Research in Autism Spectrum Disorders 49 (2018) 47–55

Contents lists available at ScienceDirect

Research in Autism Spectrum Disorders

journal homepage: www.elsevier.com/locate/rasd

Differences in verbal and nonverbal IQ test scores in children with autism spectrum disorder

Sabrina N. Grondhuis[a,*], Luc Lecavalier[b], L. Eugene Arnold[c], Benjamin L. Handen[d], Lawrence Scahill[e], Christopher J. McDougle[f], Michael G. Aman[g]

[a] *Millsaps College, Jackson, MS, United States*
[b] *The Ohio State University, Columbus, OH, United States*
[c] *Ohio State University Wexner Medical Center, Columbus, OH, United States*
[d] *University of Pittsburgh School of Medicine, Pittsburgh, PA, United States*
[e] *Emory University, Marcus Autism Center, Atlanta, GA, United States*
[f] *Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States*
[g] *The Ohio State University, The Nisonger Center UCEDD, Columbus, OH, United States*

A R T I C L E I N F O

Number of reviews completed is 2

Keywords:
Autism spectrum disorder
ASD
Intelligence
IQ
Language
Assessment

A B S T R A C T

Background: Intelligence tests are a fundamental component of diagnostic assessments for children with suspected autism spectrum disorder (ASD). There are many assessments available, but scores across tests may not be comparable in children with ASD.

Method: Eighty children (68 boys) age 4–14 years with ASD completed the Stanford-Binet Intelligence Scales, 5th Edition (SB5), a verbal measure of intelligence, and the Leiter International Performance Scale – Revised (Leiter-R), a nonverbal measure of intelligence.

Results: Although discrepancies went in both directions, we found significantly higher mean scores on the Leiter-R than SB5 (9.6 point difference). Children younger than 8 years had more pronounced discrepancies (13.8 points vs. 3.5 points for > 8 years), and children with less-developed language skills had greater discrepancies (13.1 points vs. 5.8 points for higher language skills).

Conclusion: This suggests these IQ tests are not interchangeable and language demands may produce different results that could impact clinician interpretation. Both clinicians and researchers should be aware of the likely impact of adopting primarily verbal vs. nonverbal tests when assessing children with ASD, especially those with less language.

1. Introduction

Autism spectrum disorder (ASD) is a heterogeneous disorder with about 50% classified as having below average intellectual functioning (Centers for Disease Control & Prevention, 2014). Because intellectual ability is highly predictive of long-term outcomes in children with ASD, IQ testing is a fundamental component of assessment in this population (Klaiman, Fernandez-Carriba, Hall, & Saulnier, 2015). The Diagnostic and Statistical Manual for Mental Disorders, 5th edition (DSM-5; American Psychiatric Association, 2013) requests that clinicians document the presence or absence of intellectual disability (ID) in the diagnostic evaluation of children and adults with ASD. Knowledge of the person's IQ is also often helpful both in clinical care and in research.

Results of IQ testing for children with ASD provide a baseline measurement for future comparisons, may be essential to determine

* Corresponding author.
E-mail address: grondsn@millsaps.edu (S.N. Grondhuis).

https://doi.org/10.1016/j.rasd.2018.02.001
Received 22 September 2017; Received in revised form 17 January 2018; Accepted 1 February 2018
1750-9467/

appropriate school placement, and often inform treatment planning. For example, test results may guide clinical decisions on the selection of treatment programs and advise parents and educators on supports needed for the child. Intelligence tests may be based on different theories of intelligence and thus vary in their focus, degree of dependence on language, and target age group. Further, core symptoms of ASD, such as restricted and repetitive behaviors, can impede administration of the test, making testing challenging and limiting the utility of the result (Baum, Shear, Howe, & Bishop, 2015). Because of this variability, available IQ tests may not produce congruent results; further, scores may be impacted by child characteristics and symptomatology that are independent of intelligence.

Most intelligence tests (e.g., Wechsler Intelligence Scales for Children, 5th edition [WISC-V]; Wechsler, 2014; Stanford-Binet Intelligence Scales, 5th edition [SB5]; Roid, 2003) include assessment of multiple domains (e.g., quantitative, working memory, processing speed) in order to evaluate several aspects of intelligence. Scores in these different areas are combined to produce composite indices of intellectual ability. Questions on these types of IQ tests are usually verbally administered and answers are either verbal or involve task completion. Implicit in the ASD core symptom of social communication deficits (DSM-5; American Psychiatric Association, 2013) is the frequent co-occurrence of language problems. Language difficulties may range from relatively mild, with delays in meeting language milestones, to as severe as being functionally nonverbal (American Psychiatric Association, 2013; Tager-Flusberg, 2015). Thus, reliance on language for assessing IQ may be problematic and could potentially result in underestimation of the individual's cognitive abilities. This was demonstrated in a report by Coolican et al. (2008) that children with ASD had weaker verbal performance compared to nonverbal sections on the SB5 (2012), and by findings of other researchers (Mayes & Calhoun, 2003) using different intelligence assessments. Nonverbal IQ tests, such as the Leiter International Performance Scale–Revised (Leiter-R; Roid & Miller, 1997) and the Test of Nonverbal Intelligence, Fourth Edition (TONI-4; Brown, Sherbenou, & Johnsen, 2010), offer viable alternatives to verbal IQ tests. These tests do not require verbal answers, are less reliant on verbal instructions, and can provide reliable and valid results on cognitive performance across a variety of domains (Roid & Miller, 1997).

Intelligence tests are designed to be interchangeable because the tests strive to measure general cognitive abilities, albeit not always using identical approaches. This reciprocal element is for ease of comparison and to prevent retest effects from overexposure to a single instrument. While results on different IQ tests may be interchangeable in typically developing children, they may not be equivalent in children with ASD. For example, Nader, Courchesne, Dawson, & Soulières (2016) compared the WISC-IV (Wechsler, 2003) with Raven's Progressive Matrices (Raven et al., 1998) and found no difference in performance for typical children. However, in children with ASD, scores on Raven's Progressive Matrices were significantly higher than on the WISC-IV full scale.

The Stanford-Binet Intelligence Scales (SB) and Leiter International Performance Scale (Leiter) have also been compared in this respect. In 1983, Reeve, French, & Hunter(1983) compared the Leiter (Leiter, 1979) and the SB Form L-M (Terman & Merrill, 1972). No significant discrepancies were found in typically developing children. Conversely, Grondhuis and Mulick (2013) evaluated the SB5 and Leiter-R in children between 3 and 12 years of age with autistic disorder or pervasive developmental disorder not otherwise specified (PDD-NOS). They found substantial differences between tests. Children in their sample scored higher (by 21.91 points) on the Leiter-R than on the SB5 (SD = 15.72); children younger than 6 years displayed greater discrepancies between tests (M = 28.56 points) than those 6 years and older (M = 16.58 points). The sample was small (N = 47), and it only included children with two of the DSM-IV-TR subtypes (autistic disorder or PDD-NOS), warranting further study.

The current study used data from a randomized trial of risperidone alone versus risperidone plus parent training in children with ASD conducted by the Research Units on Pediatric Psychopharmacology (RUPP) Autism Network (Aman et al., 2009; Scahill et al., 2012). Here, we compare the SB5 and the Leiter-R, two IQ assessments frequently used in children with ASD, and we also explore differences across age, sex, and communication ages. Our hypotheses included:

1. Mean Leiter-R IQs will be significantly higher than SB5 IQs for these participants.
2. Younger children will display greater differences between Leiter-R and SB5 IQ, as their ASD-related communication deficits will be relatively more pronounced than in older participants.
3. Division by Communication age on the Vineland Adaptive Behavior Scales (Vineland; Sparrow, Balla, & Cicchetti, 1984) will show greater discrepancies between IQ tests, favoring the Leiter-R for children having lower communication ages.

2. Methods

2.1. Background

The three-site study of 124 children was conducted by the RUPP Autism Network between 2004 and 2007. The Institutional Review Boards at Indiana University, Ohio State University, and Yale University approved the study and written informed consent was obtained from parents prior to participation. Eligible subjects were randomly assigned in a 2:3 ratio to risperidone alone or to combined treatment (risperidone plus parent training) for 24 weeks (Aman et al., 2009; Scahill et al., 2012). Subjects were required to be healthy and medication-free by baseline of the study (Scahill et al., 2009). Participants were 4 to 14 years of age, had DSM-IV-TR (American Psychiatric Association, 2000) diagnoses of autistic disorder, Asperger's disorder, or PDD-NOS, IQs ≥ 35, and displayed severe disruptive behavior (18 ≥ on the Irritability subscale of the Aberrant Behavior Checklist and a Clinical Global Impression – Severity score ≥4). Although these children were diagnosed using DSM-IV-TR criteria, for ease of communication we shall refer to their diagnoses as ASD.

2.2. Instruments

2.2.1. Leiter International Performance Scale–Revised

The Leiter-R is an individually administered nonverbal test of IQ (standard score range: 30–170) with normative data covering ages 2 years, 9 months through 20 years, 11 months (Roid & Miller, 1997; Roid, Pomplun, & Martin, 2009). It is often used to assess individuals with communication deficits, hearing impairments (Sparrow & Davis, 2000), or with a primary language other than English (Athanasiou, 2000). In this study, four subtests (Figure Ground, Form Completion, Sequential Order, and Repeated Patterns) were administered to derive the Brief IQ (BIQ) score. The full-scale administration would have included a total of six subtests, but, to lessen participant burden, the BIQ format was used. Thus, all comparisons in this study used the Leiter BIQ.

2.2.2. Stanford-Binet Intelligence Scales, 5th edition

The SB5 is also an individually administered IQ test encompassing five subtests (Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual Spatial Processing, and Working Memory; Roid, 2003) for individuals from age 2 to 85 years (standard score range: 40–160). The SB5 assesses each subtest once for the Verbal scales and once for the Nonverbal scales, which do not require verbal answers (Coolican, Bryson, & Zwaigenbaum, 2008). Although Nonverbal items do not require verbal *responses*, directions are administered verbally, which could undermine performance in children with language impairment. Thus, the SB5 nonverbal section is not a completely nonverbal IQ measure. In this study, the SB5 included full-scale IQ (FSIQ), nonverbal IQ (NVIQ), and verbal IQ (VIQ) to offer a comprehensive evaluation. While the NVIQ may be an obvious comparison, the SB5 FSIQ is commonly used in clinical settings, and thus we believed understanding performance differences between verbal and nonverbal assessments was important. The VIQ was included because of the heterogeneous presentation of ASD; not all children diagnosed with ASD have comorbid language difficulties and varying cognitive profiles in this population are indeed very common (Coolican et al., 2008).

2.2.3. Vineland Adaptive Behavior Scales

The Vineland (Sparrow, Balla, & Cicchetti, 1984) assesses four domains of adaptive behavior: Motor Skills, Daily Living Skills, Socialization, and Communication. The scales were normed on a nationally representative sample and have high reliability (range 0.93–.99). Caregivers rate behaviors within the four areas on a scale of 0 (behavior not performed) to 2 (behavior performed on a regular basis), which are then calculated into standardized scores and age equivalencies for each domain. Although the Vineland is not a structured language assessment, we used the Communication domain to divide participants into lower- and higher-functioning language groups to study effects of language ability. The Vineland evaluates communication from a lifespan prospective and thus includes a variety of items geared towards both nonverbal and verbal communication, and receptive and expressive communication. Although there is an association between Vineland communication and IQ in general (e.g., Klin et al., 2007; Perry, Flanagan, Dunn Geier, & Freeman, 2009), this is contentious and some believe the relationship between the two constructs is not consistent (Abbeduto & Short, 1994). For instance, persons with ASD displayed greater deficits in the Communication domain than peers with ID matched for IQ (Perry et al., 2009). This suggests that adaptive behavior and IQ are not synonymous, but instead evaluate similar, yet inherently different concepts. The Vineland was independent of the two IQ tests, and this enabled us to avoid using either the SB5 or Leiter for partitioning IQ, which would have introduced a confound.

2.3. Participants

Of 124 children in the intervention trial, 110 were assessed with the Leiter-R, 86 with the SB5, and 80 participants completed both tests. Because some participants did not complete all sections (NVIQ, VIQ, or FSIQ) on the SB5, sample sizes differ across subtests depending on the section assessed (range 78–80; see Table 1). In our primary analyses of the Leiter-R BIQ vs. SB5 FSIQ, the sample

Table 1
Summary of Means and Standard Deviations on Leiter-R Brief IQ and SB5 FSIQ Comparisons.

Comparison		*Leiter-R: M (SD)*	*SB5: M (SD)*
BIQ vs. FSIQ (n = 80)		78.33 (19.83)	68.74 (20.93)
Age			
	Younger Group (< 8 years; n = 47)	83.19 (16.70)	69.34 (18.40)
	Older Group (≥ 8 years; n = 33)	71.39 (22.04)	67.88 (24.36)
Sex			
	Males (n = 68)	77.74 (20.08)	68.85 (21.39)
	Females (n = 12)	81.68 (18.83)	68.08 (18.98)
Level of Communication			
	Vineland AE < 54 months (n = 41)	74.27 (21.27)	61.15 (19.21)
	Vineland AE ≥ 54 months (n = 39)	82.59 (17.46)	76.72 (19.87)
SB5 Section			
	BIQ vs. *SB5* NVIQ (n = 78)	78.68 (19.69)	72.96 (20.51)
	BIQ vs. *SB5* VIQ (n = 78)	78.68 (19.69)	68.54 (20.98)

Note: BIQ = Brief IQ on *Leiter-R*; FSIQ = Full scale IQ on *SB5*; NVIQ = Nonverbal IQ on *SB5*.
VIQ = Verbal IQ on SB5; Vineland = Vineland Adaptive Behavior Scale; AE = Age Equivalency.

contained 68 boys (85%), was predominantly Caucasian (73.8%), and ranged in age from 4.1 to 13.8 years ($M = 7.8$ years, $SD = 2.3$ years). We split the sample at 8 years, not only because that age integer was closest to the mean age of the sample, but also because literature suggests that IQ demonstrates relative stability by the age of 8 years in children with ASD (Mayes & Calhoun, 2003a). The younger group (age < 8 years) accounted for 58.8% of the sample ($n = 47$), whereas there were 33 participants (41.3%) in the older group (age ≥ 8 years). For analyses using the Communication age equivalency on the Vineland, a mean split at 54 months (4.5 years) was used. Research indicated early language development was a crucial predictor of later success, and those with developed speech by age 5 had better outcomes than children who did not (Howlin & Moss, 2012). A split at 54 months allowed for nearly equal groups due to the sample's median (41 children [51.3%] Vineland Communication age equivalency < 54 months). Fifty-four months is also a close approximation to 5 years.

Children were evaluated with a systematic assessment procedure by an experienced team at each site using DSM-IV-TR criteria. The diagnosis of ASD was supported by parental interview with the Autism Diagnostic Interview-Revised (ADI-R; Lord, Rutter, & Le Couteur, 1994). Forty-seven children (58.8%) were diagnosed with autistic disorder, 7 (8.8%) with Asperger's disorder, and 26 (32.5%) with PDD-NOS.

2.4. Analytic plan

We examined demographic and clinical characteristics of the 80 participants with both Leiter-R and SB5 scores versus the 44 subjects without both test scores. The 80 participants with *both* IQ measures were significantly older than the 44 children with partial data (7.8 years in our sample versus 6.6 years in the total sample; $t = 2.679$, $p = .008$). The following statistics were computed for each comparison of interest (Leiter-R BIQ vs either SB5 FSIQ, NVIQ, or VIQ): (1). 2 by 2 analyses of variance (ANOVAs) as a function of the following variables: (a) IQ test (Leiter-R and SB5) and age (below or equal to and above 8 years) and (b) IQ test (Leiter-R and SB5) and Vineland Communication age equivalency (below or equal to and above 54 months). (2). We ran a hierarchical regression predicting scores on Leiter-R by SB5 in the first block followed by age (as a continuous variable) in the second block to determine additional variance accounted for by age. (3). We computed Pearson's correlation between Leiter-R and SB5; and, (4). We derived deviation scores between the two IQ tests, and descriptive categorical switches from the Leiter-R to the SB5. All pairs of data were included in the analyses, regardless of position within an individual test's range (e.g., at test floor) to contrast performance across measures and better understand real-world implications. All analyses were conducted using IBM SPSS Statistics 22 (IBM Corp, 2013).

3. Results

3.1. Primary comparison using the leiter-R brief IQ and SB5 full scale IQ

The mean Leiter-R BIQ score was 78.3 ($SD = 19.8$, range 36–119) and the mean SB5 FSIQ was 68.7 ($SD = 20.9$, range 40–116; see Table 1), a mean difference of 9.6 ($SD = 15.2$) between the two measures. Beyond the 9.6 point group difference, there were many cases for whom the differences were more extreme. For example: one participant's FSIQ from the SB5 was 30 points higher than the Leiter-R, whereas another participant had a Leiter-R score 47 points higher than the FSIQ on the SB5 (see Fig. 1.). Large differences (greater than the SD of 15 points) occurred in four cases with higher SB5 FSIQ scores and 22 cases with higher Leiter-R scores. Further, 41 children (51.3%) would have been placed in a higher descriptive IQ category (e.g., moving from moderate intellectual disability to mild intellectual disability) if assessed with the Leiter-R rather than the SB5, and nine children (11.3%) would have been assigned to a higher IQ category if assessed with the SB5. In all, only 37.5% ($n = 30$) did *not* change IQ category based on the actual scores received on the two instruments.

The ANOVA used for the age analyses [(Leiter-R vs. SB5) × 2 (younger vs. older)] demostrated a significant interaction effect, $F(1, 78) = 10.06$, $p = .002$, partial $\eta^2 = .114$, which indicates that younger and older children showed different measurement variance between the two instruments. There was a simple main effect for age that showed children younger than 8 years had larger discrepancy between tests (mean test difference of 13.8 points, $SD = 15.0$; $p < .001$) than children above 8 years (mean test difference of 3.5, $SD = 13.6$; $p = .163$; see Fig. 2.). There was also a simple main effect for the Leiter-R, as there were substantial differences in performance by age group (mean age group difference of 11.8 points, $SD = 4.3$, $p = .008$); no simple effect was found for the SB5. There was no significant main effect for age. The correlation between the Leiter-R and SB5 was strong at $r = .73$ ($p < .001$), with the SB5 accounting for 52.6% of Leiter-R variance, $F(1, 78) = 86.44$, $p < .001$. When age was added into the regression analysis the correlation between the tests increased to 0.77; variance accounted for by age was 7.2%, $F(1, 77) = 13.88$, $p < .001$.

When the sample was dichotomized at 4.5 years by Vineland Communication age, a significant interaction was found between IQ test and Communication age equivalency, $F(1,78) = 4.80$, $p = .03$, partial $\eta^2 = .058$; see Fig. 3. This showed test performance was influenced by Communication age and IQ test used. Simple main effects analyses showed the lower Communication age group (< 4.5 years) had significant differences between tests (mean test difference = 13.1 points, $SD = 2.3$, $p < .001$) as did the higher Communication age group (mean test difference = 5.9, $SD = 2.4$, $p = .015$). There was also a simple main effect for the SB5; children in the older Communication age group received significantly higher scores on the SB5 than the younger Communication age group (mean difference between Communication groups = 15.5, $SD = 4.4$, $p = .001$) while the differences between age groups on the Leiter-R failed to reach significance ($p = .060$).

We partitioned the sample by Communication age and by IQ (< 70; ≥ 70) and conducted Fisher's exact tests. With the Leiter, 16 children with Communication age less than 4.5 years had an IQ < 70 and 25 had an IQ ≥ 70. Conversely, of those with higher

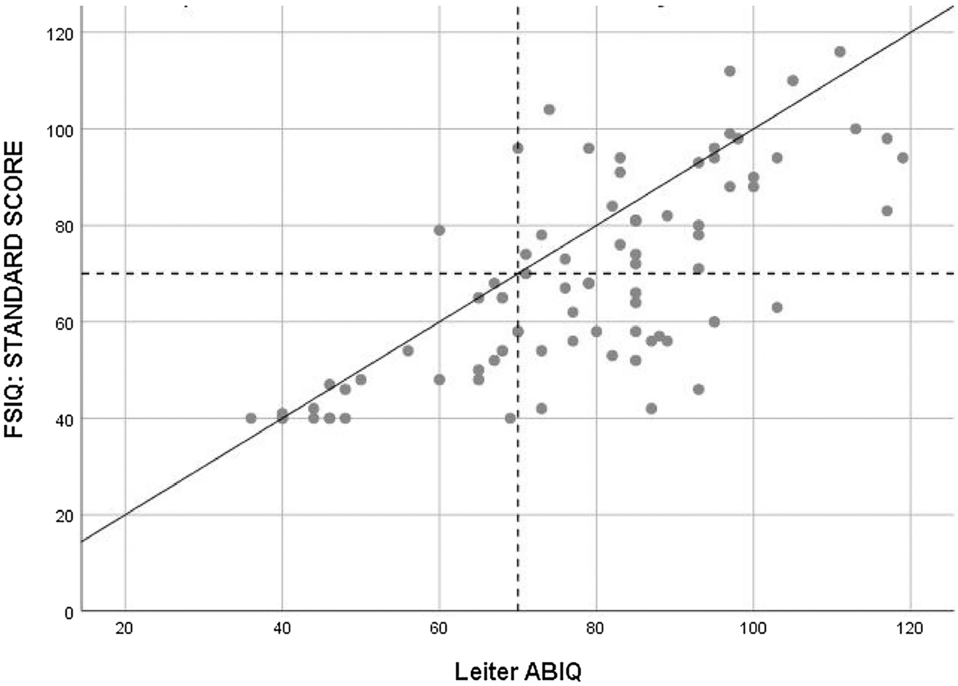

Fig. 1. Scatter plot of individual scores on the Leiter-R and Stanford-Binet.
Note: The diagonal line above indicates perfect agreement between tests. Data above the line reflect cases who scored higher on the Stanford-Binet; data below the line reflect cases who had higher Leiter-R scores. The vertical and horizontal dashed lines are placed at IQ = 70 to demarcate above and below intellectual disability. Thus, upper right and lower left quadrants show agreement on categorical classifications. The upper left and lower right quadrants show categorical disagreement.

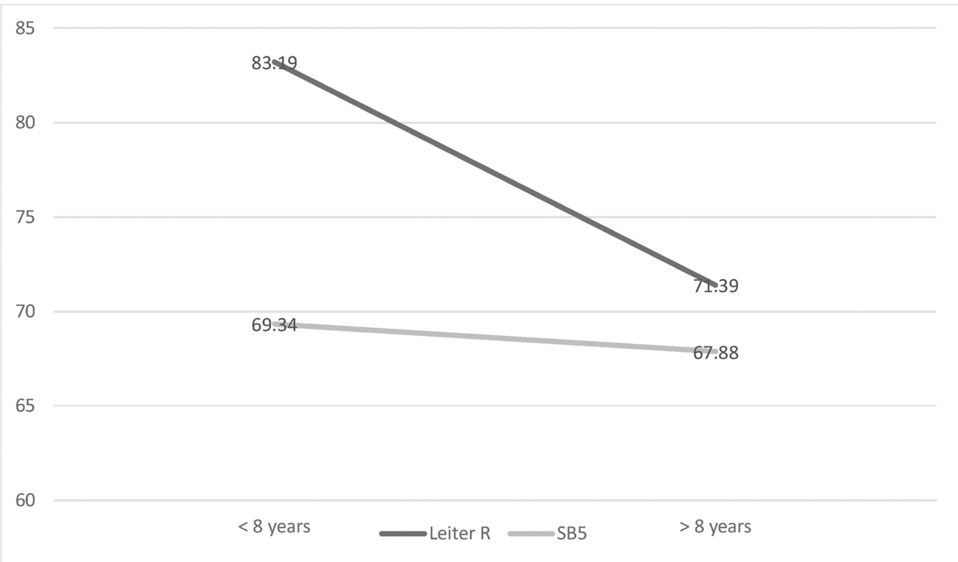

Fig. 2. Test Differences by Age Groups.

Communication ages, 7 had an IQ < 70 and 32 were ≥70; p = .049. Likewise, a Fisher's exact test for Communication age equivalency was also significant based on FSIQ SB5 scores (p = .042); 28 children with lower Communication ages received IQs < 70 and 17 had IQs ≥ 70. There were 17 Communication ages greater than 4.5 years that received an IQ < 70 and 22 had an IQ ≥ 70. This indicates a non-random difference for how participants are classified with ID (IQ < 70) or non-ID using their Leiter-R scores or ID or non-ID based on their SB5 scores when including their Vineland Communication age equivalency split at 4.5 years of age; participant's membership in the high or low age equivalency category altered the chances of being in the ID or non-ID categories for both tests. Vineland Communication age failed to account for variance above and beyond what the SB5 contributed to the Leiter-R, $F(1, 77)$ = .87, p = .354.

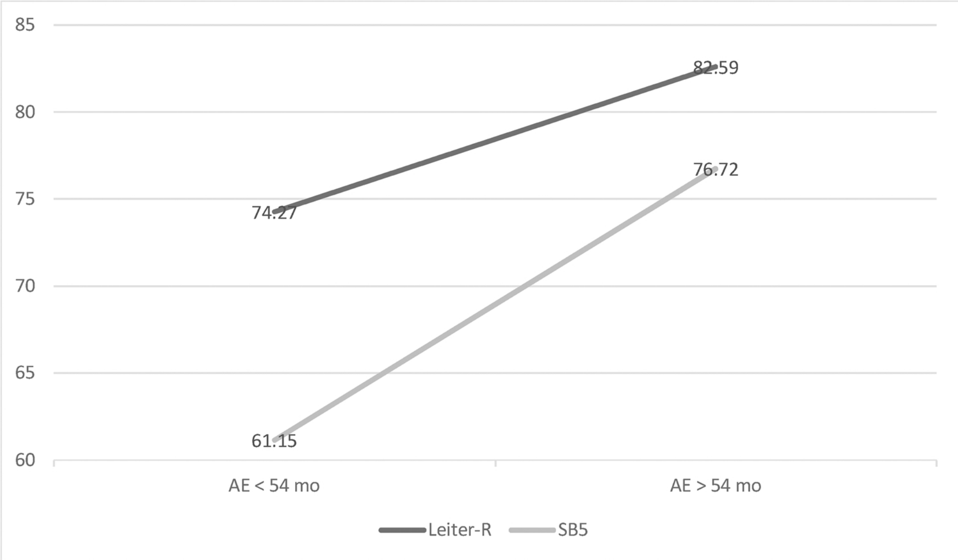

Fig. 3. Test Differences by Vineland Communication Age Equivalency.
Note: AE = Vineland Communication Age Equivalency

3.2. Comparison using the leiter-R BIQ and SB5 verbal vs. nonverbal scores

Analyses using SB5 VIQ and NVIQ were conducted to determine whether either showed greater agreement than the FSIQ. Deviations between the Leiter-R and SB5 for each subscale appear in Table 2. The NVIQ showed the greatest similarity ($M = 5.7$ points, $SD = 14.8$) between Leiter-R and SB5. Not surprisingly, the VIQ showed the least similarity ($M = 10.1$, $SD = 16.6$), with the FSIQ close behind ($M = 9.7$, $SD = 15.3$).

Additional analyses of the SB5 subsections were generally consistent with the Leiter-R BIQ vs. FSIQ analyses. All Pearson correlations between the two intelligence assessments were positive and significant at the $p < .01$ level (NVIQ $r = .73$; VIQ $r = .67$). The interactions for age comparing performances of the younger and older groups did reach significance for both domain scores (NVIQ $F[1,76] = 10.62$, $p = .002$, partial $\eta^2 = .12$; VIQ $F[1,76] = 7.54$, $p = .008$, partial $\eta^2 = .09$), as did the additional amounts of variance accounted for by adding age into the hierarchical model (NVIQ $r^2 = .081$, $F[1,75] = 15.63$, $p < .001$; VIQ $r^2 = .08$, $F[1,75] = 12.15$, $p = .001$). The main effect for age was not significant in the original analysis and failed to reach significance when using the subsections, although both suggested a trend (SB5 NVIQ, $F[1,76] = 3.08$, $p = .084$, partial $\eta^2 = .04$; SB5 VIQ, $F[1,76] = 3.30$, $p = .073$, partial $\eta^2 = .04$).

One analysis using the SB5 subsections did not completely concur with the Leiter-R BIQ and SB5 FSIQ comparisons. This concerned the interaction between IQ test and Vineland Communication. The FSIQ analysis reached significance, as did the comparison when using NVIQ, $F(1, 76) = 6.66$, $p = .012$, partial $\eta^2 = .08$, but *not* VIQ, $F(1,76) = 1.39$, $p = .251$, partial $\eta^2 = .02$.

4. Discussion

Our findings suggest the Leiter-R and SB5 do not provide comparable IQ scores for many children with ASD. The 9.6 point mean discrepancy between tests is only a small part of the story. Indeed, the extreme variations (one participant received a score 30 points higher on the SB5 and another had a score 47 points higher on the Leiter-R) suggest that discrepancies can happen in clinical practice based on instrument used. Importantly, *the test differences altered the IQ category for 62.5% of the children*. This could have important practical implications for access to school services or provision of interventions, and choice of IQ test could influence outcomes in clinical research.

Table 2
Score Discrepancy between *Leiter-R* and *SB5*.

Comparison *Leiter-R* vs. *SB5*	Mean Discrepancy	SD	Greatest Discrepancy for *Leiter-R*	Greatest Discrepancy for *SB5*
BIQ vs. FSIQ	9.67	15.31	47	30
BIQ vs. NVIQ	5.72	14.82	41	26
BIQ vs. VIQ	10.09	16.6	50	36

Note: BIQ = Brief IQ on *Leiter-R*; FSIQ = Full scale IQ on *SB5*; NVIQ = Nonverbal IQ on *SB5*; VIQ = .
Verbal IQ on SB5. Greatest deviation indicates the maximum score difference between assessments.

When considering which factors made score discrepancies more pronounced, it appears age of the children came into play in multiple ways. Children under 8 years of age had significantly higher mean IQs, on both IQ tests, than children over 8 years of age. We presume this was an artifact of the testing paradigm: only the brighter children at a young age would have been testable on the SB5 and therefore would have had scores on both tests. Conversely, a higher proportion of children with intellectual deficits at the older age would be able to complete both tests (due to the fact mental age increases with chronological age, even if ID is present). More importantly, there was a simple main effect for age that indicated scores on the two intelligence assessment were significantly different for the younger group only (13.8 point mean discrepancy for children < 8 years vs. 3.5 for children ≥ 8 years). It is likely younger children had larger deficiencies in communication relative to their general cognitive functioning, allowing their performance on nonverbal assessments like the Leiter-R to appear statistically and clinically better when compared to their performance on the SB5. It appears once these children begin to acquire additional language with age, their cognitive presentation may become more balanced, resulting in a smaller discrepancy between measures.

In this primary analysis there was no main effect for age. It is difficult to know whether to attach importance to this finding, as it could have been a product of the recruitment efforts and ascertainment of ASD for participants in the study. The age main effect finding was consistent when analyses were conducted for SB5 NVIQ or VIQ, although both suggested a trend. This could be attributed to the ease of demands at the beginning portion of the Leiter that enables younger children to perform better. Our analyses indicate clinicians who evaluate youth with ASD should be even more cognizant of their assessment options, as the two tests appear to yield greater difference in outcomes with younger children. This may be critically important for those who participate in initial diagnostic assessments, as the stakes seem to be highest for these young children when determining baseline ability levels and corresponding intervention recommendations.

The comparisons using the Vineland Communication age offered additional insight. The significant interaction demonstrated children with poorer communication displayed greater discrepancies (mean difference = 13.1) between the two IQ tests than children with better communication (mean difference = 5.8), although both simple effects were significant indicating disparate performance on the assessments regardless of Communication ability. Another simple main effect from this interaction showed performance on the SB5 was different for the Communication age groups, and children with higher Communication ages had significantly higher IQ scores on the SB5 than did those with lower communication age. No such pattern was found for the Leiter-R. We found similar significant results using the NVIQ subscale, but there was no interaction when using the VIQ. The Communication age equivalency failed to contribute additional variance above that which the SB5 FSIQ contributed to predicting Leiter-R score in our regression model. This suggests the SB5 sufficiently took communication abilities into account when measuring intelligence, and therefore scores on the SB5 were directly related to use of language. Across SB5 subtypes, children in the higher Communication age category achieved SB5 scores closer to their Leiter-R score. Furthermore, the Fisher's exact test that compared children with high and low Communication age and receiving IQ scores in the ID or non-ID range for either test indicated language ability was the overarching determinant of IQ test performance for these children. The IQ split at 70 was an exercise to approximate the ID cut-off using solely the IQ scores provided by these assessments, as that is the focus on this manuscript, and we acknowledge that information about adaptive behavior must also be taken into account to truly determine the presence of ID in real-world settings.

The SB5 FSIQ and the Leiter-R BIQ discrepancy of 9.7 points was very close to the comparison using the SB5 VIQ, which produced a discrepancy of 10.1. The NVIQ discrepancy was substantially smaller at only 5.7 points, which suggests the NV portion of the SB5 may be an option if a nonverbal IQ test is not available at the time of assessment. However, 25 children in our sample (31%) received Leiter-R BIQ scores at least 10 points higher than their SB5 NVIQ scores, so this path should not be a default choice as task-related language demands still appear to adversely impact performance. That unequal pattern of results (NVIQ > VIQ) is precisely what has been reported for people with ASD in the past (Mayes & Calhoun, 2003b). Additionally, while one might speculate that our results could be influenced by the sample's comorbid disruptive behavior disorders, we do not think this is the case. According to a comprehensive review of the literature (Kaat & Lecavalier, 2013), IQ has not been shown to be a significant predictor of disruptive behaviors. Verbal children and those with IQs > 70 were reported to have more disruptive behaviors than nonverbal children (Witwer & Lecavalier, 2010; Gadow, DeVincent, Pomeroy, & Azizian, 2005), which is interesting since our sample had disruptive behaviors by design and represented the full spectrum of IQ.

Although this investigation appeared to be similar to the study by Grondhuis and Mulick (2013), there were important differences. Findings from both studies agreed that the Leiter-R and SB5 produced different outcomes, but Grondhuis and Mulick (2013) observed a much bigger discrepancy (mean = 20.9). The sample in Grondhuis and Mulick was a clinical sample of convenience from a single clinic, whereas the larger sample in the present study was compiled for research and is more geographically diverse. It is likely that the assessors in the clinic deliberately adopted the Leiter scale for the purposes of elucidating higher performance scores if such were present. In other words, psychometricians in the Grondhuis and Mulick report may have been guided by their clinical instincts, leading to a sample that was intrinsically biased (i.e., the clinicians may have selected the Leiter because they suspected—correctly— that the children had relatively underdeveloped verbal abilities).

A strength of this study was the administration of both IQ assessments to most of the children enrolled in the RUPP study, as opposed to having two different samples of test completers. This allowed for direct comparisons of verbal and nonverbal abilities through much of the sample and increased statistical power. Our sample was well-characterized, had ASD diagnoses determined by licensed clinicians and with ADI-R confirmation, and the sample had a broad range of IQ (Leiter-R BIQ range 36–119, SB5 FSIQ range 40–116). Our findings highlight the discrepancies such children demonstrate depending on IQ test choice; while this is a phenomenon that may be anecdotally known in the clinical community, it is helpful to confirm in the literature as it relates to youth with ASD.

One limitation was having only the BIQ for the Leiter-R, and the full Leiter-R would have added strength to the study. Also, new third editions of both the Vineland Adaptive Behavior Scales and the Leiter International Performance Scale have since been released

(Vineland-3; Sparrow et al., 2016Sparrow, Cicchetti, & Saulnier, 2016; Leiter-3; Roid, Miller, Pomplun, & Koch, 2013). We are confident our findings are still instructive, despite the use of the Leiter-R, as the Leiter-3 was based on the *same underlying constructs* in its development and Leiter-3 nonverbal IQ is highly correlated with the Leiter-R BIQ (r = .78; Roid et al., 2013). Additionally, although the Leiter-R was published in 1997 and our sample was recruited between 2004 and 2007, we do not believe the time between norming and data collection were responsible for our differences. Flynn (2007) estimated that population IQ gains are only about 3 points per decade, and thus the 9.6 mean point discrepancy with higher Leiter-R scores was three times larger than what would be expected by only the Flynn effect. In any case, the passage of time should have affected performance on both IQ tests. Given the extreme variations in scores found in our participants we believe the differences seen would remain clinically meaningful regardless.

As previously mentioned, our sample of 80 participants was not fully representative of all 124 participants in the RUPP study or of the ASD population at large because our sample included children with comorbid disruptive behavior disorders. Although not all children with ASD are diagnosed with a comorbid disruptive behavior disorder, some studies (e.g., Witwer & Lecavalier, 2010) reported such a comorbidity in as many as 75% of the sample assessed, with every study participant endorsing at least one symptom of disruptive behavior. This suggests disruptive symptoms are at least commonly seen among youth with ASD. Another limitation, in terms of informing us about the situation for children with DSM-5 ASD, was our use of DSM-IV-TR diagnostic criteria. Although there is considerable overlap between DSM-IV-TR and DSM-5 criteria for autism spectrum disorder(s), it is possible that not all participants of this study would have been captured by current diagnostic criteria. Finally, although it would have been a strength to have similar data from children *without* ASDs to determine if our observations are limited to ASD, we simply did not have that luxury.

4.1. Implications

Individuals with ASD are heterogeneous by nature of the diagnostic criteria. This should be taken into account when assessing children for IQ, as reliance on any single test is not appropriate. Our results indicate that assessing intelligence using both traditional and nonverbal IQ tests might afford a more nuanced profile of results that could better capture level of functioning and needed educational supports. Equally important, the choice of tests for subject matching and subject characterization in ASD research should be carefully considered, as the indiscriminant mixing of nonverbal and verbal tests can lead to distorted impressions of group composition and comparability of groups.

Conflict of interest

Author C has received research funding from Curemark, Forest, Lilly, Neuropharm, Novartis, Noven, Shire, Supernus, Roche, and YoungLiving (as well as NIH and Autism Speaks), has consulted with Gowlings, Neuropharm, Organon, Pfizer, Sigma Tau, Shire, Tris Pharma, and Waypoint, been on advisory boards for Arbor, Ironshore, Novartis, Noven, Otsuka, Pfizer, Roche, Seaside Therapeutics, Sigma Tau, Shire, and received travel support from Noven. Author D has received funding from Curemark, Lilly, Roche, National Institute of Mental Health, National Institute of Aging, National Institute of Child Health and Human Development, and Autism Speaks. Author E serves as a consultant for the following research organizations: Bracket, Neuren, Roche, Shire, and Supurnus Pharmaceuticals. Author G has received research contracts, consulted with, served on advisory boards, or done investigator training for AMO Pharma, CogState, Inc.; CogState Clinical Trials, Ltd.; Coronado Biosciences; Forest Research; Hoffman-La Roche; Johnson and Johnson; Lumos Pharma, MedAvante, Inc.; Ovid Pharma; ProPhase LLC; and Supernus Pharmaceuticals. Authors A, B, and F declare that they have no conflict of interest.

Acknowledgements

We are thankful to the Research Units on Pediatric Psychopharmacology (RUPP) Autism Network who generously provided us with the data to make these analyses possible. Although no research funding was obtained by authors for this study, RUPP received funding from the National Institute of Mental Health through grants U10MH66768, U10MH66766, and U10MH66764. We also thank Karen Bearss for past contributions and comments on drafts of the manuscript.

References

Abbeduto, L., & Short, K. (1994). Relation between language comprehension and cognitive functioning in persons with mental retardation. *Journal of Developmental and Physical Disabilities, 6*(4), 347–369.

Aman, M. G., McDougle, C. J., Scahill, L., Handen, B., Arnold, L. E., Johnson, C., & Sukhodolsky, D. D. (2009). Medication and parent training in children with pervasive developmental disorders and serious behavior problems: Results from a randomized clinical trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 48*(12), 1143–1154.

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

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

Athanasiou, M. S. (2000). Current nonverbal assessment instruments: A comparison of psychometric integrity and test fairness. *Journal of Psychoeducational Assessment, 18*, 211–229.

Baum, K. T., Shear, P. K., Howe, S. R., & Bishop, S. L. (2015). A comparison of WISC-IV and SB-5 intelligence scores in adolescents with autism spectrum disorder. *Autism, 19*(6), 736–745.

Brown, L., Sherbenous, R. J., & Johnsen, S. K. (2010). *TONI-4, Test of nonverbal intelligence, Pro-Ed.* .

Centers for Disease Control and Prevention (2014). Prevalence of autism spectrum disorder among children aged 8 years – autism and development disabilities monitoring network, 11 sites, United States. *Morbidity and Mortality Weekly Report, 63*, 1–21.

Coolican, J., Bryson, S. E., & Zwaigenbaum, L. (2008). Data on the Stanford-Binet Intelligence Scales (5th ed.) in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 38*, 190–197.

Flynn, J. R. (2007). *What is intelligence?: Beyond the Flynn effect.* Cambridge University Press.

Gadow, K. D., Devincent, C. J., Pomeroy, J., & Azizian, A. (2005). Comparison of DSM-IV symptoms in elementary school-age children with PDD versus clinic and community samples. *Autism, 9*(4), 392–415.

Grondhuis, S. N., & Mulick, J. A. (2013). Comparison of the Leiter International performance scale-revised and the Stanford-Binet intelligence scales, in children with autism spectrum disorders. *American Journal on Intellectual and Developmental Disabilities, 118*(1), 44–54.

IBM Corp (2013). *IMB SPSS statistics for windows, version 22.0.* Armonk, NY: IBM Corp.

Kaat, A. J., & Lecavalier, L. (2013). Disruptive behavior disorders in children and adolescents with autism spectrum disorders: A review of the prevalence, presentation, and treatment. *Research in Autism Spectrum Disorders, 7*(12), 1579–1594.

Klaiman, C., Fernandez-Carriba, S., Hall, C., & Saulnier, C. (2015). Assessment of autism across the lifespan: A way forward. *Current Developmental Disorder Reports, 2*, 84–92.

Klin, A., Saulnier, C. A., Sparrow, S. S., Cicchetti, D. V., Volkmar, F. R., & Lord, C. (2007). Social and communication abilities and disabilities in higher functioning individuals with autism spectrum disorders: The Vineland and the ADOS. *Journal of Autism and Developmental Disorders, 37*(4), 748–759.

Leiter, R. G. (1979). *Leiter international performance scale, Instruction manual.* Chicago, IL: Stoelting.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659–685.

Mayes, S. D., & Calhoun, S. L. (2003a). Ability profiles in children with autism: Influence of age and IQ. *Autism, 6*(4), 65–80.

Mayes, S. D., & Calhoun, S. L. (2003b). Analysis of WISC-II, Stanford-Binet: IV, and academic achievement test scores in children with autism. *Journal of Autism and Developmental Disorders, 33*(3), 329–341.

Nader, A. M., Courchesne, V., Dawson, M., & Soulières, I. (2016). Does WISC-IV underestimate the intelligence of autistic children? *Journal of Autism and Developmental Disorders, 46*(5), 1582–1589.

Perry, A., Flanagan, H. E., Geier, J. D., & Freeman, N. L. (2009). Brief report: The Vineland adaptive behavior scales in young children with autism spectrum disorders at different cognitive levels. *Journal of Autism and Developmental Disorders, 39*(7), 1066–1078.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales.* Oxford, UK: Oxford Psychologists Press.

Reeve, R. R., French, J. L., & Hunter, M. (1983). Brief report: A validation of the Leiter International performance scale with kindergarten children. *Journal of Consulting and Clinical Psychology, 31*(3), 458–459.

Roid, G. H., & Miller, L. J. (1997). *Leiter international performance scale-revised.* Wood Dale. IL: Stoelting Col.

Roid, G. H., Pomplun, M., & Martin, J. J. (2009). *Nonverbal intellectual and cognitive assessment with the leiter international performance scale – revised (Leiter-R). Practitioner's guide to assessing intelligence and achievement*265–290.

Roid, G. H., Miller, L. J., Pomplun, M., & Koch, C. (2013). *Leiter international performance scale, (Leiter-3).* Los Angeles: Western Psychological Services.

Roid, G. H. (2003). *Stanford-Binet intelligence scales* (5th ed.). Itasca, IL: Riverside.

Scahill, L., Aman, M. G., McDougle, C. J., Arnold, L. E., McCracken, J. T., Handen, B., ... Swiezy, N. (2009). Testing the combined effects of medication and behavioral intervention in children with pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 39*(5), 707–720.

Scahill, L., McDougle, C. J., Aman, M. G., Johnson, C., Handen, B., Bearss, K., ... Stigler, K. A. (2012). Effects of risperidone and parent training on adaptive functioning in children with pervasive developmental disorders and serious behavioral problems. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(2), 136–146.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland adaptive behavior scales (Survey ed.).* Circle Pines, MN: American Guidance Service.

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland-3: vineland adaptive behavior scales.* PsychCorp.

Tager-Flusberg, H. (2015). Defining language impairments in a subgroup of children with autism spectrum disorder. *Science China Life Science, 58*, 1044–1052.

Terman, L. M., & Merrill, M. A. (1972). *Stanford binet intelligence test, form L-M.*

Wechsler, D. (2003). *Wechsler intelligence scale for children, 4th edition (WISC-IV).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2014). *Wechsler intelligence scale for children* (5th edition). Bloomington, MN: Pearson (WISC-V).

Witwer, A. N., & Lecavalier, L. (2010). Validity of comorbid psychiatric disorders in youngsters with autism spectrum disorders. *Journal of Developmental and Physical Disabilities, 22*(4), 367–380.