J Autism Dev Disord (2015) 45:2704–2720
DOI 10.1007/s10803-015-2432-7

ORIGINAL PAPER

The Autism Diagnostic Observation Schedule, Toddler Module: Standardized Severity Scores

Amy N. Esler[1] · Vanessa Hus Bal[2,3,8] · Whitney Guthrie[4] · Amy Wetherby[5] · Susan Ellis Weismer[6] · Catherine Lord[7]

Published online: 2 April 2015

Abstract Standardized calibrated severity scores (CSS) have been created for Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2) Modules 1–4 as a metric of the relative severity of autism-specific behaviors. Total and domain CSS were created for the Toddler Module to facilitate comparison to other modules. Analyses included 388 children with ASD age 12–30 months and were replicated on 435 repeated assessments from 127 children with ASD. Compared to raw scores, associations between total and domain CSS and participant characteristics were reduced in the original sample. Verbal IQ effects on Social Affect-CSS were not reduced in the replication sample. Toddler Module CSS increases comparability of ADOS-2 scores across modules and allows studies of symptom trajectories to extend to earlier ages.

Keywords Autism spectrum disorder · Autism diagnostic observation schedule · Severity · Toddlers · Social affect · Restricted and repetitive behavior

Electronic supplementary material The online version of this article (doi:10.1007/s10803-015-2432-7) contains supplementary material, which is available to authorized users.

✉ Amy N. Esler
esle0007@umn.edu

[1] Department of Pediatrics, University of Minnesota, 717 Delaware St SE, Minneapolis, MN 55414, USA

[2] Department of Psychiatry, University of California San Francisco, 1550 4th Street, San Francisco, CA 94158, USA

[3] Department of Psychology, University of Michigan, Ann Arbor, MI, USA

[4] Department of Clinical Psychology, Autism Institute, Florida State University, 1940 N. Monroe Street Suite 72, Tallahassee, FL 32303, USA

[5] Department of Psychology, Autism Institute, College of Medicine, Florida State University, 1940 N. Monroe Street Suite 72, Tallahassee, FL 32303, USA

[6] Waisman Center, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, WI 53705, USA

[7] Center for Autism and the Developing Brain, Weill Cornell Medical College, 21 Bloomingdale Rd., White Plains, NY 10605, USA

[8] Present Address: Department of Psychiatry, University of California San Francisco, San Francisco, CA, USA

Introduction

The development of early screening and diagnostic tools for autism spectrum disorder (ASD) has allowed diagnoses to occur at younger ages (Dawson and Bernier 2013; Guthrie et al. 2013; Woolfenden et al. 2012). The Toddler Module of the Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2; Lord et al. 2012a) has demonstrated high levels of reliability and validity as a diagnostic tool for ASD in children age 12–30 months (Guthrie et al. 2013; Luyster et al. 2009). However, social communication and behavioral patterns in children who develop ASD can be variable early in the second year of life (Bryson et al. 2007; Landa et al. 2007; Ozonoff et al. 2010; Werner and Dawson 2005). An important clinical use of the Toddler Module of the ADOS-2 is to identify concerns in need of continued monitoring (Lord et al. 2012b). At the research level, the Toddler Module may aid in understanding ASD symptom trajectories beginning as soon as children are walking independently (Lord et al. 2012a).

Over the past decade, much research has focused on the development of younger siblings of children with ASD ("infant siblings") to better understand onset, risk, and

underlying biological mechanisms (e.g., Ozonoff et al. 2011). Studies of infant siblings suggest variable and complicated developmental trajectories (Bryson et al. 2007; Landa et al. 2007; Messinger et al. 2013; Ozonoff et al. 2014). Regardless of diagnostic outcome, infant siblings tend to score higher (i.e., more symptomatic) on the ADOS-2 and lower on measures of developmental skills, compared to infants who do not have a family history of ASD (Messinger et al. 2013; Ozonoff et al. 2014).

It is well known that ASD is a highly heterogeneous disorder, and this hinders our understanding of its causes and recommended courses of treatment (State and Levitt 2011). The behavioral heterogeneity of ASD is in part due to differences in level of intellectual impairment, language impairment, and co-occurring challenging behaviors and/or diagnoses (e.g., ADHD, anxiety). These factors affect how ASD symptoms manifest but are not core symptoms themselves. To try to control for these influential and complex variables, researchers are looking for ways to increase homogeneity in ASD symptom presentation to better understand underlying causes, developmental course, and treatment effects.

In research and in practice, the ADOS-2 frequently is used to diagnose and describe ASD symptoms. A *calibrated severity score* (CSS) was created for Modules 1 through 4 to estimate overall level of ASD symptoms relative to others with ASD of the same age and language level (Gotham et al. 2009; Hus and Lord 2014). The CSS was created in response to the need for a metric of severity that is as independent as possible of participant variables of intellectual ability, language, and age. Compared to raw total scores, the CSS was less influenced by verbal language level, especially for Modules 1–3—where verbal IQ accounted for 43 % of the variance in raw scores, it accounted for only 10 % of the variance in the CSS. The CSS also has more uniform distributions across age/language level groups. These results were replicated by de Bildt et al. (2011) and Shumway et al. (2012) in independent samples, with a similar pattern of reduced association with verbal IQ for the CSS.

On the other hand, ASD symptoms may best be measured by domain rather than in aggregate (Shumway et al. 2012). Separate calibrated severity scores were developed for Social Affect and Restricted, Repetitive Behavior domains of the ADOS-2 to provide a clearer picture of ASD symptom severity (Hus et al. 2014; Hus and Lord 2014). Several potential uses for domain CSS have been identified, including studying whether the two domains have distinct trajectories or respond differently to intervention; increasing phenotypic homogeneity by clustering individuals according to similar levels of severity in each domain; and using a CSS to control statistically for differences in one domain while focusing on the other. There is a need for standardized tools to further define and characterize severity, to improve reliability of ratings across sites and clinicians, and to increase comparability across research samples (Weitlauf et al. 2014).

At the time that overall and domain CSS were created, large datasets using the Toddler Module of the ADOS-2 were not available to be included in analyses. Thus, a CSS could not be calculated for children who received the Toddler Module. Researchers have tried to overcome this limitation in various ways. For example, a CSS could not be generated in infant sibling and intervention studies until 36 months for many children (e.g., Messinger et al. 2013; Ozonoff et al. 2014). Other studies (Guthrie et al. 2013; Venker et al. 2014) attempted to capture symptom severity by applying Module 1 CSS to the Toddler Module. However, as the authors acknowledged, the CSS developed for Module 1 cannot be directly applicable to the Toddler Module due to differences in coding criteria and items comprising the algorithms for the respective modules. Application of CSS for the ages addressed by the Toddler Module, 12–30 months, may help us better understand developmental trajectories indicative of risk, especially because they provide a continuous scale of presence and severity of ASD symptoms across development, into the other four modules. A Toddler Module CSS would allow longitudinal comparisons of symptom severity potentially from the earliest point of concern, and may improve understanding of how ASD symptoms emerge, relatively independent of language abilities.

A note on terminology: in the recently revised ADOS-2 (Lord et al. 2012d), the CSS was renamed the Comparison Score. However, here, we maintain use of the term "CSS" to refer to the standardized severity scores to facilitate comparisons to the studies by Gotham et al. (2009), Hus et al. (2014), and Hus and Lord (2014), which this manuscript seeks to replicate.

The purpose of the present research is to develop ADOS-2 Toddler Module total and domain CSS to expand the continuous metric of ASD symptom severity to younger ages. We hypothesize that the Toddler Module CSS will be less affected by child characteristics and demographics than raw scores. However, because the Toddler Module covers a more restricted age and IQ range than Modules 1–4, we were interested to see whether the CSS would result in reductions in the influence of age and IQ to the extent demonstrated in Modules 1–4. To achieve this aim, this study employed methods from Gotham and colleagues' (2009) development of the total CSS for modules 1–3 and from Hus et al.'s (2014) and Hus and Lord's (2014) development of calibrated domain scores for Social Affect and Restricted, Repetitive Behaviors.

Methods

Participants

The sample consisted of 388 *individual children* eventually diagnosed with ASD. Repeated assessments were performed on 127 children, yielding a total of 823 *assessments,* where "assessment" is defined as contemporaneous Toddler Module data and a best estimate clinical diagnosis. The child's most recent diagnosis was used for the purposes of the current study. Mothers in the repeated assessments group had more education ($\chi^2 = 15.19$, $p < .001$). Although, at the group level, children with one assessment did not differ significantly from children with repeated assessments in age, gender, race, verbal IQ, or nonverbal IQ, significant differences in these variables emerged when children were grouped based on the Toddler Module algorithm received. Among children who used fewer than five words during the Toddler Module or were between the ages of 12 and 20 months (i.e., 12–20/Nonverbal algorithm), children in the repeated assessment group tended to be slightly younger than the single assessment group (21.52 vs. 22.65 months, $p < .01$), and they had higher nonverbal IQs (86.65 vs. 79.12, $p < .001$). Children between age 21 and 30 months who used five or more single words during the Toddler Module (i.e., Some Words 21–30 algorithm) showed more differences: children with repeated assessments were slightly older than children with one assessment (25.78 vs. 24.82 months, $p < .01$) and had higher verbal mental ages (23.78 vs. 19.35, $p < .001$), verbal IQ (88.87 vs. 78.84, $p < .01$), nonverbal mental ages (25.66 vs. 22.84, $p < .001$), and nonverbal IQ (96.42 vs. 89.27, $p < .01$). These differences were likely due to referral biases; for example, children with repeated assessments were more likely to be infant siblings (28 % of the repeated assessment group versus 10 % of the single assessment group, $\chi^2 = 27.90$) who may have enrolled in a research study prior to showing developmental concerns.

In creating the CSS for Modules 1–4, repeated assessment data were retained in the analyses (Gotham et al. 2009; Hus and Lord 2014). However, because of the differences between children seen once and children seen longitudinally described above, a subsample was used for standardization of calibrated severity scores that eliminated repeated assessment data. This subsample of 388 children with ASD (hereafter termed "original sample") contained data from all children with one assessment, and one assessment was randomly selected for children with repeated assessments. A replication sample then was created using children with ASD with repeated assessments, excluding the 388 children in the original sample, to further validate calibrated severity scores.

Original Sample

Chronological ages in the original sample represented the recommended age range for the Toddler Module, 12–30 months. Ethnicities represented in the dataset included 8 % African American (N = 30), 2 % Asian American (N = 7), 71 % Caucasian (N = 276), 6 % Hispanic (N = 24), 0.3 % Native American (N = 1), and 9 % Biracial (N = 36). Males comprised 83 % of the dataset (N = 323) and females comprised 17 % of the sample (N = 65). Thirteen percent reported maternal education at the graduate or professional level, 52 % had a bachelor's degree or some college, and 30 % reported completing high school or less (4 % did not report education level). Contemporaneous verbal IQ data was available for 274 children (71 % of the original sample) and nonverbal IQ for 329 (85 %) (see Table 1 for sample description).

The dataset represented combined data from four sites: the University of Michigan Autism and Communication Disorders Center, Florida State University (FSU), the University of Minnesota, and the University of Wisconsin-Madison. The majority (N = 211) came from FSU and were recruited from pediatric primary care physicians through the FIRST WORDS® Project, a prospective, longitudinal study of a general population screening program for communication delays and ASD. Children from the University of Michigan (N = 84) consisted of (a) consecutive referrals of children from 12 to 30 months of age to the clinic, (b) children from University of Michigan projects studying early development of children with communication delays and/or at risk for ASD, and (c) children participating in various treatment studies. The original validation sample for the Toddler Module (Luyster et al. 2009) was included in this dataset. Children from the University of Wisconsin (N = 58) were participants in a longitudinal study of language development in children with ASD starting at age 24 months and seen annually until age 5; and children from the University of Minnesota (N = 35) consisted of consecutive clinic referrals of children age 12–30 months and participants in research studying early language and motor markers in children at risk for ASD (siblings or history of prematurity). Sites also differed in their ability to include blind examiners for children seen longitudinally; however, each site implemented procedures to reduce bias. Blind evaluators were used every 6 months at the University of Michigan; at FSU, diagnostic evaluations were reviewed and confirmed by an additional, experienced clinician; and at the University of Minnesota, a subset of Toddler Module ADOS-2 administrations were observed and co-coded by a research-reliable examiner who was not aware of participants' previous performance. (Children at UW-Madison

Table 1 Sample description: ASD cases included in creation of calibrated severity score

	12–20/nonverbal				Some words 21–30			
	N	Mean	SD	Range	N	Mean	SD	Range
Age	272	22.27	3.93	12–30	116	25.22	2.69	21–30
VIQ	189	56.57	19.82	5–118	85	81.67	21.92	31–133
NVIQ	232	80.51	19.64	24–145	97	91.26	18.21	54–141
VMA	196	11.82	4.26	1–29	85	20.95	6.28	8–38
NVMA	234	18.07	4.21	5–32	97	23.90	4.76	14–38
ADI-R SA	154	9.90	4.26	1–18	49	9.69	4.82	0–19
ADI-R RRB	153	4.82	2.30	0–11	49	5.16	2.95	0–12
ADI-R RPI	–	–	–	–	11	2.91	2.17	0–6
ADI-R Total	154	22.40	6.80	3–37	49	20.02	8.03	3–35
ADOS-SA	272	14.47	3.54	3–20	116	11.94	4.25	0–21
ADOS-RRB	272	4.21	2.06	0–8	116	2.89	1.59	0–6

VIQ verbal IQ, *NVIQ* nonverbal IQ, *VMA* verbal mental age, *NVMA* nonverbal mental age, *ADI-R* autism diagnostic interview-revised, *SA* social affect, *RRB* restricted, repetitive behavior, *RPI* reciprocal peer interaction, *ADOS* autism diagnostic observation schedule

were seen annually starting at age 2 and thus received the Toddler Module only once.)

Replication Sample

Analyses of CSS distributions and their relative independence from verbal and age variables were repeated using data from children with repeated assessments (see Table 2 for sample description). Data from FSU, University of Michigan, and University of Minnesota included repeated assessments. The replication dataset included assessments from 127 individual children with ASD. Assessments included in the original analyses were removed from the replication dataset, resulting in a final sample of 435 ASD *assessments*. Ethnicities represented in the dataset included 6 % African American (N = 27 assessments), 2 % Asian American (N = 8), 70 % Caucasian (N = 306), 4 % Hispanic (N = 19), 12 % Biracial (N = 52), and 5 % Other or race not specified (N = 23). Males comprised 88.5 % of the dataset (N = 385 assessments) and females comprised 11.5 % (N = 50 assessments). Twenty-five percent reported maternal education at the graduate or professional level; 22 % reported completing high school or less.

Of the 127 children with ASD with repeated assessments, 46 had two or three Toddler Module assessments, 52 had four or five assessments, 26 had between 6 and 10 assessments, and three had between 11 and 15 assessments. Children with four or more assessments tended to be participants who were showing communication and/or ASD concerns and were participating in treatment studies at FSU

Table 2 Sample description: ASD cases included in replication sample

	12–20/nonverbal				Some words 21–30			
	N	Mean	SD	Range	N	Mean	SD	Range
Age	285	21.52	4.34	12–30	150	25.73	2.65	21–30
VIQ	169	57.35	18.95	13–118	76	89.26	23.19	42–141
NVIQ	168	87.46	18.46	19–128	75	96.42	16.29	54–128
VMA	169	12.01	4.45	3–33	76	23.57	7.09	8–34
NVMA	169	18.17	4.43	6–35	75	25.30	5.31	13–35
ADI-R SA	98	10.09	4.20	1–18	21	9.43	3.89	3–17
ADI-R RRB	97	4.61	2.13	0–9	28	5.14	3.01	1–12
ADI-R RPI	–	–	–	–	7	4.14	1.57	2–6
ADI-R Total	98	24.66	6.45	9–37	28	22.00	7.43	7–35
ADOS-SA	285	13.74	3.99	0–20	150	10.49	4.70	0–22
ADOS-RRB	285	3.62	2.00	0–6	150	2.39	1.40	0–6

VIQ verbal IQ, *NVIQ* nonverbal IQ, *VMA* verbal mental age, *NVMA* nonverbal mental age, *ADI-R* autism diagnostic interview-revised, *SA* social affect, *RRB* restricted, repetitive behavior, *RPI* reciprocal peer interaction, *ADOS* autism diagnostic observation schedule

or the University of Michigan, or children participating in a study of early diagnosis of ASD at the University of Michigan where participants were seen on a monthly basis. Participants in the monthly study at the University of Michigan were showing communication delays and/or risk for ASD, or were infant siblings. As a side note, we were not concerned about practice effects for children with repeated assessments, because although children may become familiar with particular tasks (e.g., the bath routine), ADOS-2 scores and classifications are based on spontaneous initiations and responses rather than performance on specific tasks (Lord et al. 2012d).

Measures and Procedure

The ADOS is a clinician-administered, standardized observation designed to elicit social communication and restricted, repetitive behaviors related to ASD (Lord et al. 2000). Four original modules are each tailored to an individual's language level and age to control for the effects of language level on social communication and play behaviors. The second edition of the ADOS (ADOS-2; Lord et al. 2012a, d) adds a Toddler Module for children age 12–30 months with language skills ranging from no verbal language to single words and simple phrases. Toddlers must be walking independently, and a nonverbal mental age of at least 12 months is recommended. The Toddler Module follows the structure of Module 1, which is designed for language levels ranging from nonverbal to single words and simple phrases. Module 1 activities, child behavioral descriptions, and scoring criteria were modified based on developmental expectations for toddlers.

The Toddler Module algorithm contains separate domain categories of Social Affect and Restricted, Repetitive Behaviors and a single total score to determine classification. Separate algorithms are provided based on age and language level: all children age 12–20 months, and children age 21–30 months who produce fewer than five words during the ADOS-2, receive the 12–20/Nonverbal 21–30 algorithm, and children age 21–30 months who produce five or more words during the ADOS-2 receive the Some Words 21–30 months algorithm. Clinical cut-off scores are grouped within levels of concern for ASD, acknowledging the diagnostic uncertainty inherent in very young children due to significant developmental variability or confounding conditions (e.g., intellectual disability, language impairment). Research classifications with cut-points for ASD and nonspectrum also are available (Luyster et al. 2009).

We examined the sensitivity of Toddler Module research classifications and concern ranges for our samples, and results were similar to those reported in the original validation study (Luyster et al. 2009). Using the research cutoffs of a total score of 12 for 12–20/Nonverbal and 10 for Some Words 21–30, sensitivity in the original sample was .94 for children who received the 12–20/Nonverbal 21–30 algorithm and .88 for children receiving the Some Words 21–30 algorithm. Sensitivity in our replication sample was 0.88 for the 12–20/Nonverbal 21–30 group and 0.71 for the Some Words 21–30 group. In the original sample, 82.2 % fell within the moderate-to-severe concern range, 14.4 % fell into the mild-to-moderate range, and 3.4 % fell into the little-to-no concern range. In the replication sample here, 72.2 % fell within the moderate-to-severe range, 19.1 % were in the mild-to-moderate range, and 8.7 % were in the little-to-no concern range.

In the current study, the ADOS-2 Toddler Module was conducted as part of a clinic or research evaluation. A similar battery of assessment measures was used across sites and projects. The University of Michigan, University of Minnesota, and University of Wisconsin-Madison administered the Toddler Autism Diagnostic Interview-Revised (Toddler ADI-R; Kim and Lord 2012; Lord et al. 1994) to inform diagnosis; children seen at FSU were given a developmental history interview and parent-report measures of ASD symptoms. Children at all sites received psychometric measures of cognitive and adaptive development, including Mullen Scales of Early Learning (MSEL; Mullen 1995), and Vineland Adaptive Behavior Scales, 2nd edition (Vineland-II; Sparrow et al. 2005). Additionally, language skills were assessed at the Universities of Michigan, Minnesota, and Wisconsin-Madison using the Preschool Language Scales (PLS, 4th and 5th editions; Zimmerman et al. 2002, 2011) and/or MacArthur-Bates Communication Development Inventories, 2nd edition (Fenson et al. 1993). Diagnostic distinctions of autism and non-autism ASD were made at the Universities of Michigan and Wisconsin-Madison; at FSU and University of Minnesota, subcategories were not assigned, and children meeting criteria for DSM-IV diagnoses of Autistic Disorder, PDD-NOS, or Asperger's Disorder were given a best estimate diagnosis of ASD. To be consistent with DSM-5 (APA 2013), and because clinical subcategories have been found to be unstable over time (e.g., Lord et al. 2006), unreliable across clinicians, and not representative of meaningful differences in symptom presentation (Lord et al. 2012c), children with any autism spectrum diagnosis were grouped into one ASD category for the present analyses.

Clinic-referred patients received oral feedback and a written report without financial compensation. Participants recruited only for the purpose of research received financial compensation and a written summary of evaluation results. Institutional Review Boards at the University of Michigan, FSU, University of Minnesota, and University of Wisconsin-Madison approved all procedures related to this project.

Site differences emerged in demographic and child variables. Differences in child variables across sites were expected due to differences in recruitment patterns and study design across sites. We viewed these site differences as beneficial to the purpose of this study, as we sought to include children with varied levels of impairment and symptom characteristics. The University of Wisconsin sample generally was older, had lower verbal skills, and showed greater impairment in IQ and ADOS-2 scores than children from other sites. Families in the FSU sample self-identified as more racially and ethnically diverse than families from other sites. See Supplemental Tables 1 and 2 for further details on site differences.

Development and Analysis of Toddler Module Overall and Domain CSS

The current study followed the procedures used in developing total and domain calibrated severity scores for ADOS-2 modules 1 through 4 (Gotham et al. 2009; Hus et al. 2014; Hus and Lord 2014). Calibrated severity scores were created by dividing the pool of children with ASD into narrowly defined age and language cells, and standardizing raw total scores from the Toddler Module algorithms within these cells.

Development of the Total-CSS

Children were separated into groups based on the Toddler Module algorithm received. Cells were not equal; as expected, we found relatively few children age 21–30 months who had a large single word vocabulary who were eventually diagnosed with ASD. Within the two developmental cells, distributions of Total, Social Affect (SA) domain, and Restricted, Repetitive Behavior (RRB) domain scores were generated separately for every 1-month age group. Next, age groups with similar score distributions were collapsed to create the fewest number of age- and language-level-determined "calibration cells." In the end, distributions were highly similar across ages within the 12–20/Nonverbal group and the Some Words 21–30 group. Thus, two calibration cells resulted, corresponding to the Toddler Module algorithms.

In creating the CSS for Modules 1–4 (Gotham et al. 2009; Hus and Lord 2014), Total-raw scores within calibration cells were mapped onto a 10-point severity rating scale based on percentiles of Total-raw scores corresponding to each ADOS-2 diagnostic classification. Lower calibrated severity scores were associated with fewer social communication and repetitive behavior concerns. Scores 1–3 represented nonspectrum classifications, 4–5 represented ASD classifications, and 6–10 represented autism classifications. Similarly, for the Toddler Module, a CSS of 1–3 was set to represent Total-raw scores falling within the little-to-no concern range, scores of 4–5 represented scores in the mild-to-moderate concern range, and 6–10 represented scores falling within the moderate-to-severe concern range. Toddler Module concern range thresholds were determined by the algorithm relevant to each calibration cell. The range of Total-raw scores corresponding to each point on the CSS metric was determined by percentiles of available data associated with each CSS point within a concern range, resulting in the Total-CSS.

Development of Domain CSS

Because there are not separate SA and RRB cut-offs for ADOS-2 classifications, the percentiles used for mapping the overall Total scores were used to inform mapping of raw SA and RRB totals to each respective domain CSS. As with Modules 1–4, raw RRB domain scores were mapped onto CSS values of 5–10, due to the limited range of the RRB raw total (Hus et al. 2014; Hus and Lord 2014). Because concern ranges were not available to anchor CSS for SA and RRB domains, mappings were adjusted for the SA-CSS so that, for each of the algorithm groups, at least 90 % of children in the moderate-to-severe concern range received an SA-CSS greater than or equal to 6. For children in the 12–20/Nonverbal group, sensitivity was 94.8 %; in the Some Words 21–30 group, sensitivity was 90 %. Also, 100 % of children in the mild-to-moderate concern range in both groups received an SA-CSS of 4 or higher, and none of these children received an SA-CSS score above 7. As with Modules 1–4, a goal of 80 % sensitivity was set for the RRB-CSS, due to expected lower sensitivity in detecting repetitive behaviors within the limited time and contexts of an ADOS-2 administration (Hus et al. 2014). This goal was attained for each algorithm group: 85.7 % of children in the moderate-to-severe range in the 12–20/Nonverbal group, and 88.8 % of children in the moderate-to-severe range in the Some Words 21–30 group, received an RRB-CSS of 6 or higher. Similarly, over 80 % of children in the mild-to-moderate concern range received an RRB-CSS of 5 or higher across both algorithm groups. Table 3 shows the raw score range corresponding to each CSS point within each calibration cell.

To ensure that scores 6–10 correspond to approximate fifths of the ASD participants who scored in the moderate-to-severe concern range, roughly 20 % of participants in the moderate-to-severe group should receive any individual score from 6 to 10. This was generally the case in our dataset: for the Total-CSS, percentages across scores 6 through 10 ranged from 18.5 to 22.3 %, SA-CSS ranged from 14.1 to 21.9 %; and RRB-CSS ranged from 15.4 to 20.1 %.

Table 3 Mapping of ADOS-2 total and domain scores onto CSS

Toddler module concern range	CSS	Raw totals					
		Overall total		SA domain		RRB domain	
		12–20/NV	SW 21–30	12–20/NV	SW 21–30	12–20/NV	SW 21–30
Little-to-no	1	0–2	0–3	0–2	0–1	0	0
	2	3–5	4–6	3–4	2–3	–	–
	3	6–9	7	5–6	4–5	–	–
Mild-to-moderate	4	10–11	8–9	7–9	6–8	–	–
	5	12–13	10–11	10	9–10	1–2	1
Moderate-to-severe	6	14–16	12	11–12	11	3	2
	7	17–18	13–15	13–14	12–13	4	3
	8	19–21	16–17	15–16	14–15	5	4
	9	22–23	18–20	17–18	16–18	6	5
	10	24–28	21–28	19–20	19–22	7–8	6

CSS calibrated severity score, *12–20/NV* toddler module algorithm for children age 12–20 and nonverbal children, *SW 21–30* toddler module algorithm for children age 21–30 months who used single words, *SA* social affect, *RRB* Restricted, repetitive behavior

Analyses conducted by Gotham et al. (2009), Hus et al. (2014), and Hus and Lord (2014) were repeated with this Toddler Module dataset. Distributions of raw and calibrated severity scores were compared to assess whether CSS distributions across age/language cells were more uniform than raw score distributions. Linear regression models were analyzed to compare the relative independence of CSS and raw totals from child characteristics. Potential predictors were entered into a structured hierarchical regression model, in which Block 1 included verbal and nonverbal IQs and mental ages (which are known to affect the expression of ASD symptoms and for which we hoped to limit the effect on ADOS-2 scores through the CSS; Bishop et al. 2006; Lord and Spence 2006), and Block 2 included gender, maternal education, and race (variables that could affect ASD symptoms but that often have had non-significant effects when IQ and mental age variables are controlled; Gotham et al. 2009). Only model R^2 are reported, because interpretation of the meaning of these individual coefficients is limited by multicollinearity. For all regression models, Cohen's f^2 was computed; f^2 of .02, .15, and .35 reflect small, medium, and large effect sizes, respectively (Cohen 1988). Significant predictors were then entered into Forward Stepwise models to determine the relative contributions of these individual variables to raw scores and CSS. These analyses then were replicated using Toddler Module non-overlapping assessments from children with repeated measure data to further validate the CSS. Finally, several assessments with longitudinal data were chosen to exemplify various patterns of severity change over time across diagnostic groups. Analyses were completed using SPSS Version 21 and 22.

Results

Comparing Distributions of Raw Totals and CSS

Distributions of Toddler Module raw Total, Social Affect, and Restricted, Repetitive Behavior scores were generated for each age/language cell (Fig. 1a, c, e) and compared to the distributions of CSS for each cell (Fig. 1b, d, f). Distributions of CSS showed increased comparability across the two groups. There was a non-significant trend for children in the older, verbal group to have lower Total-CSS compared to the nonverbal and younger group ($t = 1.90$, $p < .058$); the difference between groups is within 0.5 point and similar to mean CSS distributions for Modules 1–4 (Gotham et al. 2009; Hus and Lord 2014). Children in the Some Words 21–30 group had lower SA-CSS than children in the Nonverbal/12–20 group ($t = 4.40$, $p < .001$). We tolerated this difference, because Toddler ADI-R scores and IQ scores suggested a level of greater impairment in children in the Nonverbal/12–20 group. Adjusting the SA-CSS to be more equal between groups could have misrepresented true differences in severity. Differences in RRB-CSS were not significant. Means and standard deviations of CSS and raw scores are listed by age/language cell in Table 4.

As expected, site differences in CSS were present. No significant differences were found for children who used five or more words during the ADOS-2. Among nonverbal children, the University of Wisconsin sample had significantly higher Total-CSS ($F = 12.31$, $p < .001$), SA-CSS ($F = 5.86$, $p < .001$), and RRB-CSS ($F = 17.15$, $p < .001$) than children from the University of Michigan or FSU. Children from the University of Minnesota also had

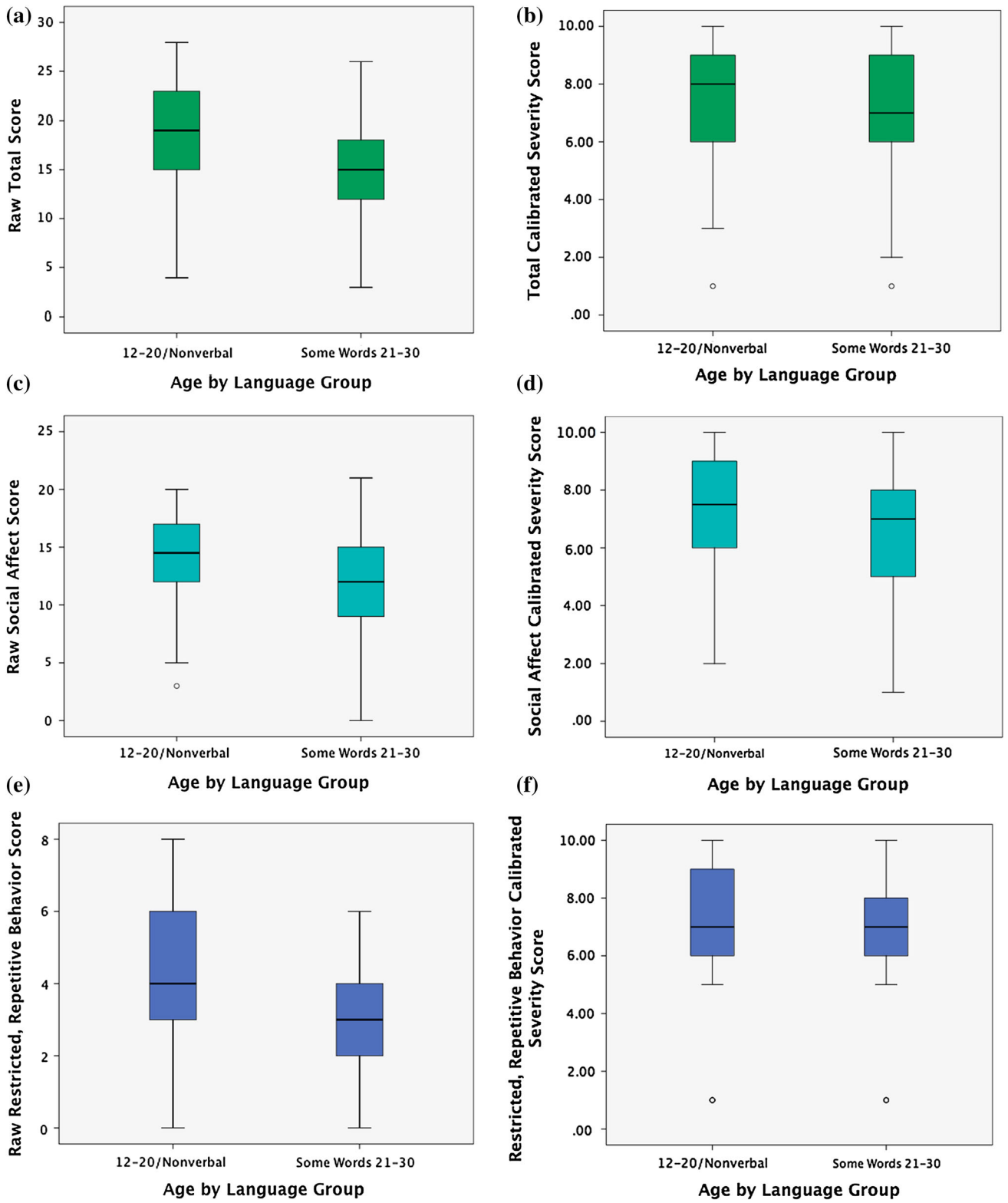

Fig. 1 Original sample. **a** Distributions of raw total scores by age/language cells. **b** Distributions of calibrated severity scores by age/language cells. **c** Distributions of raw Social Affect scores by age/language cells. **d** Distributions of calibrated severity Social Affect scores by age/language cells. **e** Distributions of raw Restricted/Repetitive Behavior scores by age/language cells. **f** Distributions of calibrated severity Restricted/Repetitive Behavior scores by age/language cells

Table 4 Raw score and CSS means and standard deviations by age/language cell (ASD assessments only)

Algorithm	Total-raw			Total-CSS		SA-raw		SA-CSS		RRB-raw		RRB-CSS	
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
12–20/NV	272	18.68	4.66	7.44	1.86	14.47	3.54	7.44	1.84	4.21	2.06	7.16	2.04
SW 21–30	116	14.83	4.79	7.03	2.09	11.94	4.25	6.51	2.07	2.89	1.59	6.76	1.91

CSS calibrated severity score, *12–20/NV* toddler module algorithm for children age 12–20 and nonverbal children, *SW 21–30* toddler module algorithm for children age 21–30 months who used single words, *SA* social affect, *RRB* restricted, repetitive behavior

higher RRB-CSS than children from the University of Michigan ($p < .01$) (Supplemental Table 2).

Correlations Between Domain and Total CSS

Correlation results were very similar to those of Modules 1–4 (Hus et al. 2014; Hus and Lord 2014). Associations between SA- and RRB-CSS were significant but weak ($r = .28$). Correlations between both SA- and RRB-CSS and Total-CSS were strong, but the correlation between SA-CSS and Total-CSS was stronger ($r = .90$) than for RRB-CSS and Total-CSS ($r = .59$), due to the greater proportion of SA items contributing to the Total-CSS than RRB items.

Relative Independence of CSS from Participant Characteristics

Using the original sample of children with ASD (N = 388), linear regression analyses were performed separately for dependent variables of total and domain CSS and raw scores to examine whether participant characteristics such as age and IQ would be less associated with CSS than they were with raw scores.

Predictors of Total-Raw and Total-CSS

Using the full model, 30.5 % of the variance in Toddler Module Total-raw was explained. No individual predictor was statistically significant, but multicollinearity was high for IQ and mental age variables. Verbal IQ showed a trend ($p = .063$) as a predictor of Total-raw scores. For Total-CSS, the full model accounted for 20.1 % of the variance, and no variables emerged as significant predictors. This represents a reduction in the influence of child characteristics from an f^2 of .44 for Total-raw to an f^2 of .25 for Total-CSS.

Although no single predictor was statistically significant, because the models were significant, individual predictors were entered into Forward Stepwise models to assess the individual contribution of each variable (see Supplemental Table 3). For Total-raw scores, verbal IQ accounted for the majority of the variance (26.4 %), while nonverbal mental age contributed an additional 3.0 %. All other variables were excluded from the model, indicating they were not significant. In the Forward model predicting Total-CSS, verbal IQ again accounted for the majority of the variance (15.7 %), and nonverbal IQ explained 3.1 %. These results reflect a reduction in the influence of verbal IQ from a large effect size ($f^2 = .36$) for Total-raw to a medium effect size ($f^2 = .19$) for Total-CSS.

Predictors of SA-Raw and SA-CSS

For the SA domain, child characteristics in the full model accounted for 23.6 % of the variance in SA-raw scores, and again, only verbal IQ showed a trend for significance ($p = .063$). In contrast, 19.3 % of the variance in SA-CSS was explained by child characteristics, with verbal IQ showing a trend for significance ($p = .077$). Thus, the influence of child characteristics was reduced from an f^2 of .31 for SA-raw to an f^2 of .24 for SA-CSS.

Again, because the models were significant, individual predictors were entered into Forward Stepwise models. For SA-raw, verbal IQ contributed the greatest proportion of the variance, (19.3 %), while nonverbal mental age accounted for 3.5 %. For SA-CSS, verbal IQ explained 16.2 % of the variance, while nonverbal mental age explained 2.1 %. All other variables were excluded from both models. The CSS for SA modestly reduced the influence of verbal IQ from an f^2 of .21 for SA-raw to an f^2 of .19.

Results from Forward Stepwise models are presented in Supplemental Table 3.

Predictors of RRB-Raw and RRB-CSS

For the RRB domain, child characteristics accounted for 17.6 % of the variance in RRB-raw, and no predictors emerged as significant. For RRB-CSS, child characteristics accounted for 11.4 %, and nonverbal IQ demonstrated a trend as a predictor of RRB-CSS ($p = .058$). The influence of child characteristics was reduced from an f^2 of .21 for RRB-raw to an f^2 of .13 for RRB-CSS.

All predictors were entered into Forward Stepwise models, and only verbal IQ emerged as a predictor of RRB-raw, accounting for 15.4 % of the variance. For RRB-CSS,

verbal IQ and nonverbal IQ were statistically significant but explained small proportions of the variance in RRB-CSS (7.1 and 1.8 %, respectively). Thus, the influence of verbal IQ was reduced from an f^2 of .18 for RRB-raw to an f^2 of .08 for RRB-CSS.

Replication with Repeated Assessment Data

Comparisons, Correlations, and Relative Independence of Raw Scores and CSS

In mapping raw total scores onto a 10-point calibration scale, raw scores corresponding to each calibrated severity score were highly similar across original and replication samples, with no shifts in range greater than one raw score point (e.g., whereas a CSS of 8 corresponded to raw total scores of 19–21 for the Nonverbal/12–20 group in the original sample, the range was 19–20 for the replication sample). The original CSS map was therefore used for analyses with the replication sample.

Distributions of total and domain raw scores and CSS are presented in Fig. 2. Distributions of Total-CSS showed increased comparability across the two groups in the replication sample in contrast to raw total scores. However, the trend of the Some Words 21–30 group having lower CSS than the Nonverbal/12–20 group was exaggerated and more significant in this replication sample. Children in the Some Words 21–30 group had significantly lower Total-CSS ($t = 3.71$, $p < .001$), SA-CSS ($t = 6.46$, $p < .001$), and SA-RRB ($t = 2.19$, $p = .029$) compared to the Nonverbal/12–20 group. In general, mean CSS were lower in the replication sample than in the original sample (see Table 5). This difference is likely due to recruitment effects and the fact that the University of Wisconsin sample, which was generally older and less cognitively able, was not included in the replication sample. As a result, the repeated assessment sample had higher verbal and nonverbal skills and included a higher proportion of children who were in treatment studies and/or assessed prior to developing clear ASD concerns compared to the original sample.

Linear regression analyses were repeated with the replication sample, with Forward Stepwise models performed where appropriate. Results of Forward Stepwise regressions are presented in Supplemental Table 4. The full model accounted for 41.8 % of the variance in Total-raw scores, and verbal IQ emerged as a significant predictor. The same model accounted for 30.6 % of the variance in Total-CSS, and verbal IQ remained a significant predictor. This represents a reduction in the influence of child characteristics from an f^2 of .72 for Total-raw to an f^2 of .44 for Total-CSS. Because there was only one significant predictor of Total-raw and Total-CSS, Forward Stepwise models were not run.

For Social Affect, the full model accounted for 40.2 % of the variance in SA-raw, and verbal IQ and maternal education level emerged as significant predictors. The same model accounted for 38.0 % of the variance in SA-CSS. Verbal IQ was a significant predictor of SA-CSS, and maternal education level showed a trend for significance ($p = .052$). The influence of child characteristics was slightly reduced from an f^2 of .67 for SA-raw to an f^2 of .61 for SA-CSS. Next, verbal IQ and maternal education level were entered into Forward Stepwise models. For SA-raw, verbal IQ explained 36.7 % of the variance, and maternal education was excluded from the model, indicating it was not significant. For SA-CSS, verbal IQ explained 34.5 % of the variance, and maternal education again was excluded. Effect sizes remained large ($f^2 = .58$ for SA-raw and $f^2 = .53$ for SA-CSS).

For Restricted and Repetitive Behaviors, the full model accounted for 16.3 % of the variance in RRB-raw and 12.2 % of the variance in RRB-CSS. In this case, gender emerged as a small but statistically significant predictor of RRB-raw with slightly higher scores for males; no variable was a significant predictor of RRB-CSS. The influence of child variables showed a small reduction from an f^2 of .19 for RRB-raw to an f^2 of .14 for RRB-CSS. As only one variable emerged as a predictor of RRB-raw, Forward Stepwise models were not performed.

Case Summaries

Four children with longitudinal data were selected to illustrate the utility of the Toddler Module CSS for examining early patterns of ASD symptoms and their trajectories over time. CSS by chronological age are plotted in Fig. 3, with ADOS-2 module and raw score displayed for each time point. See Table 6 for child characteristics at first and last assessment.

Case 1

'Henry' is a clinic-referred male who showed a stable and severe pattern of ASD symptoms. Henry was diagnosed with ASD at 17 months and enrolled in full-time applied behavior analysis (ABA) intervention at 18 months. At 17 months, he rarely initiated social interaction, rarely vocalized, and typically communicated using physical means (use of other's body, giving objects). He engaged in frequent complex mannerisms, visual sensory exploration, and repetitive spinning of objects. After entering ABA, Henry markedly improved in structural communication and began using vocalizations and words to request. His relatively lower SA-CSS after initiating intervention reflected improvements in pairing eye contact with requests, using words and phrases for a variety of pragmatic

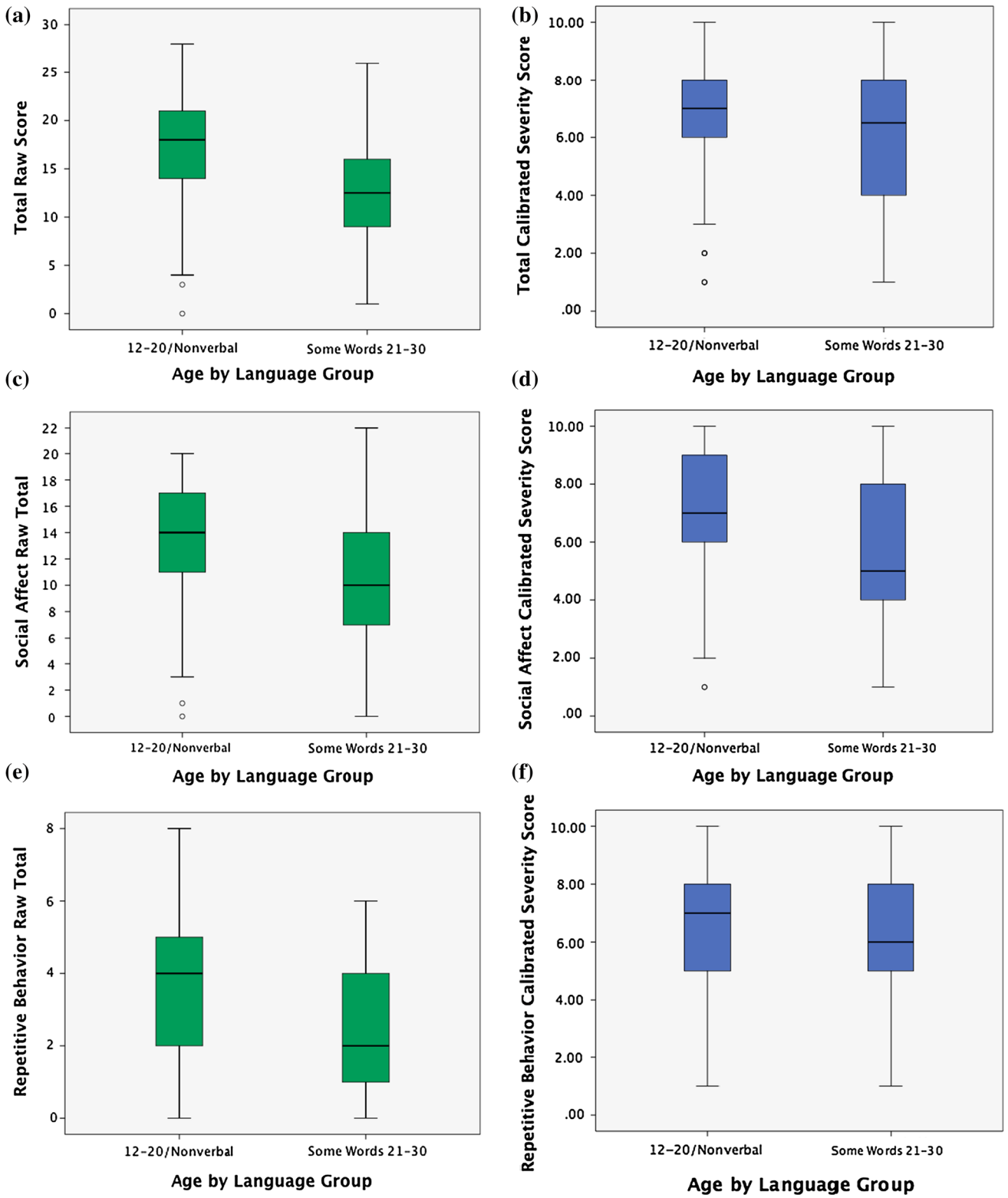

Fig. 2 Replication sample. **a** Distributions of raw total scores by age/language cells. **b** Distributions of calibrated severity scores by age/language cells. **c** Distributions of raw Social Affect scores by age/language cells. **d** Distributions of calibrated severity Social Affect scores by age/language cells. **e** Distributions of raw Restricted/Repetitive Behavior scores by age/language cells. **f** Distributions of calibrated severity Restricted/Repetitive Behavior scores by age/language cells

Table 5 Raw score and CSS means and standard deviations by age/language cell (replication sample ASD assessments only)

Algorithm	Total-raw			Total-CSS		SA-raw		SA-CSS		RRB-raw		RRB-CSS	
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
12–20/NV	285	17.36	4.92	6.94	2.02	13.74	3.99	7.08	2.04	3.62	2.00	6.61	1.99
SW 21–30	150	12.89	5.26	6.12	2.46	10.49	4.70	5.70	2.28	2.39	1.40	6.17	1.91

CSS calibrated severity score, *12–20/NV* toddler module algorithm for children age 12–20 and nonverbal children, *SW 21–30* toddler module algorithm for children age 21–30 months who used single words, *SA* social affect, *RRB* restricted, repetitive behavior

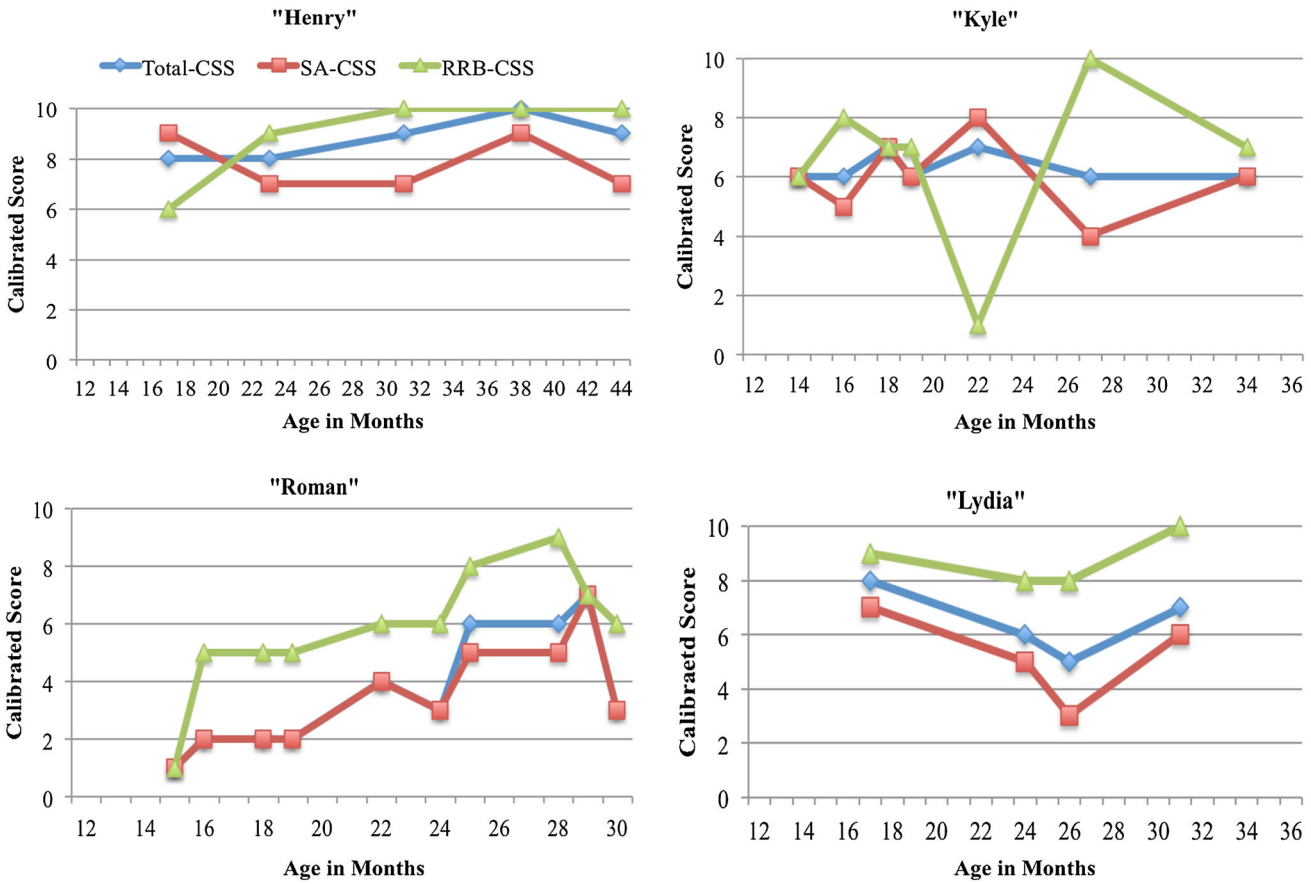

Fig. 3 Case summaries of longitudinal total and domain calibrated severity scores

Table 6 Case summary characteristics

	Demographics		First assessment			Last assessment				
	Gender	Race	Age (mos)	VIQ	NVIQ	Age (mos)	VIQ	NVIQ	ADOS Module	Diagnosis
Henry	Male	White	17	34	101	44	100	108	2	Autism
Kyle	Male	White	14	69	98	34	129	113	2	Autism
Roman	Male	Hispanic	15	90	104	30	94	93	2	PDD-NOS
Lydia	Female	White	14	68	96	31	87	104	1	Autism

purposes, and initiating and responding to social interactions more frequently, albeit still inconsistently. His RRB-CSS showed a stable pattern of frequent engagement in repetitive sensory and motor behaviors, stereotyped speech, and repetitive uses of objects. Difficulties interrupting these behaviors affected interaction quality and rapport.

Case 2

'Kyle,' who was seen as part of a clinical research study on early diagnosis, showed a pattern of moderate and stable severity. Kyle was given a best estimate diagnosis of Autistic Disorder at 14 months. He produced five or more words during each Toddler Module administration. In social communication on the Toddler Module, Kyle showed persistent atypical use of eye contact, facial expressions, and gestures, although he did initiate joint attention, point to, and show objects to some degree. He also engaged in unusual sensory interests and exhibited preoccupations/repetitive uses of objects; however, he showed fluctuations in restricted/repetitive behaviors between 19 and 34 months of age, and his sensory interests decreased over time. Kyle experienced a significant increase in verbal skills starting at 22 months, and these skills remained above average from this point.

Case 3

'Roman' is a male with an older brother with ASD, seen as part of a study on early diagnosis. Roman showed a less severe pattern of ASD symptoms overall and ultimately was assigned a DSM-IV diagnosis of PDD-NOS at 24 months. Roman showed few deficits in social interaction and communication early on, although he consistently engaged in mild preoccupations and repetitive uses of objects, reflected in his RRB-CSS trajectory. His Toddler Module scores were mainly in the little-to-no concern range until he developed phrase speech. At that time, mild deficits in social overtures and responses and inconsistent use of eye contact and gestures were observed. After age 24 months, he also began engaging in complex mannerisms. These behaviors led to corresponding increases in domain and total CSS. He scored just under the ASD range at his final appointment, when he showed fewer complex mannerisms and improvements in use of facial expressions, gestures, and showing. Best estimate diagnosis remained PDD-NOS.

Case 4

'Lydia' is a clinic-referred female with an older sister with ASD. Her parents sought an evaluation at 14 months due to concerns about social communication and motor development (an ADOS-2 was not given until she was walking, which occurred at 17 months). Lydia was diagnosed with ASD at 17 months and immediately enrolled in full-time ABA intervention. She was diagnosed with absence seizures at 28 months. She showed a moderate and stable pattern of ASD symptoms over time. Throughout her assessments, Lydia showed deficits in social communication involving limited eye contact and use of gestures or pointing. She consistently shared enjoyment but in a limited number of ways; for example, she frequently smiled and brought toys over to her parents' laps but did not orient the toys or initiate joint attention to distal objects. After initiating intervention, improvements were observed in Lydia's structural language and use of words for a variety of pragmatic purposes, pairing eye contact with social overtures, and participating in structured play. However, she continued to show a high level of repetitive uses of objects and stereotyped speech after starting intervention. Separating the SA-CSS and RRB-CSS trajectories illustrates the relative improvement in social communication variables compared to restricted, repetitive behaviors.

Discussion

As with the CSS for Modules 1–4, the Toddler Module CSS resulted in more uniform distributions across age and language level compared to raw total and domain scores. The CSS was less influenced by child characteristics not specific to ASD, including verbal IQ, than raw total and domain scores. In the original sample, verbal IQ was a significant predictor of raw and domain scores; however, its influence was reduced for CSS compared to raw scores. For Total scores, verbal IQ was reduced from accounting for 26.4 % of the variance in Total-raw to 15.7 % of the variance in Total-CSS. For SA scores, verbal IQ explained 19.3 % of SA-raw, and this was modestly reduced to 16.2 % for SA-CSS. In the case of RRB, verbal IQ accounted for 15.4 % of the variance in RRB-raw and 7.1 % of RRB-CSS. Nonverbal mental age exerted a small but statistically significant influence on Total and SA raw scores and CSS, and nonverbal IQ emerged as statistically significant, accounting for small amounts of the variance in RRB-raw and RRB-CSS. The amount of variance explained by these nonverbal cognitive variables was reduced for RRB-CSS and SA-CSS, but not for Total-CSS. Furthermore, mean Toddler CSS were comparable across Toddler algorithms and to CSS means for Modules 1–4 (de Bildt et al. 2011, Gotham et al. 2009, Hus et al. 2014; Hus and Lord 2014; Shumway et al. 2012), supporting the utility of using these scores for comparisons of children with ASD across modules using cross-sectional data.

Total and domain CSS decreased the influence of verbal IQ less for the Toddler Module than they had for Modules 1–3 (Gotham et al. 2009; Hus et al. 2014). However, it is likely that the behaviors measured by the Toddler Module are less separable from developmental and verbal levels than those measured by later modules. Early measures of verbal skills (such as the Mullen Scales) include items which overlap with ADOS-2 SA items. Thus, the fact that

the influence of child characteristics was not substantially reduced for the Toddler Module, particularly for SA-CSS, is not surprising.

Our replication sample, which included only repeated assessment data, yielded slightly less encouraging results. We observed a similar pattern of reduced influence of verbal IQ on Total-CSS and RRB-CSS compared to raw scores, with the influence of verbal IQ not substantially reduced for SA-CSS compared to SA-raw. Furthermore, children in the Some Words 21–30 group had significantly lower raw scores and CSS than children in the 12–20/Nonverbal group, which was not the case in the original sample. This result could be related to sampling; a larger proportion of the replication dataset consisted of children from prospective studies. Infant siblings accounted for 29 % of the replication sample overall and 31 % of children in the Some Words 21–30 group, compared to 14 and 12 % for children in the original sample. It is reasonable to assume the prospective nature of the studies involving repeated assessments led to some children being seen while ASD symptoms were first emerging. Our original sample included children seen for a single assessment, including the older, more severely affected Wisconsin sample and a higher proportion of children who were clinic-referred. It will be important to replicate these findings in samples with a variety of research- and clinic-referred populations, including younger siblings of children with ASD as well as more clinic referrals, to inform us about the diagnostic and treatment utility of the CSS. Our case examples provide illustration of how the CSS may be used to track development; however, they were not selected to represent overall longitudinal trends for children with ASD. Furthermore, the finding of lower CSS in children in the Some Words 21–30 group in the replication sample underscores the need for care in drawing diagnostic conclusions for young children without significant language impairment, as symptoms may not be as pronounced on the ADOS-2 in these children.

Toddler Module calibrated severity scores should be especially useful in studies examining changes in the behavioral phenotype of ASD over time. Domain CSS may contribute to studies seeking to identify early behavioral patterns that predict ASD risk prior to the emergence of the full disorder. For example, the presence of repetitive behaviors at 12 months has been identified as a key predictor of diagnosis (Ozonoff et al. 2008; Wolff et al. 2014), and changes in repetitive behaviors between age 2 and 3 were a predictor of adult outcomes in a longitudinal study of individuals starting at age 2 through adulthood (Anderson et al. 2014). The Toddler Module RRB-CSS is now available to examine ASD symptoms independent of social communication symptoms. As with other modules, Toddler calibrated severity scores may also be especially useful for studies that examine relationships between genetic or neurobiological markers and dimensional behavioral features of ASD.

There is an emerging evidence base for preventative intervention programs for infants at risk for ASD (Green et al. 2013; Steiner et al. 2013). These programs enroll children as young as 8 months of age due to their risk status as younger siblings of children with diagnosed ASD. There is a need for objective measures of changes and improvements in ASD symptoms for very young children in response to intervention. Toddler Module calibrated severity scores provide a means to track ASD symptoms starting as soon as children are walking, allowing for examination of long-term outcomes for children. However, researchers should be cautioned that the ADOS-2 is a diagnostic measure, and its purpose is to detect core symptoms in ASD in social communication, play, and repetitive behaviors. If children truly move out of a diagnosis of ASD, then the CSS should reflect this trajectory. However, for children with established diagnoses of ASD, calibrated severity scores designed to capture severity of core symptoms may not be expected to abate in the same way that measures of anxiety or ADHD symptoms may show improvement in response to treatment (Hus and Lord 2014). The two children in our case examples who initiated full-time ABA intervention prior to 18 months showed a pattern of some reduction in Social Affect severity but little reduction in severity of repetitive behaviors. However, conclusions cannot be drawn from these anecdotal examples, and more work is needed to examine the utility of the CSS for measuring an individual's response to intervention. Although there is a practical need for tools to measure progress in core symptoms of ASD, it is not recommended that the CSS be used in isolation in making funding or eligibility decisions for intervention.

We reiterate the caution stated in previous studies in which calibrated severity scores for the ADOS-2 were developed (Gotham et al. 2009; Hus et al. 2014; Hus and Lord 2014) and described within the ADOS-2 manual: Toddler Module calibrated severity scores should not be interpreted as an overall measure of a child's level of impairment. These scores are one marker of severity of ASD symptoms, as measured by the ADOS-2, relative to other children with ASD at the same age and language level. Calibrated severity scores provide one piece of information in determining a child's need for supports. Additional assessment of cognitive development, language, adaptive skills, and internalizing and externalizing behaviors is needed to develop a comprehensive picture of a child's needs.

Limitations

As stated earlier, due to the variability in sample sources for this dataset, results may be influenced by recruitment effects. In order to achieve a dataset of very young children large enough to conduct our analyses, data from several different studies with different recruitment patterns were combined. Our dataset consists of consecutive clinic referrals, community-based samples, and participants recruited for a variety of treatment studies and studies specific to high-risk infants. Clinic-referred samples contain potential bias, in that there is evidence that young children with significant delays in developmental skills are more likely to be referred for diagnostic evaluation (De Giacomo and Fombonne 1998; Stone et al. 1994). Moreover, clinic-referred patients under 30 months who are *not* language delayed may be more likely to have significant ASD symptoms, accounting for their early referral (Luyster et al. 2009). Both of these issues may result in a score distribution at the higher end of the range of ADOS-2 scores. On the other hand, children followed prospectively may have initiated research participation before symptoms had clearly manifested, which may have resulted in lower scores on the ADOS-2 compared to a clinic-referred group. An acknowledged limitation of our replication sample is that it was not independent from our original sample; one assessment from children with repeated assessments was randomly selected for inclusion in the original sample, and the replication sample consisted of the remaining assessments from children with multiple assessments only. Results from analyses with the replication sample also should be interpreted with caution due to the known differences in our sample in characteristics of children seen multiple times compared to children seen once.

Toddler Module calibrated severity scores show promise as a tool for behavioral phenotyping of ASD in very young children. Our analyses did not include an examination of patterns of total and domain CSS for children who received nonspectrum diagnoses or for children who were determined to be typically developing. This information is important, as patterns of typical development in very young children can be variable. As practitioners and researchers focus on identifying ASD at younger and younger ages, there is concern that increased awareness of ASD and the push for earlier diagnosis has sometimes led to mislabeling of typical variations in development as ASD (Gnaulati 2013). It will be important to understand the degree of overlap between the dimensions of social communication and repetitive behaviors across children with ASD, other nonspectrum conditions, and typical development. Initial work in this area has been done with toddlers using raw ADOS-2 total and domain scores, and distinct trajectories were identified for children with ASD and those with typical development or other nonspectrum developmental disorders (Chawarska et al. 2009; Lord et al. 2012b). A future direction of our work is to replicate these trajectories of ASD symptoms using the Toddler Module CSS in children with and without ASD.

Conclusion

The current study extends findings of calibrated severity scores for the ADOS-2 Modules 1-4 to the Toddler Module to increase comparability of scores across time, age, and module. Toddler calibrated severity scores are less influenced by verbal level and thus should provide a better metric of ASD symptom severity than raw total and domain scores. However, although this effect was reduced, it was not eliminated, and researchers and clinicians will need to be aware that scores on the Toddler Module are likely to be higher for children with significant language delays. As with Module 1–4 calibrated severity scores, Toddler calibrated severity scores should be replicated in large independent samples to further explore their reliability and clinical utility.

Acknowledgments This research was supported by a University of Minnesota Grant-in-Aid Award to A.E., Lohr Fellowship and Rackham Graduate School Fellowship to V.H.B., Dennis Weatherstone Predoctoral Fellowship to W.G.; in part by NICHD R01HD065272, NIDCD R01DC007462, and CDC U01DD000304 to A.W.; National Institutes of Health, NIDCD R01 DC007223 to S.E.W., and National Institute of Mental Health Grants R01MH081873 and RC1MH089721 to C.L. We gratefully acknowledge all of the families who participated in this research.

Conflict of interest C. Lord and W. Guthrie receive royalties for the ADOS; profits from this study were donated to charity.

References

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Anderson, D. K., Liang, J. W., & Lord, C. (2014). Predicting young adult outcome among more and less cognitively able individuals with autism spectrum disorders. *Journal of Child Psychology and Psychiatry, 55*(5), 485–494.

Bishop, S. L., Richler, J., & Lord, C. (2006). The structure of autism symptoms as measured by the autism diagnostic observation schedule. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence, 12*(4–5), 247–267.

Bryson, S. E., Zwaigenbaum, L., Brian, J., Roberts, W., Szatmari, P., Rombough, V., & McDermott, C. (2007). A prospective case series of high-risk infants who developed autism. *Journal of Autism and Developmental Disorders, 37*(1), 12–24. doi:10.1007/s10803-006-0328-2.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Psychology Press.

Dawson, G., & Bernier, R. (2013). A quarter century of progress on the early detection and treatment of autism spectrum disorder. *Development and Psychopathology, 25*(4), 1455–1472. doi:10.1017/s0954579413000710.

de Bildt, A., Oosterling, I. J., van Lang, N. D. J., Sytema, S., Minderaa, R. B., van Engeland, H., & de Jonge, M. V. (2011). Standardized ADOS scores: Measuring severity of autism spectrum disorders in a Dutch sample. *Journal of Autism and Developmental Disorders, 41*(3), 311–319. doi:10.1007/s10803-010-1057-0.

De Giacomo, A., & Fombonne, E. (1998). Parental recognition of developmental abnormalities in autism. *European Child and Adolescent Psychiatry, 7*(3), 131–136.

Fenson, L., Resznick, S., Thal, D., Bates, E., Hartung, J., Pethick, S., et al. (1993). *The MacArthur communicative development inventories: User's guide and technical manual*. San Diego: Singular Publishing Group.

Gnaulati, E. (2013). *Back to normal: Why ordinary childhood behavior is mistaken for ADHD, bipolar disorder, and autism spectrum disorder*. Boston: Beacon Press.

Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(5), 693–705.

Green, J., Wan, M. W., Guiraud, J., Holsgrove, S., McNally, J., Slonims, V., & Johnson, M. (2013). Intervention for infants at risk of developing autism: A case series. *Journal of Autism and Developmental Disorders, 43*(11), 2502–2514.

Guthrie, W., Swineford, L. B., Nottke, C., & Wetherby, A. M. (2013). Early diagnosis of autism spectrum disorder: Stability and change in clinical diagnosis and symptom presentation. *Journal of Child Psychology and Psychiatry, 54*(5), 582–590. doi:10.1111/jcpp.12008.

Hus, V., Gotham, K., & Lord, C. (2014). Standardizing ADOS domain scores: Separating severity of social affect and restricted and repetitive behaviors. *Journal of Autism and Developmental Disorders*, *44*(10), 2400–2412.

Hus, V., & Lord, C. (2014). The autism diagnostic observation schedule, Module 4: Revised algorithm and standardized severity scores. *Journal of Autism and Developmental Disorders*, *44*(8), 1996–2012.

Kim, S. H., & Lord, C. (2012). New autism diagnostic interview-revised algorithms for toddlers and young preschoolers from 12 to 47 months of age. *Journal of Autism and Developmental Disorders, 42*(1), 82–93. doi:10.1007/s10803-011-1213-1.

Landa, R. J., Holman, K. C., & Garrett-Mayer, E. (2007). Social and communication development in toddlers with early and later diagnosis of autism spectrum disorders. *Archives of General Psychiatry, 64*(7), 853–864.

Lord, C., Luyster, R., Gotham, K., & Guthrie, W. (2012a). *Autism diagnostic observation schedule, 2nd edition (ADOS-2) Manual (Part II): Toddler module*. Torrence, CA: Western Psychological Services.

Lord, C., Luyster, R., Guthrie, W., & Pickles, A. (2012b). Patterns of developmental trajectories in toddlers with autism spectrum disorder. *Journal of Consulting and Clinical Psychology, 80*(3), 477–489. doi:10.1037/a0027214.

Lord, C., Petkova, E., Hus, V., Gan, W. J., Lu, F. H., Martin, D. M., & Risi, S. (2012c). A multisite study of the clinical diagnosis of different autism spectrum disorders. *Archives of General Psychiatry, 69*(3), 306–313. doi:10.1001/archgenpsychiatry.2011.148.

Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of General Psychiatry, 63*(6), 694–701.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H, Jr, Leventhal, B. L., DiLavore, P. C., & 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*(3), 205–223.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012d). *Autism diagnostic observation schedule, 2nd edition (ADOS-2) manual (Part I): Modules 1–4*. Torrence, CA: Western Psychological Services.

Lord, C., Rutter, M., & Lecouteur, A. (1994). Autism diagnostic interview-revised—A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659–685. doi:10.1007/bf02172145.

Lord, C., & Spence, S. J. (2006). Autism spectrum disorders: Phenotype and diagnosis. In S. O. Moldin & J. L. R. Rubenstein (Eds.), *Understanding autism: From basic neuroscience to treatment* (pp. 1–23). Boca Raton, FL: CRC Press. doi:10.1201/9781420004205.ch1

Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., & Lord, C. (2009). The Autism diagnostic observation schedule-Toddler Module: A new module of a standardized diagnostic measure for autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(9), 1305–1320. doi:10.1007/s10803-009-0746-z.

Messinger, D., Young, G. S., Ozonoff, S., Dobkins, K., Carter, A., Zwaigenbaum, L., & Sigman, M. (2013). Beyond Autism: A Baby Siblings Research Consortium study of high-risk children at three years of age. *Journal of the American Academy of Child and Adolescent Psychiatry, 52*(3), 300–308. doi:10.1016/j.jaac.2012.12.011.

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

Ozonoff, S., Macari, S., Young, G. S., Goldring, S., Thompson, M., & Rogers, S. J. (2008). Atypical object exploration at 12 months of age is associated with autism in a prospective sample. *Autism, 12*(5), 457–472.

Ozonoff, S., Young, G. S., Belding, A., Hill, M., Hill, A., Hutman, T., & Iosif, A. M. (2014). The broader autism phenotype in infancy: When does it emerge? *Journal of the American Academy of Child and Adolescent Psychiatry, 53*(4), 398–407. doi:10.1016/j.jaac.2013.12.020.

Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., & Stone, W. L. (2011). Recurrence risk for autism spectrum disorders: A Baby Siblings Research Consortium study. *Pediatrics, 128*(3), E488–E495. doi:10.1542/peds.2010-2825.

Shumway, S., Farmer, C., Thurm, A., Joseph, L., Black, D., & Golden, C. (2012). The ADOS calibrated severity score: Relationship to phenotypic variables and stability over time. *Autism Research, 5*(4), 267–276. doi:10.1002/aur.1238.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland adaptive behavior scales* (2nd ed.). Circle Pines, MN: American Guidance Service.

State, M. W., & Levitt, P. (2011). The conundrums of understanding genetic risks for autism spectrum disorders. *Nature Neuroscience, 14*(12), 1499–1506. doi:10.1038/nn.2924.

Steiner, A. M., Gengoux, G. W., Klin, A., & Chawarska, K. (2013). Pivotal response treatment for infants at-risk for autism spectrum disorders: A pilot study. *Journal of Autism and Developmental Disorders, 43*, 91–102.

Stone, W. L., Hoffman, E. L., Lewis, S. E., & Ousley, O. Y. (1994). Early recognition of autism: Parental reports vs clinical observation. *Archives of Pediatrics and Adolescent Medicine, 148*(2), 174–179.

Venker, C. E., Ray-Subramanian, C. E., Bolt, D. M., & Ellis Weismer, S. (2014). Trajectories of autism severity in early childhood. *Journal of Autism and Developmental Disorders, 44*(3), 546–563. doi:10.1007/s10803-013-1903-y.

Weitlauf, A. S., Gotham, K. O., Vehorn, A. C., & Warren, Z. E. (2014). Brief Report: DSM-5 "Levels of Support": A comment on discrepant conceptualizations of severity in ASD. *Journal of Autism and Developmental Disorders, 44*(2), 471–476. doi:10.1007/s10803-013-1882-z.

Werner, E., & Dawson, G. (2005). Validation of the phenomenon of autistic regression using home videotapes. *Archives of General Psychiatry, 62*(8), 889–895.

Wolff, J. J., Botteron, K. N., Dager, S. R., Elison, J. T., Estes, A. M., Gu, H., et al. (2014). Longitudinal patterns of repetitive behavior in toddlers with autism. *Journal of Child Psychology and Psychiatry*, *55*(8), 945–953.

Woolfenden, S., Sarkozy, V., Ridley, G., & Williams, K. (2012). A systematic review of the diagnostic stability of Autism Spectrum Disorder. *Research in Autism Spectrum Disorders, 6*(1), 345–354. doi:10.1016/j.rasd.2011.06.008.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool language scale, fourth edition (PLS-4)*. San Antonio, TX: Harcourt Assessment.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool language scale, fifth edition (PLS-5)*. San Antonio, TX: Pearson.