Received: 28 September 2022 | Accepted: 19 January 2023

DOI: 10.1002/aur.2897

RESEARCH ARTICLE

# Consistency and reliability of automated language measures across expressive language samples in autism

**Heather MacFarlane**[1] | **Alexandra C. Salem**[1] | **Steven Bedrick**[2] | **Jill K. Dolata**[3,4] | **Jack Wiedrick**[5] | **Grace O. Lawley**[6] | **Lizbeth H. Finestack**[7] | **Sara T. Kover**[8] | **Angela John Thurman**[9,10] | **Leonard Abbeduto**[9,10] | **Eric Fombonne**[1]

[1]Department of Psychiatry, Oregon Health & Science University, Portland, Oregon, USA

[2]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA

[3]Department of Pediatrics, Oregon Health & Science University, Portland, Oregon, USA

[4]School of Communication Sciences and Disorders, Pacific University, Forest Grove, Oregon, USA

[5]Biostatistics & Design Program, Oregon Health & Science University, Portland, Oregon, USA

[6]Computer Science and Electrical Engineering, Oregon Health & Science University, Portland, Oregon, USA

[7]Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, Minnesota, USA

[8]Department of Speech and Hearing Sciences, University of Washington, Seattle, Washington, USA

[9]MIND Institute, University of California Davis Health, Sacramento, California, USA

[10]Department of Psychiatry and Behavioral Sciences, University of California Davis Health, Sacramento, California, USA

**Correspondence**
Heather MacFarlane, Department of Psychiatry, Oregon Health & Science University, Portland, OR, USA.
Email: macfarlh@ohsu.edu

**Funding information**
National Institutes of Health, Grant/Award Numbers: UL1TR001860, P50HD103526, R01HD074346, R01DC012033; Simons Foundation, Grant/Award Number: SFARI 383668

## Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with substantial clinical heterogeneity, especially in language and communication ability. There is a need for validated language outcome measures that show sensitivity to true change for this population. We used Natural Language Processing to analyze expressive language transcripts of 64 highly-verbal children and young adults (age: 6–23 years, mean 12.8 years; 78.1% male) with ASD to examine the validity across language sampling context and test-retest reliability of six previously validated Automated Language Measures (ALMs), including Mean Length of Utterance in Morphemes, Number of Distinct Word Roots, C-units per minute, unintelligible proportion, um rate, and repetition proportion. Three expressive language samples were collected at baseline and again 4 weeks later. These samples comprised interview tasks from the Autism Diagnostic Observation Schedule (ADOS-2) Modules 3 and 4, a conversation task, and a narration task. The influence of language sampling context on each ALM was estimated using either generalized linear mixed-effects models or generalized linear models, adjusted for age, sex, and IQ. The 4 weeks test-retest reliability was evaluated using Lin's Concordance Correlation Coefficient (CCC). The three different sampling contexts were associated with significantly ($P < 0.001$) different distributions for each ALM. With one exception (repetition proportion), ALMs also showed good test-retest reliability (median CCC: 0.73–0.88) when measured within the same context. Taken in conjunction with our previous work establishing their construct validity, this study demonstrates further critical psychometric properties of ALMs and their promising potential as language outcome measures for ASD research.

## Lay Summary

Autistic individuals often demonstrate communication differences that traditional clinical measures and language tests cannot fully capture. Using language transcripts from 64 children and young adults, we establish the performance consistency across 4 weeks of six automated language outcome measures and discuss the language sampling context's effect on such measures. This methodology could provide a rigorous, objective, and accessible way to evaluate individual language profiles and measure their change over time.

**KEYWORDS**
autism, automated measures, communication, expressive language, natural language processing

Heather MacFarlane and Alexandra C. Salem contributed equally to this study.

## INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by differences in social communication and interaction, paired with the presence of restricted and repetitive patterns in behaviors, interests, and activities (American Psychiatric Association, 2013). Although many studies have examined communication in autism, language patterns in ASD are remarkably variable and can be hard to quantify (Meir & Novogrodsky, 2020).

Standardized clinical measures have been created to assess a specific target population with regard to age, dialect, and language level. Some commonly used examples include: the Clinical Evaluation of Language Fundamentals, Preschool- 3rd edition (children aged three to 6 years; Wiig et al., 2020), the Preschool-Language Scale, 5th edition (children birth to age eight; Lyons, 2021), and the Clinical Evaluation of Language Fundamentals, 5th edition (a version for five- to eight-year-olds, with another version for children and adolescents aged nine to 21; Wiig et al., 2013). All of these were created for monolingual speakers of Generally American English and lack similarity to real-world interactions; therefore, standardized language assessment instruments administered in clinical practice lack ecological validity for assessing natural conversation behavior (Costanza-Smith, 2010). Such measures often provide a single aggregate score, which reduces an individual's language ability to one number and can obscure floor effects or differences in communication patterns by subsuming a whole language profile into one overall "outcome" (Hilvert et al., 2020). Indeed, while language sampling is a recommended best practice in clinical diagnostic evaluations, many clinicians report barriers to its clinical use (Pavelko et al., 2016). Many authors have thus called for using expressive language samples for analyzing language use and measuring outcomes in developmentally diverse groups (Barokova & Tager-Flusberg, 2020; Costanza-Smith, 2010; Tager-Flusberg et al., 2009).

Analysis of expressive language samples can more accurately describe a child's language ability and conversational skills by providing information about specific domains of strength and weakness and with greater correspondence to performance in real-world contexts. This approach may be particularly helpful in analysis of autistic communication styles, in which there is sometimes a parent-reported qualitative difference in communication despite typical performance related to grammatical development (Dolata et al., 2022; Volden & Phillips, 2010). Standardized measures do not typically assess a person's use of echolalia, scripted language, repetitive speech, neologisms, or pragmatic conversational difficulties, which can be a component of autistic linguistic output; these characteristics would be easier to observe in a natural language sample. Expressive language sampling (ELS) has been used in studies to develop outcome measures (Abbeduto et al., 2020; Berry-Kravis et al., 2013; Thurman et al., 2021) and to examine the effects of context and sample length on language measures (Heilmann et al., 2010; Kover et al., 2012).

Tager-Flusberg et al. (2009) first called for a set of expressive language outcome measures to evaluate the efficacy of interventions. They wanted to address unresolved ambiguities in this field by encouraging a standardized approach using a common set of measures to allow for the comparison of findings across studies. Barokova and Tager-Flusberg (2020) reasserted this need, calling for outcome measures that can be generated from natural language samples: measures that are easy to obtain, psychometrically sound, and sensitive to change. Abbeduto et al. (2020) also argued for the need for standardization of the interactions that are used as the bases for the collection of expressive language samples. Outcome measures—such as the automated language measures (ALMs) discussed below—derived from expressive language samples and collected under standardized conditions have been shown to successfully differentiate diagnostic groups, including fragile X syndrome (FXS) and Down syndrome (DS) (Abbeduto et al., 2020; Berry-Kravis et al., 2013; Shaffer et al., 2020; Thurman et al., 2021).

Automated language measures (ALMs) are measures of expressive language that are automatically calculated on transcribed speech samples using Natural Language Processing (NLP) methodology—a branch of computer science that integrates computational linguistics with machine learning to understand human language. Computational methods bring many potential advantages to the analysis of language in scientific and clinical contexts. Two particularly notable advantages are (1) efficiency (in that they facilitate the automated analysis of language samples that would be prohibitively large for an unassisted human to analyze) and (2) reliability (a given algorithm's analysis will be consistent across time, to a degree that is challenging to achieve when relying solely on human annotations and observations) (Ratner & MacWhinney, 2016). Additionally, and crucially, ALMs can also be built to quantify aspects of language that are difficult to operationalize into practically-usable measures, such as echolalia, talkativeness, relative use of um and uh, intelligibility, and diversity of vocabulary.

Unlike most standardized clinical measures, expressive language measures and ALMs can be used across a wide range of ages and language levels, and can be derived from samples of different lengths collected from in-person clinical contexts (Channell et al., 2018; Heilmann et al., 2010; Kover et al., 2012; Tager-Flusberg et al., 2009), as well as remotely collected from telehealth visits and video call interactions (Butler et al., 2022). Some authors have suggested that short language samples provide reliable data for these analyses (Heilmann et al., 2010), and this has been confirmed in recent work (Pavelko et al., 2020; Wilder & Redmond, 2022). In our

prior work, we established the discriminant and convergent validity of ALMs by comparing typically developing and ASD groups on seven different ALMs, thus establishing their construct validity (Lawley et al., 2022; Salem et al., 2021). However, it remains to be determined whether these ALMs additionally have cross-context consistency, or the short-term reliability required of outcome measures.

When using ELS as the basis for outcome measures, the sampling context must be taken into account. Kover et al. (2012) found a differential effect of context for measures of Mean Length of Utterance (MLU), fluency, and attempted speech among children with FXS, DS, and typical development (TD): all participants, regardless of diagnosis, talked more during a conversation task than a narration task. This echoes the earlier finding by Kover and Abbeduto (2010) in their study of adolescents with FXS plus ASD, FXS only, and DS. Several studies have found that MLU tends to be lower in conversation tasks than narration tasks among neurodiverse adolescents and young adults (Abbeduto et al., 1995; Kover & Abbeduto, 2010; Miles et al., 2006). Thus, sampling context is an important variable in expressive language measurements.

The Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000), a commonly-administered standardized measure for autism diagnostic evaluation, has also been used as a context for language sample analysis because many of the assessment's probes are designed to motivate expressive language from verbal participants (Kover et al., 2014; Suh et al., 2014; Tager-Flusberg et al., 2009). However, only three prior studies have examined cross-context language use between the ADOS and another ELS context. Martin et al. (2012) analyzed perseveration in boys with FXS (with and without co-occurring ASD), DS, and TD across a narrative task and the social interaction activities of the ADOS. Kover et al. (2014) analyzed spontaneous expressive language for ASD participants across different play contexts, including the ADOS play activities. Hilvert et al. (2020) analyzed a predefined set of language measures in boys with ASD (with and without co-occurring FXS) across a semi-structured conversation and several activities from Modules 2 and 3 of the ADOS. Kover et al. (2014) and Hilvert et al. (2020) both reported that all participants were less talkative in the ADOS versus the comparison language sampling context, pointing to the importance of accurately describing the sampling context in such studies.

An important step for establishing the validity and reliability of these language outcome measures is to test their replicability over a short test-retest time frame with the assumption that significant development is unlikely to occur over a period of a few weeks, especially for neurodivergent individuals. Abbeduto et al. (2020) and Thurman et al. (2021) collected narrative and conversation language samples from individuals 6 to 23 years of age with FXS or DS twice, 4 weeks apart, and found no practice effects and strong test-retest reliability in both sampling contexts for expressive language measures of talkativeness, vocabulary, syntax, utterance planning, and articulation quality although the psychometrics were somewhat stronger for older and more developmentally advanced participants. However, no equivalent test-retest studies have been conducted with an autistic sample.

The goal of this exploratory study is to establish and compare the short-term test-retest reliability and consistency of six ALMs across different time points and sampling contexts for children, adolescents, and young adults on the autism spectrum. This takes us further toward our end goal of validating meaningful language outcome measures for autistic individuals by establishing the psychometric reliability of these measures. We have two specific aims: first, to evaluate the consistency of a set of valid ALMs across sampling contexts and methodologies; second, to evaluate the test-retest reliability of these ALMs within a repeated sampling context over short periods of time.

# METHODS

## Participants

The sample for the current study was drawn from a larger sample of native English-speaking individuals with ASD, FXS, and DS, aged between six and 23 years, who were recruited as part of a multi-site study evaluating the utility of expressive language sampling (ELS) as a source of outcome measures (Abbeduto et al., 2020; Hoffmann et al., 2022; Thurman et al., 2021). A total of 13 participants with ASD were reported by their caregivers to speak a language in addition to English; but only one participant was described as fluent in their other language. Data was collected in three separate waves for all groups: time point 1 (T1; baseline), time point 2 (T2; occurring approximately 4 weeks after T1), and time point 3 (T3; occurring approximately 1 year after T1). Only participants with ASD were included in this study. Aim 1 of our work utilizes language samples from T1 only while Aim 2 utilizes language samples from both T1 and T2. Data was collected at three participating sites: University of California, Davis; University of Minnesota, Twin Cities, and University of Washington.

Out of 81 autistic participants enrolled in the study, 17 were excluded: one who did not meet ASD criteria on the ADOS-2 and subsequently withdrew from the study, 13 who used single words or phrase speech and therefore participated in Modules 1 or 2 of the ADOS-2, and three who did not complete all three language sampling tasks. The 64 participants who had valid data on the three instruments used to sample expressive language (see below) formed the baseline (T1) sample. All participants provided documentation of a clinical diagnosis of ASD

upon entering the study and exceeded the ASD classification threshold on the ADOS-2 at either T1 or T2. The baseline sample included 50 males (78.1%). Nine participants identified as Hispanic or Latino (14%). A total of 55 participants received Module 3 of the ADOS-2 and nine participants received Module 4. Sample characteristics are shown in Table 1.

To generate the test-retest reliability data, all participants were assessed twice: at baseline (T1), and again about 4 weeks later (mean interval: 28.4 days (*SD*: 4.6)). Four participants did not have usable data at T2 (one participant withdrew from the study after T1, one participant canceled their T2 visit due to a medication change, and two participants had missing data for the ELS tasks), resulting in a slightly smaller sample (N = 60) for the T1-T2 test-retest analyses. The four participants lost at T2 did not differ in any statistically meaningful way on any background characteristics from those included at T2.

This study was approved by the Institutional Review Board at each participating university site. Informed written consent was obtained from the parent or legal guardian or the adult youth (when appropriate) prior to participation. Assent was obtained from each participant (when appropriate). The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and international committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

**TABLE 1** Sample characteristics.

| Measure | Mean (*SD*) | Range |
|---|---|---|
| Age (in years) | 12.8 (3.9) | 6.2–23.4 |
| FSIQ | 94.4 (22.2) | 40–132 |
| NVIQ | 93.6 (23.0) | 22–131 |
| VIQ | 92.5 (26.0) | 18–138 |
| Vineland ABC | 78.7 (18.1) | 7–128 |
| Vineland Communication | 85.2 (19.0) | 35–125 |
| Vineland Socialization | 76.0 (16.2) | 24–124 |
| Vineland Daily Living Skills | 85.2 (19.3) | 35–132 |
| ADOS CSS | 6.7 (2.1) | 2–10 |
| **Race** | | ***N*** |
| Asian | | 2 |
| Native Hawaiian/Pacific Islander | | 1 |
| White | | 46 |
| Other | | 4 |
| More than one race | | 11 |
| *Yearly income* | | |
| <$50,000 | | 4 |
| $50,000–$100,000 | | 22 |
| $100,000–$150,000 | | 17 |
| >$150,000 | | 20 |

*Note*: Table reports data collected at baseline (T1). One participant did not meet ASD criteria at T1; however, they did meet ASD criteria at retest (as sometimes happens; Janvier et al., 2022) and thus were included in the study. All multiracial participants included "White" as one of their races. One participant was missing yearly income data.
Abbreviations: ABC, adaptive behavior composite; ADOS CSS, autism diagnostic observation schedule calibrated severity score; FSIQ, full scale IQ; NVIQ, nonverbal IQ; SD, standard deviation; VIQ, verbal IQ.

## Data

### Instruments

*ASD symptom severity*

We administered the Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2) according to the standard procedures (Lord et al., 2012). All administrations were recorded for transcription. The ADOS-2 comprises a series of activities that provide the opportunity to observe behaviors reflecting the core characteristics of ASD. The ADOS-2 has five modules, each designed for individuals with different developmental and/or language levels. The module administered to any given participant was selected according to the ADOS-2 manual guidelines. The ADOS-2 was administered by a research-reliable examiner, who scored the participant's behavior in real time. The Calibrated Severity Score (CSS) was used to estimate severity of ASD symptoms.

*Cognitive ability*

At T1, participants' cognitive ability was assessed using the Stanford-Binet Intelligence Scales, Fifth Edition (SB-5) (Roid, 2003). This test yields full-scale IQ (FSIQ), nonverbal IQ (NVIQ), and verbal IQ (VIQ) scores. Mean IQ scores are 100 in the normalizing sample, with a standard deviation of 15.

*Adaptive behavior*

Parents or caregivers completed the Vineland Adaptive Behavior Scales, Second Edition (VABS-II) (Sparrow et al., 2012). The VABS-II was normed for individuals aged three to 21 years. We used the standardized scores (mean: 100; *SD*: 15) of the Adaptive Behavior Composite Score.

### Expressive language samples

At each time point, three expressive language samples were collected in three different contexts—the ADOS-2, a conversation (CON) task, and a narration (NAR) task—with the order of administration randomized across participants. These particular CON and NAR procedures have been used in previous studies across a range of ages and abilities (Abbeduto et al., 1995; Berry-Kravis et al., 2013; Channell et al., 2018; Finestack & Abbeduto, 2010; Finestack et al., 2013; Hoffmann et al., 2022; Kover & Abbeduto, 2010; Kover et al., 2012;

Murphy & Abbeduto, 2007). The task procedures were designed to be naturalistic while ensuring reasonable standardization of materials, content of the talk, and examiner behavior. Although the 2nd edition of the ADOS (ADOS-2) was administered in this study, for brevity, this context is referred to as the "ADOS" throughout the text.

*Autism diagnostic observation schedule (ADOS)*

The ADOS context combined three activities from the ADOS administration: the Emotions, Social Difficulties and Annoyance, and Friends, Relationships, and Marriage interviews. These structured segments focus on the participant's way of describing their emotions, their perceptions of social difficulties, their understanding of the nature of personal and social relationships, why someone might want to engage in such relationships, and what the participant's role might be in those relationships. We selected these three activities of the ADOS for several reasons. First, following the measure's manualized procedures, these activities occurred after other brief ADOS activities, allowing the clinician to develop rapport with the participant before engaging in conversation. This set of activities also has high standardization of examiner questions across administrations, leading to good comparability between participants. They are also some of the few ADOS tasks without physical prompts, such as toys or books, and involve a purely linguistic exchange. The examiner uses scripted interview questions that are open-ended and designed to facilitate participant response. Follow-up probes are used at the examiner's discretion to ensure sufficient responses are obtained. Although not equivalent to a naturalistic conversation between family or peers, these tasks function as a guided conversation that can be collected in a lab setting and thus provide a stable sample for analysis. All three interviews occur in both Module 3 and Module 4 administrations and, when combined, have a comparable length in minutes to the conversation and narration tasks included in the study. The average length of the combined ADOS tasks was 10.6 minutes (*SD*: 3.3 min).

*Conversation (CON)*

In the conversation task, the examiner engaged the participant in talk on a variety of predetermined topics (e.g., school, family, hobbies) according to guidelines that specify the order of topics and the ways in which topics are introduced and maintained. The conversation begins with a topic that the parent or guardian has previously indicated is one that the participant would enjoy sharing, thereby ensuring maximum comfort with the interaction and avoiding any topics that could lead to frustration. The remaining "standard" topics are personally relevant and familiar and include topics such as friends, families, pets, school, and work. To ensure age-appropriateness, slightly different sets of topics are used for children and adolescents relative to adults (e.g., school is a useful topic for the former, but not the latter). The procedures are otherwise identical for participants of different ages. In general, the script that the examiner follows minimizes their own participation, maximizes the participant's contribution, and avoids frequent use of examiner language that would constrain the amount or complexity of participant talk (e.g., yes-no questions). The conversation is ideally brought to a close by the examiner after 12 min, although for consistency in length of samples only the first 10 min are transcribed as there was variability in conversation length across participants (e.g., examiners did not abruptly halt a conversation if the participant was in the middle of discussing a topic). The average length of the Conversation task was 10.0 min (*SD*: 0.6 min).

Two sets of topics (versions A and B) were created for children and adolescents and two for adults, which made it possible to present alternate versions in test and retest administrations for any given participant. Assignment of version for T1 and T2 was randomly determined across participants. A participant who received version A at T1 received version B at T2 and vice versa. Further details about the conversation task can be found in (Abbeduto et al., 2020).

*Narration (NAR)*

In the narration task, the participant tells the story in a wordless picture book. Examiner prompts and responses are scripted. The procedure begins with the examiner asking the participant to look at the book to get a sense of the story, but without talking about it. The examiner controls the turning of the pages so that the participant reviews each pair of pages for eight to 10 s. The participant then tells the story page by page, with page turning controlled by the examiner, with five to 7 s spent per page. As in the conversation task, the examiner follows a script that minimizes their own participation, maximizes the participant's contribution, and avoids examiner language that would constrain the participant's talk. Prompts are largely limited to the first page, thus the examiner provides minimal scaffolding. The administration is untimed but typically takes 10–15 min to administer and yields narratives of 3–8 min in length for TD children (Kover et al., 2012).

We used two books from Mercer Mayer's "frog" series: "Frog Goes to Dinner" and "Frog, Where Are You?". The books depict events that can be described at different levels of detail and abstraction, from the physical acts of story characters to their intentions and emotional reactions, as well as offering the potential for description of anticipated events. The validity of this narration procedure across a range of age and developmental levels was described previously in Abbeduto et al. (2020). We found previously that these two books yield expressive language samples that do not differ on the dependent variables of interest for individuals with FXS (Kover et al., 2012), making it possible to present

alternate versions in test and retest administrations for any given participant. Each of the two books include 16 page spreads. The scripts used for the two books are identical. Assignment of version to T1 and T2 was randomly determined across participants. The average length of the narration task was 7.0 min (*SD*: 2.3 min). Further details about the narration task can be found in (Abbeduto et al., 2020).

Conversation and NAR were administered by examiners trained to predetermined levels of administration fidelity (90% or higher), as described previously in Abbeduto et al. (2020). After training, fidelity was assessed on 16 randomly selected administrations of CON and NAR with ASD participants, stratified across administration sites. The mean fidelity score was 98% (range 89%–100%). Of the samples reviewed for fidelity, only one NAR (89%) fell slightly below the 90% threshold established a priori for fidelity. Manuals for CON and NAR are available at https://ctscassist.ucdmc.ucdavis.edu/ctscassist/surveys/?s=W9W99JLMNX. Included are procedures for administration, training, and assessment of fidelity.

## Transcription

All three types of expressive language samples were audio-recorded using digital recorders. These samples were transcribed by highly trained assistants following transcription procedures developed previously, which have been shown to yield adequate levels of inter-transcriber reliability (Abbeduto et al., 1995; Channell et al., 2018; Kover et al., 2012). The transcription process involved a first draft by a primary transcriber, feedback by a secondary transcriber, and final editing by the primary transcriber, as described in Abbeduto et al. (2020) and Thurman et al. (2021). Transcription was guided by Systematic Analysis of Language Transcripts (SALT; Miller et al., 2015). SALT is a computer program that allows standard and user-defined analyses of transcripts prepared as text files according to well-established conventions in child language research. Additional conventions have been added over the years based on the unique characteristics of our study participants and the contexts in which we sample their language (e.g., for segmentation of utterances, compound words, and proper nouns).

In preparing the transcripts, talk was segmented into Communication-units (C-units). A C-unit is defined as an independent clause with associated modifiers, including dependent clauses (Loban, 1976), though in practice, non-clausal utterances such as sentence fragments and elliptical responses also constitute C- units (Miller et al., 2015). The C-unit provides a more accurate measure of language ability than does segmentation into full utterances for speakers beyond a developmental level of 3 years (Abbeduto et al., 1995). Unintelligible speech is marked by "XX", as in, "I went to the store. And I bought XX." Transcribers were required to achieve agreement with a gold standard transcription, with different a priori levels established for different dimensions of the transcription process (e.g., segmentation to C-units, number of morphemes); agreement was required to be at least 70%–80% depending on the particular aspect of agreement with the gold standard (e.g., segmentation into C-units, presence of a disfluency). Transcribers were blind to diagnosis, which time point the sample was from, and results of other measures completed by the participant. Each of the three participating sites transcribed the samples they collected. Inter-transcriber agreement, across 18 samples (four ADOS-2 samples, seven CON samples, and six NAR samples), was observed to be 82% for utterance segmentation, 96% for identification of partly or fully unintelligible C-units, 94% for identification of C-units containing mazes, 83% for identification of the exact number of morphemes in each C-unit, and 86% for the exact number of words in each C-unit.

The computational workflow for the present study is designed to operate on transcripts that have been prepared with a minimal amount of additional manual annotation. Specifically, it assumes that manual SALT annotation has not been performed. Therefore, to stay consistent with our previous processing pipeline which used unannotated transcripts (MacFarlane et al., 2022; Salem et al., 2021), a data preparation step was performed in which the manual annotations were removed and then replaced with automatically-produced annotations using AutoSALT, a software tool previously developed by this group (Gorman et al., 2015), thus experimentally simulating a scenario in which raw, unannotated transcripts were used. AutoSALT analyses unannotated transcripts and automatically performs a useful subset of the SALT morphological annotation tasks (in particular, identification of morpheme and suffix clusters for complex words). In previous work, AutoSALT was found to perform this task with a very high degree of accuracy (98.9%, evaluated at the token level; Gorman et al., 2015) and the resulting calculations of SALT-derived metrics (e.g., Mean Length of Utterance in Morphemes) produced nearly identical results to those computed using manually annotated transcripts. The only post-hoc change needed for this analysis was the addition of activity labels to the ADOS transcripts, as they were not included in the original transcripts, but are necessary for isolating the three combined ADOS tasks using the AutoSALT software. These annotations were added by trained research staff (from the Oregon Health & Science University team), who had over 90% labeling agreement.

## Automated language measures

A total of six outcome measures were generated from the transcripts. Of the six, five ALMs were generated as described in Salem et al. (2021): Mean Length of

Utterance in Morphemes (MLUM; calculated on all complete, fluent, and intelligible C-units), Number of Distinct Word Roots (NDWR; counted on all complete, fluent, and intelligible C-units), unintelligible proportion (number of partially or fully unintelligible C-units divided by the total number of C-units), C-units per minute (CPM; number of attempted communication units per minute), and repetition proportion (number of child words that are repeated in a set of two or more from the examiner's immediately preceding turn, divided by the total number of child words). A fluent utterance is one that does not contain any disfluencies, such as false starts, repetitions, fillers, and stutters. Finally, we calculated um rate (total number of "ums" divided by the total number of intelligible words) as described by Lawley et al. (2022). Lawley et al. (2022) calculated um rate and um ratio (number of "ums" divided by number of "ums" and "uhs"). Of the fillers a participant said, a higher um ratio indicates that they said more "ums" than "uhs" while a lower um ratio indicates that they said more "uhs" than "ums". They found that while both um rate and um ratio significantly differentiated children with ASD from typically developing children, um ratio had a significant effect of sex with boys showing a lower um ratio than girls. Therefore, we only included um rate. Previously, in comparisons of participants with and without ASD, we confirmed the discriminant validity of the six ALMs presented here (Lawley et al., 2022; Salem et al., 2021). For this new language sample, because time was marked with whole minute markers rather than exact time alignments, CPM was calculated just within the minutes marked in an activity. We verified the discriminant validity of this "trimmed" CPM; for a detailed description, see Supplemental Information, Methods section.

## Statistical analyses

To evaluate consistency across sampling contexts, we first calculated means, standard deviations, and ranges of each ALM for each of the three contexts (ADOS, CON, and NAR) at T1. We estimated a series of generalized linear mixed-effects models (GLMMs) or generalized linear models (GLMs) for each ALM (Dobson & Barnett, 2018; Nelder & Wedderburn, 1972) using maximum-likelihood estimation. We based the GLMMs on previous work by Lawley et al. (2022), Sonderegger et al. (2018), and Gorman et al. (2016). MLUM, NDWR, and CPM were estimated using GLM because the systematic between-individual variance across sampling contexts and time points was so low in magnitude that participant random effects were unidentifiable. We applied cluster-robust standard errors to each of the GLM models (Huber, 1967; White, 1982), fitting the GLM under the assumption that the three context-bound measurements on each individual at the two time points were actually independent but then correcting for the effects of this assumption by adjusting the values of the standard errors to account for non-independence (Liang & Zeger, 1986); the confidence intervals reported here are now properly adjusted for that deviation. The three proportion ALMs—um rate, unintelligible proportion, and repetition proportion—were estimated using GLMM, the preferred model for non-independent data, given that the effect estimates (and not just the standard errors) are also adjusted for the covariance structure. In each model, we treated the ALM as the response variable and we used a fixed effect of context. Models were estimated with and without age, sex, and IQ as additional fixed effects. Results of the unadjusted and adjusted models were nearly identical for the effect of sampling context, so only the adjusted models are presented here as the inclusion of covariates assures the results are more generalizable to other similar samples. In each of the GLMMs we also included a random effect of participant.

For the GLMs—MLUM, NDWR, and CPM—we used the gamma family with log link function. For the GLMMs—um rate, unintelligible proportion, and repetition proportion—we used the binomial family with logit link function and fit the model using mean-variance adaptive Gauss-Hermite quadrature with 15 integration points. Specifically, we split the proportion into counts of occurrences of yes/no and fed those counts into the function: for instance, we split um rate into count of occurrences of "ums" and count of occurrences of other fillers or words (excluding "ums"). We also rescaled the continuous predictor variables for numerical stability reasons by dividing age by 10 and IQ by 100.

To compare between the three sets of sampling contexts, we used two sets of dummy coding contrasts. First we defined ADOS as the reference (intercept term), and compared it to CON and NAR. Then we defined NAR as the reference, and compared it to ADOS and CON. For each set of contrasts, we report the estimate, standard error, t- or z-score, and $p$-value. We did not perform post-hoc multiple-testing adjustments on the contrasts as the $p$-values from the linear models themselves are an accurate test of our primary question: whether any of the contexts is significantly different from any of the others. Furthermore, the well-known Tukey post-hoc test is based on balanced independent-samples ANOVA theory and is only a rough approximation to generalized linear model results. While more sophisticated options are available for GLMMs (e.g., the Kenward-Roger degrees-of-freedom approximation (Kenward & Roger, 1997)), our study lacks a large enough sample size to confidently interpret any such results.

To evaluate test-retest reliability for all sampling contexts between T1 and T2, we calculated Lin's Concordance Correlation Coefficient (CCC), a reproducibility index which evaluates the agreement between two readings of the same measure at different times by scoring variation from the concordance line (Lin, 1989). CCC is

a metric which describes how closely a new set of observations (e.g., those taken at a later time) reproduces the original set of observations on the same subjects. While Intraclass Correlation Coefficients (ICC) depend on the assumptions of a specific class of repeated-measures ANOVA models, which were not met in this sample, CCC is a popular agreement index which is not contingent upon those ANOVA assumptions being met (Chen & Barnhart, 2008). Moreover, ICC is a correlation among exchangeable observations, whereas CCC measures linear agreement between paired observations; it is unclear whether our test-retest observations are exchangeable (Berchtold, 2016). A major advantage of this measure is that it directly compares each participant against themselves at a later time point, rather than comparing the whole group aggregate against itself. This approach accommodates the heterogeneity of autism. We performed z-transformation, as is universally recommended for correlation coefficients, and used a confidence level of 0.95 for calculation of confidence intervals.

Interpretation of Lin's CCC is dependent upon the observations being measured. The practical use of the CCC is as a metric describing how closely one set of observations reproduces another set of observations on the same subjects, and in mathematical form it comprises two multiplicative terms: the Pearson correlation between the two sets of observations and a "lack-of-bias" index (ranging from 0 to 1) that measures how similar in mean and variance the two sets are, something the Pearson correlation itself does not measure. A value of 1 indicates a perfect reproduction, and a value near 0 indicates either no correlation or extreme lack of metric agreement (or both) between the two sets. Under the reasonable assumptions that variance properties are similar (as would be expected for the same group of subjects measured close in time) and the measure is consistent (approximately targeting the same quantity each time), the primary reliability concerns are whether the measure at a single occasion is precise and unaffected by irrelevant differences between the measurement occasions. For our purposes, the two assessments should not differ from one another by more than about a half standard deviation, which implies that the lack-of-bias index should be relatively large at approximately 0.9, and the linear association should explain a strong majority of the variance in the paired relationship, implying a Pearson correlation of approximately 0.8 or better. Thus, a CCC value of $0.9 \times 0.8 = \sim 0.7$ would indicate a measure with good properties all around. Note also that the CCC is a lower bound on both components of the coefficient, so if CCC = 0.7 then neither component can be smaller than 0.7 in value.

We did not perform any correction for multiple comparisons because this is an exploratory study examining how the metrics behave in a small convenience sample; we do not assert the generalizability of our findings beyond the present sample. A $p$-value of $<0.05$ was retained as a level of statistical significance. All analyses were performed using R statistical computing software version 4.0.0. (RCoreTeam, 2017).

# RESULTS

## Consistency across context

Means, *SD*, and ranges for each ALM in each context are summarized in Table 2.

**TABLE 2** Distributions of ALMs across three language sampling contexts and at two time points 4 weeks apart.

| | **Mean (*SD*) [range]** | | | | | |
|---|---|---|---|---|---|---|
| | **ADOS** | | **CON** | | **NAR** | |
| | **T1** | **T2** | **T1** | **T2** | **T1** | **T2** |
| MLUM | 5.9 (1.7) | 5.9 (1.8) | 6.9 (1.6) | 7.0 (1.6) | 8.8 (2.9) | 9.1 (2.6) |
| | [2.1–9.2] | [2.2–10.6] | [3.2–10.5] | [3.5–12.8] | [4.4–16.1] | [4.4–16.7] |
| NDWR | 199 (96) | 205 (118) | 248 (81) | 241 (70) | 151 (61) | 147 (53) |
| | [59–423] | [52–578] | [104–439] | [89–410] | [68–378] | [59–312] |
| Um rate | 0.012 (0.013) | 0.012 (0.016) | 0.011 (0.011) | 0.011 (0.013) | 0.005 (0.008) | 0.005 (0.009) |
| | [0–0.058] | [0–0.077] | [0–0.039] | [0–0.068] | [0–0.032] | [0–0.042] |
| Unintell prop | 0.029 (0.035) | 0.025 (0.031) | 0.027 (0.029) | 0.023 (0.031) | 0.025 (0.041) | 0.021 (0.032) |
| | [0–0.14] | [0–0.159] | [0–0.109] | [0–0.167] | [0–0.216] | [0–0.171] |
| CPM | 10.0 (3.4) | 9.7 (3.9) | 11.4 (4.3) | 10.8 (3.5) | 8.7 (2.8) | 8.9 (3.1) |
| | [3.7–20.5] | [3.4–21.2] | [4.8–24.5] | [4.8–19.9] | [3.8–16.3] | [3.2–19.1] |
| Repetition prop | 0.036 (0.026) | 0.028 (0.022) | 0.022 (0.022) | 0.018 (0.016) | 0.002 (0.003) | 0.002 (0.004) |
| | [0–0.135] | [0–0.088] | [0–0.101] | [0–0.069] | [0–0.012] | [0–0.017] |

Abbreviations: ADOS, autism diagnostic observation schedule; ALM, automated language measure; CON, conversation task; CPM, C-units per minute; MLUM, mean length of utterance in morphemes; NAR, narration task; NDWR, number of distinct word roots; SD, standard deviation; T1, time point 1; T2, time point 2.

**TABLE 3** Impact of language sampling context on the results of six automated language measures (ALMs).

| | Unintell prop | | | | CPM | | | | Repetition prop | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Est.** | **SE** | ***z*-value** | ***p*-value** | **Est.** | **SE** | ***t*-value** | ***p*-value** | **Est.** | **SE** | **z-value** | **p-value** |
| CON versus ADOS | −0.03 | 0.10 | −0.28 | 0.78 | 0.16 | 0.03 | 5.44 | **<0.001** | −0.58 | 0.06 | −9.75 | **<0.001** |
| NAR versus ADOS | −0.14 | 0.12 | −1.11 | 0.27 | −0.11 | 0.04 | −2.71 | **<0.01** | −2.93 | 0.17 | −16.83 | **<0.001** |
| CON versus NAR | 0.11 | 0.12 | 0.86 | 0.39 | 0.27 | 0.04 | 6.94 | **<0.001** | 2.35 | 0.18 | 13.39 | **<0.001** |
| | **MLUM** | | | | **NDWR** | | | | **Um rate** | | | |
| | **Est.** | **SE** | ***t*-value** | ***p*-value** | **Est.** | **SE** | ***t*-value** | ***p*-value** | **Est.** | **SE** | ***z*-value** | ***p*-value** |
| CON versus ADOS | 0.15 | 0.03 | 5.28 | **<0.001** | 0.22 | 0.05 | 4.63 | **<0.001** | −0.19 | 0.07 | −2.71 | **<0.01** |
| NAR versus ADOS | 0.40 | 0.04 | 10.31 | **<0.001** | −0.27 | 0.05 | −4.91 | **<0.001** | −0.92 | 0.10 | −9.01 | **<0.001** |
| CON versus NAR | −0.25 | 0.04 | −6.95 | **<0.001** | 0.48 | 0.04 | 11.43 | **<0.001** | 0.74 | 0.10 | 7.22 | **<0.001** |

*Note*: Models were adjusted for age, sex, and IQ. A larger *t*- or *z*- value demonstrates a stronger effect of context. Directionality of significant effects by context for each ALM is as follows. MLUM: NAR > CON > ADOS; NDWR: CON > ADOS > NAR; Um rate: ADOS > CON > NAR; CPM: CON > ADOS > NAR; Repetition proportion: ADOS > CON > NAR.

Abbreviations: ADOS, autism diagnostic observation schedule; CON, conversation task; CPM, C-units per minute; MLUM, mean length of utterance in morphemes; NAR, narration task; NDWR, number of distinct word roots; SE, standard error.

Detailed results of the GLMs and GLMMs can be found in Tables S3–S6. Table 3 summarizes the main results of the paired contrasts between the three tasks at T1. Um rate, unintelligible proportion, and repetition proportion all showed their highest rates in the ADOS, followed by CON, and then NAR. MLUM showed its highest rate in NAR, followed by CON, and then the ADOS. NDWR and CPM showed their highest rates for CON, followed by the ADOS, and then NAR. Unintelligible proportion is the only ALM which did not reach statistical significance and thus showed little dependence upon language sampling context. As an additional analysis, we included number of total fluent and intelligible words as an offset in the NDWR model to examine the effect of sample length on this count metric. The direction of differences between contexts did not change, but the context comparison of ADOS versus CON did lose significance in the scaled model.

The effect of sampling context on the performance of ALMs was similar in unadjusted and adjusted models (see Supplemental Information, Results section for full model results). There was a significant effect of age for MLUM and NDWR, with both increasing with increasing age. Adjusting for age had no effect on the magnitude of the context association with the other four ALMs. IQ was significantly associated with all ALMs except CPM. MLUM, NDWR, and um rate increased with increasing IQ, whereas unintelligible proportion and repetition proportion decreased with increasing IQ. Sex was not significant for any model. Even after accounting for age, sex, and IQ, *p*-values for the context coefficients remained smaller than any other coefficients in the model, and the magnitude of the context effects were not appreciably attenuated after adjustment. Thus, adjusting for age, sex, and IQ did not alter our assessment of the context effects on ALMs.

To further illustrate cross-context consistency Figures 1 and 2 provide a visual representation of one ALM across contexts, as an example. Figure 1 shows the participant distribution of CPM for each of the three sampling contexts at T1. Figure 2 illustrates the correlation of CPM between each pair of sampling contexts at T1.

## Test-retest reliability

Lin's CCC for T1 and T2 are presented in Table 4. Repetition proportion performed the worst with CCC estimates in the low range (0.17 to 0.40). Of the five remaining ALMs, MLUM, NDWR, um rate, and CPM performed the best. The distributions of CCCs across language contexts are highly consistent between time points. Excluding the poorly-performing repetition proportion, the distribution of the ALMs' test-retest reliability estimates are very similar for the ADOS (range: 0.57–0.81; median: 0.78), CON (range: 0.53–0.88; median: 0.73), and NAR (range: 0.67–0.80; median: 0.75).

# DISCUSSION

In this study, we tested the consistency across language sampling contexts and reliability across time points of previously validated ALMs in a new sample of autistic participants with strong spoken language skills (i.e., participants were capable of producing complex, fluent utterances). We found that, with the exception of unintelligible proportion, these measures have significantly different distributions across the sampling contexts of conversation, narration, and ADOS interview tasks. We have established that over a short period of time (4 weeks), all ALMs except repetition proportion demonstrated good test-retest reliability when measured in the same context. Overall, the ALMs presented here are very consistent across time points but are more variable across contexts. Language features elicited in one context are

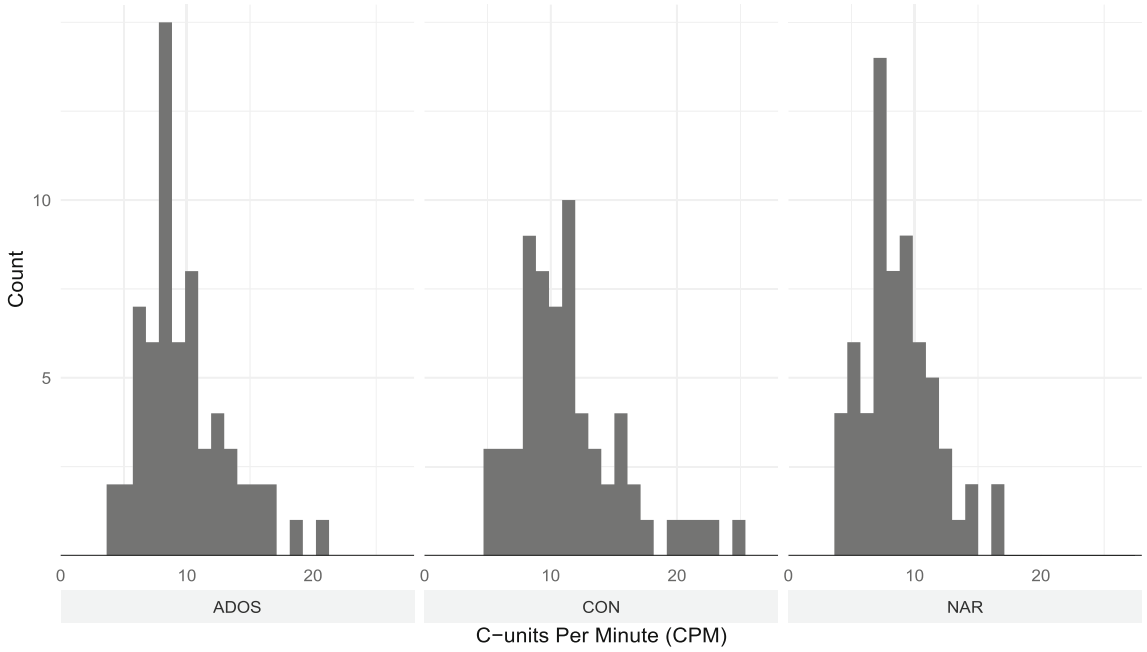


**FIGURE 1** Distribution of C-units per minute (CPM) across contexts (T1).

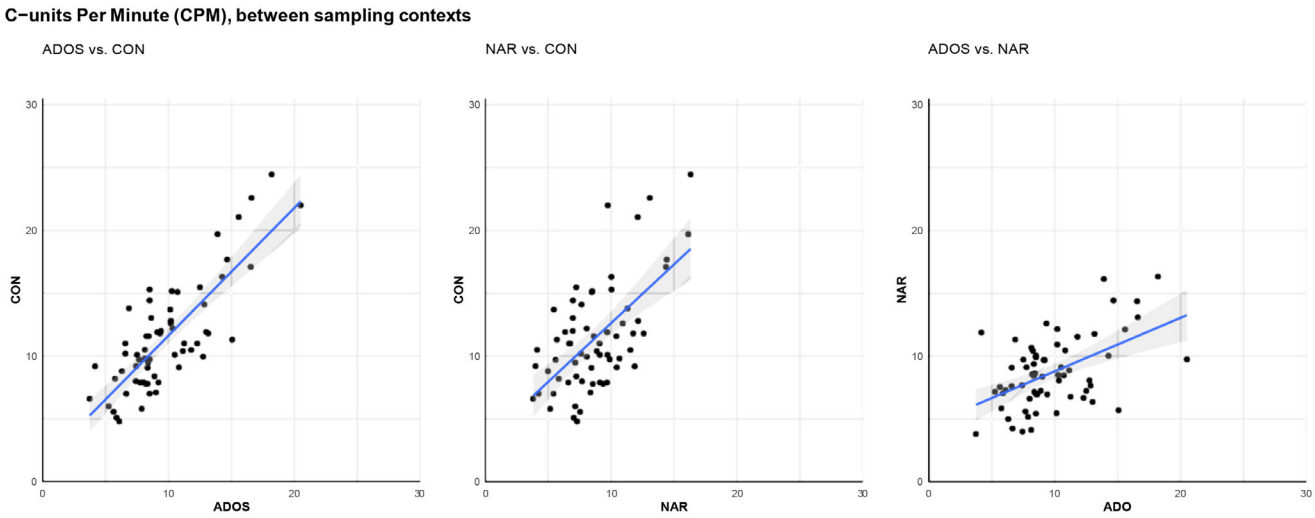


**FIGURE 2** Correlation of C-units per minute (CPM) between three contexts (T1).

**TABLE 4** Reliability of six automated language measures (ALMs) at two time points for three language sampling contexts.

| CCC (95% CI) | | | |
|---|---|---|---|
| **Measure** | **ADOS** | **CON** | **NAR** |
| MLUM | 0.81 (0.70–0.88) | 0.73 (0.58–0.83) | 0.75 (0.61–0.84) |
| NDWR | 0.78 (0.67–0.86) | 0.88 (0.80–0.92) | 0.80 (0.70–0.88) |
| Um rate | 0.73 (0.60–0.83) | 0.60 (0.41–0.73) | 0.76 (0.63–0.85) |
| Unintell prop | 0.57 (0.38–0.72) | 0.53 (0.33–0.69) | 0.67 (0.52–0.79) |
| CPM | 0.78 (0.65–0.86) | 0.85 (0.78–0.90) | 0.68 (0.52–0.80) |
| Repetition prop | 0.40 (0.17–0.58) | 0.39 (0.17–0.56) | 0.17 (−0.08–0.40) |

Abbreviations: ADOS, autism diagnostic observation schedule; CCC, concordance correlation coefficient; CI, confidence interval; CON, conversation task; CPM, C-units per minute; MLUM, mean length of utterance in morphemes; NAR, narration task; NDWR, number of distinct word roots.

not valid as measures of performance in any other context.

## Consistency across context

Sampling context was shown to have a substantial effect on five of the six outcome measures we examined with generalized linear modeling and generalized linear mixed effects modeling. Variation in the means of ALMs across sampling context was more pronounced for MLUM, NDWR, um rate, CPM, and repetition proportion, and less so for unintelligible proportion. To a large extent, dependence of language characteristics on the particular features of the language sampling task was to be expected because human language will naturally vary across different settings (e.g., a formal presentation versus a casual family dinner). However, the findings have implications for later use of ALMs as outcome measures in developmental and treatment research. Change in language features must be evaluated on language samples collected in comparable and standardized contexts in order to be meaningfully interpreted. NDWR is included in this study as a raw count, rather than as a rate metric, following prior studies from this group (MacFarlane et al., 2022; Salem et al., 2021). Although it has been suggested that this measure should be scaled by sample length, when comparing a measure of vocabulary across diverse contexts the unit of normalization can be the task itself, rather than a word or time count. Scaling by sample length may obscure true differences in talkativeness and vocabulary. However, the cross-context differences observed here indicate that comparing NDWR across contexts may not be an appropriate use of this measure as it is currently calculated.

Though the ALMs are significantly different between sampling contexts, this does not mean they are unreliable as outcome measures. ALMs may be consistently different between contexts but internally consistent within contexts, such that we could always predict MLUM to be higher in narrative tasks versus a conversation task versus an ADOS administration, for example. Because the test-retest reliability demonstrated in our study was very satisfactory, the observed differences across contexts for a given ALM cannot be interpreted as reflecting measurement error and therefore the consistent contextual differences are in fact true differences between the contexts. This points to the importance of explicitly defining the language sampling context used in a study of such outcome measures, as different measures may better elicit different aspects of language ability. In prior research (MacFarlane et al., 2022), we found that the prediction of ASD status using voice and language measures was significantly affected by task indices such as length of task and how many words and utterances were spoken, with more accurate predictions occurring for shorter samples. A similar influence could be partially contributing to the effect of conversational context on the ALM results we see here, since different contexts have differing activity length and differing levels of participant talkativeness. Previous work from Abbeduto et al. (1995) and Kover et al. (2012) offer further explanations for observed variations between contexts. For example, the narration task is more likely to be focused on character mental states, which require the use of complement clauses, thereby producing a higher MLU (as seen here compared to ADOS and CON).

## Test-retest reliability

For ALMs to be psychometrically-sound measures for treatment research, both reliability and discriminant validity of ALMs must be established as well as their ability to capture change over time. We established the discriminant validity of these six ALMs in prior work, showing that they differentiated between youth with and without autism (MacFarlane et al., 2022; Salem et al., 2021). In the present study we demonstrated, using a classic test-retest reliability paradigm with two measurements separated by 4 weeks, that satisfactory levels of reliability were obtained for four ALMs (MLUM, NDWR, um rate, CPM) with no evidence that reliability was strongly influenced by the sampling context. Thus, our results do not indicate that choice of a particular task or context should increase reliability or optimize measurement properties of these ALMs. Repetition proportion and unintelligible proportion, however, did not show strong test-retest correlations and may not be optimal outcome measures in this aspect. The results should be regarded as preliminary, and whether or not they would extend to other language sampling contexts than those used in our study remains to be examined. Evaluation of the ALMs' sensitivity to change will require comparing measurements between two time points that are further apart; we are currently undertaking that work.

Given that "performance" on the measures in this case is not a specific score but rather a quantified output of natural language, the differences that do exist in the ALMs between T1 and T2 are very likely a reflection of normal language variability that would occur in any testing situation. This work is conducted with data from natural language samples; there will always be an inherent variability of language which is difficult to account for using NLP methodology such as is employed here. However, by making assumptions about the strength of the correlation and similarity of variance between repeated assessments on the same participants, using the mathematics of the CCC itself we can translate the test-retest reliability properties into a detectable standardized effect size in a straightforward way. For example, if the correlation is an acceptable 0.8 then large effect sizes of around 1 can be detected with reasonable power (e.g., 80%) in paired samples of modest size (e.g., $n = 12$) using a

typical test-retest trial design. Single-subject designs are unlikely to capture very large effect sizes (e.g., >3) with good confidence, but larger trials assessing 50 participants or more could detect effect sizes on the order of 0.4 or less assuming correlation of 0.8 or better in the repeated measures.

## Potential advantages of ALMs

The associations with context were robust to the effects of age, sex, and IQ, showing that sampling context alone accounts for most differences in these six ALMs. Our sample had a wide distribution of age and IQ and when we adjusted the context analysis on those variables, the effect of context on the ALM scores remained practically unchanged. Sex was not a significant covariate in any model. These findings have important implications: they suggest a relative independence of the ALMs from demographic variables. If confirmed by other studies, it could allow these ALMs to be used as a tool for participants with a wide distribution of age and cognitive abilities. Further work needs to be done to validate these measures in populations with less expressive language ability.

An advantage of these ALMs is that, with the exception of NDWR, all measures were normalized by activity duration—either through averaging over C-units, or through calculating a proportion—which can be done over any length sample. Measures which use time as an input variable can bias outcome measures of language ability because samples may be of different lengths. Measures which are independent of length of sample are more widely applicable and useful. Some ALMs (um rate, unintelligible proportion, and repetition proportion) do have a low frequency of occurrence. However, one feature of proportions is that they tend to be less variable when the event occurrence is rare, so they may still have the potential to be diagnostic as a relative rate comparison as long as the sampling window is large enough. Thus, the clinical use of low-frequency outcome measures can be seen as "more expensive" because they require longer sampling. Therefore, using longer language samples that are administered as part of the typical diagnostic process (such as the ADOS) makes these measures convenient to use. Although the low frequency may make some measures "expensive", they still provide useful information and thus are worth pursuing in exploratory studies.

Interestingly, the ALMs which were more able to discriminate between ASD and non-ASD in previous work—um ratio (a slightly different but comparable measure of disfluency, as discussed in Methods), unintelligible proportion, and CPM, (MacFarlane et al., 2022; Salem et al., 2021)—are also the most consistent across sampling contexts. This result should be further explored by introducing a non-ASD comparison group to future context analysis studies.

In previous work repetition proportion has consistently underperformed compared to other ALMs (MacFarlane et al., 2022; Salem et al., 2021); here this measure continues to be less consistent and reliable than any other. Our finding that unintelligible proportion is robust to the effect of sampling context could mean that it is an independent measure—a feature not altered by the individual in response to different language activities. In addition, it is the second-least reliable ALM on retest, suggesting that it is much more impacted by influences other than context itself.

## Clinical relevance

Our finding that the ADOS showcases a lower level of participant expressive language skill (seen as a lower MLUM, higher um rate, and higher repetition proportion) than the non-ADOS language samples is consistent with results from both Hilvert et al. (2020) and Kover et al. (2014) who found that autistic children are less talkative and produce utterances of lesser syntactic complexity in the ADOS compared to conversational and play settings. While the ADOS may not demonstrate maximum ability, the ADOS context employed here still provides usefully comparable results to the conversation and narration contexts. Furthermore, we previously showed the discriminant validity of these ALMs in the same ADOS context (MacFarlane et al., 2022; Salem et al., 2021). If this result is confirmed by future work, we may find that additional clinical language elicitation tasks are unnecessary if the ADOS language sample is sufficient for analysis. Given the widespread use and routine collection of the ADOS, it would be a convenient sample for language research. At the same time, however, it is important to note that the present results are based on specific probes that are included only in Modules 3 and 4 of the measure; it remains unclear the extent to which the other ADOS modules or activity segments will provide comparable results. Of the three contexts analyzed here, the narration task produced the fewest distinct words, lowest um rate and speaking rate, and the fewest repetitions. Therefore, though the preferred context depends largely on the goals of a particular study and which outcome measures are targeted, it may not be the preferred sampling context to use when trying to elicit more natural, conversational language from participants.

Using a single aggregate score for clinical measures can be problematic due to floor effects and an erasure of the varying abilities of an individual (Hilvert et al., 2020). A fuller language profile, as described here with expressive language sampling, offers nuanced assessments of a child's communication strengths and weaknesses. In particular, measures of disfluency, intelligibility, and talkativeness are promising areas for the development of language outcome measures due to their consistency across diagnostic groups, as shown previously (MacFarlane et al., 2022; Salem et al., 2021). Previous work from this group has allowed for a more robust understanding of overall language use in participants

with FXS and DS (Abbeduto et al., 2020; Thurman et al., 2021), and this study paves the way for further analysis of autistic participants.

## Limitations and future directions

We acknowledge several limitations of our study. Participants were all English speakers capable of using complex and fluent multiword utterances and most had IQs in the typical range. Studies that include more individuals with cultural, linguistic, and cognitive diversity will broaden the scope of impact for this work. All language samples were collected in a clinical setting and therefore may not accurately reflect a child's natural language or full communicative ability, as could be seen in more naturalistic settings. Although sex was never significantly correlated with any of the measures in any context, the male bias in our study sample resulted in a small number of female participants and in a corresponding lack of statistical power to accurately examine this variable. However, model-estimated sex effects were uniformly small in magnitude for all ALMs except unintelligible proportion, suggesting that lack of statistical power is not the most likely explanation for the lack of significance of sex coefficients in our models. Additional work is needed to more thoroughly establish the construct validity of these ALMs. Another limitation lies in the reliance of our language analysis methodology on manual transcription of recorded language samples, a costly and labor-intensive process. However, we expect progresses in automatic speech recognition methodology will eventually bypass this inconvenient step.

Future steps of this work include investigating ALM sensitivity to true change by examining language patterns over all three time points of the study. Introducing comparison groups reflecting other developmental differences (e.g., intellectual disability, developmental language disorder) would allow for further exploration of the performance of these ALMs across contexts and diagnoses. These ALMs can be used as building blocks to develop higher-level language outcome measures that can capture more complicated aspects of language. Eventually, such a set of ALMs may help build clinical tools which could be used diagnostically.

## ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health under award R01DC012033, R01HD074346, P50HD103526, UL1TR001860, and by the Simons Foundation under award SFARI 383668. We gratefully acknowledge the children and their families who participated in the studies.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

*Heather MacFarlane* https://orcid.org/0000-0003-3834-2340
*Alexandra C. Salem* https://orcid.org/0000-0002-0645-8472
*Steven Bedrick* https://orcid.org/0000-0002-0163-9397
*Jill K. Dolata* https://orcid.org/0000-0002-2231-5543
*Grace O. Lawley* https://orcid.org/0000-0002-8265-6411
*Lizbeth H. Finestack* https://orcid.org/0000-0002-5300-9282
*Sara T. Kover* https://orcid.org/0000-0001-8299-7585
*Angela John Thurman* https://orcid.org/0000-0003-4220-7897
*Leonard Abbeduto* https://orcid.org/0000-0002-2311-7194
*Eric Fombonne* https://orcid.org/0000-0002-8605-3538

## REFERENCES

Abbeduto, L., Benson, G., Short, K., & Dolish, J. (1995). Effects of sampling context on the expressive language of children and adolescents with mental retardation. *Mental Retardation*, *33*(5), 279–288.

Abbeduto, L., Berry-Kravis, E., Sterling, A., Sherman, S., Edgin, J. O., McDuffie, A., Hoffmann, A., Hamilton, D., Nelson, M., Aschkenasy, J., & Thurman, A. J. (2020). Expressive language sampling as a source of outcome measures for treatment studies in fragile X syndrome: Feasibility, practice effects, test-retest reliability, and construct validity. *Journal of Neurodevelopmental Disorders*, *12*(1), 1–23.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-V*. American Psychiatric Association.

Barokova, M., & Tager-Flusberg, H. (2020). Commentary: Measuring language change through natural language samples. *Journal of Autism and Developmental Disorders*, *50*(7), 2287–2306.

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*, 1–7. https://doi.org/10.1177/2059799116672875

Berry-Kravis, E., Doll, E., Sterling, A., Kover, S. T., Schroeder, S. M., Mathur, S., & Abbeduto, L. (2013). Development of an expressive language sampling procedure in fragile X syndrome: A pilot study. *Journal of Developmental and Behavioral Pediatrics*, *34*(4), 245–251.

Butler, L., La Valle, C., Schwartz, S., Palana, J. B., Liu, C., Peterman, N., Shen, L., & Tager-Flusberg, H. (2022). Remote natural language sampling of parents and children with autism spectrum disorder: Role of activity and language level. *Frontiers in Communication*, *7*, 1–11. https://doi.org/10.3389/fcomm.2022.820564

Channell, M. M., Loveall, S. J., Conners, F. A., Harvey, D. J., & Abbeduto, L. (2018). Narrative language sampling in typical development: Implications for clinical trials. *American Journal of Speech-Language Pathology*, *27*(1), 123–135.

Chen, C.-C., & Barnhart, H. X. (2008). Comparison of ICC and CCC for assessing agreement for data without and with replications. *Computational Statistics & Data Analysis*, *53*(2), 554–564.

Costanza-Smith, A. (2010). The clinical utility of language samples. *Perspectives on Language Learning and Education*, *17*(1), 9–15.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*(3), 837–845.

Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models* (fourth ed.). Chapman and Hall/CRC.

Dolata, J. K., Suarez, S., Calamé, B., & Fombonne, E. (2022). Pragmatic language markers of autism diagnosis and severity. *Research in Autism Spectrum Disorders*, *94*, 101970.

Finestack, L. H., & Abbeduto, L. (2010). Expressive language profiles of verbally expressive adolescents and young adults with down syndrome or fragile X syndrome. *Journal of Speech, Language, and Hearing Research*, *53*(5), 1334–1348.

Finestack, L. H., Sterling, A. M., & Abbeduto, L. (2013). Discriminating down syndrome and fragile X syndrome based on language ability. *Journal of Child Language*, *40*(1), 244–265.

Gorman, K., Bedrick, S., Kiss, G., Morley, E., Ingham, R., Mohammed, M., Papadakis, K., & van Santen, J. (2015). Automated morphological analysis of clinical language samples. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 108–116, Denver, Colorado. Association for Computational Linguistics.

Gorman, K., Olson, L., Hill, A. P., Lunsford, R., Heeman, P. A., & van Santen, J. P. H. (2016). Uh and um in children with autism spectrum disorders or language impairment. *Autism Research*, *9*(8), 854–865.

Heilmann, J., Nockerts, A., & Miller, J. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech, and Hearing Services in Schools*, *41*, 393–404.

Hilvert, E., Sterling, A., Haebig, E., & Friedman, L. (2020). Expressive language abilities of boys with idiopathic autism spectrum disorder and boys with fragile X syndrome + autism spectrum disorder: Cross-context comparisons. *Autism & Developmental Language Impairments*, *5*, 1–16. https://doi.org/10.1177/2396941520912118

Hoffmann, A., Thurman, A. J., Sterling, A., Kover, S. T., Finestack, L., Berry-Kravis, E., Edgin, J. O., Drayton, A., Fombonne, E., & Abbeduto, L. (2022). Analysis of a repetitive language coding system: Comparisons between fragile X syndrome, autism, and down syndrome. *Brain Sciences*, *12*(5), 575. https://doi.org/10.3390/brainsci12050575

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 5.1, 221–234.

Janvier, D., Choi, Y. B., Klein, C., Lord, C., & Kim, S. H. (2022). Brief report: Examining test-retest reliability of the autism diagnostic observation schedule (ADOS-2) calibrated severity scores (CSS). *Journal of Autism and Developmental Disorders*, *52*(3), 1388–1394. https://doi.org/10.1007/s10803-021-04952-7

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997.

Kover, S. T., & Abbeduto, L. (2010). Expressive language in male adolescents with fragile X syndrome with and without comorbid autism. *Journal of Intellectual Disability Research*, *54*(3), 246–265.

Kover, S. T., Davidson, M. M., Sindberg, H. A., & Ellis Weismer, S. (2014). Use of the ADOS for assessing spontaneous expressive language in young children with ASD: A comparison of sampling contexts. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2221–2233.

Kover, S. T., McDuffie, A., Abbeduto, L., & Brown, W. T. (2012). Effects of sampling context on spontaneous expressive language in males with fragile X syndrome or down syndrome. *Journal of Speech, Language, and Hearing Research*, *55*(4), 1022–1038.

Lawley, G. O., Bedrick, S., MacFarlane, H., Dolata, J. K., Salem, A. C., & Fombonne, E. (2022). "Um" and "Uh" usage patterns in children with autism: Associations with measures of structural and pragmatic language ability. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-022-05565-4

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*(1), 255–268.

Loban, W. (1976). Language development: Kindergarten through grade twelve. NCTE Committee on Research Report No. 18.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., & Rutter, M. (2000). The autism diagnostic observation schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule–2nd edition (ADOS-2)*. Western Psychological Corporation.

Lyons, M. (2021). Preschool language scale – 5. In F. R. Volkmar (Ed.), *Encyclopedia of autism spectrum disorders*. Springer. https://doi.org/10.1007/978-3-319-91280-6_979

MacFarlane, H., Salem, A. C., Chen, L., Asgari, M., & Fombonne, E. (2022). Combining voice and language features improves automated autism detection. *Autism Research*, *15*(7), 1288–1300.

Martin, G. E., Roberts, J. E., Helm-Estabrooks, N., Sideris, J., Vanderbilt, J., & Moskowitz, L. (2012). Perseveration in the connected speech of boys with fragile X syndrome with and without autism Spectrum disorder. *American Journal on Intellectual and Developmental Disabilities*, *117*(5), 384–399.

Meir, N., & Novogrodsky, R. (2020). Syntactic abilities and verbal memory in monolingual and bilingual children with high functioning autism (HFA). *First Language*, *40*(4), 341–366.

Miles, S., Chapman, R., & Sindberg, H. (2006). Sampling context affects MLU in the language of adolescents with down syndrome. *Journal of Speech, Language, and Hearing Research*, *49*(2), 325–337.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2015). *Assessing language production using SALT software: A clinician's guide to language sample analysis*. SALT Software, LLC.

Murphy, M. M., & Abbeduto, L. (2007). Gender differences in repetitive language in fragile X syndrome. *Journal of Intellectual Disability Research*, *51*(5), 387–400.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Pavelko, S. L., Owens, R. E., Jr., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, *47*(3), 246–258.

Pavelko, S. L., Price, L. R., & Owens, R. E. (2020). Revisiting reliability: Using sampling utterances and grammatical analysis revised (SUGAR) to compare 25- and 50-utterance language samples. *Language, Speech, and Hearing Services in Schools*, *51*(3), 778–794.

Ratner, N. B., & MacWhinney, B. (2016). Your laptop to the rescue: Using the child language data exchange system archive and CLAN utilities to improve child language sample analysis. *Seminars in speech and language*, *37*(2), 74–84.

RCoreTeam. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roid, G. H. (2003). *The Stanford-Binet intelligence scales* (Fifth ed.). Riverside Publishing.

Salem, A. C., MacFarlane, H., Adams, J. R., Lawley, G. O., Dolata, J. K., Bedrick, S., & Fombonne, E. (2021). Evaluating atypical language in autism using automated language measures. *Scientific Reports*, *11*, 10968. https://doi.org/10.1038/s41598-021-90304-5

Shaffer, R. C., Schmitt, L., John Thurman, A., Abbeduto, L., Hong, M., Pedapati, E., Dominick, K., Sweeney, J., & Erickson, C. (2020). The relationship between expressive language

sampling and clinical measures in fragile X syndrome and typical development. *Brain Sciences*, *10*(2), 66. https://doi.org/10.3390/brainsci10020066

Sonderegger, M., Wagner, M., & Torreira, F. (2018). Quantitative methods for linguistic data. http://people.linguistics.mcgill.ca/~morgan/qmld-book/

Sparrow, S. S., Cicchetti, D., & Balla, D. A. (2012). *Vineland adaptive behavior scales* (Second ed.). American Psychological Association.

Suh, J., Eigsti, I.-M., Naigles, L., Barton, M., Kelley, E., & Fein, D. (2014). Narrative performance of optimal outcome children and adolescents with a history of an autism spectrum disorder (ASD). *Journal of Autism and Developmental Disorders*, *44*(7), 1681–1694.

Tager-Flusberg, H., Rogers, S., Cooper, J., Landa, R., Lord, C., Paul, R., Rice, M., Stoel-Gammon, C., Wetherby, A., & Yoder, P. (2009). Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. *Journal of Speech, Language, and Hearing Research*, *52*(3), 643–652.

Thurman, A. J., Edgin, J. O., Sherman, S. L., Sterling, A., McDuffie, A., Berry-Kravis, E., Hamilton, D., & Abbeduto, L. (2021). Spoken language outcome measures for treatment studies in Down syndrome: Feasibility, practice effects, test-retest reliability, and construct validity of variables generated from expressive language sampling. *Journal of Neurodevelopmental Disorders*, *13*(1), 13. https://doi.org/10.1186/s11689-021-09361-6

Volden, J., & Phillips, L. (2010). Measuring pragmatic language in speakers with autism spectrum disorders: Comparing the children's communication checklist–2 and the test of pragmatic language. *American Journal of Speech-Language Pathology*, *19*(3), 204–212. https://doi.org/10.1044/1058-0360(2010/09-0011)

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

Wiig, E. H., Secord, W., & Semel, E. M. (2020). *CELF preschool 3: Clinical evaluation of language fundamentals-preschool*. Pearson, Incorporated.

Wiig, E. H., Secord, W. A., & Semel, E. (2013). *Clinical evaluation of language fundamentals: CELF-5*. Pearson.

Wilder, A., & Redmond, S. M. (2022). The reliability of short conversational language sample measures in children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*, *65*(5), 1939–1955.

## SUPPORTING INFORMATION

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

**How to cite this article:** MacFarlane, H., Salem, A. C., Bedrick, S., Dolata, J. K., Wiedrick, J., Lawley, G. O., Finestack, L. H., Kover, S. T., Thurman, A. J., Abbeduto, L., & Fombonne, E. (2023). Consistency and reliability of automated language measures across expressive language samples in autism. *Autism Research*, *16*(4), 802–816. https://doi.org/10.1002/aur.2897