Published in final edited form as:

Res Autism Spectr Disord. 2017 September ; 41-42: 57–65. doi:10.1016/j.rasd.2017.08.002.

Comparison of Parent Report and Direct Assessment of Child Skills in Toddlers

Lauren E. Miller, M.S.^a, Kayla A. Perkins, B.A.^{a,1}, Yael G. Dai, B.A.^a, and Deborah A. Fein, Ph.D.^a

^aDepartment of Psychological Sciences, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, CT 06269, USA

Abstract

Background—There are unique challenges associated with measuring development in early childhood. Two primary sources of information are used: parent report and direct assessment. Each approach has strengths and weaknesses, particularly when used to identify and diagnose developmental delays. The present study aimed to evaluate consistency between parent report and direct assessment of child skills in toddlers with and without Autism Spectrum Disorder (ASD) across receptive language, expressive language, and fine motor domains.

Method—109 children were evaluated at an average age of two years; data on child skills were collected via parent report and direct assessment. Children were classified into three groups (i.e., ASD, Other Developmental Disorder, or Typical Development) based on DSM-IV-TR diagnosis. Mixed design ANOVAs, with data source as a within subjects factor and diagnostic group as a between subjects factor, were used to assess agreement. Chi square tests of agreement were then used to examine correspondence at the item level.

Results—Results suggested that parent report of language and fine motor skills did not significantly differ from direct assessment, and this finding held across diagnostic groups. Item level analyses revealed that, in most cases of significant disagreement, parents reported a skill as present, but it was not seen on direct testing.

Conclusions—Results indicate that parents are generally reliable reporters of child language and fine motor abilities in toddlerhood, even when their children have developmental disorders such as ASD. However, the fullest picture may be obtained by using both parent report and direct assessment.

Corresponding Author: Lauren E. Miller, M.S., Department of Psychological Sciences, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, CT 06269, USA; lauren.miller@uconn.edu; Phone: (513) 884-4019; Fax: (860) 486-3680.

¹**Present Address:** Rhode Island Consortium for Autism Research & Treatment, Bradley Hospital, 1011 Veteran's Memorial Parkway, East Providence, RI 02915, USA

Conflicts of Interest

Deborah Fein is part owner of the M-CHAT-R, LLC, which receives royalties from companies that incorporate the M-CHAT-R into commercial products and charge for its use. Data reported in the current article is from the freely available paper version of the M-CHAT-R. Lauren Miller, Kayla Perkins, and Yael Dai declare that they have no potential or competing conflicts of interest.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

parent report; direct assessment; toddlers; child ability; autism spectrum disorder

Introduction

Early detection of developmental delays, including autism spectrum disorder (ASD), has been shown to facilitate earlier intervention and better outcomes (Orinstein et al., 2014; Rogers & Vismara, 2008). These findings have precipitated greater focus on improving routine developmental screening and evaluation rates for at risk young children and the general population. There are, however, unique challenges associated with obtaining accurate developmental data in early childhood, as toddlers tend to respond differently across contexts (Sachse & Von Suchodoletz, 2008). Currently, two primary methods are used to evaluate child development: parent report and direct assessment (Luyster, Kadlec, Carter, & Tager-Flusberg, 2008; Nordahl-Hansen, Kaale, & Ulvund, 2014). There is limited consensus regarding which method offers the best picture of child ability, as each approach has strengths and limitations.

Parents are an important source of information regarding child skills deficits and atypical behaviors, as they are uniquely positioned to observe and interact with children across various situations (Sasche & Von Suchodoletz, 2008). Thus, they may provide data regarding child development that could otherwise not be measured in a clinical setting. Parent report, too, is not subject to issues with child motivation and cooperation that frequently occur in testing situations (Nordahl-Hansen et al., 2014). Furthermore, parent report measures are an increasingly attractive option for detecting developmental delays, as they are quick, easy to use, and cost effective relative to formal evaluation (Nordahl-Hansen et al., 2014; Sasche & Von Suchodoletz, 2008). Based on these strengths, the pediatric healthcare system is moving toward greater involvement of parents in the process of identifying early developmental delays (Feldman et al., 2005; Nordahl-Hansen, Kaale & Ulvund, 2013). Specifically, routine developmental screening via parent report is increasingly being used to identify at risk children, in accordance with American Academy of Pediatrics recommendations (Emerson, Morrell, & Neece, 2016; Johnson & Myers, 2007).

However, concerns have been raised about the accuracy of parent report (Ozonoff et al., 2011; Tomasello & Mervis, 1994; Zapolski & Smith, 2013). Although they are often keen observers of their child's early development, parents generally lack expertise in evaluating developmental milestones, sometimes making it difficult to report reliably (Nordahl-Hansen et al., 2014). Alternatively, parents may attend more to challenging or unusual behaviors, thus introducing bias into their reporting (Zapolski & Smith, 2013). In addition, parents may overestimate child abilities because of a reluctance to acknowledge that their child has a delay (Ozonoff et al., 2011). As parent report necessarily reflects a parent's perception of child functioning, it is considered subjective rather than objective (Sachse & Von Suchodoletz, 2008).

In contrast, standardized tests administered by a trained tester are, by definition, objective. Direct testing should thus offer an unbiased picture of child development, as each child is evaluated in a highly similar (i.e., standard) manner, and data is based on observations made by a professional with expertise and experience in assessing early development (Nordahl-Hansen et al., 2014; Sachse & Von Suchodoletz, 2008). However, in an unfamiliar clinical setting, children tend to behave differently than they may in familiar settings. Issues of behavioral noncompliance and poor attention and motivation commonly seen in young children may influence test results, potentially limiting validity (Nordahl-Hansen et al., 2014). Particularly when evaluating language skills, children may produce fewer utterances when outside of their everyday communicative activities, thus limiting the validity of formal assessment. In addition, formal developmental and diagnostic testing can be time intensive and cost ineffective (Sasche & Von Suchodoletz, 2008). Furthermore, some critics suggest that standardized tests are inappropriate for children with ASD and global delays, as they often measure skills that are too advanced for the child (Luyster et al., 2008). Thus, it is unclear whether direct testing, long regarded as the gold standard, best estimates child abilities in very young children, particularly those with developmental deficits.

To address concerns with both parent report and direct assessment as outlined above, efforts to determine the ideal approach to assessing child development have focused on evaluating agreement between these two primary sources of information. However, the existing literature relies on a variety of methodological approaches and definitions of ‘parent report’ and ‘direct assessment.’ Reliability of parent recall of early developmental milestones and health events, as compared to medical records, indicates that parents are generally good reporters of gross motor milestones (e.g., age at first steps) and medical outcomes (e.g., birth weight, illnesses during early infancy), but that their ability to recall early language milestones is lower (Majnemer & Rosenblatt, 1994; Pless & Pless, 1995). Most often, though, parent report has been used to refer to parent completed screening tools or developmental checklists, which have then been compared to standardized developmental evaluation measures (Bodnarchuk & Eaton, 2004; Luyster et al., 2008; Nordahl-Hansen et al., 2014; Sachse & Von Suchodoletz, 2008; Voigt et al., 2007) or, in the case of ASD symptoms, home videos (Ozonoff et al., 2011). Findings suggest that parents are adequate reporters of child language ability (Nordahl-Hansen et al., 2014), although agreement between parent report and direct assessment is stronger for speech production than for comprehension (Luyster et al., 2008; Sachse & Von Suchodoletz, 2008; Tomasello & Mervis, 1994). This discrepancy may be due to the generally low reliability of measures of early receptive language (Sachse & Von Suchodoletz, 2008). Alternatively, for children with ASD, discrepancies between parent report and direct testing may be an artifact of their difficulty generalizing language across contexts (Nordahl-Hansen et al., 2014). Parents demonstrate good concordance with trained assessors when evaluating gross motor milestones (Bodnarchuk & Eaton, 2004), yet limited research is available on agreement for fine motor skills in early childhood.

The present study aims to evaluate consistency between two sources of information, parent report and direct assessment, when measuring child development in three domains: receptive language, expressive language, and fine motor skills. Here, we define ‘parent report’ as information obtained from a parent during a structured interview, the Vineland Adaptive

Behavior Scales, Second Edition (Vineland), which we then compare to results of direct testing using the Mullen Scales of Early Learning (Mullen). Although the basis of information about a child's development in a parent report is different from that in direct assessment, in that the Vineland aims to assess what skills a child uses in his or her daily life and the Mullen aims to measure a child's competence, both approaches yield similar information at a content level. A similar method of comparing parent report and direct testing of child language skills using the Vineland and Mullen was employed by Luyster et al. (2008) to study development in toddlers with ASD. We expand on this approach by investigating agreement between parent report and direct testing of early developmental outcomes in children with ASD, other developmental disorders, and typical development. The aims of the current study are threefold:

1. We examine consistency between Vineland and Mullen scores in the domains of receptive language, expressive language, and fine motor skills. Based on prior research showing good agreement between parent report and direct testing of language production, but somewhat weaker agreement when assessing language comprehension (Luyster et al., 2008; Nordahl-Hansen et al., 2014; Sachse & Von Suchodoletz, 2008; Tomasello & Mervis, 1994), we expected to find a similar pattern in our data. We also hypothesized that parent report of fine motor skills would be generally consistent with standardized assessment, as similar findings exist for gross motor functioning (Bodnarchuk & Eaton, 2004).
2. We assess the impact of child diagnosis (i.e., ASD, other developmental disorders, or typical development) on consistency between Vineland and Mullen scores. As children with ASD often have difficulty generalizing skills and behavior across contexts (Nordahl-Hansen et al., 2014), we hypothesized that agreement between parent report and direct testing would be weaker for children with ASD as compared to those toddlers with other developmental delays or typical development.
3. We explore agreement between parent report and direct assessment at the individual item level in order to evaluate parent reporting of particular developmental skills. Although we consider these analyses exploratory, we expected to see the greatest discrepancies for items measuring receptive language skills, based on prior research findings (see Aim 1).

Methods

Participants

Participants were 109 children drawn from a larger federally funded project at the University of Connecticut. The aims of the original study focused on validating the Modified Checklist for Autism in Toddlers, Revised (M-CHAT-R), a population based screening tool used to detect ASD in young children (Robins et al., 2014). All participants screened positive on the M-CHAT-R and follow-up phone interview, indicating risk for ASD, between the ages of 16 and 30 months and received a developmental evaluation at the average age of two years. Exclusion criteria included significant sensory or motor impairments (e.g., blindness, severe cerebral palsy) that would negatively impact the participant's ability to complete testing, as

well as documented receptive and expressive language functioning (i.e., Mullen scores) below a 12 month level, which resulted in insufficient item level data for comparative analyses. Participants were also excluded if they had an older sibling with an ASD diagnosis, as infant sibling research has shown that having another child with ASD impacts parent report of developmental concerns (McMahon et al., 2007; Ozonoff et al., 2009).

Participants were classified into three groups based on DSM-IV-TR diagnosis. All diagnoses were assigned based on clinical best estimate judgment of symptoms, incorporating behavioral observation, developmental history, and testing data. The ASD group ($n = 28$) was composed of children with Autistic Disorder ($n = 15$) and Pervasive Developmental Disorder–Not Otherwise Specified ($n = 13$). The other developmental disorder group ($n = 57$) was composed of children with global developmental delay ($n = 34$), developmental language disorder ($n = 22$), and reactive attachment disorder ($n = 1$). The typically developing group ($n = 24$) was composed of children who were either typically developing ($n = 13$) or who exhibited minor delays that were insufficient to qualify for a DSM-IV-TR diagnosis at the time of evaluation ($n = 11$). Participant demographic characteristics are summarized in Table 1.

Procedures

Children who screened positive on the M-CHAT-R and follow-up phone interview were offered a free developmental and diagnostic evaluation, which was completed by a licensed clinical psychologist or developmental pediatrician and a doctoral student in clinical psychology. During the evaluation, a clinician conducted a developmental history and clinical interview with the child's parent using the Vineland and a structured interview for the diagnosis of ASD designed by the research team. The child's cognitive level and autism diagnostic status and severity were assessed using the Mullen and Autism Diagnostic Observation Schedule, Generic (ADOS; Lord, Risi, & Lambrecht, 2000), respectively. Evaluations lasted approximately three hours, including a feedback session in which diagnosis and recommendations were reviewed with the child's parent, with a comprehensive written report to follow. All children were evaluated in their primary language; 98 children were evaluated in English, and 11 were evaluated in Spanish.

The University of Connecticut Institutional Review Board (IRB) approved this study, which was carried out in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments (i.e., 2000 revision). At the time of screening, parents were given an information sheet describing the larger original study. Consent for participation in the research project was indicated by completion of the M-CHAT-R, as a waiver of written consent was granted by the IRB. Written informed consent was obtained from parents at the time of evaluation, prior to inclusion in the current study.

Measures

The *Vineland Adaptive Behavior Scales, Second Edition: Survey Interview Form* (Vineland; Sparrow, Cicchetti, & Balla, 2005) is a semi-structured caregiver interview that assesses adaptive behaviors (i.e., how a child functions in his or her daily life) in the domains of socialization, communication, daily living, and motor skills. Age equivalent scores on the

receptive and expressive language and fine motor subscales were used in the current study. Internal consistency on the Vineland, as measured by split half reliability, is good to excellent, ranging from .81 to .96 for the subscales used and ages tested in the current study. Inter-rater reliability is fair, averaging .70 for subscales for the normative sample aged birth to six years (Sparrow et al., 2005).

The *Mullen Scales of Early Learning* (Mullen; Mullen, 1995) is a developmental assessment of cognitive, motor, and language abilities in young children. The current study used age equivalent scores on the receptive and expressive language and fine motor scales. Average estimates of internal consistency for the Mullen are satisfactory, ranging from .75 to .83 across all scales, and inter-rater reliability is considered strong, ranging from .91 to .99 (Mullen, 1995).

The *Autism Diagnostic Observation Schedule, Generic* (ADOS; Lord et al., 2000) is a semi-structured observational assessment designed to measure symptoms of ASD in toddlerhood through adulthood. The ADOS includes four separate modules based on an individual's expressive language level and chronological age. The current study used Module 1, designed for pre-verbal children and those with single words. Inter-rater reliability on the ADOS is considered good across all domains: social (.93), communication (.84), social communication (.92), and RRBs (.82) (Lord et al., 2000). ADOS scores and observations were included in diagnostic decision making, but were not primary outcome variables of interest in the current study.

Participant performance on primary study measures (i.e., Vineland and Mullen) is summarized in Table 2.

Analysis

All statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 22.0 (IBM Corporation, 2013). Primary analyses included mixed design analysis of variance (ANOVA), with data source (i.e., parent report, direct assessment) as a within subjects factor and diagnostic group (i.e., ASD, other developmental disorder, and typical development) as a between subjects factor, to examine consistency between parent report and direct assessment of child ability. Separate mixed design ANOVAs were run for each developmental domain. The decision to use a mixed design ANOVA was based on the need to compare differences between groups split on two factors: a within subjects factor in which all participants, serving as their own matched pair, were measured in two conditions (i.e., sources of information), and a between subjects factor in which participants were classified separately based on DSM-IV-TR diagnosis. This analytic approach allowed us to examine consistency between parent report and direct testing, while also allowing for investigation of the impact of child diagnosis on agreement between the Vineland and Mullen. Assumptions of normality, homogeneity of variances, and sphericity were met, and no significant outliers were identified in our sample.

Secondary analyses included Chi square tests of agreement on individual matched pairs of items from both primary study measures, to determine agreement at the level of specific developmental skills. In cases where assumptions of Chi square testing were violated due to

small sample sizes (i.e., less than five cases in a contingency table cell), Fisher's Exact test was used. In preparation for these item level analyses, corresponding items on the Vineland and Mullen measuring similar developmental skills were selected. Item scores on the Vineland were recoded to 0 (fail) or 1 (pass) by considering raw scores of 1 or 2 as passes, indicating that a child at least "sometimes" performed the measured skill, as this suggests emerging competence. Raw scores of 0 were maintained as 0 (fail), as this suggests lack of competence. Similarly, Mullen raw item scores were recoded to 0 (fail) or 1 (pass) to match scoring thresholds in corresponding Vineland items. For example, Vineland fine motor item 13, "turns book or magazine pages one by one," corresponds to Mullen fine motor item 14, "turns pages in a book." On this particular Mullen item, a raw score of 0 indicates that a child cannot turn pages in a book, a raw score of 1 indicates that a child turns pages in a book several at a time, and a raw score of 2 indicates that a child turns pages in a book one at a time. Thus, for this item to be consistent with the Vineland, a raw score of 0 or 1 on the Mullen would be recoded as 0 (fail), as the child *cannot* turn pages in a book one by one, and a raw score of 2 would be recoded as 1 (pass), as the child *can* turn pages one at a time. Given the small sample sizes in the ASD and typical development groups, item level analyses were conducted on the full sample instead of separately for each diagnostic group.

An alpha level of .05 was adopted for all statistical tests.

Results

Effect of Demographics on Agreement

Difference scores were calculated by subtracting Vineland subscale age equivalent scores from corresponding Mullen scale age equivalent scores (see Tables 3 and 4 for descriptions of item content). One-way ANOVAs were then run to determine whether the language in which an evaluation was conducted (English or Spanish) had an effect on agreement. Results indicated no significant main effect of evaluation language on agreement between data sources (parent report vs. direct assessment) in the receptive language domain, $F(1, 107) = .086$, $p = .770$, $\eta^2 < .001$, expressive language domain, $F(1, 107) = 3.572$, $p = .061$, $\eta^2 = .032$, or fine motor domain, $F(1, 107) = 1.907$, $p = .170$, $\eta^2 = .018$.

To assess whether maternal education level (high education = some college, Bachelor's degree, advanced degree; low education = some high school, high school diploma/test of general education development (GED), vocational/technical school) impacted agreement in each domain, a second set of one-way ANOVAs was conducted. Results indicated no significant main effect of maternal education on agreement in the receptive language domain, $F(1, 106) = .083$, $p = .774$, $\eta^2 < .001$, expressive language domain, $F(1, 106) = .545$, $p = .462$, $\eta^2 = .005$, or fine motor domain, $F(1, 106) = .345$, $p = .558$, $\eta^2 = .003$.

Effect of Data Source on Domain Scores

Receptive language ability—A main effect of data source was not significant, $F(1, 106) = .765$, $p = .384$, $\eta_p^2 = .007$, suggesting that parents did not significantly differ in their ratings of child receptive language ability compared to direct testing. Additionally, no interaction effect was found between source of information and diagnostic group, $F(2, 106)$

$= .518, p = .597, \eta_p^2 = .010$. A significant main effect of diagnostic group was found in the expected direction, $F(2, 106) = 18.013, p < .001, \eta_p^2 = .254$, such that scores were lower for children in the ASD group than for children in the DD ($p = .047$) and TD ($p < .001$) groups (Table 2). Receptive language scores were also significantly lower for children in the DD group compared to those in the TD group ($p < .001$).

Expressive language ability—A main effect of data source was not significant, $F(1, 106) = .067, p = .796, \eta_p^2 = .001$. Furthermore, an interaction between source of information and diagnostic group was not significant, $F(2, 106) = 1.044, p = .356, \eta_p^2 = .019$. Results did reveal a significant main effect of diagnostic group, $F(2, 106) = 31.467, p < .001, \eta_p^2 = .373$. Expressive language scores were higher for children in the TD group than for children in the ASD ($p < .001$) and DD ($p < .001$) groups, but scores did not differ significantly between children in the ASD and DD groups ($p = .156$) (Table 2).

Fine motor ability—A main effect of data source was not significant, although results did trend in that direction, with parents reporting slightly higher fine motor abilities than seen on direct assessment, $F(1, 106) = 3.880, p = .051, \eta_p^2 = .035$. No significant interaction was found between source of information and diagnostic group, $F(2, 106) = .063, p = .939, \eta_p^2 = .001$. Results revealed a significant main effect of diagnostic group, $F(2, 106) = 7.421, p = .001, \eta_p^2 = .123$, with fine motor scores higher for children in the TD group than for children in the ASD ($p = .001$) and DD ($p = .004$) groups (Table 2). Scores did not differ significantly between children in the ASD and DD groups ($p = .640$).

Item-Level Comparison of Agreement

To determine agreement at the item level, a series of Chi square tests of agreement were performed on individual matched item pairs across diagnostic groups (Table 3). Table 4 describes the direction of item level agreement and disagreement between parents and direct assessment. Percent agreement in Table 3 is defined as the sum of “Both” and “Neither” in Table 4.

Overall, item level analyses revealed somewhat mixed findings. Percent agreement on items assessing basic abilities (e.g., “Listens and looks”) was strong (Table 3). However, there are key limitations to interpreting our kappa values: nearly all scores for easy items accrued in one quadrant of the Chi square contingency table, indicating that both parents and direct testing reported that a child could perform the skill. This contributed to misleadingly low kappa values that do not accurately reflect the strength of agreement between data sources.

For items assessing more advanced comprehension skills (e.g., following instructions), there was some disagreement between parent report and direct testing. For following one- and two-step directions, approximately equal numbers of children could perform the tasks on parent report only or testing only (Table 4). However, on other items measuring more complex developmental skills on receptive language, expressive language, and fine motor domains, there was significant disagreement (Table 3). For these items, parents mostly reported that the child had the skill, but it was not seen on direct testing (i.e., parents reported that the child could name three or 10 objects, recognize three or more body parts,

identify pictures of named objects, turn pages in a book, stack four blocks, and use a hand-wrist twisting motion more often than seen on direct assessment) (Table 4).

Discussion

The current study aimed to evaluate consistency between two primary sources of information about early childhood development, parent report and direct assessment. Specifically, we examined agreement between the Vineland, a parent interview, and the Mullen, a standardized developmental measure, across the domains of receptive language, expressive language, and fine motor skills. To further understand the influence of child diagnosis on agreement, we assessed consistency between data sources in a sample of two-year-old children with ASD, other developmental disorders, and typical development. Finally, we explored consistency at the individual item level to determine if certain skills are more subject to reporter bias, or disagreement. Overall, our results suggest that parent report of child ability does not differ significantly from direct assessment, and this finding is generally stable across diagnostic groups. Taken together, these findings suggest that both parent report and direct testing are appropriate measures of child developmental functioning.

We first aimed to investigate agreement between parent report and direct assessment of child receptive language, expressive language, and fine motor skills. On ANOVA, no significant main effect of data source was found, suggesting that evaluation of overall skills in each domain did not differ based on source of information (i.e., parent report vs. direct testing). This finding is largely consistent with prior research in this area, particularly when assessing language production and gross motor functioning (Bodnarchuk & Eaton, 2004; Luyster et al., 2008; Nordahl-Hansen et al., 2014; Sachse & Von Suchodoletz, 2008). However, contrary to our hypothesis, we did not find weaker agreement when assessing language comprehension ability. Instead, in our sample, parent report of receptive language skills was roughly equivalent to scores demonstrated on direct testing (Table 2). This finding contrasts with prior evidence of relatively weak agreement between parent report and direct assessment of language understanding in young children with ASD (Luyster et al., 2008; Nordahl-Hansen et al., 2014), as well as in late talkers and typically developing toddlers (Sachse & Von Suchodoletz, 2008). It is possible that our reliance on the Vineland, a semi structured parent interview, instead of a parent report checklist (e.g., MacArthur-Bates Communicative Development Inventories, which was used in Luyster et al., 2008, Nordahl-Hansen et al., 2014, and Sachse & Von Suchodoletz, 2008) contributed to improved consistency. Even so, our data showing good agreement suggests that parents are generally reliable reporters of both child expressive and receptive language abilities.

Although no significant disagreement between parent report and direct assessment of fine motor skills was found on ANOVA, results indicated an effect of data source trending toward significance, such that parents reported slightly higher fine motor skills than seen on direct assessment. However, as shown in Table 2, this discrepancy was quite small, with Vineland age equivalent scores only approximately one month higher than scores on the Mullen. This small inconsistency could result from parents assuming that a child can perform a developmentally age appropriate motor task without having actually observed it, or it could be a result of child unwillingness to perform these tasks in the evaluation setting

due to disinterest in testing materials, frustration, or inability to comprehend testing demands. Fairly limited literature exists on agreement between parent report and direct testing of fine motor skills in early childhood, and our hypothesis in this domain was based on findings suggesting good consistency when evaluating gross motor skills. It is possible that parents attend more to whole body (i.e., gross motor) movements, such as sitting, standing, and walking, than finger and hand (i.e., fine motor) movements, as acquisition of gross motor milestones are seen as key signs of whether or not a toddler is developing typically (Bodnarchuk & Eaton, 2004). This may result in less accurate parent reporting of fine motor skills. However, it is equally likely that early developmental testing of fine motor functioning is particularly subject to problems of child noncompliance, as successful testing of emerging fine motor skills depends on a child's interest in and willingness to manipulate testing stimuli.

We then aimed to evaluate the impact of child diagnosis on consistency between parent report and direct assessment. As expected, we found that, across data sources, children in the ASD group were lower functioning than those in the typical development group, with large effect sizes, with toddlers with other developmental disorders generally falling in between (Table 2). However, on ANOVA, we did not find a significant interaction between data source (i.e., parent report vs. direct testing) and diagnostic group, suggesting that a child's diagnostic status did not impact agreement. This finding contradicted our hypothesis, as, based on the tendency of children with ASD to have difficulty generalizing skills across contexts (Nordahl-Hansen et al., 2014), we predicted greater disagreement between parent report and direct testing for children with ASD. Given the low functional level (i.e., language skills developed at a 12 month age equivalence) of children in our ASD group, it is possible that they are not yet experiencing the characteristic difficulties with generalizing skills often seen in older children with ASD (Nordahl-Hansen et al. (2014) examined four-year-olds). Overall, our findings suggest that even parents of children with developmental delays, including ASD, can report accurately on their child's functioning, further supporting the utility of parent report measures for the identification of delays.

We finally aimed to explore agreement between data sources at the individual item level, to determine if specific developmental skills are more sensitive to reporter bias. Although these analyses were exploratory, we hypothesized that we would find greater disagreement for items measuring receptive language ability, given prior research indicating weaker agreement when assessing speech understanding (Luyster et al., 2008; Nordahl-Hansen et al., 2014; Sachse & Von Suchodoletz, 2008; Tomasello & Mervis, 1994). Overall, item level analyses revealed somewhat mixed findings. That is, for items measuring basic skills, even those requiring comprehension of language, results suggested very good agreement, with most parents and clinicians (i.e., through direct assessment) reporting that a child could perform these skills. Yet, for slightly more challenging receptive language items, specifically those assessing a child's ability to follow instructions, there was some disagreement, even though the majority of parents and clinicians agreed that a child could follow one-step instructions and could not follow two-step instructions. For these particular skills, the percentage of only parents or only testing indicating that a child could perform the items was roughly equivalent, indicating no systematic pattern of disagreement (Table 4). A more systematic pattern of disagreement emerged for items tapping more complex skills across all

three functional domains. In these cases, more parents endorsed a child's ability to perform a skill than seen on direct testing, for almost all items.

The patterns of disagreement shown in our item level analyses may reflect underestimation of abilities by testing, overestimation by parents, or differences in child behavior across settings. That is, if a clinician documented that a child performed a skill on direct testing of that skill, then the child clearly has the skill in his or her repertoire. If, however, the child did not perform the item on direct assessment, the parent may still be accurate in reporting that the behavior occurs at home. Alternatively, this disagreement could be related to unexamined child (e.g., age, temperament) or evaluator (e.g., gender, clinician-child interaction style) characteristics. Overall, these findings reflect the challenges associated with assessing child development in the toddler age range. That is, while basic skills appear to be easy to quantify using either parent report or direct testing, accurate evaluation of more complex skills and behaviors is likely more dependent on the standardized measure used, particularly whether or not its stimuli are attractive and attention grabbing for young children, as well as a parent's ability to observe and recollect very specific behaviors (e.g., stacking four blocks) reliably.

Limitations and Future Directions

Due to several limitations, the results of the current study should be interpreted with some caution. Most notably, due to small within groups numbers, item level analyses were conducted on the full sample, limiting our ability to explore patterns of data source agreement or disagreement at the item level within certain clinical populations, such as children with ASD. As such, our item level analyses should be considered exploratory, and future research is needed to examine specific skills that are under or over reported, as well as the influence of child diagnosis on reporting of particular developmental skills. Additionally, as noted previously, specific child and evaluator characteristics, as well as comprehensive parent demographic information (e.g., age), were not directly measured and therefore could not be controlled for in analyses. Due to the number of different clinicians involved in the present study, evaluator characteristics may play a role. Future research should examine the impact of evaluator characteristics on agreement between parent report and direct testing, as these variables may influence the way in which parents, and children, respond in the evaluation setting.

Furthermore, as participants in the current project were drawn from a larger study validating the M-CHAT-R (Robins et al., 2014), our findings may not generalize to all typically developing children, particularly those without a history of any developmental concerns. Given the goals of the larger M-CHAT-R validation study, only children who screened positive on the M-CHAT-R and follow up interview, thus indicating risk for ASD or another developmental disorder, were evaluated. We did not evaluate children who screened negative on the M-CHAT-R, nor did we actively recruit a typically developing sample for comparison. In addition, because some children were excluded from analyses due to very low performance on the Mullen, the current study is not representative of all lower functioning toddlers and does not capture the full range of child abilities across the autism spectrum. Results should thus be generalized with some caution.

In addition, our measure of parent report (Vineland) is a clinician-administered parent interview, rather than a parent checklist done independently. Our findings may not generalize to parent rating scales, including developmental screening tools and child behavior checklists. Further research investigating the role of format (i.e., interview or rating scale) on patterns of disagreement between sources of information on child development, particularly for children with ASD, is indicated. Our standardized developmental measure (Mullen), too, has its limitations. Although widely used in research with young children with ASD, the Mullen was normed in the 1980s and is thus outdated compared to other early developmental tests. Its testing stimuli also reflect its age, and the measure provides only a limited assessment of each core area of development. On the whole, standardized testing only allows for a snapshot of a child at one time point, thereby limiting the ability to capture the full range of a child's functioning. In addition, due to the current study's focus on child capabilities, we did not evaluate agreement on ASD symptoms, which is an important area for further study.

Finally, when a parent reports that a child has not yet attained a skill, but that skill is evidenced on testing, it is clear that the parent is in error. However, when a parent reports that a child has a skill, yet the skill is not seen on direct assessment, there is no way to determine if the parent's report is correct without employing more invasive research and evaluation techniques, such as extended home videos. We also did not systematically ascertain whether a child's behavior and performance during the evaluation was typical for his or her behavior at home. It is important to consider data both from parent report and direct testing when evaluating child development, particularly for young children with potential delays.

Implications

To our knowledge, this is one of few studies investigating agreement between parent report and direct assessment of child ability, particularly at the item level, in a broad sample of children with ASD, other developmental disorders, and typical development. Taken together, the results of the present study suggest that parents are generally a reliable source of information regarding child language and motor abilities, as we found good overall consistency between Vineland and Mullen scores. Additionally, parent sociodemographic factors, such as level of education and language spoken, as well as child diagnostic status, do not appear to significantly impact the accuracy of reporting. Although our findings suggest that parents are generally reliable reporters of child ability in toddlerhood, given the challenges associated with assessing skills in early development, the fullest picture of child functioning may be obtained by using both sources of information. Thus, results argue for the combined utility of parent report and direct assessment in creating an accurate composite of child behavior. As such, it appears that parents can, and should, be utilized to facilitate early detection and diagnosis of developmental delays, including ASD. As the healthcare climate shifts to further incorporate parent reporting into pediatric preventative care visits, particularly through routine developmental screening, it is increasingly necessary to improve our understanding of the accuracy of parent report of child ability and, perhaps more importantly, to identify and address any barriers (i.e., parent, child, or clinician characteristics) to the utility of parent reporting.

Acknowledgments

This study was funded by grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) R01 HD039961 and the Maternal and Child Health Bureau (MCHB) R40 MC00270. Neither the NICD nor the MCHB had any involvement in the study design; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to submit the article for publication.

References

- Bodnarchuk JL, Eaton WO. Can parent reports be trusted? Validity of daily checklists of gross motor milestone attainment. *Applied Developmental Psychology*. 2004; 25:481–490.
- Emerson ND, Morrell HE, Neece C. Predictors of Age of Diagnosis for Children with Autism Spectrum Disorder: The Role of a Consistent Source of Medical Care, Race, and Condition Severity. *Journal of Autism and Developmental Disorders*. 2016; 46(1):127–138. [PubMed: 26280401]
- Feldman HM, Dale PS, Campbell TF, Coburn DK, Kurs-Lasky M, Rockette HE, Paradise JL. Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development*. 2005; 76(4):856–868. [PubMed: 16026501]
- IBM Corporation. IBM SPSS statistics for Windows, version 22.0. Armonk, NY: Author; 2013.
- Johnson CP, Myers SM. Identification and evaluation of children with autism spectrum disorders. *Pediatrics*. 2007; 120(5):1183–1215. [PubMed: 17967920]
- Lord C, Risi S, Lambrecht L. 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]
- Luyster RJ, Kadlec MB, Carter A, Tager-Flusberg H. Language assessment and development in toddlers with autism spectrum disorders. *Journal of autism and developmental disorders*. 2008; 38(8):1426–1438. [PubMed: 18188685]
- Majnemer A, Rosenblatt B. Reliability of parental recall of developmental milestones. *Pediatric Neurology*. 1994; 10:304–308. [PubMed: 8068156]
- McMahon CR, Malesa EE, Yoder PJ, Stone WL. Parents of children with autism spectrum disorders have merited concerns about their later-born infants. *Research and Practice for Persons with Severe Disabilities*. 2007; 32(2):154–160.
- Mullen, EM. Mullen Scales of Early Learning. Bloomington, MN: Pearson; 1995.
- Nordahl-Hansen A, Kaale A, Ulvund S. Inter-rater reliability of parent and preschool teacher ratings of language in children with autism. *Research in Autism Spectrum Disorders*. 2013; 7:1391–1396.
- Nordahl-Hansen A, Kaale A, Ulvund SE. Language assessment in children with autism spectrum disorder: Concurrent validity between report-based assessments and direct tests. *Research in Autism Spectrum Disorders*. 2014; 8(9):1100–1106.
- Orinstein AJ, Helt M, Troyb E, Tyson KE, Barton ML, Eigsti IM, Naigles L, Fein DA. Intervention for optimal outcome in children and adolescents with a history of autism. *Journal of Developmental and Behavioral Pediatrics: JDBP*. 2014; 35(4):247–256. [PubMed: 24799263]
- Ozonoff S, Iosif AM, Young GS, Hepburn S, Thompson M, Colombi C, ... Rogers SJ. Onset patterns in autism: Correspondence between home video and parent report. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2011; 50(8):796–806. [PubMed: 21784299]
- Ozonoff S, Young GS, Steinfeld MB, Hill MM, Cook I, Hutman T, ... Sigman M. How early do parent concerns predict later autism diagnosis? *Journal of Developmental and Behavioral Pediatrics*. 2009; 30:367–375. [PubMed: 19827218]
- Pless CE, Pless IB. How well they remember: The accuracy of parent reports. *Archives of Pediatric and Adolescent Medicine*. 1995; 149(5):553–558.
- Robins DL, Casagrande K, Barton M, Chen CA, Dumont-Mathieu T, Fein D. Validation of the Modified Checklist for Autism in Toddlers, Revised with Follow-up (M-CHAT-R/F). *Pediatrics*. 2014; 133(1):37–45. [PubMed: 24366990]
- Rogers SJ, Vismara LA. Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child & Adolescent Psychology*. 2008; 37(1):8–38. [PubMed: 18444052]

- Sachse S, Von Suchodoletz W. Early identification of language delay by direct language assessment or parent report? *Journal of Developmental and Behavioral Pediatrics*. 2008; 29:34–41. [PubMed: 18300723]
- Sparrow, SS., Cicchetti, DV., Balla, DA. *Vineland adaptive behavior scales*. 2. Circle Pines, MN: American Guidance Service; 2005.
- Tomasello M, Mervis CB. The instrument is great, but measuring comprehension is still a problem. *Monographs of the Society for Research in child Development*. 1994; 59(5):174–179.
- Voigt RG, Llorente AM, Jensen CL, Kennard Fraley J, Barbaresi WJ, Heird WC. Comparison of the validity of direct pediatric developmental evaluation versus developmental screening by parent report. *Clinical Pediatrics*. 2007; 46(6):523–529. [PubMed: 17579105]
- Zapolski TCB, Smith GT. Comparison of parent versus child-report of child impulsivity traits and prediction of outcome variables. *Journal of Psychopathology and Behavioral Assessment*. 2013; 35:301–313. [PubMed: 24039341]

Highlights

- Parent report of child receptive and expressive language and fine motor skills is generally consistent with direct assessment.
- Child diagnosis does not influence consistency between parent report and direct testing.
- When disagreement between sources exists, parents are more likely to report a skill as present than seen on direct testing.

Table 1

Participant Demographic Characteristics

Variable	ASD Group (<i>n</i> = 28)	DD Group (<i>n</i> = 57)	TD Group (<i>n</i> = 24)
<i>Age at Evaluation (M (SD))</i>	24.71 (4.74)	23.75 (4.07)	23.58 (4.84)
<i>Gender</i>			
Male (%)	67.9	61.4	50.0
Female (%)	32.1	38.6	50.0
<i>Race/Ethnicity</i>			
Caucasian (%)	60.7	50.9	45.8
African American (%)	14.3	8.8	4.2
Asian/Pacific Islander (%)	7.1	0	12.5
American Indian (%)	0	1.8	0
Biracial (%)	3.6	0	4.2
Hispanic/Latino (%)	14.3	36.7	33.3
Missing data (%)	0	1.8	0
<i>Maternal Education</i>			
Some high school (%)	14.3	26.3	4.2
High school diploma/GED (%)	10.7	19.3	8.3
Vocational/technical school (%)	0	7.0	12.5
Some college (%)	50.0	19.3	25.0
Bachelor's degree (%)	17.9	15.8	33.3
Advanced degree (%)	7.1	10.5	16.7
Missing data (%)	0	1.8	0
<i>Evaluation in Spanish (n)</i>	0	8	3

Note. ASD = autism spectrum disorder; DD = other developmental disorder; TD = typically developing.

Table 2

Average Vineland and Mullen Age-Equivalent Scores

Measure	ASD Group (<i>n</i> = 28)			DD Group (<i>n</i> = 57)			TD Group (<i>n</i> = 24)		
	<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range
<i>Receptive Language</i>									
Vineland	13.30	10.84	1–43	15.96	5.72	3–26	21.63	6.99	1–34
Mullen	12.96	4.43	3–26	16.54	5.45	9–28	23.14	5.16	15–34
<i>Expressive Language</i>									
Vineland	12.50	5.14	5–26	15.02	4.70	6–30	22.00	6.41	10–34
Mullen	13.50	2.43	8–18	14.65	4.15	5–24	21.71	5.97	15–37
<i>Fine Motor</i>									
Vineland	19.46	6.81	6–28	20.54	6.06	9–38	24.26	5.93	15–41
Mullen	18.64	3.20	13–26	19.53	4.60	10–34	23.12	5.02	15–33

Note. ASD = autism spectrum disorder; DD = other developmental disorder; TD = typically developing.

Table 3

Chi Square Analyses of Agreement on Item Pairs

Item Description	Agreement (%)	χ^2	<i>p</i>	κ	Φ or OR
<i>Receptive Language</i>					
Listens and looks	97.2		.999	-.012	-
Recognizes name	91.7		.007	.358	16.2
Understands "no"	93.6		.173	.189	8.4
Follows one-step instructions	75.2	19.191	< .001	.419	.42
Recognizes 3+ body parts	62.4	17.746	< .001	.318	.40
Follows two-step instructions	77.1	16.898	< .001	.393	.39
Identifies pictures of named objects	69.7	23.150	< .001	.413	.46
<i>Expressive Language</i>					
Names 3+ objects	58.7	15.743	< .001	.252	.38
Names 10+ objects	76.1		.001	.216	-
<i>Fine Motor</i>					
Turns pages in book one by one	69.7		.052	.160	3.5
Stacks four blocks vertically	77.1	29.534	< .001	.516	.52
Uses hand-wrist twisting motion	82.6		< .001	.462	19.5

Note. Where Chi square statistics are absent, Fisher's Exact *p*-values are reported due to violations of Chi square assumptions. A significant *p*-value indicates significant disagreement between methods. For items using Fisher's Exact test, an odds ratio (OR) was used as a measure of effect size, except in cases where one cell in the contingency table equaled 0.

Table 4

Categorization of Evaluator Response by Item

Item Description	Evaluator Reporting Child's Ability to Perform Item			
	Neither (%)	Parent Only (%)	Testing Only (%)	Both (%)
<i>Receptive Language</i>				
Listens and looks	0.0	1.8	0.9	97.2
Recognizes name	2.6	2.6	5.5	89.0
Understands "no"	0.9	2.8	3.7	92.7
Follows one-step instructions	18.3	11.0	13.8	56.9
Recognizes 3+ body parts	36.7	35.8	1.8	25.7
Follows two-step instructions	63.3	10.1	12.8	13.8
Identifies pictures of named objects	42.2	26.6	3.7	27.5
<i>Expressive Language</i>				
Names 3+ objects	42.2	41.3	0.0	16.5
Names 10+ objects	71.6	23.9	0.0	4.6
<i>Fine Motor</i>				
Turns pages in book one by one	6.4	25.7	4.6	63.3
Stacks four blocks vertically	26.6	14.7	8.3	50.5
Uses hand-wrist twisting motion	71.6	14.7	2.8	11.0