Does the Factor Structure of IQ Differ Between the Differential Ability Scales (DAS-II) Normative Sample and Autistic Children?

Caitlin C. Clements, Marley W. Watkins, Robert T. Schultz, and Benjamin E. Yerys

The Differential Abilities Scales, 2nd edition (DAS-II) is frequently used to assess intelligence in autism spectrum disorder (ASD). However, it remains unknown whether the DAS-II measurement model (e.g., factor structure, loadings), which was developed on a normative sample, holds for the autistic population or requires alternative score interpretations. We obtained DAS-II data from 1,316 autistic individuals in the Simons Simplex Consortium and 2,400 individuals in the normative data set. We combined ASD and normative data sets for multigroup confirmatory factor analyses to assess different levels of measurement invariance, or how well the same measurement model fit both data sets: "weak" or metric, "strong" or scalar, and partial scalar if full scalar was not achieved. A weak invariance model showed excellent fit (Confirmatory Fit Index [CFI] > 0.995, Tucker Lewis Index [TLI] > 0.995, root mean square error of approximation [RMSEA] < 0.025), but a strong invariance model demonstrated a significant deterioration in fit during permutation testing (all p's<0.001), suggesting measurement bias, meaning systematic error when assessing autistic children. Fit improved significantly, and partial scalar invariance was achieved when either of the two spatial subtest (Recall of Designs or Pattern Construction) intercepts was permitted to vary between the ASD and normative groups, pinpointing these subtests as the source of bias. The DAS-II appears to measure verbal and nonverbal—but not spatial—intelligence in autistic children similarly as in normative sample children. These results may be driven by Pattern Construction, which shows higher scores than other subtests in the ASD sample. Clinicians assessing autistic children with the DAS-II should interpret verbal and nonverbal reasoning composite scores over the spatial score or General Composite Ability. ***Autism Res** 2020, 13: 1184–1194.* © 2020 International Society for Autism Research, Wiley Periodicals, Inc.

Lay Summary: The Differential Abilities Scales, 2nd edition (DAS-II) is a popular intelligence quotient (IQ) test for assessing children with autism. This article shows that the DAS-II spatial standardized scores should be interpreted with caution because they hold a different meaning for autistic children. Verbal and nonverbal reasoning scores appear valid and to hold the same meaning for those with and without autism spectrum disorder.

Keywords: autism spectrum disorders; autistic disorder; intelligence; educational psychology; factor analysis; psychometrics

Introduction

Intellectual disability (ID) commonly co-occurs with autism spectrum disorder (ASD): approximately 50% of autistic individuals meet criteria for ID [Charman et al., 2011; Loomes, Hull &Mandy, 2017]. To assess ID in school-age autistic children, clinicians frequently use the DAS-II (Differential Ability Scales, 2nd Edition, Elliott, 2007a) to measure cognitive ability. However, it remains unknown whether the DAS-II functions similarly in autistic and neurotypical children [Wicherts, 2016]. The DAS-II measurement model (i.e., the relationship between subtests and the latent constructs of verbal, nonverbal, and spatial intelligence which is described by the factor structure, factor loadings, covariances, etc.) was developed with a nationally representative normative sample and has never been tested in a large autistic sample to our knowledge. If the DAS-II measurement model fails to hold for autistic children, alternative methods and score interpretations will be needed for measuring cognitive ability and informing ID assessments.

Research has shown that the measurement models of some intellectual assessments perform differently in some subgroups. For example, the DAS-II measurement model showed small differences for a sample of African Americans [Trundt, Keith, Caemmerer, & Smith, 2018], the WISC-IV measurement model showed differences for a sample with attention-deficit/hyperactivity disorder [Thaler et al., 2015],

From the The Children's Hospital of Philadelphia, Center for Autism Research, Philadelphia, Pennsylvania (C.C.C., R.T.S., B.E.Y.); Psychology Department, University of Pennsylvania, Philadelphia, Pennsylvania (C.C.C.); Department of Educational Psychology, Baylor University, Waco, Texas (M.W.W.); Department of Pediatrics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania (R.T.S.)

Received August 23, 2019; accepted for publication February 11, 2020

Address for correspondence and reprints: Caitlin C. Clements, Psychology Department, University of Pennsylvania, 425 S. University Ave, Philadelphia, PA, 19104.
E-mail: clements@sas.upenn.edu

Published online 29 February 2020 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/aur.2285

and a factor analysis of the WAIS-R, WAIS-III, WISC-R, and WISC-III in a sample of high functioning autism identified a "social context" factor not present in the normative sample [Goldstein et al., 2008]. When a measurement model performs differently in a particular subgroup, this suggests that measurement bias affects scores for individuals in that subgroup such that their measured scores do not reflect their true scores on the latent trait (e.g., nonverbal intelligence) in the same way that scores for the normative group do, whether driving measured scores up or down [Reynolds & Lowe, 2009]. Please note that throughout this article, the terms "nonverbal intelligence quotient" or NVIQ are used instead of fluid reasoning (g_f) for consistency with DAS-II nomenclature [Elliott et al., 2018, p. 372].

Clinicians have long discussed "IQ splits" in individuals with ASD, and recent research lends more support to this observed phenomenon. Siegel, Minshew, and Goldstein [1996] initially reported that in 45 high-functioning autistic individuals, 36% of participants showed unusually large differences (i.e., 12 IQ points in standard scale of mean 100, standard deviation 15) between their nonverbal IQ and verbal IQ scores (20% NVIQ > verbal IQ [VIQ], 16% VIQ > NVIQ). Many other researchers reported similar data, and an analysis of the largest known sample of DAS-II data on autistic children (n = 2,110; the Simons Simplex Consortium [SSC]) confirmed the "splits" finding with 32% of individuals showing DAS-II Early Years NVIQ > VIQ discrepancies of at least 16 points, and 20% showing the same discrepancy on DAS-II School Age [Nowell, Schanding, Kanne, & Goin-Kochel, 2015]. At present, it is unclear whether these "splits" reflect true differences between verbal and nonverbal intelligence, or are better attributed to measurement bias due to a poor fit of the DAS-II measurement model in autistic children. This question can be answered by testing measurement invariance.

Measurement invariance is a method to determine whether an assessment such as the DAS-II measures the same latent construct with the same precision in multiple populations. In other words, it tests whether the observed test score of an individual—who has a certain true score on the latent construct—is independent of that individual's group membership [Thompson, 2016]. Different levels of measurement invariance are tested sequentially with increasing strictness. At the first level, the *same confirmatory factor model* is fit to each group separately. This level of invariance merely demonstrates that the same model can be fit to each group but does not rule out measurement bias in the relationship between one group's test scores and true ability. At the second or "weak" factorial invariance level, configural invariance, a multigroup model is fit to the combined data sets; this model requires that the *same items load on the same factors* for each group but imposes no between-group constraints on factor loadings or any other parameters. At the third level, also referred to as "weak" factorial invariance, *factor loadings are constrained* to be equal in both groups but no other between-group constraints are imposed. At the fourth level, scalar or "strong" factorial variance is required to conclude that between-group differences in mean scores are entirely due to true group differences in latent abilities and not measurement bias. Scalar invariance requires *equality between groups on intercepts* and permits estimation of differences between group factor means by no longer setting factor means equal to zero as in metric and configural invariance. In one final level, residual or "strict" invariance, *residuals are constrained* to be equal in both groups. However, this level of factorial invariance is not necessary; it is widely accepted that scalar or "strong" invariance is sufficient for the use of a measure with a particular population, such as autistic children. If scalar invariance is achieved between the autistic and normative samples, then it can be concluded that group differences in nonverbal, verbal, and spatial intelligence scores reflect true group differences in ability. If scalar invariance is not achieved, then group differences might be due to measurement bias and artifacts rather than true differences in intelligence; thus an autistic child's DAS-II score would be biased compared to the normative sample.

The objective of this study is to determine whether DAS-II scores are biased for autistic children.

Methods

Participants

The ASD sample was drawn from the SSC, which was a multisite study of 2,110 children ages 4–18 years who met gold-standard diagnostic criteria for ASD. Participants completed a comprehensive diagnostic and behavioral testing battery that included the DAS-II School Age core subtests. For additional information on SSC data collection, recruitment, diagnoses, and inclusion criteria, see Fischbach & Lord, 2010. SSC participants were included in the present study if they had a DAS-II School Years subtest score (n = 1,316; see Table 1). Over 90% of participants had complete data on all six core DAS-II subtests.

The control sample consisted of the nationally representative DAS-II School Age normative sample ages 6–17 years (n = 2,400; see Table 1) and was provided by Pearson, publisher of the DAS-II. For additional information on this sample, see the DAS-II Technical Manual [Elliott, 2007b].

This study was approved by The Children's Hospital of Philadelphia Institutional Review Board and adheres to the legal requirements of the United States.

Data Analysis

Missing data. Eight of 2,400 individuals in the normative data set (Standardization data from the Differential

Table 1. Participant demographics

	Normative sample	ASD sample	Normative sample with complete data	ASD sample with complete data
N	2400	1316	2388	1197
% Male[a]	50.0	87.4	50.0	87.9
Age in years, mean [SD]	12.0 [3.5]	10.5 [3.7]	12.0 [3.5]	10.5 [3.7]
DAS-II Global Composite Ability	99.9 [15.2]	94.3 [19.8]	99.9 [15.2]	94.4 [19.7]
DAS-II Nonverbal Composite	99.8 [14.8]	93.4 [19.1]	99.8 [14.8]	95.0 [18.6]
DAS-II Verbal Composite	100.0 [15.1]	92.9 [22.6]	100.0 [15.1]	93.0 [22.5]
DAS-II Spatial Composite	99.8 [14.9]	95.1 [18.2]	99.9 [14.9]	96.3 [18.1]
Subtest matrices (n)	50.2 [10.2]	46.7 [12.3]	50.2 [10.2]	47.5 [12.1]
Pattern construction (s)	50.0 [9.9]	48.9 [11.6]	50.0 [9.9]	49.6 [11.7]
Recall of designs (s)	50.0 [9.9]	45.6 [11.6]	50.0 [9.9]	46.4 [11.4]
sequential and quantitative reasoning (n)	50.2 [10.3]	45.8 [12.9]	50.2 [10.3]	46.9 [12.6]
Verbal similarities (v)	50.2 [9.9]	46.2 [13.9]	50.2 [9.9]	46.3 [13.8]
Word definitions (v)	50.1 [9.8]	45.5 [14.7]	50.1 [9.8]	45.5 [14.7]

Note. All DAS-II values show mean [SD] of the standard score; lowercase letters in parentheses denote composite in which subtest is scored.
[a]Missing for 56 individuals with ASD.

Ability Scales-II (DAS-II). Copyright 1998, 2000, 2004, 2007 NCS Pearson, Inc. and Colin D. Elliott. Normative data. Copyright 2007 NCS Pearson, Inc. Used with permission. All rights reserved.) were missing data on one subtest. The ASD sample showed significantly more missing data (119 of 1,316 participants). While each nonverbal and spatial subtest had data from >99% of ASD participants, both verbal subtests were missing for 8.1% of participants (n = 106). Data were not missing at random: the 106 participants with verbal subtest missingness showed substantially lower verbal abilities on other measures (Verbal Communication score on the Autism Diagnostic Interview—Revised, $t(120) = -7.08$, $P < 0.001$, mean missing = 19.0, mean nonmissing = 16.3) and module selected for the Autism Diagnostic Observation Schedule, which is based on language level and age ($\chi^2(3) = 586.0$, $P < 0.001$). The ASD sample showed a very wide ability range with and without these 106 participants, and in fact the range of General Composite Ability remained the same (40–167).

All analyses were conducted on the full data sets that included all participants, including those missing subtest score(s) which were imputed by Full Information Maximum Likelihood, following the guidelines provided by Newman [2014]. Then, in an effort to explore any bias introduced by the missing data from 119 ASD participants, we conducted sensitivity analyses to determine whether meaningful differences resulted. First, we reconducted analyses excluding participants with missing data (i.e., listwise deletion). Second, we adjusted imputed values by subtracting and adding arbitrary values (implemented with the mice package in R [van Buuren & Groothuis-Oudshoorn, 2011]), then we reconducted analyses with the new data sets. Third, we tested the base oblique model in the ASD data set alone using an auxiliary variable related to verbal communication: the parent report ADI-R (Autism Diagnostic Interview—Revised) verbal communication total score. Auxiliary variable analysis and Full Information Maximum Likelihood were implemented in Mplus v8.2 [Muthén & Muthén, 1998].

Confirmatory factor analysis. First, we determined the base model for invariance testing by fitting the same confirmatory factor model separately to the normative data and to the ASD data to ensure the most basic measurement model fit both samples. In selecting the target model, we consulted the DAS-II Technical Manual, which reported two models. The first model, a correlated three-factor ("oblique") model, uses the six core subtests, similar to our data set [Elliott, 2007b, p. 159]. The correlated three-factor model allows correlations between the three factors (verbal, nonverbal reasoning, and spatial) and does not include a higher-order general (g) factor (Fig. 1). The Technical Manual describes a second model, the higher-order model, which uses both the six core subtests and the less frequently used six diagnostic subtests [Elliott, 2007b, p. 157]. In the higher-order model, the six core subtests load onto three factors (verbal, nonverbal reasoning, and spatial), which in turn load onto a general (g) factor; the diagnostic subtests load onto three separate factors that in turn load onto g (Fig. S1). Of note, for the six-core subtest battery, the Technical Manual reports fit statistics for the correlated three-factor and not the higher-order model. The Technical Manual does not describe fitting the higher-order model to the six-core subtest battery alone, which is most commonly used clinically and in our ASD data set. Note that we use the classical definition of the term "higher-order model" to refer to the model in Figure S1, which is sometimes called by the name of the more general category to which it belongs, "hierarchical model."

In addition, we fit bifactor models demonstrated by previous research to fit the normative data [e.g., Canivez & McGill, 2016; Dombrowski, Golay, McGill, & Canivez,

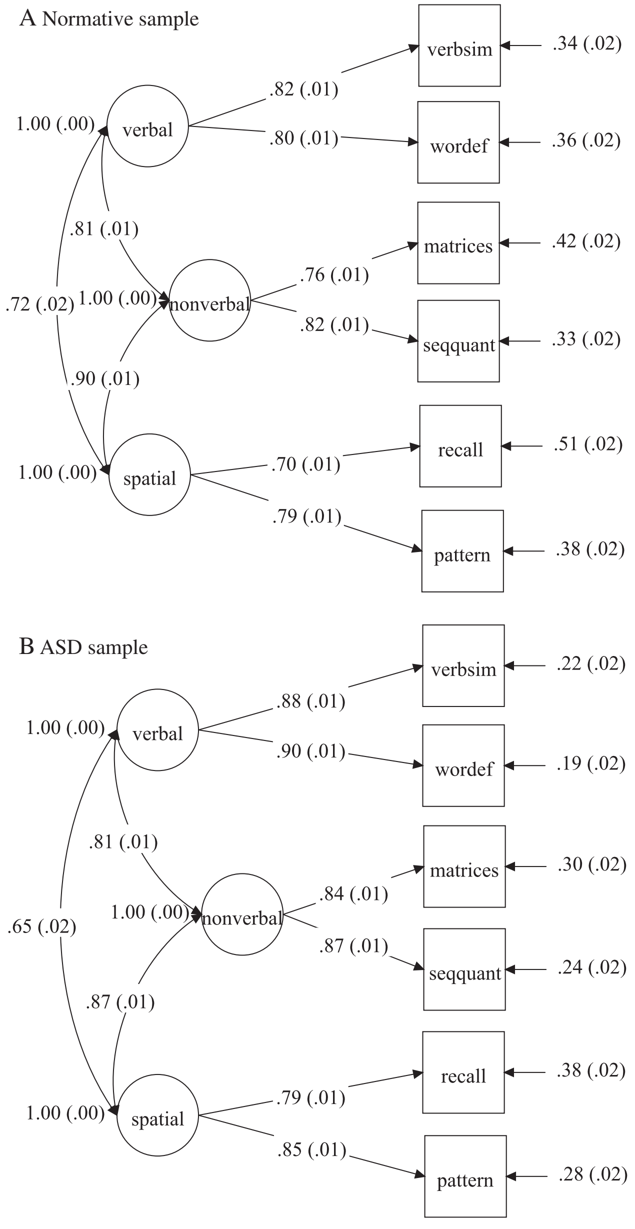

Figure 1. Correlated three-factor model for (a) normative and (b) ASD samples. VerbSim, verbal similarities; WordDef, word definitions; Pattern, pattern construction; Recall, recall of designs; SeqQuant, sequential and quantitative reasoning.

2018; Dombrowski, McGill, Canivez, & Peterson, 2019]. A bifactor model includes the general factor and group factors (i.e., verbal, nonverbal, and spatial) and assumes that the general factor is orthogonal to the group factors. Note that we use the classical definition of the term "group factor" to refer to verbal, nonverbal, and spatial factors (sometimes referred to as specific factors). We fit two bifactor models: a three-factor bifactor model with verbal, nonverbal, and spatial group factors and *g* as suggested by Canivez and McGill [2016], and a two-factor bifactor model with verbal and spatial group factors, and the two nonverbal reasoning subtests loading directly on *g* instead of a nonverbal factor (Fig. S2) as reported by Dombrowski et al. [2018]. In the bifactor models, we fixed correlations between all factors at zero, and fixed equality between the two factor loadings on each group factor, which decreases the number of parameters being estimated and thus allows model identification. Finally, we also fit a simple unidimensional model that allowed the six subtests to load directly on *g*.

Measurement invariance. Next, we combined the normative and ASD data sets into one multigroup data set. We used the best model established in the previous step to test sequentially stricter levels of measurement invariance: configural, metric (weak), scalar (strong), then residual (strict). If invariance was not achieved, we ran partial invariance tests to identify the locus of misfit.

Comparisons between measurement invariance models were made in accordance with recommendations by Jorgensen, Kite, Chen, and Short [2018] to assess statistical significance rigorously via permutation testing, rather than cutoffs established by Chen [2007], which have inconsistent Type I error rates. In each permutation, group membership was randomly assigned; a distribution was built from 1,000 replications then used to determine whether true group membership differed significantly from what would be expected under the null hypothesis, as evidenced by the size of change among fit indices during the replications. We rejected models with $P < 0.05$ on multiple fit indices in favor of the simpler model in the comparison.

All primary factor analyses were implemented in Mplus version 8.2 [Muthén & Muthén, 1998]. Permutation testing, sensitivity analyses, and all remaining analyses were implemented in R version 3.5.2 (R Core Team, 2018) using packages lavaan [Rosseel, 2012] and semTools [Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2019]. All models were estimated with maximum likelihood with robust standard errors (implemented with MLR) due to significant non-normality of every subtest in both data sets according to the Shapiro–Wilk test (all $W > 0.95$, all $P < 0.02$).

Results

Confirmatory Factor Analysis

First we fit a correlated three-factor model as reported in the Technical Manual for the six-subtest core battery. The model demonstrated excellent fit with the data, as expected. See Tables 2 and 3, and Table S1 for complete fit statistics for all models and intersubtest correlations. The higher-order model yielded a factor loading >1.0 of the nonverbal factor on *g* for both the normative (1.005) and ASD (1.045) data sets (Table S1). These results suggest that

Table 2. DAS-II subtest correlations

	Matrices	Pattern construction	Recall of designs	Sequential and quantitative reasoning	Verbal similarities	Word definitions
Matrices	1	0.54	0.48	0.62	0.50	0.49
Pattern construction	0.62	1	0.55	0.58	0.46	0.44
Recall of designs	0.58	0.67	1	0.51	0.43	0.42
Sequential and quantitative reasoning	0.72	0.63	0.57	1	0.54	0.54
Verbal similarities	0.58	0.46	0.45	0.62	1	0.65
Word definitions	0.57	0.47	0.46	0.64	0.79	1

Note. The upper set of correlations depicts the normative data set; the lower set depicts the ASD data set.

the nonverbal factor contributes no specific variance. In other words, the general factor absorbs all variance in the nonverbal factor. We next attempted to fit a three-factor bifactor model that allows subtests to load on both group and general factors. The three-factor bifactor model did not converge for either the normative or ASD data sets. After removing individuals with missing data, however, the model converged for the normative data set and yielded a factor loading of 1.00 for the nonverbal factor on *g*, indicating persistence of nonverbal factor variance issues. Additionally, the three-factor bifactor model did not converge at all for the ASD data set. A bifactor model with two group factors (verbal and spatial) and *g* loaded by all six core subtests (i.e., the nonverbal subtests did not form a general factor) converged for both data sets and showed excellent fit. The unidimensional model demonstrated poor fit for each group (TFI and CLI < 0.94 for both ASD and Norm groups; Table 3) and did not merit further exploration.

An acceptable base solution must be adequate in terms of both model fit and psychological interpretation [Jöreskog, 1969]. The higher-order and bifactor models with three first-order and group factors, respectively, were psychologically interpretable but produced improper solutions or failed to converge [Diamantopoulos & Siguaw, 2003], making them inappropriate for consideration as base models for invariance testing. The three-factor oblique model and the two-factor bifactor model without a separate nonverbal group factor both exhibited acceptable model fit although the oblique model showed slightly better fit than the two-factor bifactor model on many indices (χ^2, *P*, Confirmatory

Table 3. Model fit statistics

Model	*df*	χ^2	*P*	CFI	TLI	RMSEA (90%)	SRMR	AIC
Unidimensional								
Normative	9	385.0	0.000	0.934	0.890	0.132 (0.121–0.143)	0.040	101,304.5
ASD-SSC	9	537.6	0.000	0.864	0.774	0.211 (0.196–0.227)	0.060	56,529.0
Correlated three-factor (Oblique)								
Normative	6	4.9	0.555	1.000	1.000	0.000 (0.000–0.024)	0.005	100,913.6
ASD-SSC	6	13.3	0.038	0.998	0.995	0.030 (0.007–0.053)	0.009	55,941.3
Higher-order								
Normative	6	4.9	0.555	1.000	1.000	0.000 (0.000–0.024)	0.005	100,913.6
ASD-SSC	6	13.3	0.038	0.998	0.995	0.030 (0.007–0.053)	0.009	55,941.3
Bifactor, three-factor								
Normative[a]	9	23.3	0.001	0.997	0.996	0.026 (0.013–0.039)	0.042	100,518.0
ASD-SSC	9	—	—	—	—	—	—	—
Bifactor, two-factor								
Normative[a]	7	5.0	0.654	1.000	1.001	0.000 (0.000–0.020)	0.005	100,911.8
ASD-SSC	7	22.4	0.002	0.996	0.992	0.041 (0.023–0.060)	0.013	55,949.8
Measurement invariance								
Configural	12	18.8	0.094	0.999	0.998	0.017 (0.000–0.032)	0.007	156,854.9
Metric	15	24.7	0.054	0.999	0.998	0.019 (0.000–0.031)	0.016	156,856.0
Scalar	18	133.7	<0.001	0.988	0.980	0.059 (0.050–0.068)	0.035	156,968.7
Partial scalar: Spatial[b]	17	29.6	0.030	0.999	0.998	0.020 (0.006–0.032)	0.019	156,857.3
Partial scalar: Nonverbal[b]	17	130.9	<0.001	0.988	0.979	0.060 (0.051–0.070)	0.035	156,967.8
Partial scalar: Verbal[b]	17	131.3	<0.001	0.988	0.979	0.060 (0.051–0.070)	0.033	156,968.3
Partial scalar, strict: Spatial[b]	23	52.7	<0.001	0.997	0.996	0.026 (0.018–0.035)	0.015	156,874.4

Abbreviations: AIC, Akaike information criterion; SRMR, standardized root mean square residual.

[a] Results from n = 2388 with participants with missingness excluded; model did not converge with data set with missing data (n = 2400).

[b] The intercept of one subtest on the respective spatial, nonverbal, or verbal factor is free to vary between groups; model fit is identical in the two models of the factor's two subtests varying.

Table 4. Unstandardized factor means and subtest intercepts, by model

	Configural	Metric	Scalar	Partial scalar[a]	Partial scalar[b]	Partial scalar, strict[a]
Factor means	Fixed at 0	Fixed at 0	Free	Free	Free	Free
Factor loadings	Free	Invariant	Invariant	Invariant	Invariant	Invariant
Factor intercepts	Free	Free	Invariant	5/6 invariant	5/6 invariant	5/6 invariant
Residuals	Free	Free	Free	Free	Free	Invariant
Normative						
Verbal	0	0	0	0	0	0
Nonverbal	0	0	0	0	0	0
Spatial	0	0	0	0	0	0
Matrices (n)	50.2	50.2	50.3	50.3	50.3	50.3
Pattern construction (s)	50.0	50.0	50.5	50.0	50.0	50.0
Recall of designs (s)	50.0	50.0	49.3	50.0	50.0	50.0
Sequential and quantitative reasoning (n)	50.2	50.2	50.1	50.1	50.1	50.1
Verbal similarities (v)	50.2	50.2	50.3	50.3	50.3	50.3
Word definitions (v)	50.1	50.1	50.0	50.0	50.0	50.0
ASD						
Verbal	0	0	−0.64	−0.64	−0.64	−0.65
Nonverbal	0	0	−0.50	−0.50	−0.50	−0.51
Spatial	0	0	−0.34	−0.14	−0.63	−0.14
Matrices (n)	46.6	46.6	50.3	50.3	50.3	50.3
Pattern construction (s)	48.9	48.9	50.5	50.0	53.8	50.0
Recall of designs (s)	45.6	45.6	49.3	46.6	50.0	46.6
Sequential and quantitative reasoning (n)	45.7	45.7	50.1	50.1	50.1	50.1
Verbal similarities (v)	45.5	45.4	50.3	50.3	50.3	50.3
Word definitions (v)	44.6	44.6	50.0	50.0	50.0	50.0

Note. Lowercase letters denote factor onto which subtest loads. See Table S2 for standardized loadings and correlations. See Table S3 for unstandardized intercepts and means for other partial scalar invariance models.

[a]The recall of designs intercept was freed to vary between groups.

[b]The pattern construction intercept was freed to vary between groups.

Fit Index [CFI], Tucker Lewis Index [TLI], standardized root mean square residual, and for ASD only, root mean square error of approximation [RMSEA]), particularly for the ASD data set. Although the two-factor bifactor model could reasonably be selected as the base model, the correlated three-factor model was identified by the publisher, and its results are more easily interpreted by clinicians because the three factors translate directly to the three DAS-II composite

Table 5. Models compared with permutation testing on multiple fit indices

	χ^2	CFI	RMSEA	TLI	AIC	SRMR
Model Comparison						
Configural vs. baseline						
Delta	20.2	0.999	0.019	0.998	156,854.9	0.006
P value	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999
Metric vs. configural						
Delta	7.1	<0.001	0.002	<0.001	1.09	0.005
P value	0.13	0.12	0.048	0.064	0.13	0.099
Scalar[a] *vs. metric*						
Delta	118.7	−0.011	0.041	−0.017	112.7	0.013
P value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Partial scalar[a] *vs. metric*						
Delta	5.3	0.000	0.001	0.000	1.3	0.001
P value	0.069	0.066	0.048	0.051	0.069	0.016
Partial scalar[a] *vs. scalar*[a]						
Delta	−113.4	0.010	−0.040	0.017	−111.4	−0.012
P value	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999
Partial scalar[a] *vs. strict partial scalar*[a]						
Delta	29.1	−0.002	0.008	−0.002	17.1	0.003
P value	<0.001	<0.001	<0.001	<0.001	<0.001	0.1

Note. More complex model being tested appears first. Permutation testing executed using the permuteMeasEq function in the semTools R package.

Abbreviations: AIC, Akaike information criterion; SRMR, standardized root mean square residual.

[a]The recall of designs intercept was freed to vary between groups.

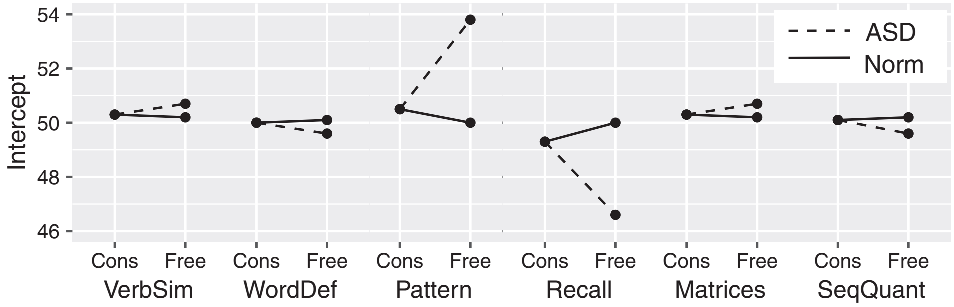

Figure 2. Change in intercept when no longer constrained equal between groups. Cons, constrained; VerbSim, verbal similarities; WordDef, word definitions; Pattern, pattern construction; Recall, recall of designs; SeqQuant, sequential and quantitative reasoning.

scores of verbal, nonverbal reasoning, and spatial, whereas the two-factor bifactor model lacks the nonverbal reasoning factor. Consequently, the correlated three-factor model was chosen as the base model to test measurement invariance between the ASD and normative groups. The final model was tested with the ADI-R verbal communication auxiliary variable, and results did not change meaningfully.

Measurement Invariance

Full invariance. We used the correlated three-factor model to test measurement invariance between the ASD and normative groups. Results indicated that configural and metric invariance were achieved (see Tables 3 and 5). Scalar invariance was not achieved: on all fit indices, permutation testing showed a significant deterioration in fit (all *P*s < 0.001, Table 5). Traditional metrics also provided evidence of poor scalar model fit: CFI, TLI, and RMSEA showed change beyond acceptable limits and RMSEA rose over the 0.05 threshold [Chen, 2007].

Partial invariance. Partial scalar invariance was assessed by allowing single subtest intercepts to vary between groups. We observed little change in the model when verbal factor subtests (word definitions or verbal similarities) or nonverbal factor subtests (matrices or sequential and quantitative reasoning) were allowed to vary, suggesting that the model easily accommodates equality between ASD and normative intercepts on these subtests; group differences in verbal and nonverbal factor scores are due to true group differences in verbal and nonverbal abilities, not bias. This pattern was not true for the subtests loading on the spatial factor (pattern construction and recall of designs). When either of these intercepts was freed to vary between groups, the model fit improved significantly on all indices. This partial scalar invariance model (i.e., with the recall of designs intercept freed) was then tested for partial strict invariance or holding residuals equal between groups. Partial strict invariance was not achieved (five of six fit indices with *P*'s ≤ 0.01; Table 5).

A closer look at the full scalar model revealed that the Recall of Designs subtest intercept was 49.3 when held equal between groups; when freed to vary between groups, the Recall of Design intercept was 50.0 for the normative group and 46.6 for the ASD group (Table 4, Fig. 2), suggesting that autistic children are expected to have a lower Recall of Designs score than neurotypical children with the same true spatial ability. The opposite pattern was observed for the other spatial subtest, Pattern Construction: when freed, the intercept was 50.0 for the normative group and *increased* to 53.8 for autistic children, indicating that they have a *higher* Pattern Construction score than neurotypical children of the same ability. For comparison, the four verbal and nonverbal subtest intercepts showed much smaller changes, and remained within 0.6 points of the normative group intercept when freed (Table S3). Unlike the verbal and nonverbal factors, spatial factor group differences are not only due to true group differences in spatial ability; some of the difference is also due to measurement bias. For additional data and factor loadings, see Tables S1–S3.

Factor mean differences. As expected, we observed mean between-group differences on all three factors (Table 4). Autistic children showed unstandardized factor scores that were 0.64, 0.50, and 0.34 lower than normative verbal, nonverbal, and spatial scores, respectively. Unfortunately, we can interpret only the direction, not the size, of these mean differences because they were obtained with the scalar model, which showed a poor fit with the data. The mean factor differences changed in the partial scalar models, but the direction always remained the same.

Missing data. Sensitivity analyses conducted with adjustments to imputed values showed no meaningful differences from primary measurement invariance analyses (i.e., minimal or zero change in fit indices, factor loadings, means, or intercepts).

Discussion

Our findings indicate that the DAS-II School Age measures verbal and nonverbal intelligence in autistic children similarly to how it measures these constructs in neurotypical children, but the same is not true of spatial intelligence. Weak measurement invariance (metric and configural) was achieved for the DAS-II in a multigroup confirmatory factor analysis using a correlated three-factor model, but strong (scalar) measurement invariance was not achieved. Without scalar invariance, group mean differences in DAS-II scores do not reflect true group differences in intelligence alone but also unique aspects due to being autistic (i.e., measurement bias). Since partial scalar invariance was achieved only when the spatial subtest intercepts were free to vary, we attribute failed scalar invariance to group bias or artifacts in the spatial subtests.

Interpreting DAS-II Spatial Subtest Scores for Children With ASD

The two spatial subtests showed large changes in intercepts, in opposite directions, when the intercepts were free to vary between groups. The Recall of Designs intercept for autistic children fell 3.4 points below the normative intercept, while the Pattern Construction intercept rose 3.8 points above the normative intercept. These results indicate that for each subtest, an autistic child's score is expected to be below or above, respectively, the score of a neurotypical child with the same true spatial ability. Simply put, Recall of Designs underestimates an autistic child's true ability, and Pattern Construction overestimates it. The large differences in opposite directions for the spatial subtests should not be interpreted as "cancelling each other out" because it is likely that different (and unknown) proportions of each subtest's change are due to measurement bias. Although some methods exist for quantifying bias [Nye & Drasgow, 2011], they are more readily applied to unidimensional models than to our three factor model.

The Pattern Construction subtest may be driving the problematic fit: the average autistic participant performed much better on this subtest than on any other. On average, the ASD sample scored around 46 points on all other subtests (45.5–46.7 points; Table 1), but almost three points higher on Pattern Construction (48.9 points). In contrast, the normative sample showed nearly identical mean scores on all subtests (50.0–50.2 points). Put another way, the normative sample showed a 0 point difference between Pattern Construction and Recall of Designs, while the autistic sample showed a 3.3 point difference on these two spatial subtests. These *different patterns* may explain why the normative model did not fit the ASD data to achieve strong measurement invariance. Consequently, the spatial score *does not hold the same meaning* for children from the ASD and normative samples. For autistic children, the spatial subtests may be tapping different abilities.

The failed measurement invariance is not attributable to group mean differences. As expected, the ASD group showed average lower scores on every subtest, and every factor. Clinicians administering the DAS-II to autistic children might consider placing more emphasis on the verbal and nonverbal reasoning composite scores instead of the spatial or composite GCA (General Conceptual Ability). Historically, some ASD clinicians and researchers have relied upon the SNC (special nonverbal composite) instead of the GCA because the SNC excludes the verbal composite. The logic is that verbal subtests may be poor indicators of intelligence of an autistic person, given the communication difficulties inherent in the diagnosis. However, our results suggest that the spatial score, not the verbal score, poses validity issues. We suggest that clinicians avoid interpreting the SNC and GCA and instead defer to the verbal and nonverbal reasoning standardized scores when utilizing the DAS-II. For example, an autistic child with a true spatial intelligence of 95 could record a DAS-II spatial composite score of 92, or 98; their true spatial intelligence could be overestimated or underestimated, depending on the pattern of their Pattern Construction and Recall of Designs subtest scores. Since it is not possible at this time to quantify and predict how each autistic child's true spatial ability would be misrepresented by the DAS-II Spatial Composite score, we recommend avoiding interpretation of the DAS-II Spatial composite score for autistic children, and consequently their SNC and GCA scores.

Implications for "IQ Splits" in ASD

These results suggest that the oft discussed autistic verbal–nonverbal "IQ splits" are likely to be real, and not an artifact of the DAS-II functioning differently in autistic children than normative sample children. The ASD IQ splits refer to differences between the verbal and nonverbal reasoning scores and do not include the spatial score. Even when such studies of IQ splits have used the DAS-II, such as Nowell et al.'s [2015] investigation of splits in the present ASD data set, the authors analyzed only the verbal and nonverbal composite scores, not the spatial composite score. The verbal and nonverbal composite scores reflect true differences in verbal and nonverbal abilities, according to the partial scalar invariance achieved in the present analysis.

Issues with Modeling the Nonverbal Reasoning Factor

Surprisingly, with the six core subtests, we were unable to fit properly the higher-order factor model that the publisher emphasizes. The published documentation only provides higher-order model results for the infrequently used full battery of six core and six diagnostic subtests. The problem in fitting the higher-order model to the six core subtests lay in the nonverbal factor loading entirely onto the general factor and providing no specific variance. This issue resurfaced when we attempted to fit a three-factor bifactor model, which differs from the higher-order model in that the general factor is orthogonal to the group factors and not permitted to correlate with them. Both nonverbal subtests loaded directly onto *g*, not the nonverbal factor. The issues were even more salient in the ASD data set, where the nonverbal factor showed an even higher and more improbable loading (1.045) onto the general factor in the higher-order model, and the three-factor bifactor model failed to converge at all. Thus, the issue of the nonverbal factor not existing independently of *g* seems intrinsic to the DAS-II and not specific to a particular data set. Eliminating either the general factor (correlated three-factor model) or the nonverbal factor (two-factor bifactor model) resolved the convergence issue and the resulting models showed excellent fit. We are not the first to report that the nonverbal factor may be absorbed entirely by the general factor [Dombrowski et al., 2018], and that second-order factors may provide little additional specific variance over and above *g* [Canivez & McGill, 2016; Dombrowski et al., 2019]. However, it merits mention that when additional DAS-II subtests enter into the model, such as all 20 subtests, other groups have replicated the publishers' reported higher-order model [Dombrowski et al., 2019; Keith, Low, Reynolds, Patel, & Ridley, 2010].

Limitations

The primary limitation of this study concerns the depth at which we can understand the bias. The partial invariance methods used here allow us to identify which factor(s) shows bias, and the directionality of the bias for each subtest. We cannot, however, transform differences in intercepts to differences in DAS-II subtest points and suggest a correction. We also do not know why the bias occurs in these particular subtests. Future research to answer these questions would involve an item-level analysis of differential item functioning between the normative and autistic samples.

A second limitation concerns the missing verbal subtest data in the ASD data set, which was systematically missing for individuals with lower verbal abilities on other auxiliary verbal variables. Much autism research excludes individuals with low verbal abilities [Russell et al., 2019], and we wanted our results to generalize to this very understudied population. Thus we included individuals with missing verbal data in the analyses, and the missing data may have affected model fit. To address this limitation, we reran all invariance analyses twice: with complete cases only and with imputed missing data for these subtests. In both alternative analyses, we found no meaningful change in results.

Finally, the SSC autistic sample used in this analysis, while large and diverse in terms of race and ethnicity, includes only simplex individuals, meaning individuals with no first degree relatives with ASD. If simplex ASD is found to be qualitatively different than multiplex ASD (where ASD is present in one or more first degree relatives), then these results may not generalize to multiplex ASD. At present, this limitation does not cause concern because no studies have identified significant differences in the pattern of cognitive abilities between simplex and multiplex ASD, to our knowledge.

Future Directions

Measurement invariance for autistic individuals has not been investigated in other IQ assessments, such as the Wechsler or Stanford–Binet scales, to our knowledge. Our DAS-II findings suggest that such future analyses may be important. Furthermore, future studies might test measurement bias in commonly used ASD measures by sex as larger data sets of females with ASD become available; measurement invariance can be detected with as few as 200 participants per group [Finch & French, 2016]. Finally, DAS autistic norms could be developed to improve interpretability of the spatial subtest scores for autistic populations.

Conclusions

The DAS-II Spatial standardized score should be interpreted with caution for autistic children. This score likely includes measurement bias or artifacts present for autistic children that are absent in the normative sample children. The verbal and nonverbal reasoning standardized scores do hold the same meaning for both autistic and neurotypical children, according to these results from the largest samples analyzed to date.

Acknowledgments

The authors wish to thank Tiffany Ryan-D'Urso and Jen Bertollo for their assistance with this project. We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to phenotypic data on SFARI Base. Approved researchers can obtain the Simons Searchlight population dataset described in this study (https://www.sfari.org/resource/simons-simplex-collection) by applying at https://base.sfari.org. The authors report no conflicts of interest.

Abbreviations

ASD	autism spectrum disorder
CFI	Confirmatory Fit Index
DAS-II	Differential Abilities Scales, 2nd Edition
GCA	General Conceptual Ability
ID	intellectual disability
IQ	intelligence quotient
NVIQ	nonverbal IQ
RMSEA	root mean square error of approximation
SNC	special nonverbal composite
SSC	Simons Simplex Consortium
TLI	Tucker Lewis Index
VIQ	Verbal IQ

References

Canivez, G. L., & McGill, R. J. (2016). Factor structure of the Differential Ability Scales–Second Edition: Exploratory and hierarchical factor analyses with the core subtests. Psychological Assessment, 28(11), 1475–1488.

Charman, T., Pickles, A., Simonoff, E., Chandler, S., Loucas, T., & Baird, G. (2011). IQ in children with autism spectrum disorders: data from the Special Needs and Autism Project (SNAP). Psychological Medicine, 41(3), 619–627.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal, 14(3), 464–504.

Diamantopoulos, A., & Siguaw, J. A. (2003). Introducing LISREL: A guide for the uninitiated. Thousand Oaks, CA: Sage.

Dombrowski, S. C., Golay, P., McGill, R. J., & Canivez, G. L. (2018). Investigating the theoretical structure of the DAS-II core battery at school age using Bayesian structural equation modeling. Psychology in the Schools, 55(2), 190–207.

Dombrowski, S. C., McGill, R. J., Canivez, G. L., & Peterson, C. H. (2019). Investigating the theoretical structure of the Differential Ability Scales—Second Edition through hierarchical exploratory factor analysis. Journal of Psychoeducational Assessment, 37(1), 91–104.

Elliott, C. D. (2007a). Differential Ability Scales (2nd ed.). San Antonio, TX: Harcourt Assessment.

Elliott, C. D. (2007b). Differential Ability Scales (2nd ed.): Introductory and technical handbook. San Antonio, TX: Harcourt Assessment.

Elliott, C.D., Salerno, J. D., Dumont, R., & Willils, J. O. (2018). The DifferentialAbility Scales–Second Edition. In D. P. Flanagan & E. M. McDonough (Eds.),Contemporary intellectual assessment: Theories, tests, and issues (4th ed., pp.360–382). Guilford.

Finch, W. H., & French, B. F. (2016). Quantifying the influence of partial scalar invariance on mean comparisons: two proposed effect sizes. International Journal of Quantitative Research in Education, 3(4), 292–313.

Fischbach, G. D., & Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron, 68(2), 192–195.

Goldstein, G., Allen, D. N., Minshew, N. J., Williams, D. L., Volkmar, F., Klin, A., & Schultz, R. T. (2008). The structure of intelligence in children and adults with high functioning autism. Neuropsychology, 22(3), 301–312.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183–202.

Jorgensen, T. D., Kite, B. A., Chen, P. Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. Psychological Methods, 23(4), 708–728.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2019). semTools: Useful tools for structural equation modeling. R package version 0.5-2. Retrieved from https://CRAN.R-project.org/package=semTools

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Ability Scales–II: Consistency across ages 4 to 17. Psychology in the Schools, 47(7), 676–697.

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What isthe male-to-female ratio in autism spectrum disorder? A systematic review andmeta-analysis. Journal of the American Academy of Child & Adolescent Psychiatry, 56(6), 466–474.

Muthén, L. K., & Muthén, B. O. (1998). Mplus user's guide (Eighth ed.). Los Angeles, CA: Muthén & Muthén.

Newman, D. A. (2014). Missing data: Five practical guidelines. Organizational Research Methods, 17, 372–411.

Nowell, K. P., Schanding, G. T., Kanne, S. M., & Goin-Kochel, R. P. (2015). Cognitive profiles in youth with Autism Spectrum Disorder: An investigation of base rate discrepancies using the Differential Ability Scales—Second Edition. Journal of Autism and Developmental Disorders, 45(7), 1978–1988.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. Journal of Applied Psychology, 96(5), 966–980.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), Handbook of school psychology (4th ed., pp. 332–374). Hoboken, NJ: Wiley.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1–36 http://www.jstatsoft.org/v48/i02/

Russell, G., Mandy, W., Elliott, D., White, R., Pittwood, T., & Ford, T. (2019). Selection bias on intellectual ability in autism research: A cross-sectional review and meta-analysis. Molecular Autism, 10(1), 9.

Siegel, D. J., Minshew, N. J., & Goldstein, G. (1996). Wechsler IQ profiles in diagnosis of high-functioning autism. Journal of Autism and Developmental Disorders, 26(4), 389–406.

Thaler, N. S., Barchard, K. A., Parke, E., Jones, W. P., Etcoff, L. M., & Allen, D. N. (2015). Factor structure of the Wechsler Intelligence Scale for Children: Fourth Edition in children with ADHD. Journal of Attention Disorders, 19(12), 1013–1021.

Thompson, M. S. (2016). Assessing measurement invariance of scales using multiple-group structural equation modeling. In

K. Schweizer & C. DiStefano (Eds.), Principles and methods of test construction: Standards and recent advances (pp. 218–244). Boston, MA: Hogrefe.

Trundt, K. M., Keith, T. Z., Caemmerer, J. M., & Smith, L. V. (2018). Testing for construct bias in the Differential Ability Scales: A comparison among African American, Asian, Hispanic, and Caucasian children. Journal of Psychoeducational Assessment, 36(7), 670–683.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1–67 https://www.jstatsoft.org/v45/i03/

Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. The Clinical Neuropsychologist, 30(7), 1006–1016.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1: Supporting information.