Journal of
Neurodevelopmental Disorders

RESEARCH Open Access

# Within-task variability on standardized language tests predicts autism spectrum disorder: a pilot study of the Response Dispersion Index

Abby E. Hare-Harris[1*], Marissa W. Mitchel[2], Scott M. Myers[2], Aaron D. Mitchel[3], Brian R. King[4], Brittany G. Ruocco[5], Christa Lese Martin[2], Judy F. Flax[5] and Linda M. Brzustowicz[5]

## Abstract

**Background:** Qualitatively atypical language development characterized by non-sequential skill acquisition within a developmental domain, which has been called developmental *deviance* or *difference*, is a common characteristic of autism spectrum disorder (ASD). We developed the Response Dispersion Index (RDI), a measure of this phenomenon based on intra-subtest scatter of item responses on standardized psychometric assessments, to assess the within-task variability among individuals with language impairment (LI) and/or ASD.

**Methods:** Standard clinical assessments of language were administered to 502 individuals from the New Jersey Language and Autism Genetics Study (NJLAGS) cohort. Participants were divided into four diagnostic groups: unaffected, ASD-only, LI-only, and ASD + LI. For each language measure, RDI was defined as the product of the total number of test items and the sum of the weight (based on item difficulty) of test items missed. Group differences in RDI were assessed, and the relationship between RDI and ASD diagnosis among individuals with LI was investigated for each language assessment.

**Results:** Although standard scores were unable to distinguish the LI-only and ASD/ASD + LI groups, the ASD/ASD + LI groups had higher RDI scores compared to LI-only group across all measures of expressive, pragmatic, and metalinguistic language. RDI was positively correlated with quantitative ASD traits across all subgroups and was an effective predictor of ASD diagnosis among individuals with LI.

**Conclusions:** The RDI is an effective quantitative metric of developmental deviance/difference that correlates with ASD traits, supporting previous associations between ASD and non-sequential skill acquisition. The RDI can be adapted to other clinical measures to investigate the degree of difference that is not captured by standard performance summary scores.

**Keywords:** Developmental difference, Autism spectrum disorder, Language impairment, Intra-subtest scatter

## Background

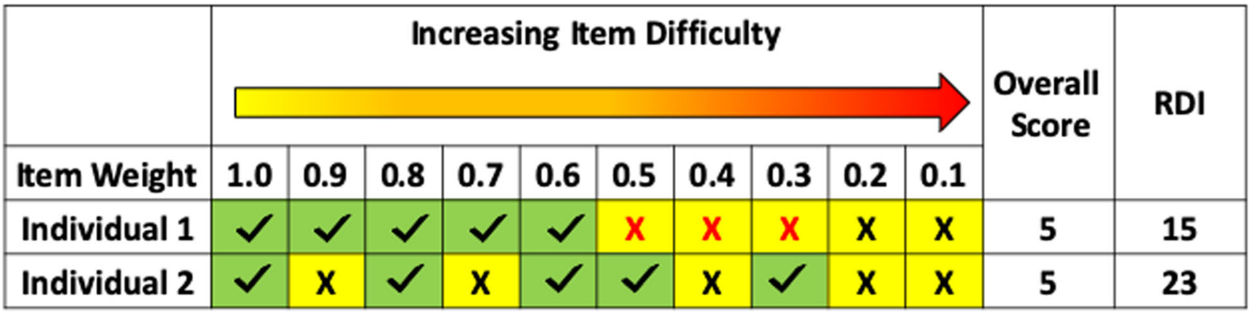
In the early twentieth century, Gesell et al. observed that typical development is methodical, sequential, timed, and therefore largely predictable [1, 2]. This principle is the basis for using developmental milestones and tests as markers of neuromaturation. Gesell's emphasis on patterns and sequences of development set the stage for Piaget, who also emphasized the importance of the *sequence* of staged maturation, while recognizing the role of active experience and social interaction in development [3]. Developmental delay is defined by acquisition of skills in the typical sequence but at a slower rate and, in many cases, with a lower overall developmental ceiling. Delay within a developmental domain results in a measurable performance deficit relative to age norms, and early childhood deficits may be temporary or they may persist. Because the sequence of skill attainment is

* Correspondence: ahare@bloomu.edu
[1]Department of Biological and Allied Health Sciences, Hartline Science Center, Bloomsburg University, 400 East Second St, Bloomsburg, PA 17815, USA
Full list of author information is available at the end of the article

preserved, developmental delay results in a profile of abilities within one or more developmental domains that is similar to the performance of a younger typically developing child. Conversely, the concept of qualitatively atypical development characterized by non-sequential skill acquisition within a developmental domain (with or without developmental delay), resulting in a more widely scattered profile of ability, has been described as *developmental deviance* [4–6] or *developmental difference* [7]. In this case, a child attains more difficult skills within a developmental sequence without having accomplished easier tasks in the sequence. In contrast to developmental delay, this pattern of development results in a profile of abilities that is not commonly found in typically developing children of any age. Although developmental delay is routinely quantified and represented by a rate, quotient, or standard score, atypicality due to non-sequential development is typically described qualitatively [5, 6]. Despite the emphasis on the importance of the sequence of developmental progression by Gesell and Piaget, among others, it remains largely unmeasured in clinical practice. The phenomenon of non-sequential, qualitatively atypical development is manifested on psychometric tests as unusual or inconsistent response patterns to items within subtest scales, resulting in increased intra-subtest scatter or within-task variability [8–10]. It would be useful to quantify this phenomenon to allow investigation of its utility for making clinical diagnoses, prognosticating, and advancing understanding of typical and disordered development.

Qualitative atypicality of language development is a prominent characteristic of autism spectrum disorder (ASD). For example, atypical sequences of skill attainment in the areas of communication and socialization have been reported in individuals with ASD [11]. Echolalia, jargon, and other unusual, non-developmental semantic and syntactic error patterns are also more frequent among children with ASD than typically developing children or children with intellectual disability [12, 13]. Even among individuals with ASD who have higher verbal skills, the language profile that emerges in childhood and persists into adulthood is characterized by unevenness, including poor comprehension relative to apparent expressive language abilities, semantic processing anomalies despite normal performance on vocabulary tests, and idiosyncratic word usage despite relatively intact articulation and syntax [14]. This pattern of language development that is often observed in individuals with ASD is generally thought to be distinct from that found in developmental language disorder [15, 16]. Although "developmental language disorder" (DLD) is currently the preferred term for unexplained language problems in children, much of the pertinent research has focused on "specific language impairment" (SLI) [17]. Unlike DLD, the diagnosis of SLI traditionally required a normal non-verbal IQ and, in some cases, a large discrepancy between verbal and non-verbal ability (a criterion lacking in validity) [17]. Because much of the past research used the SLI definition, we have maintained the use of the term "SLI" when referring to those studies, although the conclusions likely apply to the current "DLD" terminology. SLI, defined as persistent deficits in language skills in the absence of broader cognitive impairment or hearing loss, affects up to 7% of young children [18]. In contrast to ASD, SLI is not clearly associated with prominent non-sequential milestone attainment, and children with SLI can be viewed as exhibiting more of a delayed pattern of development [19, 20]. Despite the distinctions between ASD and SLI, there is also a notable overlap in the language profiles of some individuals with SLI and ASD, particularly in pragmatic language difficulties [21, 22] and structural language impairments [23–26]. Furthermore, language development in preschool children with ASD is frequently observed to be qualitatively atypical as well as delayed [27]. These observations highlight the heterogeneity of both disorders and the potential value of a quantitative metric for assessment of non-sequential skill acquisition for understanding language development and distinguishing among clinical disorders.

Standardized psychometric tests are routinely used in clinical practice to evaluate cognition, language, and other aspects of development. Although these metrics are effective for identifying specific developmental delays and persistent deficits, performance summary scores do not capture atypical response patterns indicative of non-sequential development [8]. Many standardized assessments are arranged hierarchically, using a series of items of graded difficulty. An individual's raw score is determined by establishing a basal, or initial sequence of a certain number of correct answers, and a ceiling, based on a stoppage rule (e.g., a certain number of consecutive items answered incorrectly). Typically developing individuals and those with developmental delay are expected to pass all items to a level of maximum capacity and then fail all items beyond that point (ceiling), with some normal variability occurring around the items at the approximate ability level of the individual being tested [8]. However, individuals with non-sequential development exhibit a more scattered pattern of incorrect answers that is not captured in the standard score. As illustrated in Fig. 1, two individuals with the same overall score (number of correct responses) on a particular subtest may appear to be impaired to a similar extent in that domain; however, at the item response level, the distribution of incorrect answers within a subtest might tell a different story. Increased within-task variability, or intra-subtest scatter, could result from answering some of the easier items incorrectly while answering harder items correctly (as in Fig. 1) or a tendency toward failure on



| | Increasing Item Difficulty | | | | | | | | | | Overall Score | RDI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item Weight | 1.0 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 | | |
| Individual 1 | ✓ | ✓ | ✓ | ✓ | ✓ | X | X | X | X | X | 5 | 15 |
| Individual 2 | ✓ | X | ✓ | X | ✓ | ✓ | X | ✓ | X | X | 5 | 23 |

$$\text{RDI}_{\text{Ind 1}} = (0.5 + 0.4 + 0.3 + 0.2 + 0.1) \times 10 = \mathbf{15}$$

$$\text{RDI}_{\text{Ind 2}} = (0.9 + 0.7 + 0.4 + 0.2 + 0.1) \times 10 = \mathbf{23}$$

**Fig. 1** Distribution of incorrect responses to a clinical measure that is arranged hierarchically and demonstration of RDI calculation. Individual 1 represents a participant who displays developmental delay and reaches a testing ceiling (depicted with red X's). Individual 2 represents a participant who displays developmental difference and does not reach a testing ceiling. The overall score is the total number of correct responses (check marks). Weights for each test item are defined as the percentage of typically developing individuals who responded correctly. The Response Dispersion Index (RDI) calculation is shown for individuals 1 and 2

individual items after long runs of correct responses. The degree of dispersion of correct responses, or within-task variability, warrants investigation as a means of measuring aspects of performance that cannot be captured by currently available standardized subtest performance scores.

Several measures of within-task variability have been applied to clinical intelligence metrics as indicators of cognitive dysfunction, with mixed results in distinguishing individuals with various brain insults or psychiatric disorders from controls [8, 28]. Although few studies have assessed within-task variability on direct measures of language ability quantitatively, one study found that response patterns for children with ASD differed significantly from those of typically developing children and children with non-specific developmental delays on several measures of syntactic development [13]. In this pilot study, we quantified atypical language development resulting from non-sequential skill acquisition by measuring dispersion of item-level responses (within-task variability) on standardized language measures and examined this metric in groups of individuals with ASD with and without language impairment (LI), individuals with LI alone, and controls without ASD or LI. We further assessed the association between within-task variability and the degree of quantitative ASD traits as measured by the Social Responsiveness Scale (SRS) [29].

## Methods

### Participants

This study is a retrospective study of 502 individuals (154 families) from the New Jersey Language and Autism Genetics Study (NJLAGS) cohort [30–32]. All NJLAGS families were initially ascertained for at least one individual with ASD (with no known genetic cause), at least one other family member with a language impairment, and additional unaffected family members willing to participate. All participants (mean age = 52 years old; 53% male), regardless of family membership, were included in this study as individual subjects. The average age of participants with ASD was 13 years old, and the average age of those with language impairment was 24 years old. Participants were primarily recruited from the New Jersey area with the following ethnicity breakdown: 75% White/Caucasian, 5% African American, 1% Asian, < 1% Pacific Islander, 6% more than one race, and 12% unknown/unspecified. Since all language measures are standardized for the English language, all participants were required to have English as their primary language. The study was approved by the Rutgers University Institutional Review Board and the Institutional Review Boards at Geisinger Health System and Bloomsburg University.

### Assessments

All individuals were evaluated using a comprehensive testing battery that included measures of oral language, autistic traits, and cognitive ability. Six subtests of the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4) [33, 34] were included: Word Structure, Recalling Sentences, Formulating Sentences, Word Classes (Expressive and Receptive), Concepts and Following Directions, and Word Definitions. Five subtests of the Comprehensive Assessment of Spoken Language (CASL) [35] were included in the testing battery: Nonliteral Language, Ambiguous Sentences, Meaning from Context, Pragmatic Judgment, and Inference. All assessments were administered using standard basal and ceiling rules. Standard scores for the individuals > 21 years old were determined using the normative values for individuals who are 21 years old. Language impairment (LI) was defined as having a core standard score of ≤ 85 on the CELF-4. Alternatively, an individual without a CELF-4 core standard score qualified as having LI if he/she scored at least 1 SD below the standard mean on 60% of all language measure subtests administered based on age of subject. A total of 19 individuals were removed for having > 10% missing data.

All individuals with ASD met criteria for autism on the following measures: (1) Autism Diagnostic Interview–Revised [36], (2) Autism Diagnostic Observation Schedule [37], and (3) DSM-IV criteria. The autistic traits of all participants, including individuals without ASD, were assessed using the SRS [29]. The SRS is a 65-item questionnaire that assesses the severity of ASD traits using a quantitative scale.

## Diagnostic groups

Since the language tests used in the present study were normed on a pediatric population (up to 21 years old), we focused our phenotypic analyses on individuals ≤ 21 years old (total $n = 187$). These individuals were assigned to four diagnostic groups: (1) individuals with an ASD diagnosis, who did not meet criteria for LI (ASD-only; $n = 17$); (2) individuals who met criteria for LI, but did not have an ASD diagnosis (LI-only; $n = 28$); (3) individuals with an ASD diagnosis who also met criteria for LI (ASD + LI; $n = 27$); and (4) individuals who did not meet criteria for LI or ASD (unaffected; $n = 115$). We also assessed group differences when the cohort was stratified by ASD diagnosis. There were 44 individuals with an ASD diagnosis (ASD+) and 133 without an ASD diagnosis (ASD−). Means and standard deviations for all subtests, IQ, and age for individuals ≤ 21 years old are listed in Table 1.

**Table 1** Means and standard deviation of standard scores for CELF and CASL subtests

| Test | | Unaffected | | ASD-only | | LI-only | | ASD + LI | |
|---|---|---|---|---|---|---|---|---|---|
| | | *N* < 21 | Mean (SD) | *N* < 21 | Mean (SD) | *N* < 21 | Mean (SD) | *N* < 21 | Mean (SD) |
| | | *N* all | | *N* all | | *N* all | | *N* all | |
| CELF | Concepts and Following Directions | 68 | 10.31 (2.47) | 9 | 10.56 (2.19) | 25 | 4.60 (2.52) | 21 | 3.57 (2.84) |
| | Formulated Sentences | 108 | 11.33 (2.30) | 17 | 9.71 (2.34) | 28 | 5.43 (2.92) | 29 | 4.07 (3.32) |
| | | 332 | 11.51 (2.17) | 18 | 9.72 (2.27) | 46 | 5.46 (3.01) | 32 | 3.91 (3.25) |
| | Recalling Sentences | 115 | 10.53 (2.44) | 19 | 8.16 (2.81) | 27 | 5.33 (2.11) | 32 | 3.50 (3.17) |
| | | 348 | 10.04 (2.62) | 20 | 7.85 (3.07) | 47 | 4.77 (2.47) | 35 | 3.46 (3.15) |
| | Word Structure | 34 | 9.71 (2.34) | 6 | 8.67 (2.50) | 16 | 4.19 (2.56) | 11 | 4.27 (3.32) |
| | Word Definitions | 45 | 12.62 (2.52) | 9 | 11.56 (1.81) | 3 | 5.67 (0.58) | 10 | 5.20 (4.37) |
| | | 278 | 12.96 (2.31) | 9 | 11.56 (1.81) | 23 | 7.61 (3.65) | 13 | 5.08 (4.25) |
| | Word Classes-Total | 78 | 11.44 (2.70) | 12 | 9.92 (1.68) | 12 | 6.08 (1.93) | 19 | 5.37 (3.37) |
| | | 311 | 11.67 (2.13) | 13 | 9.31 (2.72) | 31 | 6.26 (2.56) | 22 | 5.00 (3.32) |
| CASL | Ambiguous Sentences | 59 | 102.20 (15.30) | 9 | 90.22 (19.27) | 7 | 80.14 (6.44) | 10 | 77.50 (12.42) |
| | | 286 | 97.73 (13.68) | 10 | 87.90 (19.60) | 25 | 75.12 (8.82) | 13 | 76.77 (13.92) |
| | Inference | 67 | 101.37 (11.30) | 7 | 83.29 (24.20) | 12 | 78.50 (16.51) | 19 | 70.58 (20.61) |
| | | 70 | 101.17 (11.75) | 7 | 83.29 (24.20) | 12 | 78.50 (16.51) | 19 | 70.58 (20.61) |
| | Pragmatic Judgment | 115 | 103.17 (10.88) | 21 | 88.29 (16.70) | 27 | 79.33 (12.96) | 28 | 65.93 (19.75) |
| | | 343 | 98.21 (11.09) | 22 | 86.09 (19.28) | 47 | 79.77 (14.61) | 31 | 64.97 (20.28) |
| | Non-literal Language | 85 | 106.01 (12.92) | 13 | 96.08 (17.11) | 17 | 85.53 (14.52) | 22 | 70.68 (21.22) |
| | | 311 | 98.98 (13.84) | 14 | 92.07 (22.25) | 36 | 77.56 (17.07) | 25 | 68.80 (21.85) |
| | Meaning from Context | 56 | 103.46 (18.54) | 9 | 95.44 (15.88) | 7 | 83.00 (7.26) | 8 | 72.00 (17.67) |
| | | 283 | 102.19 (13.17) | 10 | 91.00 (20.53) | 25 | 79.36 (11.90) | 11 | 97.27 (18.47) |
| Age | Years | 115 | 10.94 (4.55) | 17 | 11.06 (5.89) | 28 | 8.59 (3.59) | 27 | 9.92 (4.00) |
| | | 369 | 29.73 (19.17) | 24 | 30.08 (18.55) | 53 | 27.83 (16.80) | 37 | 28.00 (18.72) |
| SRS | Total raw score | 76 | 26.36 (24.31) | 13 | 77.85 (31.92) | 15 | 57.93 (38.06) | 23 | 90.13 (34.04) |
| | | 279 | 30.09 (25.45) | 20 | 87.00 (34.69) | 28 | 60.39 (41.43) | 30 | 94.27 (32.85) |
| IQ | PIQ | 105 | 106.18 (11.73) | 16 | 96.38 (17.82) | 26 | 91.77 (14.59) | 27 | 84.41 (21.22) |
| | | 337 | 108.93 (12.24) | 18 | 97.17 (17.71) | 46 | 92.41 (13.44) | 29 | 88.48 (15.65) |

Due to the small number of individuals ≤ 21 years old, the statistical power of our sample was limited when stratifying the cohort by both ASD and LI diagnoses. Although our standard measures used normative samples not exceeding 21 years old, there was no evidence of ceiling effects in this sample when administered to ages beyond the normative data. Therefore, we also assessed group differences among the entire cohort regardless of participant age among the four diagnostic groups: (1) ASD-only ($n = 24$), (2) LI-only ($n = 53$), (3) ASD + LI ($n = 37$), and (4) unaffected ($n = 369$). We also assessed group differences when the cohort was stratified into two ASD diagnostic groups: (1) ASD+ ($n = 61$) and (2) ASD− ($n = 422$). Means and standard deviations for all subtests, IQ, and age for the entire cohort are listed in Table 1.

### Response Dispersion Index

In order to quantify the dispersion of incorrect answers for each diagnostic group, we adapted a measure of item-response distributions from VanMeter and colleagues [11], who utilized an equation for calculating *inefficiency* developed by Hallenbeck et al. (1965) [9]. In this method, *inefficiency* is defined as the product of the sum of the weights of the items missed and the peak item passed.

$$\text{Inefficiency} = \sum (\text{weight of items missed}) \times \text{peak item passed}$$

However, this measure assumes that test items get progressively more difficult and that testing is discontinued after a traditional ceiling is met. Because many of the language subtests in this study have variable start/stop rules and do not necessarily increase in difficulty in a linear fashion, we used the total number of subtest items instead of the peak item passed. For subtests that did have ceiling rules, any items that were not administered because the subject had already reached a ceiling were counted as incorrect responses. Any items that were not administered because they occurred before the start point for the subject's age were counted as correct responses. Our modified metric, which we have named the Response Dispersion Index (RDI), therefore is defined as the product of the sum of the weights of the items missed and the total number of subtest items.

$$\text{RDI} = \sum (\text{weight of items missed}) \times \#\text{subtest items}$$

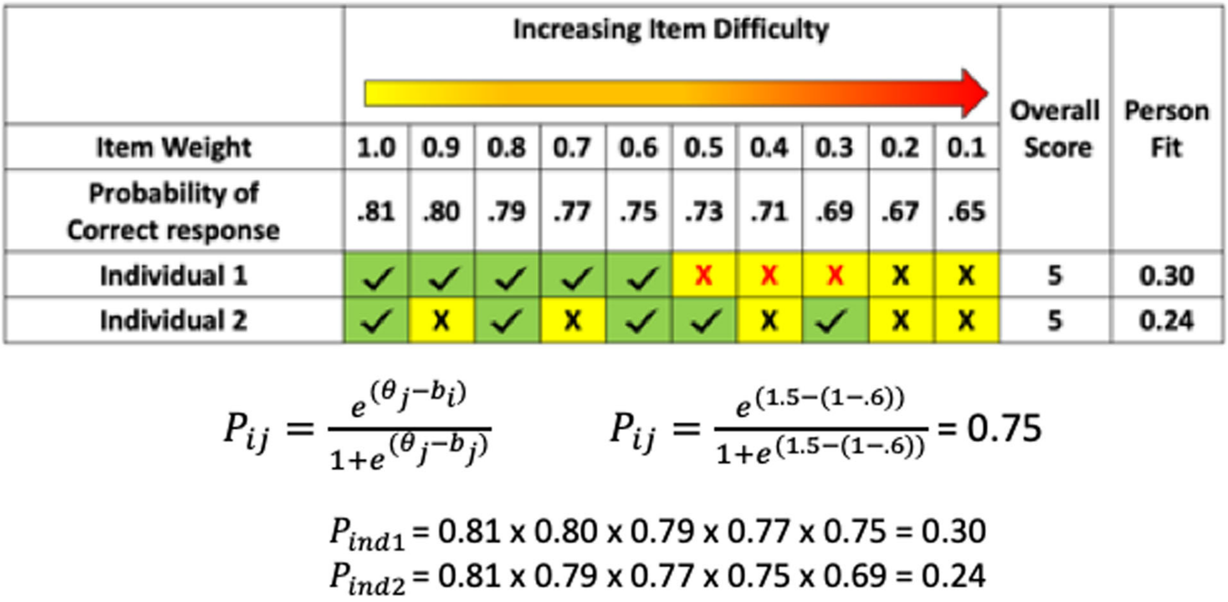
We defined the weight of each item missed as the percentage of unaffected family members who correctly answered the test item. For multivariate items (i.e., those in which it is possible to receive partial credit rather than a binary correct/incorrect response), the weight is defined as the sum of the products of the percentage of unaffected family members (≤ 21 years old) who correctly answered the test item and response value for each possible item response. We calculated the RDI for all participants for each CELF and CASL subtest.

$$\text{Weight} = \sum_{1-n} (\%\text{unaffected correct response} \times \text{response value})$$

The weight of each test item decreases as the difficulty level increases, resulting in higher RDI scores for individuals who exhibit a greater degree of within-task variability, as illustrated by our previous hypothetical example (Fig. 1). Individual 1 reached a standard testing ceiling, and the weights for the last 5 items are totaled, resulting in an RDI score of 15. However, despite obtaining the same overall score of 5, individual 2 had a more atypical pattern of incorrect items of varying difficulty (items 2, 4, 7, 9, and 10) and did not reach a testing ceiling. This pattern of incorrect responses results in an RDI score of 23, indicating a greater degree of within-task variability than individual 1 (Fig. 1).

### Rasch person-fit model

In order to validate our RDI metric, we compared the diagnostic classification accuracy of the RDI with the classification accuracy of a one-parameter Rasch model that is a derivative of the Item Response Theory Model utilized by Godber et al. [8]. This model provides an alternative measure of within-task scatter to assess the degree of dispersion in each language assessment. While other measures of within-task scatter do not take item difficulty into account, the Rasch model calculates the likelihood that the overall subtest score is a true measure of an individual's ability given the difficulty of the test items the individual answered correctly [8, 38, 39]. It does so by comparing the observed accuracy for each test item to the accuracy expected based on the individual's overall ability level ($\Theta$) relative to the difficulty of the items that the individual answered correctly. In this case, the difficulty level of each item was defined as the percentage of unaffected individuals who answered the item incorrectly. The model assumes that the overall score of a subtest is indicative of a person's ability and uses this to predict which weighted test items the individual should have answered correctly. Then, the model assesses the "fit" between this predicted pattern and the individual's actual pattern of item responses. Since the language tests investigated here assume that individuals with poorer language ability would obtain lower standard scores, a decreased model fit should be indicative of developmental difference. Consider again our hypothetical example in Fig. 2. For each item, the probability of obtaining a correct answer is calculated using the same ability and item difficulty level for both individuals. The person-fit for each individual is then defined as the product of the probability of a correct response for each item that the individual answered correctly. For example,



| | Increasing Item Difficulty | | | | | | | | | | Overall Score | Person Fit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item Weight | 1.0 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 | | |
| Probability of Correct response | .81 | .80 | .79 | .77 | .75 | .73 | .71 | .69 | .67 | .65 | | |
| Individual 1 | ✓ | ✓ | ✓ | ✓ | ✓ | X | X | X | X | X | 5 | 0.30 |
| Individual 2 | ✓ | X | ✓ | X | ✓ | ✓ | X | ✓ | X | X | 5 | 0.24 |

$$P_{ij} = \frac{e^{(\theta_j - b_i)}}{1+e^{(\theta_j - b_j)}} \qquad P_{ij} = \frac{e^{(1.5-(1-.6))}}{1+e^{(1.5-(1-.6))}} = 0.75$$

$$P_{ind1} = 0.81 \text{ x } 0.80 \text{ x } 0.79 \text{ x } 0.77 \text{ x } 0.75 = 0.30$$
$$P_{ind2} = 0.81 \text{ x } 0.79 \text{ x } 0.77 \text{ x } 0.75 \text{ x } 0.69 = 0.24$$

**Fig. 2** Demonstration of Rasch calculation. Individual 1 represents a participant who displays developmental delay and reaches a testing ceiling (depicted with red X's). Individual 2 represents a participant who displays developmental difference and does not reach a testing ceiling. The overall score is the total number of correct responses (check marks). Weights for each test item are defined as the percentage of typically developing individuals who responded correctly. A sample calculation for determining the probability of a correct response is shown for item 5, and the overall probability of the observed distribution of responses is shown for individuals 1 and 2

individual 1 answered the first 5 items correctly, so the person-fit metric is the product of the probabilities for the first 5 items, or 0.30. Individual 2, on the other hand, answered items 1, 3, 5, 6, and 8 correctly, so their person-fit score is 0.24, indicating a more atypical response pattern (Fig. 2).

We calculated person-fit for all CELF-4 and CASL subtests using the *psych* package in R [40] and used a Pearson correlation to assess the concordance between RDI and person-fit. While person-fit and RDI were expected to produce similar results, the RDI has several advantages over the Rasch model. The Rasch model assumes that the overall score of a subtest is indicative of a person's ability and uses this measure to assess the "fit" between this predicted pattern and the individual's actual pattern of item responses. However, our preliminary analysis indicated that an overall subtest score is not an accurate reflection of a persons' ability if there is a high degree of within-task scatter. Instead, the RDI eliminates the user-ability parameter (Θ), resulting in a more parsimonious, direct assessment of within-task scatter with fewer assumptions built into the model.

### Statistical methods

In order to standardize RDI scores across subtests, we first calculated aggregate RDI scores across subtests by averaging the standardized residuals (i.e., *z*-scores) of all subtests for each participant. Group differences in subtest standard scores and RDI for all language subtests were assessed using a one-way ANOVA with a post hoc Tukey analysis and Bonferroni correction for multiple testing. Due to the smaller sample size for individuals who were ≤ 21 years old, this analysis was repeated with the entire cohort to increase our statistical power (results shown in Additional file 1). Pearson correlations were used to investigate the relationship between RDI and SRS scores across all diagnostic subgroups. Logistic regression was used to investigate the ability of the RDI to predict ASD diagnosis. Specifically, we tested whether RDI scores on the CASL and CELF-4 can accurately classify participants into ASD+ or ASD− groups. These aggregate RDI values were entered into the model as a predictor variable, with diagnostic classification (ASD+ or ASD−) as the outcome variable and IQ, age, and sex as covariates.

## Results

As a group, the standard scores of individuals with ASD/ASD + LI did not differ significantly from individuals with LI-only for any language measure in the testing battery after correction for multiple testing (Table 1). Despite similar overall standard scores, visual inspection of the distribution of correct/incorrect test item responses appeared to be qualitatively different between the ASD + LI/ASD-only and LI-only groups for a number of measures. This was even more striking when comparing the LI-only group to the ASD + LI group alone.

**Table 2** Group differences in Response Dispersion Index (RDI) for ASD+ and ASD– for participants ≤ 21 years old

| | | ASD+ | | ASD– | | ASD+ vs ASD– | |
|---|---|---|---|---|---|---|---|
| | | *N* | Mean (SD) | *N* | Mean (SD) | *t* | *p* value |
| CELF | Concepts and Following Directions | 17 | .456 (1.170) | 81 | – .096 (.941) | – 2.103 | 0.038 |
| | Formulated Sentences | 38 | .409 (1.206) | 124 | – .125 (.897) | – 2.527 | 0.015 |
| | Recalling Sentences | 40 | .691 (1.244) | 128 | – .216 (.802) | – 4.340 | < 0.001 |
| | Word Structure | 13 | .549 (1.020) | 40 | – .178 (.938) | – 2.377 | 0.021 |
| | Word Definitions | 16 | .468 (1.307) | 46 | – .163 (.826) | – 1.809 | 0.086 |
| | Word Classes–Total | 26 | .527 (1.185) | 86 | – .159 (.884) | – 3.193 | 0.002 |
| CASL | Ambiguous Sentences | 14 | .582 (1.067) | 60 | – .136 (.942) | – 2.505 | 0.015 |
| | Inference | 17 | .889 (1.189) | 74 | – .204 (.835) | – 3.594 | 0.002 |
| | Pragmatic Judgment | 35 | .550 (1.058) | 122 | – .158 (.928) | – 3.850 | < 0.001 |
| | Non-literal Language | 27 | .579 (1.163) | 94 | – .166 (.887) | – 3.085 | 0.004 |
| | Meaning from Context | 12 | .446 (1.092) | 60 | – .089 (.966) | – 1.715 | 0.091 |
| | Average RDI all subtests | 44 | .775 (1.058) | 133 | – .063 (.763) | – 4.850 | < 0.001 |

In order to quantitatively characterize the differences in the item-response distributions between the diagnostic groups, we calculated the RDI for all subtests for each participant. We calculated aggregate RDI scores across subtests by averaging the standardized residuals (i.e., *z*-scores) of all subtests for each participant (Table 2). The ASD+ group had significantly higher RDI scores than ASD– group for all but two subtests (Table 2; Table 3). When the cohort was further subdivided by language impairment, individuals with ASD + LI had higher RDI scores than the LI-only group; however, this difference did not reach statistical significance. This is likely due to the limited power of our small sample size. When considering all individuals regardless of age (Table 4), the ASD + LI group had consistently higher RDI scores than the LI group for all measures of oral language, with the following subtests reaching statistical significance ($p < .001$ unless otherwise noted): CASL Non-literal Language, CASL Pragmatic Judgment, CASL Meaning from Context, CELF Recalling Sentences, CELF Word Definitions, and CELF Word Classes ($p = .003$). While not statistically significant, the ASD + LI group also exhibited consistently higher RDI scores for the remaining measures of language ability (Table 4). This trend was also consistent among individuals without LI. When compared to the unaffected group, the ASD-only group had significantly higher RDI scores ($p < .001$ unless otherwise noted) on the following subtests: CASL Ambiguous Sentences, CASL Pragmatic Judgment, CASL Non-literal Language, CASL Meaning from Context, CELF Concepts and Following Directions, CELF Word Classes ($p = .002$), CELF Formulated Sentences ($p = .002$), and CELF Recalling Sentences (Table 4).

**Table 3** Group differences in Response Dispersion Index (RDI) for ASD+ and ASD– for all participants

| Test | Subtest | ASD+ | | ASD– | | ASD+ vs ASD– | |
|---|---|---|---|---|---|---|---|
| | | *N* | Mean (SD) | *N* | Mean (SD) | *t* | *p* value |
| CELF | Concepts and Following Directions | 30 | .419 (1.079) | 93 | – .131 (.943) | – 2.649 | .009 |
| | Formulated Sentences | 48 | 1.057 (1.429) | 377 | – .135 (.844) | – 5.652 | < .001 |
| | Recalling Sentences | 52 | 1.363 (1.454) | 394 | – .180 (.761) | – 7.519 | < .001 |
| | Word Structure | 17 | .450 (.987) | 50 | – .153 (.967) | – 2.210 | .031 |
| | Word Definitions | 22 | 1.613 (1.830) | 300 | – .118 (.796) | – 4.406 | < .001 |
| | Word Classes-Total | 34 | 1.486 (1.395) | 340 | – .149 (.817) | – 6.716 | < .001 |
| CASL | Ambiguous Sentences | 23 | 1.296 (1.076) | 307 | – .097 (.925) | – 6.883 | < .001 |
| | Inference | 26 | .956 (1.032) | 80 | – .311 (.770) | – 5.759 | < .001 |
| | Pragmatic Judgment | 53 | 1.240 (1.168) | 388 | – .169 (.845) | – 8.486 | < .001 |
| | Non-literal Language | 36 | 1.601 (1.368) | 341 | – .169 (.783) | – 7.630 | < .001 |
| | Meaning from Context | 20 | 1.862 (1.478) | 305 | – .122 (.828) | – 5.943 | < .001 |

**Table 4** Group differences in Response Dispersion Index (RDI) for phenotypic groups for all participants

| Test | | Unaffected | | ASD-only | | LI-only | | ASD + LI | | | Omnibus ANOVA | Unaffected vs ASD-only | LI-only vs LI + ASD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean (SD) | *N* | Mean (SD) | *N* | Mean (SD) | *N* | Mean (SD) | *F* | *p* value | *p* value | *p* value |
| CELF | Concepts and Following Directions | 68 | − .413 (.723) | 9 | − .096 (1.089) | 25 | .638 (1.050) | 21 | .651 (1.016) | 13.180 | < .001 | < .001 | 1.000 |
| | Formulated Sentences | 331 | − .293 (.604) | 18 | .418 (1.295) | 46 | 1.006 (1.327) | 30 | 1.440 (1.386) | 69.426 | < .001 | 0.002 | 0.112 |
| | Recalling Sentences | 347 | − .331 (.568) | 20 | .483 (1.213) | 47 | .935 (1.040) | 32 | 1.913 (1.328) | 121.291 | < .001 | < .001 | < .001 |
| | Word Structure | 34 | − .548 (.610) | 6 | − .044 (.926) | 16 | .688 (1.062) | 11 | .719 (.951) | 11.608 | < .001 | 0.512 | 1.000 |
| | Word Definitions | 277 | − .247 (.596) | 9 | .081 (.656) | 23 | 1.434 (1.188) | 13 | 2.674 (1.611) | 101.749 | < .001 | 0.533 | < .001 |
| | Word Classes-Total | 310 | − .291 (.618) | 13 | .391 (.719) | 30 | 1.327 (1.120) | 21 | 2.163 (1.281) | 114.278 | < .001 | 0.005 | < .001 |
| CASL | Ambiguous Sentences | 283 | − .224 (.825) | 10 | .744 (1.019) | 24 | 1.404 (.696) | 13 | 1.721 (.947) | 51.376 | < .001 | 0.002 | 0.684 |
| | Inference | 68 | − .515 (.559) | 7 | .269 (1.047) | 12 | .849 (.794) | 19 | 1.210 (.929) | 37.411 | < .001 | 0.029 | 0.504 |
| | Pragmatic Judgment | 341 | − .298 (.691) | 22 | .829 (1.248) | 47 | .760 (1.212) | 31 | 1.532 (1.031) | 71.767 | < .001 | < .001 | < .001 |
| | Non-literal Language | 306 | − .300 (.593) | 14 | .558 (.966) | 35 | .977 (1.214) | 22 | 2.265 (1.165) | 111.968 | < .001 | < .001 | < .001 |
| | Meaning from Context | 280 | − .236 (.708) | 9 | .898 (1.220) | 25 | 1.151 (1.013) | 11 | 2.652 (1.199) | 74.799 | < .001 | < .001 | < .001 |

In order to validate our findings, we used a Rasch person-fit model as an assessment of within-task variability. As described earlier, this model calculates the likelihood that the overall subtest score is a true measure of an individual's ability given the difficulty of the test items the individual answered correctly [8, 38, 39]. Since the language tests investigated here assume that individuals with poorer language ability would obtain lower standard scores, a decreased model fit should be indicative of greater within-task variability. While this procedure is comparable to our RDI metric, by incorporating the individual ability level, the person-fit model includes a measure of developmental delay to infer within-test scatter. The person-fit metric was highly correlated with RDI (*r* range − 0.84 to − 0.98; $p < .001$) for all subtests of the CASL and all but one subtest of the CELF-4 (≤ 21 year old group shown in Table 5; entire cohort listed in Table 6; scatterplots shown in Additional file 1: Figure S1), providing a source of convergent validity for the RDI metric results.

Given the group differences between the ASD+ and ASD− groups, we next investigated the relationship between RDI and quantitative autistic traits across all individuals, regardless of diagnosis, as measured by the SRS. We found that SRS scores were positively correlated with RDI scores on all subtests except for the CELF-4 Word Structure subtest (Table 7). This was also true when considering all individuals in the NJLAGS cohort (Table 7; scatterplots shown in Additional file 1: Figure S2). In addition to correlation with SRS scores, we examined the accuracy of the RDI in predicting ASD diagnosis using a

**Table 5** Rasch model of developmental difference for CELF-4 and CASL subtests and correlation with Response Dispersion Index (RDI) for participants ≤ 21 years old

| | | ASD+ person-fit | ASD– person-fit | ASD+ vs. ASD– | Pearson's correlation person-fit vs RDI | | |
|---|---|---|---|---|---|---|---|
| Test | Subtest | Mean (SD) | Mean (SD) | *p* value | *N* | *r* | *p* value |
| CELF | Concepts and Following Directions | .53 (.231) | .73 (.255) | 0.004 | 98 | − 0.263 | 0.008 |
| | Formulated Sentences | .48 (.294) | .67 (.281) | < 0.001 | 162 | − 0.938 | < 0.001 |
| | Recalling Sentences | .47 (.290) | .68 (.240) | < 0.001 | 168 | − 0.983 | < 0.001 |
| | Word Structure | .48 (.256) | .66 (.235) | 0.022 | 52 | − 0.994 | < 0.001 |
| | Word Definitions | .36 (.207) | .55 (.216) | 0.005 | 61 | − 0.814 | < 0.001 |
| | Word Classes-Total | .54 (1.235) | 1.37 (1.173) | 0.004 | 110 | − 0.983 | < 0.001 |
| CASL | Ambiguous Sentences | .29 (.266) | .45 (.259) | 0.038 | 74 | − 0.980 | < 0.001 |
| | Inference | .34 (.248) | .57 (.206) | < 0.001 | 90 | − 0.981 | < 0.001 |
| | Pragmatic Judgment | .02 (.021) | .04 (.028) | < 0.001 | 160 | − 0.861 | < 0.001 |
| | Non-literal Language | .29 (.266) | .45 (.259) | 0.004 | 120 | − 0.981 | < 0.001 |
| | Meaning from Context | .35 (.226) | .47 (.231) | 0.105 | 71 | − 0.975 | < 0.001 |

**Table 6** Rasch model of developmental difference for CELF-4 and CASL subtests and correlation with Response Dispersion Index (RDI) for all participants

| Test | Subtest | ASD+ person-fit | ASD– person-fit | ASD+ vs. ASD– | Pearson's correlation person-fit vs RDI | | |
|---|---|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | *p* value | *N* | *r* | *p* value |
| CELF | Concepts and Following Directions | .57 (.279) | .71 (.252) | 0.010 | 123 | – 0.998 | < .001 |
| | Formulated Sentences | .45 (.297) | .80 (.232) | < .001 | 425 | – 0.945 | < .001 |
| | Recalling Sentences | .44 (.281) | .79 (.209) | < .001 | 446 | – 0.947 | < .001 |
| | Word Structure | .49 (.238) | .64 (.233) | 0.026 | 67 | – 0.994 | < .001 |
| | Word Definitions | .30 (.216) | .59 (.207) | < .001 | 322 | – 0.773 | < .001 |
| | Word Classes-Total | .45 (.267) | .78 (.185) | < .001 | 374 | – 0.996 | < .001 |
| CASL | Ambiguous Sentences | .21 (.242) | .57 (.271) | < .001 | 330 | – 0.984 | < .001 |
| | Inference | .26 (.261) | .57 (.216) | < .001 | 106 | – 0.983 | < .001 |
| | Pragmatic Judgment | .43 (.290) | .82 (.233) | < .001 | 441 | – 0.997 | < .001 |
| | Non-literal Language | .32 (.275) | .72 (.202) | < .001 | 377 | – 0.992 | < .001 |
| | Meaning from Context | .27 (.254) | .67 (.211) | < .001 | 325 | – 0.989 | < .001 |

logistic regression analysis. The aggregate RDI values were entered into the model as a predictor variable, with ASD diagnostic classification as the outcome variable. IQ, age, and sex were included as covariates. This model showed that the RDI was able to successfully predict ASD diagnosis among individuals ≤ 21 years old 81% of the time (Hosmer-Lemeshow goodness of fit: $\chi^2$ (1) = 2.518, df = 8, $p = .961$; Nagelkerke $R^2 = .379$). Aggregate RDI contributed significantly to the classification accuracy of the model ($B$(S.E.) = 1.359 (0.334), $Z_{\text{wald}} = 16.569$, $p < .001$), with the odds ratio (Exp($B$) = 3.893, 95% CI [2.023, 7.491]) indicating that as RDI increases by one standard deviation, participants were approximately 3.8 times as likely to be diagnosed with ASD.

## Discussion

Although delays in various aspects of development are routinely quantified and treated as continuously distributed variables, non-sequential skill acquisition resulting in *developmental deviance* or *developmental difference* has conventionally been treated as a qualitative, dichotomous descriptor of atypical skill profiles. In this study, instead of using a categorical approach, we developed a measure of the dispersion of item-level responses called the RDI, which is modified from Van-Meter and colleagues' *inefficiency* metric, to determine the degree of within-task variability in language skills [11]. This pilot study is the first, to our knowledge, to use a quantitative metric to investigate within-task variability of performance on standardized language tests in individuals with ASD [11].

Using the RDI, we demonstrated that individuals with ASD exhibit a higher degree of within-task variability than individuals without ASD, and individuals with ASD + LI exhibited the highest levels of within-task variability for all language measures. The RDI was positively

**Table 7** Positive correlation between Response Dispersion Index (RDI) and Social Responsiveness Scale (SRS)

| | | < 21 years old | | | All participants | | |
|---|---|---|---|---|---|---|---|
| | | *N* | Correlation | *p* value | *N* | Correlation | *p* value |
| CELF | Concepts and Following Directions | 69 | 0.381 | 0.001 | 123 | 0.358 | 0.001 |
| | Formulated Sentences | 117 | 0.251 | 0.006 | 428 | 0.344 | < 0.001 |
| | Recalling Sentences | 121 | 0.317 | < 0.001 | 450 | 0.388 | < 0.001 |
| | Word Structure | 33 | 0.256 | 0.151 | 67 | 0.312 | 0.044 |
| | Word Definitions | 47 | 0.398 | 0.006 | 323 | 0.431 | < 0.001 |
| | Word Classes-Total | 86 | 0.325 | 0.002 | 377 | 0.381 | < 0.001 |
| CASL | Ambiguous Sentences | 57 | 0.438 | 0.001 | 334 | 0.322 | < 0.001 |
| | Inference | 72 | 0.374 | 0.001 | 108 | 0.484 | < 0.001 |
| | Pragmatic Judgment | 115 | 0.303 | 0.001 | 443 | 0.373 | < 0.001 |
| | Non-literal Language | 94 | 0.320 | 0.002 | 386 | 0.409 | < 0.001 |
| | Meaning from Context | 60 | 0.280 | 0.030 | 329 | 0.407 | < 0.001 |

correlated with quantitative autistic traits (as measured by the SRS) across all members of the NJLAGS cohort regardless of ASD/LI diagnosis. We also demonstrated that the RDI is a strong predictor of ASD status, supporting previous findings that atypical, non-sequential development is associated with ASD [11, 13, 27, 41]. These findings suggest that quantifying within-task variability in language development based on item-level subtest response patterns adds important information about language in ASD that is not captured by standard performance summary scores. We have demonstrated that the RDI can be used effectively to quantify atypical, non-sequential development.

While the RDI generally differed between individuals with LI only and those with ASD across language measures, the Word Structure subtest from the CELF-4 was one exception. The RDI calculated for this subtest did not reach significance between diagnostic groups and was also the only subtest that was not significantly correlated with quantitative ASD traits across all groups. Although this could be due to the smaller sample size for this subtest, Word Structure involves a cloze procedure in which the child must complete a sentence with a targeted structure (morphology) [33]. Morphological structures assessed with this subtest include verb tense, copula and auxiliary forms, referential pronouns, and other grammatical morphemes. While there is a fairly robust literature documenting differences in the development of semantics, pragmatics, and syntax in individuals with ASD [12–15], evidence for atypical development in word-level morphology in ASD is sparse. One study found that morphological skills in children with ASD, while impaired relative to typical children, were similar to children with developmental delay [42]. Furthermore, the researchers found few differences in the order of acquisition of specific grammatical morphemes between the children with ASD and typical children. Thus, the lack of association between RDI in morphology and ASD diagnosis in the present study is consistent with previous findings that the development of morphology in children with ASD, while delayed, is not necessarily acquired in an atypical sequence.

Although this is the first study to use a quantitative metric to assess within-task variability in direct language testing, clinicians have been interested in inconsistent or unusual item-level response patterns within a test or subtest as a diagnostic variable or marker of cognitive dysfunction for many years [7–10, 28]. Similar to clinical language measures, psychometric measures of cognition assume a fixed order in skill development and, as such, are often arranged hierarchically. This assumption makes these tools effective in identifying developmental delay; however, as demonstrated by Visser et al., this methodology may under-characterize the extent of within-task variability [7]. The RDI metric used in our study eliminates this assumption by calculating the difficulty of individual test items. The weight of each test item decreases as the difficulty level increases, regardless of test item order, making the RDI metric a robust method for detecting developmental difference for a variety of assessment types.

An alternative metric for detecting within-task variability is the Rasch person-fit model based on Item Response Theory. Godber et al. successfully used this method to discriminate between children treated with cranial irradiation for acute lymphoblastic leukemia and matched healthy controls [8]. As described earlier, Rasch person-fit is a metric that estimates an individual's overall ability level by analyzing the difficulty level of correctly answered items within a psychometric assessment. The difficulty level of each item is defined as the percentage of unaffected individuals (or a normative sample, if available) who answer the item correctly. This procedure is comparable to our RDI metric, and we found that the person-fit metric was highly correlated with the RDI when applied to the language measures in the NJLAGS sample.

While person-fit and RDI produce similar results, the RDI has several advantages over the Rasch model of intra-subtest scatter. The Rasch model assumes that the overall score of a subtest is indicative of a person's ability and uses this measure to predict which weighted test items that individual should have answered correctly, assessing the "fit" between this predicted pattern and the individual's actual pattern of item responses. In contrast, the RDI metric makes no mathematical assumptions regarding the relationship between an overall subtest score and a person's ability. Furthermore, by eliminating the user-ability parameter, the RDI directly measures within-task scatter, making this metric more parsimonious and reducing the number of assumptions built into the model. This is particularly important given that the user-ability parameter is typically defined by patterns of developmental delay, not difference. In the present study, we demonstrate that developmental delay fails to adequately capture all patterns of atypical response profiles. Additionally, by using a normative sample to define item weights, the RDI can be standardized for a given psychometric assessment, making it a tool that could be adapted for use in clinical settings to provide additional information about an individual's language ability without requiring any further testing. Current and newly developed language tests could easily provide clinicians with standardized RDI scores, in addition to indices of delay, as part of an automated scoring software program.

It is important to note that the RDI does not represent a correlation between individual item responses and the degree of within-task scatter; rather, the RDI quantifies the distribution of all incorrect responses as a group. Since developmental difference is characterized by an atypical sequence of developmental milestone attainment, it is more informative to assess the overall item-response pattern as opposed to individual item responses within a given task.

### Limitations

Since the NJLAGS project was not originally designed for this type of pilot study, there are a few limitations to using this dataset for our study. First, the NJLAGS sample was ascertained for multiplex families to increase the genetic loading for language impairment in these families. As such, NJLAGS may represent a more severe language impairment cohort and may not be representative of other ASD or language cohorts. Second, as this is a retrospective study of the NJLAGS cohort, our diagnostic groups were not matched for age or IQ. However, our statistical analyses indicate that age and IQ do not have a significant effect on the relationship between ASD traits and RDI. Lastly, the weights utilized by the RDI metric were derived using the unaffected group, not an independent normative sample. Future studies are needed to standardize RDI using an unselected normative sample to define item weights. Ideally, RDI weights would be calculated at the time of development of new or revised language measures using the original standardization sample.

Another potential limitation is that the language tests used with this cohort were administered using standard basal and ceiling rules. While this is standard clinical procedure, we might expect greater variability in RDI if each subtest was administered in its entirety, since individuals exhibiting developmental difference (vs. delay) might be more likely to answer earlier items incorrectly and more difficult items beyond the ceiling correctly. However, the downside to administering the entire subtest is the amount of time it takes, limiting clinical feasibility. Since we found significant differences regardless, it is likely that the RDI is robust enough to provide clinically useful information even while following standard basal and ceiling rules. However, it may be worthwhile for future studies to investigate the impact of administering entire subtests on RDI rather than following basal and ceiling rules.

## Future directions

Standardized tests have been criticized because performance summary scores do not take the individual's pattern of the responses into account. The results of this pilot study suggest that further research is warranted to determine whether developmental difference, as measured by the RDI, is a meaningful indicator of dysfunction that is not captured by standard performance summary scores. If the association between language RDI and ASD is confirmed, it may have implications for screening, diagnosis, and advancing the understanding of language development in ASD. It would be valuable to follow children with LI, both with and without ASD, longitudinally to determine if RDI scores correlate with long-term outcomes and thereby inform prognostication. While this has never been studied specifically, there is some evidence that developmental difference may predict outcomes for children. For example, in the area of speech sound disorders, surface error patterns may reflect different underlying phonological processing deficits, which in turn may lead to divergent language and literacy outcomes. Several longitudinal studies demonstrate that children who initially present with non-developmental speech errors have poorer phonological awareness, decoding, spelling, and reading comprehension scores at follow-up compared to children who presented initially with developmentally delayed, but typical, speech sound error patterns [43, 44]. It is likely that certain types of tests will be more informative than others, and the importance of factors such as age and ability level will require exploration.

Although this study addresses only difference in performance on certain language measures and its role in ASD, the concept may apply to cognitive tests and other measures of development and should be evaluated in other neurodevelopmental disorders and across diagnostic categories. Intrasubtest scatter on IQ tests, for example, has been found to be associated with cognitive inefficiency and variability in attention or arousal, among other things, and there is evidence that it may be useful for detecting cognitive dysfunction among individuals with relatively normal profiles on IQ tests [8, 28]. The RDI may also be useful for identifying subtle but meaningful dysfunction associated with attention-deficit hyperactivity disorder (ADHD), specific learning disorders, and acquired insults such as traumatic brain injury and medical interventions such as chemotherapy, radiation therapy, and early anesthesia exposure.

While it is beyond the scope of the present study, future research is warranted to explore the underlying reasons for the atypical profile of language skills seen in individuals with ASD that is quantified by the RDI. Uneven response patterns may correlate with specific features that are commonly seen in individuals with ASD, such as differences in executive functioning [45], motivation [46], theory of mind [47], and repetitive behaviors [48]. Such features likely impact language acquisition itself as well as the ability to perform under standardized conditions, the latter of which would be influenced by task demands specific to individual language measures (e.g., working memory requirements, complexity of verbal directions). Examining the correlation between quantitative traits in general cognitive processes and RDI in language measures may shed light on which mechanisms hinder or promote efficiency in language acquisition. Similarly, comparing task demands and response patterns may identify certain features of tests themselves that are likely to influence an individual's performance.

## Conclusions

We investigated within-task variability in language skills using a quantitative, continuous metric rather than treating difference as a qualitative, dichotomous trait. This metric, the RDI, was correlated with ASD traits among the entire NJLAGS cohort studied, supporting a continuous model of developmental difference. The RDI can also be adapted to other clinical measures to investigate the degree of difference in various developmental domains; such analyses could help obtain a more comprehensive developmental profile of individuals with ASD and other neurodevelopmental disorders, which in turn might be used to predict long-term outcomes.

## Additional file

**Additional file 1: Figure S1.** Correlation between RDI and person-fit for all subtests. **Figure S2.** Correlation between RDI and SRS for all subtests. (DOCX 25986 kb)

### Abbreviations

ASD: Autism spectrum disorder; ASD−: No diagnosis of autism spectrum disorder; ASD+: Diagnosis of autism spectrum disorder; ASD + LI: Autism spectrum disorder and language impairment; ASD-only: Affected for autism spectrum disorder but not language impairment; CASL: Comprehensive Assessment of Spoken Language; CELF-4: Clinical Evaluation of Language Fundamentals, Fourth Edition; DLD: Developmental language disorder; IQ: Intelligence quotient; LI: Language impairment; LI-only: Affected for language impairment but not autism spectrum disorder; NJLAGS: New Jersey Language and Autism Genetics Study; RDI: Response Dispersion Index; SLI: Specific language impairment; SRS: Social Responsiveness Scale

### Acknowledgements

Preliminary results of this study were presented at the 2016 International Meeting for Autism Research in Baltimore, MD. We are grateful to all of the families that have participated in the NJLAGS study.

### Authors' contributions

AEH, MWM, SMM, and CLM conceived the RDI study, participated in the data analysis and interpretation of data, drafted the manuscript, and revised it for important intellectual content. LMB and JFF conceived the NJLAGS study. AEH, JFF, BG, and LMB participated in the NJLAGS study design, data collection, data cleaning, and analysis. ADM and BRK conducted the data analysis for this study. All authors read and approved the final manuscript.

### Funding

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under grant numbers RC1MH088288 (Brzustowicz), R01MH70366 (Brzustowicz), R01MH074090 (Martin/Ledbetter), and R01MH107431 (Martin). This work was also supported by a grant from the New Jersey Governor's Council for Medical Research and Treatment of Autism under grand number CAUT12APS006 (Brzustowicz). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

### Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the NIMH Data Archive repository, https://ndar.nih.gov/edit_collection.html?id=1932.

### Ethics approval and consent to participate

The study was approved by the Institutional Review Boards at Rutgers University (IRB #:13-112Mc), Geisinger Health System (IRB #: 2015-0231), and Bloomsburg University (IRB #: 2017-86-COST).

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Department of Biological and Allied Health Sciences, Hartline Science Center, Bloomsburg University, 400 East Second St, Bloomsburg, PA 17815, USA. [2]Autism & Developmental Medicine Institute, Geisinger Health System, 120 Hamm Drive, Suite 2A, Lewisburg, PA 17837, USA. [3]Psychology Department, O'Leary Center, Bucknell University, Lewisburg, PA 17837, USA. [4]Computer Science Department, Breakiron Building, Bucknell University, Lewisburg, PA 17837, USA. [5]Genetics Department, Life Sciences Building, Rutgers University, 145 Bevier Road, Piscataway, NJ 08854, USA.

Received: 12 February 2019 Accepted: 28 August 2019
Published online: 13 September 2019

### References

1. Gesell A, Halverson H, Amatruda C. The first five years of life: a guide to the study of the preschool child. New York: Harper; 1940.
2. Gesell A, Amatruda C. Developmental diagnosis: normal and abnormal child development, clinical methods and pediatric applications. 2nd ed. New York: Paul B. Hoeber; 1947.
3. Lourenco O, Machado A. In defense of Piaget's theory: a reply to 10 common criticisms. Psychol Rev. 1996;103:143–64.
4. Capute A, Palmer F. A pediatric overview of the spectrum of developmental disabilities. J Dev Behav Pediatr. 1980;1:66–9.
5. Accardo P, Accardo J, Capute A. A neurodevelopmental perspective on the continuum of developmental disabilities. In: Accardo P, editor. Capute & Accardo's neurodevelopmental disabilities in infancy and childhood: vol I neurodevelopmental diagnosis and treatment. 3rd ed. Baltimore: Brookes; 2008. p. 3–25.
6. Voigt RG, Barbaresi WJ, Colligan RC, Weaver AL, Katusic SK. Developmental dissociation, deviance, and delay: occurrence of attention-deficit-hyperactivity disorder in individuals with and without borderline-to-mild intellectual disability. Dev Med Child Neurol. 2006;48(10):831–5.
7. Visser L, Vlaskamp C, Emde C, Ruiter S, Timmerman M. Difference or delay? A comparison of Bayley-III cognition item scores of young children with and without developmental disabilities. Res Dev Disabil. 2017;71:109–19.
8. Godber T, Anderson V, Bell R. The measurement and diagnostic utility of intrasubtest scatter in pediatric neuropsychology. J Clin Psychol. 2000;56(1):101–12.
9. Hallenbeck C, Fink S, Grossman J. Measurement of intellectual inefficiency. Psychol Rep. 1965;17(2):339–49.
10. Fink S, Shontz F. Inference of intellectual inefficiency from the WAIS vocabulary subtest. J Clin Psychol. 1958;14:409–12.
11. VanMeter L, Fein D, Morris R, Waterhouse L, Allen D. Delay versus deviance in autistic social behavior. J Autism Dev Disord. 1997;27(5):557–69.
12. Lord C. Language in high-functioning adolescents with autism: questions about deviance and delay. In: Cicchetti D, Toth SL, editors. Rochester symposium on developmental psychopathology, vol 7: adolescence: opportunities and challenges. Rochester: University of Rochester Press; 1996. p. 149–65.
13. Eigsti I, Bennetto L, Dadlani M. Beyond pragmatics:morphosyntactic development in autism. J Autism Dev Disord. 2007;37:1007–23.
14. Boucher J. Research review: structural language in autistic spectrum disorder - characteristics and causes. J Child Psychol Psychiatry. 2012;53(3):219–33 [cited 2012 Mar 13] Available from: http://www.ncbi.nlm.nih.gov/pubmed/22188468.
15. Williams D, Botting N, Boucher J. Language in autism and specific language impairment: where are the links? Psychol Bull. 2008;134:944–63.
16. Whitehouse AJO, Barry JG, Bishop DVM. Further defining the language impairment of autism: is there a specific language impairment subtype? J Commun Disord. 2008;41(4):319–36.

17. Bishop DVM. Why is it so hard to reach agreement on terminology? The case of developmental language disorder (DLD). Int J Lang Commun Disord. 2017;52(6):671–80.
18. Tomblin J, Records N, Buckwalter P, Zhang X, Smith E, O'Brien M. Prevalence of specific language impairment in kindergarten children. J Speech Hear Res. 1997;40:1245–60.
19. Bishop D V, Maybery M, Wong D, Maley A, Hallmayer J. Characteristics of the broader phenotype in autism: a study of siblings using the children's communication checklist-2. Am J Med Genet B Neuropsychiatr Genet [Internet]. 2006/01/04. 2006;141B(2):117–22. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16389586
20. Leonard LB. Children with specific language impairment. 2nd ed. Cambridge: The MIT Press; 2014.
21. Bishop DVM, Frazier NC. Exploring the borderlands of autistic disorder and specific language impairment: a study using standardized diagnostic instruments. J Child Psychol Psychiatry Allied Discip. 2002;43(7):917–29.
22. Leyfer OT, Tager-Flusberg H, Dowd M, Tomblin JB, Folstein SE. Overlap between autism and specific language impairment: comparison of Autism Diagnostic Interview and Autism Diagnostic Observation Schedule scores. Autism Res. 2008;1(5):284–96 [cited 2013 Apr 14] Available from: http://www.ncbi.nlm.nih.gov/pubmed/19360680.
23. Kjelgaard MM, Tager-Flusberg H. An investigation of language impairment in autism: implications for genetic subgroups. Lang Cogn Process. 2001; 16(2/3):287–308.
24. Lewis B, Freebaim L, Hansen A, Miscimarra L, Iyengar S, Taylor H. Speech and language skills of parents of children with speech sound disorders. Am J Speech Lang Pathol. 2007;16(2):108–18.
25. Rapin I, Dunn M, Allen D, Stevens M, Fein D. Subtypes of language disorders in school-age children with autism. Dev Neuropsychol. 2009;34: 66–84.
26. Tager-Flusberg H. Defining language phenotypes in autism. Clin Neurosci Res. 2006;6(3–4):219–24.
27. Eigsti I, de Marchena A, Schuh J, Kelley E. Language acquisition in autism spectrum disorders: a developmental review. Res Autism Spectr Disord. 2011;2(5):681–91.
28. Mittenberg W, Hammeke T, Rao S. Intrasubtest scatter on the WAIS-R as a pathognomonic sign of brain injury. Psychol Assess. 1989;1(4):273–6.
29. Constantino J, Davis S, Todd R, Schindler M, Gross M, Brophy S, et al. Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. J Autism Dev Disord. 2003;33(4):427–33.
30. Bartlett CW, Hou L, Flax JF, Fermano Z, Hare A, Cheong SY, et al. A genome-scan for shared autism spectrum disorder and specific language impairment loci. In: American Society of Human Genetics. Toronto, Canada; 2011. p. Poster.
31. Flax J, Hare A, Fermano Z, Zimmerman-Bier B, Cartwright C, Bartlett C, et al. Characterizing language profiles in families ascertained for autism and specific language impairment. 2013.
32. Bartlett CW, Flax JF, Fermano Z, Hare A, Hou L, Petrill SA, et al. Gene × gene interaction in shared etiology of autism and specific language impairment. Biol Psychiatry. 2012;72(8):692–9 [cited 2013 Jun 3] Available from: http://www.ncbi.nlm.nih.gov/pubmed/22704665.
33. Semel E, Wiig E, Secord W. Clinical evaluation of language fundamentals, fourth edition (CELF-4). Toronto: The Psychological Corporation/A Harcourt Assessment Company; 2003.
34. Wiig E, Secord W, Semel E. Clinical evaluation of language fundamentals -- preschool, second edition (CELF Preschool-2). Toronto: The Psychological Corporation/A Harcourt Assessment Company; 2004.
35. Carrow-Woolfolk E. Comprehensive assessment of spoken language. MN: AGS: Circle Pines; 1999.
36. Rutter M, Le Couteur A, Lord C. The autism diagnostic interview, revised (ADI-R). Western Psychological Services: Los Angeles; 2003.
37. Gotham K, Risi S, Pickles A, Lord C. The autism diagnostic obersvation schedule: revised algorithms for improved diagnostic validity. J Autism Dev Disord. 2007;37:613–27.
38. Lord F. Small N justifies Rasch Model. In: New horizons in testing: latent trait theory and computerized adaptive testing. New York: Wiley; 1983. p. 51–61.
39. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press; 1980.
40. Revelle W. Psych: procedures for personality and psychological research. Evanston, Illinois: Northwestern University; 2016.
41. Landa RJ, Goldberg MC. Language, social, and executive functions in high functioning autism: a continuum of performance. J Autism Dev Disord. 2005;35(5):557–73.
42. Park C, Yelland G, Taffe J, Gray K. Morphological and syntactic skills in language samples of pre school aged children with autism: atypical development? Int J Speech Lang Pathol. 2012;12(2):95–108.
43. Holm A, Farrier F, Dodd B. Phonological awareness, reading accuracy and spelling ability of children with inconsistent phonological disorder. Int J Lang Commun Disord. 2008;43(3):300–22.
44. Leitão S, Fletcher J. Literacy outcomes for students with speech impairment: long-term follow-up. Int J Lang Commun Disord. 2004;39(2):245–56.
45. Corbett B, Constantine L, Hendren R, Rocke D, Ozonoff S. Examining executive functioning in children with autism spectrum disorder, attention deficit hyperactivity disorder and typical development. Psychiatry Res. 2009; 166(2–3):210–22.
46. Chevallier C, Kohls G, Troiani V, Brodkin E, Schultz R. The social motivation theory of autism. Trends Cogn Sci. 2012;16(4):231–9.
47. Loukusa S, Makinen L, Kuusikko-Gauffin S, Ebeling H, Moilanen I. Theory of mind and emotion recognition skills in children with specific language impairment, autism spectrum disorder and typical development: group differences and connection to knowledge of grammatical morphology, word-finding abilities and verbal worki. Int J Lang Commun Disord. 2014; 49(4):498–507.
48. Paul R, Chawarska K, Cicchetti D, Volkmar F. Language outcomes of toddlers with autism spectrum disorders: a two year follow-up. Autism Res. 2008;1(2): 97–107.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.