scales.

Fenson, L.; Dale, P.; Reznick, S.; Bates, E.; Pethick, S.; Hartung, J., et al. MacArthur Communicative Developmental Inventories: User's guide and technical manual. San Diego, CA: Singular/ Thomson Learning; 1993.

Gilliam, J. Gilliam Autism Rating Scale. Austin, TX: Pro-Ed; 1995.

Gotham K, Pickles A, Lord C. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. Journal of Autism and Developmental Disorders. 2009; 39:693–705. [PubMed: 19082876]

Gotham K, Risi S, Dawson G, Tager-Flusberg H, Joseph R, Carter A, et al. A replication of the Autism Diagnostic Observation Schedule (ADOS) revised algorithms. Journal of the American Academy of Child and Adolescent Psychiatry. 2008; 47:641–651.

Gotham K, Risi S, Pickles A, Lord C. The Autism Diagnostic Observation Schedule: Revised algorithms for improved diagnostic validity. Journal of Autism and Developmental Disorders. 2007; 37:613–627. [PubMed: 17180459]

Guy, W. Clinical Global Impressions. ECDEU Assessment Manual for Psychopharmacology. Rockville, MD: National Institutes of Mental Health; 1976.

Hurt, B.; Arnold, L.; Aman, M. Clinical instruments and scales in pediatric psychopharmacology. In: Martin, A.; Scahill, L.; Kratochvil, C., editors. Pediatric Psychopharmacology: Principles and Practice. 2. New York, NY: Oxford University Press; 2010.

Krug D, Arick J, Almond P. Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. Journal of Child Psychology and Psychiatry. 1980; 21:221–229. [PubMed: 7430288]

Lord C, Risi S, DiLavore P, Shulman C, Thurm A, Pickles A. Autism from 2 to 9 years of age. Archives of General Psychiatry. 2006; 63:694–701. [PubMed: 16754843]

Lord C, Risi S, Lambrecht L, Cook E Jr, Leventhal B, DiLavore P, et al. 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. 2000; 30:205–223. [PubMed: 11055457]

Meng X, Rosenthal R, Rubin D. Comparing correlated correlation coefficients. Psychological Bulletin. 1992; 111(1):172–175.

Mullen, EM. Mullen Scales of Early Learning. Circle Pines, MN: American Guidance Service; 1995.

Oosterling I, Roos S, de Bildt A, Rommelse N, de Jonge M, Visser J, et al. Improved diagnostic validity of the ADOS revised algorithms: A replication study in an independent sample. Journal of Autism and Developmental Disorders. 2010; 40:689–703. [PubMed: 20148299]

Risi S, Lord C, Gotham K, Corsello C, Chrysler C, Szatmari P, et al. Combining information from multiple sources in the diagnosis of autism spectrum disorders. Journal of the American Academy of Child and Adolescent Psychiatry. 2006; 45:1094–1103. [PubMed: 16926617]

Rutter, M.; Le Couteur, A.; Lord, C. Autism Diagnostic Interview-Revised. Los Angeles: Western Psychological Services; 2003.

Schopler, E.; Reichler, R.; Renner, B. The Child Autism Rating Scale (CARS): For diagnostic screening and classification of autism. New York: Irvington; 1986.

Schaffer D, Gould M, Brasic J, et al. A children's global assessment scale (CGAS). Archives of General Psychiatry. 1983; 40:1228–1231. [PubMed: 6639293]

Sparrow, S.; Cicchetti, D.; Balla, D. Vineland Adaptive Behavior Scales, Survey Interview Form/ Caregiver Rating Form. 2. Livonia, MN: Pearson Assessments; 2005.

Stone W, Lee E, Ashford L, Brissie J, Hepburn S, Coonrod E, et al. Can autism be diagnosed accurately in children under 3 years? Journal of Child Psychology and Psychiatry. 1999; 40:219–226. [PubMed: 10188704]

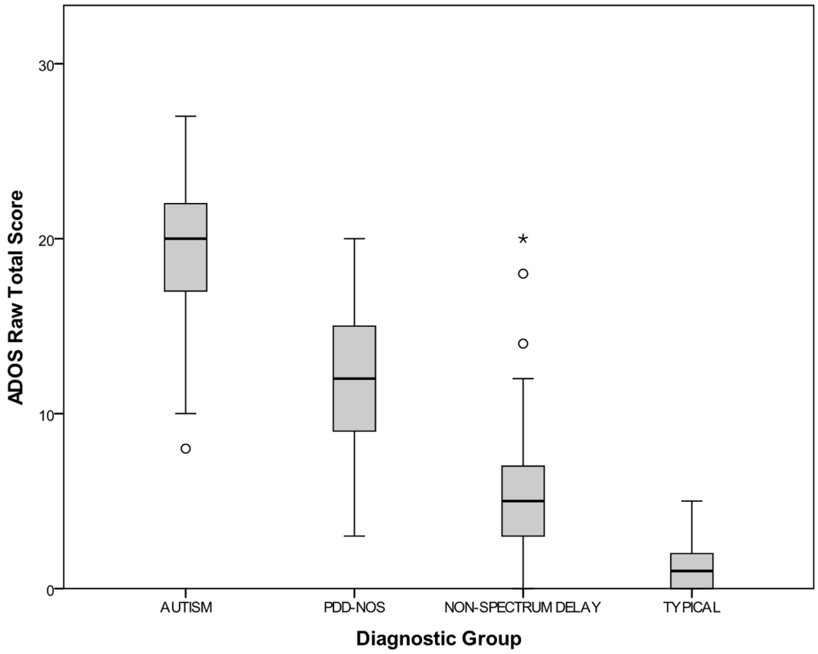

Figure 1.

ADOS Raw Scores by Diagnostic Groups: Autism (*n* = 154), PDD-NOS (*n* = 47), Non-spectrum Delay (*n* = 95), and Typically Developing (*n* = 69). Box and whiskers represent the 25th, 50th, and 75th percentiles and range, respectively.

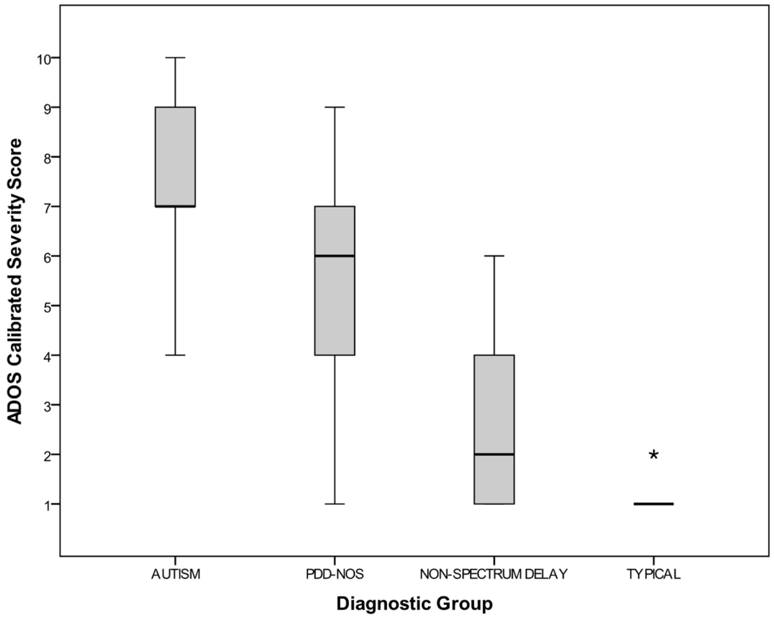

Figure 2.

ADOS Calibrated Severity Scores (CSS) by Diagnostic Groups: Autism (*n* = 154), PDD-NOS (*n* = 47), Non-spectrum Delay (*n* = 95), and Typically Developing (*n* = 69). Box and whiskers represent the 25th, 50th, and 75th percentiles and range, respectively.

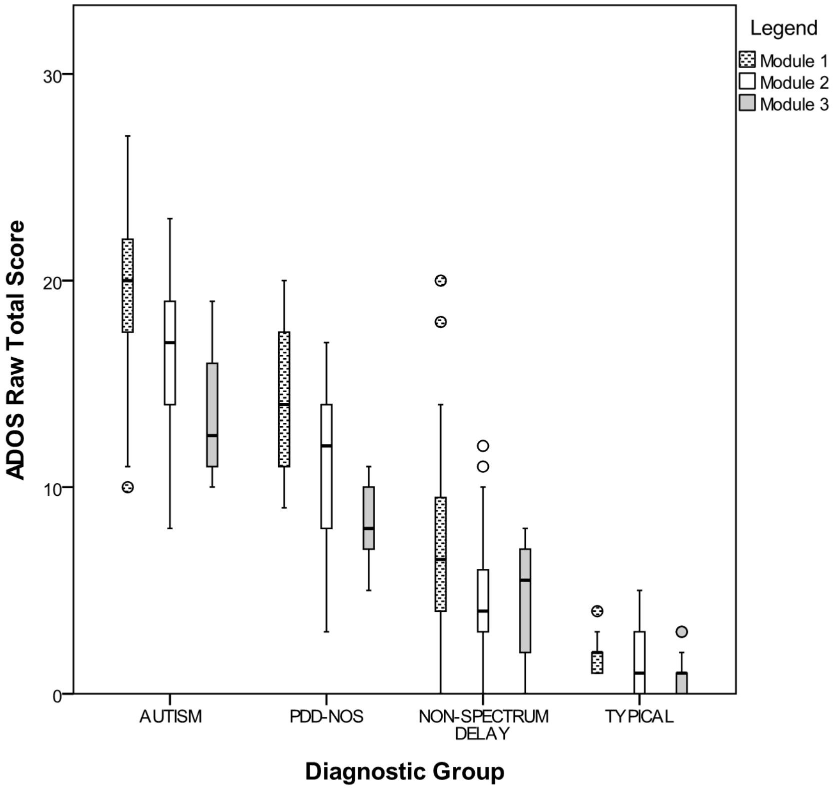

Figure 3.

ADOS Raw Scores by Module and Diagnostic Groups: Autism (*n* = 154), PDD-NOS (*n* = 47), Non-spectrum Delay (*n* = 95), and Typically Developing (*n* = 69). Box and whiskers represent the 25th, 50th, and 75th percentiles and range, respectively.

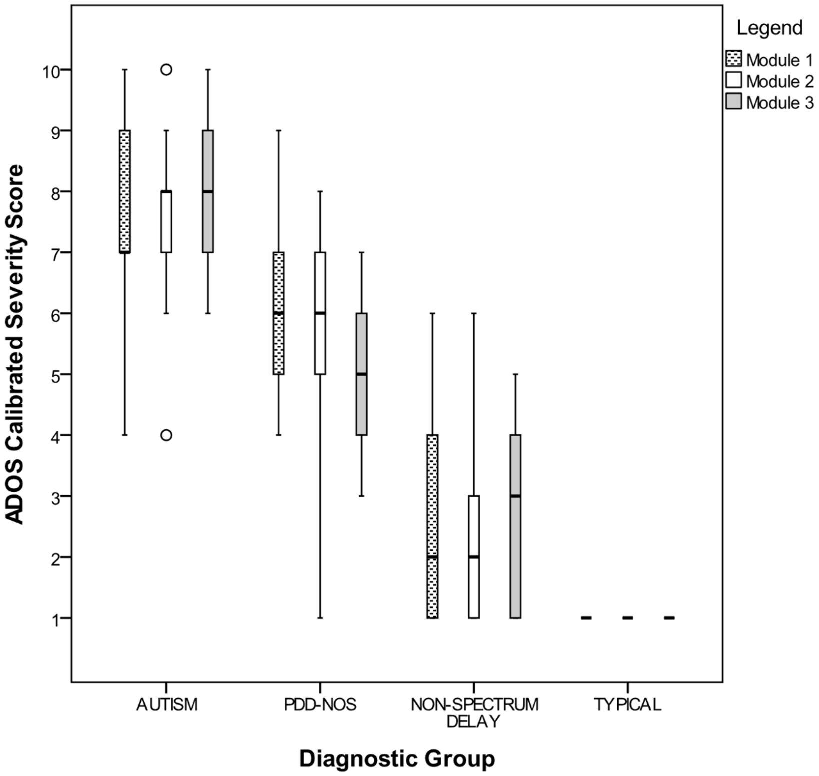

Figure 4.

ADOS Calibrated Severity Scores by Module and Diagnostic Groups: Autism (*n* = 154), PDD-NOS (*n* = 47), Non-spectrum Delay (*n* = 95), and Typically Developing (*n* = 69). Box and whiskers represent the 25th, 50th, and 75th percentiles and range, respectively.

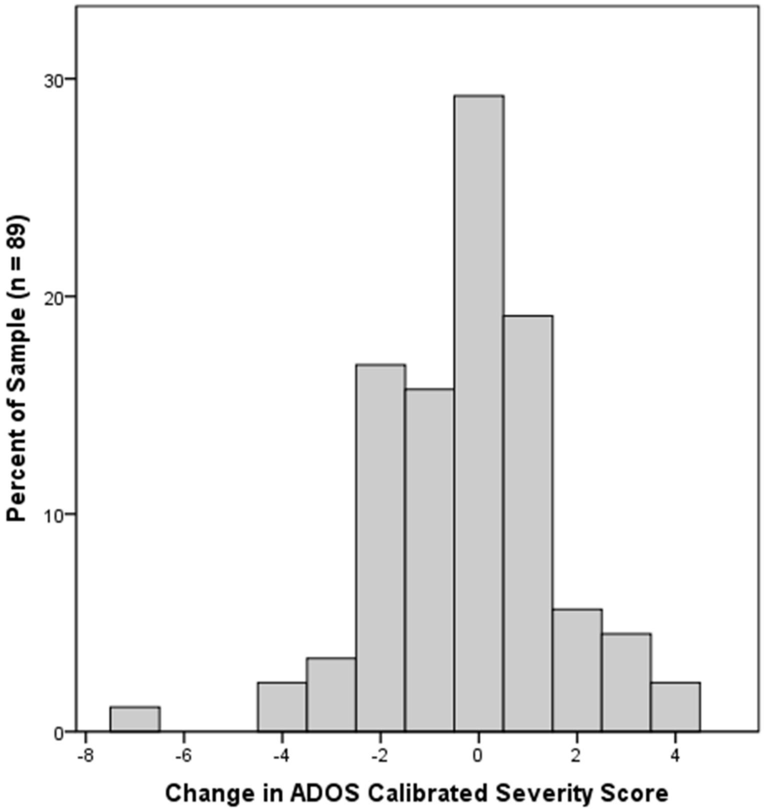

Figure 5.
Histogram of change in ADOS Calibrated Severity Score at 12–24 months.

Table 1

Sample Description

	Autism		PDD-NOS		Non-spectrum Delay		Typical	
	n	***M(SD)***	***n***	***M(SD)***	***n***	***M(SD)***	***n***	***M(SD)***
Age (months)	157	54.66(23.96)^{a}	47	51.77(21.92)^{a,b}	95	47.17(14.00)^{b}	69	46.31(18.75)^{b}
Nonverbal DQ	153	58.92(19.44)^{a}	47	79.25(18.54)^{b}	95	72.41(18.01)^{b}	69	109.34(13.76)^{c}
Verbal DQ	153	41.51(20.75)^{a}	47	71.31(22.10)^{b}	95	66.77(19.18)^{b}	69	108.37(14.85)^{c}
Vineland ABC	157	65.69(9.61)^{a}	47	75.96(7.89)^{b}	95	77.28(9.96)^{c}	69	102.62(8.95)^{d}
EOWPVT SS	44	68.75(16.72)^{a}	33	85.48(18.08)^{b}	67	83.45(17.59)^{b}	56	107.54(18.42)^{c}
PPVT SS	53	64.58(22.20)^{a}	28	85.39(20.60)^{b}	67	82.28(19.63)^{b}	49	110.82(13.41)^{c}
MCDI Understood	85	135.84(96.81)^{a}	15	196.07(137.25)^{a,b}	30	208.43(107.05)^{b}	4	381.00(26.15)^{c}
MCDI Produced	121	119.81(164.74)^{a}	36	329.33(228.30)^{b}	84	280.17(222.08)^{b}	40	539.28(140.39)^{c}

Note. Superscripts denote means that differ significantly at $p < 0.01$.; DQ = Developmental Quotient; ABC = Adaptive Behavior Composite; EOWPVT SS = Expressive One Word Picture Vocabulary Test Total Standard Score; PPVT SS = Peabody Picture Vocabulary Test Total Standard Score; MCDI = MacArthur-Bates Communicative Development Inventories.

Table 2

ADOS Results by Clinical Diagnostic Category

	Autism (n = 157)	PDD-NOS (n = 47)	Non-spectrum Delay (n = 95)	Typical (n = 69)
	***n*(%)**	***n*(%)**	***n*(%)**	***n*(%)**
Module 1	136(86)	19(40)	44(47)	9(13)
Module 2	17(11)	19(40)	45(47)	34(49)
Module 3	5(3)	9(20)	6(6)	26(38)
	***M*(*SD*)**	***M*(*SD*)**	***M*(*SD*)**	***M*(*SD*)**
ADOS Raw Score	19.44(3.75)^{a}	11.77(4.30)^{b}	5.69(3.72)^{c}	1.45(1.42)^{d}
ADOS CSS	7.64(1.39)^{a}	5.51(1.83)^{b}	2.57(1.58)^{c}	1.13(0.34)^{d}

Note. Superscripts denote means that differ significantly at $p < 0.01$. ADOS = Autism Diagnostic Observation Schedule; CSS = Calibrated Severity Score.

Table 3

Multiple Linear Regression Models for ADOS Raw Score and Calibrated Severity Score, ASD Only

	R^2	ΔF	*df*	*B*	*SE B*	β	*p*
***DV* = CSS (*n* = 177[a])**							
Step 1	0.176	18.73	2,176				
Constant				8.80	0.42		
VDQ				−0.02	0.01	−0.27	0.02
NVDQ				−0.01	0.01	−0.17	0.15
Step 2	0.182	0.317	4,172				
Constant				9.59	0.77		
VDQ				−0.02	0.01	−0.25	0.04
NVDQ				−0.02	0.01	−0.21	0.10
Age				−0.07	0.07	−0.07	0.38
Sex				−0.14	0.35	−0.03	0.68
Maternal							0.58
Education[b]				0.07	0.13	0.04	
Race[c]				0.01	0.14	0.01	0.94
***DV* = Raw Score (*n* = 177[a])**							
Step 1	0.541	103.89	2,176				
Constant				25.67	0.88		
VDQ				−0.13	0.02	−0.64	<0.001
NVDQ				−0.03	0.02	−0.11	0.21
Step 2	0.559	1.686	4,172				
Constant				28.11	1.61		
VDQ				−0.12	0.02	−0.60	<0.001
NVDQ				−0.04	0.02	−0.18	0.06
Age				−0.38	0.15	−0.13	0.02
Sex				−0.07	0.73	−0.01	0.93
Education[b]				0.20	0.27	0.04	0.47
Race[c]				0.11	0.02	0.02	0.70

Note. DV = dependent variable; CSS = calibrated severity score; VDQ = verbal developmental quotient; NVDQ = non-verbal developmental quotient. Alpha was set at $p < 0.01$ as partial correction for multiple comparisons.

[a] n reflects subjects with non-missing values for all variables.

[b] Maternal Education was dichotomized to Graduate School versus No Graduate School.

[c] Race was dichotomized to White versus Not White.