RESEARCH ARTICLE

Eliciting Language Samples for Analysis (ELSA): A New Protocol for Assessing Expressive Language and Communication in Autism

Mihaela D. Barokova, Chelsea La Valle, Sommer Hassan, Collin Lee, Mengyuan Xu, Riley McKechnie, Emily Johnston, Manon A. Krol, Jennifer Leano, and Helen Tager-Flusberg

Expressive language and communication are among the key targets of interventions for individuals with autism spectrum disorder (ASD), and natural language samples provide an optimal approach for their assessment. Currently, there are no protocols for collecting such samples that cover a wide range of ages or language abilities, particularly for children/adolescents who have very limited spoken language. We introduce a new protocol for collecting language samples, eliciting language samples for analysis (ELSA), and a novel approach for deriving basic measures of verbal communicative competence from it that bypasses the need for time-consuming transcription. Study 1 presents ELSA-adolescents (ELSA-A), designed for minimally and low-verbal older children/adolescents with ASD. The protocol successfully engaged and elicited speech from 46 participants across a wide range of ages (6;6–19;7) with samples averaging 20–25 min. The collected samples were segmented into speaker utterances (examiner and participant) using real-time coding as one is listening to the audio recording and two measures were derived: frequency of utterances and conversational turns per minute. These measures were shown to be reliable and valid. For Study 2, ELSA was adapted for younger children (ELSA-Toddler [ELSA-T]) with samples averaging 29 min from 19 toddlers (2;8–4;10 years) with ASD. Again, measures of frequency of utterances and conversational turns derived from ELSA-T were shown to have strong psychometric properties. In Study 3, we found that ELSA-A and ELSA-T were equivalent in eliciting language from 17 children with ASD (ages: 4;0–6;8), demonstrating their suitability for deriving robust objective assessments of expressive language that could be used to track change in ability over time. We introduce a new protocol for collecting expressive language samples, ELSA, that can be used with a wide age range, from toddlers (ELSA-T) to older adolescents (ELSA-A) with ASD who have minimal or low-verbal abilities. The measures of language and communication derived from them, frequency of utterances, and conversational turns per minute, using real-time coding methods, can be used to characterize ability and chart change in intervention research. ***Autism Res*** *2020, 00: 1–15.* © 2020 International Society for Autism Research and Wiley Periodicals LLC

Lay Summary: We introduce a new protocol for collecting expressive language samples, ELSA, that can be used with a wide age range, from toddlers (ELSA-T) to older adolescents (ELSA-A) with autism spectrum disorder who have minimal or low-verbal abilities. The measures of language and communication derived from them, frequency of utterances and conversational turns per minute, using real-time coding methods, can be used to characterize ability and chart change in intervention research.

Keywords: autism spectrum disorder; language; communication; assessment; measures; outcome; ELSA

Introduction

Expressive language is among the most heterogeneous characteristics in autism spectrum disorder (ASD) and one of the strongest predictors of future outcomes [Friedman, Sterling, Dawalt, & Mailick, 2019; Howlin, Goode, Hutton, & Rutter, 2004; Venter, Lord, & Schopler, 1992]. This is why language and communication are often targets of interventions for both younger and older children [e.g., Kasari, Gulsrud, Wong, Kwon, & Locke, 2010; Tager-Flusberg & Kasari, 2013]. Expressive language in ASD has often been evaluated using standardized tests or parent report measures for assessment purposes as well as for charting change over time. Although most of these measures have good psychometric properties, they are generally not ideal, particularly for use as outcome measures in ASD

From the Department of Psychological and Brain Sciences, Center for Autism Research Excellence, Boston University, Boston, Massachusetts (M.D.B., C.L.V., S.H., C.L., M.X., R.M., E.J., M.A.K., J.L., H.T.-F.); Easy Speech Pathology Clinic, Palm Desert, California (S.H.); Rosemead School of Psychology, Biola University, La Mirada, California (C.L.); 1500 Commerce Park Dr, Reston, VA 20191 (M.X.); Laboratories of Cognitive Neuroscience, Division of Developmental Medicine, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts (R.M.); Donders Institute, Radboudumc, Nijmegen, The Netherlands (M.A.K.)

Received April 14, 2020; accepted for publication August 1, 2020

Address for correspondence and reprints: Mihaela D. Barokova, Department of Psychological and Brain Sciences, Center for Autism Research Excellence, Boston University, 100 Cummington Mall, Boston, MA 02125. E-mail: barokova@bu.edu

Published online 00 Month 2020 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/aur.2380

research. Standardized tests may not have been normed on this population, and individuals with lower cognitive or language abilities are likely to show floor effects. Parent report measures are subjective and vulnerable to placebo effects when used in intervention studies [e.g., Berry-Kravis et al., 2016; Guastella et al., 2015; Masi, Lampit, Glozier, Hickie, & Guastella, 2015]. Because of these limitations, it has been argued that natural language samples (NLS) provide a better alternative for assessing expressive language in ASD [Abbeduto et al., 2020; Barokova & Tager-Flusberg, 2018; Tager-Flusberg et al., 2009]. NLS are easy to collect and are more representative of speakers' everyday language use [Brown, 1973; MacWhinney, 2007]. They are not subject to floor effects, and procedures for their elicitation can be tailored to the social and cognitive abilities of participants. NLS can be collected across the spectrum of abilities from toddlers with emerging language to minimally or low-verbal adolescents and adults, whose atypical behaviors prevent them from completing highly structured testing procedures. We present here a set of studies describing a new protocol for collecting natural language samples called eliciting language samples for analysis (ELSA) from children and adolescents with ASD across a wide range of ages and language abilities and evaluate the psychometric properties of the language and communication measures derived from them.

In ASD research, language samples have been collected across different types of protocols and contexts [e.g., Bang & Nadig, 2015; Park, Yelland, Taffe, & Gray, 2012; Schoen, Paul, & Chawarska, 2011]. However, the lack of a consistent protocol makes the comparison of results across studies difficult to interpret. For example, studies of infants and toddlers often rely on collecting language samples during free play with parents [e.g., Bang & Nadig, 2015; Casenhiser, Binns, McGill, Morderer, & Shanker, 2015; Fusaroli, Weed, Fein, & Naigles, 2019; Kaiser & Roberts, 2013]. Other studies use language samples collected during assessments such as the Autism Diagnostic Observation Schedule-2 (ADOS-2) [Lord et al., 2012; Morley, Roark, & van Santen, 2013; Park et al., 2012] or the Early Social Communication Scales [Mundy et al., 2003; Roos, McDuffie, Weismer, & Gernsbacher, 2008] or parent report measures such as the Communication and Symbolic Behavior Scales [Schoen et al., 2011; Wetherby & Prizant, 2002]. Yet other studies rely on elicitation protocols designed to assess very specific aspects of the language and communication of younger children [e.g., Drew, Baird, Taylor, Milne, & Charman, 2007; Pasco, Gordon, Howlin, & Charman, 2008]. However, there might be systematic differences in child speech based on the specifics of the elicitation protocol. For instance, Kover, Davidson, Sindberg, and Weismer [2014] found that children with ASD produced significantly more utterances during free play with the examiner than during the ADOS. Furthermore, there is no elicitation protocol that is appropriate for children and adolescents across a wide age range who have more limited verbal abilities.

Although the collection of a NLS may only take less than half an hour, the process of transcription, often using software tools such as systemic analysis of language transcripts (SALT) [Miller, Andriacchi, & Nockerts, 2011] or computerized language analysis (CLAN) [MacWhinney, 2000], is very time consuming. Thus a 10–12-min sample typically takes between 60 and 90 min to transcribe [Miller et al., 2011], which may be a limiting factor in the widespread use of NLS to assess expressive language in ASD (or other disorders). Despite the promise of automated analyses of NLS using computational approaches to speaker identification and speech and nonspeech sound segmentation such as LENA, their potential use is quite limited. While LENA has been successfully used in studies of young children with ASD [e.g., Woynaroski et al., 2017; Yoder, Oller, Richards, Gray, & Gilkerson, 2013], it is not able to segment speakers when the target participant is older [Jones et al., 2019]. Thus, for now, we must still rely on transcription to derive measures of language or communication.

Among the most widely used measures derived from transcripts in studies of children with ASD are total number of utterances, frequency of utterances per unit time, and number of conversational turns [Barokova & Tager-Flusberg, 2018; Casenhiser et al., 2015; Colle, Baron-Cohen, Wheelwright, & Van Der Lely, 2008; Hogan-Brown, Losh, Martin, & Mueffelmann, 2013; Kasari et al., 2014; Kover & Abbeduto, 2010; Kover et al., 2014; Suh et al., 2014]. Although very broad, these measures capture the speakers' general verbal communication and conversational turns also tap on back-and-forth verbal engagement. These broad measures also allow for a comparison across even the most heterogeneous participants with ASD from those who use single words to simple phrases to sentences to those who are verbally fluent, which makes them suitable candidates for outcome measures in clinical trials and treatment and intervention studies, which typically enroll participants across a wide range of ability. Other more specific measures may include, for example, mean length of utterance (MLU), number of different words (NDW) used, or number of conversational topics, but their use requires some language abilities that children with more significant language impairments may not possess. Although measures derived from NLS are informative, their psychometric properties have yet to be examined for children and adolescents with ASD. Only one study to date has validated transcript-derived measures against standardized measures of language and communication [Condouris, Meyer, & Tager-Flusberg, 2003]. However, participants in this study included verbally fluent children and adolescents, and only lexical-semantic and grammatical measures were validated, leaving open the question of the validity of the more common measures used: frequency of utterances and conversational turns.

The goals of the studies presented in this paper address key gaps in the literature on NLS in ASD. In Study 1, we present a novel language elicitation protocol, –ELSA-adolescents (ELSA-A), developed for collecting language samples from older children and adolescents including those who are minimally and low verbal. First, we evaluate the effectiveness of the protocol in eliciting language from this population. Then we introduce a novel, real-time coding approach, coding as one is listening to the audio recording of the sample, to measure frequency of utterances and conversational turns that can replace the more time-consuming transcription process, and we assess the psychometric properties of these measures. The choice of these expressive language measures is motivated by the heterogeneity in verbal ability across individuals with ASD and by the need for assessing their psychometric properties. In Study 2, we present an adaptation of the original ELSA-A protocol for use with toddlers and preschoolers, ELSA-toddler (ELSA-T), and again evaluate the psychometric properties of the same measures derived from it. Finally, in Study 3 we examine the comparability of ELSA-A and ELSA-T in eliciting language from children with ASD between 4 and 7 years old.

Study 1: ELSA-A

Almost no studies have examined the language and communication of older minimally or low-verbal children and adolescents with ASD. This is primarily due to the lack of assessment tools that can capture the heterogeneity of this population in terms of age, language ability, nonverbal IQ, and comorbid psychopathology [Kasari et al., 2014; Tager-Flusberg & Kasari, 2013]. Three studies used NLS to assess or track changes in minimally or low-verbal participants with ASD [Chiang, 2009; Kasari et al., 2014; Paul, Campbell, Gilbert, & Tsiouri, 2013]. Two of these studies collected language samples from preschoolers [Paul et al., 2013] or young school-aged children [Kasari et al., 2014] in a free play context, while Chiang [2009] collected lengthy home-based naturalistic recordings from minimally to low-verbal children and adolescents with ASD ranging more widely in age. These studies demonstrate the feasibility of using NLS with minimally verbal children and adolescents with ASD, but the lack of a common protocol precludes comparisons across studies. Indeed, there are already available and widely used language elicitation protocols through the SALT platform [Miller et al., 2011], as well as more standardized protocols that have been validated with children and adolescents with other neurodevelopmental disorders [e.g., see Abbeduto et al., 2020 for Fragile X]. However, these protocols often use free play or conversation and narration as elicitation contexts, which precludes their use with older minimally to low-verbal children and adolescents for whom free play is not developmentally appropriate and narration requires higher verbal abilities.

The measures of language and communication used in the three studies with minimally to low-verbal participants all included broad measures of ability such as total number of communicative utterances [Chiang, 2009; Kasari et al., 2014] and frequency of words per minute [Kasari et al., 2014; Paul et al., 2013]. Although useful, the psychometric properties of these measures derived from older children and adolescents with ASD, who have some degree of language impairment, have yet to be examined. Our goals, therefore, are to develop a common NLS protocol that can be used across a wide age range of minimally and low-verbal individuals with ASD and to evaluate the psychometrics of the most widely used measures within this population.

Methods

Participants

The participants in this study included 46 (12 females) children and adolescents with ASD between 6;6 and 19;7 years old, who were administered Module 1 or 2 of the ADOS-2 [Lord et al., 2012] if they were between 6 and 12 years old or Module 1 or 2 of the adapted ADOS (A-ADOS) [Bal et al., 2019] if they were 12 years and older in order to confirm their diagnosis and status as minimally verbal (ADOS-2 Mod 1 or A-ADOS Mod 1) [see Bal, Katz, Bishop, & Krasileva, 2016] or low verbal (ADOS-2 Mod 2 or A-ADOS Mod 2). In order to evaluate test–retest reliability, 10 (6 females) participants with ASD between 8;8 and 18;10 years old were also included. Four of the test–retest participants were from the original 46 and the remaining 6 were from a different project. The test–retest participants were included because they had provided two ELSA-A samples. Even though not all of them were minimally to low verbal, all of them presented with some degree of language impairment, and therefore qualified for ELSA-A collection. English was the primary language for all participants. Tables 1 and 2 provide information about the participants, including scores on standardized measures.

Procedures

Institutional Review Board (IRB) approval was obtained prior to enrolling participants. Participants' visit to the lab included a battery of standardized assessments and the collection of a natural language sample using the ELSA-A protocol. The 10 test–retest participants followed the same procedures, but they returned to the lab, on average, within 87.5 days (SD = 100 days) of their initial visit to provide a second ELSA-A sample. Because all test–retest participants were older, did not receive speech/language therapy and were minimally to low verbal or

Table 1. Demographic Information and Standardized Assessment Scores for the 46 ELSA-A Participants (Study 1)

Characteristic/assessment	*N*				
Sex	46	34 Male			
		12 Female			
Race	44	33 White			
		3 Black, African American			
		2 Asian			
		1 Hispanic, Latino, or Spanish Origin			
		5 Multiple races			
Ethnicity	44	38 Non-Hispanic			
		5 Hispanic			
		1 Prefer not to respond			
			M	SD	Range
Age (in months)	46		160.61	44.99	78:235
ADOS-2	46	36 Mod 1			
A-ADOS		10 Mod 2			
		Calibrated severity score	7.89	1.54	3:10
Leiter-R	46	Non-verbal IQ	58.89	23.33	30:115
SCQ	41	Communication score	3.37	2.71	0:9
VABS-2	46	Communication standard score	47.20	14.29	26:83

Abbreviations: A-ADOS, adapted Autism Diagnostic Observation Schedule; ADOS-2, Autism Diagnostic Observation Schedule-2; ELSA-A, eliciting language samples for analysis-adolescents; Leiter-R, Leiter International Performance Scales-Revised; SCQ, Social Communication Questionnaire; VABS-2, Vineland Adaptive Behavior Scale-2.

Table 2. Demographic Information and Standardized Assessment Scores for the 10 ELSA-A Test–Retest Participants (Study 1)

Characteristic/assessment	*N*				
Sex	10	4 Male			
	10	6 Female			
Race		6 White			
		1 Black, African American			
		3 Asian			
	10	0 Hispanic, Latino, or Spanish Origin			
Ethnicity		9 Non-Hispanic			
		1 Hispanic			
			M	SD	Range
Age (in months)	10		155.50	43.35	80:226
ADOS-2	10	7 Mod 1			
A-ADOS		1 Mod 2			
		1 Mod 3			
		1 Mod 4			
		Calibrated severity score	7.90	1.85	5:10
Leiter-R	10	Nonverbal IQ	59.70	21.81	39:115
SCQ	3	Communication score	3.00	2.65	1:6
VABS-2	10	Communication standard score	47.60	11.66	33:65

Abbreviations: A-ADOS, adapted Autism Diagnostic Observation Schedule; ADOS-2, Autism Diagnostic Observation Schedule-2; ELSA-A, eliciting language samples for analysis-adolescents; Leiter-R, Leiter International Performance Scales-Revised; SCQ, Social Communication Questionnaire; VABS-2, Vineland Adaptive Behavior Scale-2.

presented with some degree of language impairment, we did not expect change in their language over the test–retest time period.

Measures. ASD diagnoses were confirmed with the ADOS-2 or A-ADOS [Bal et al., 2019; Lord et al., 2012] and the Social Communication Questionnaire (SCQ) [Rutter, Bailey, & Lord, 2003], completed by parents. Non-verbal IQ was assessed with the Leiter International Performance Scales-Revised (Leiter-R) [Roid & Miller, 1997], and adaptive behavior was assessed using the parent questionnaire version of the Vineland Adaptive Behavior Scale-2 (VABS-2) [Sparrow, Cicchetti, & Balla, 2005].

Eliciting Language Samples for Analysis-Adolescent (ELSA-A)

A natural language sample was collected from all participants by a trained examiner using the ELSA-A protocol. ELSA-A was developed specifically for the elicitation of speech from older children and adolescents with ASD

who have some degree of language impairment, including those who are minimally to low verbal. ELSA-A takes about 20–25 min to administer and consists of eight activities, which fall into one of two categories: semi-structured play with developmentally appropriate materials and narrative (retelling the plot of movie shorts). In addition, examiners are encouraged to engage in a conversation (back and forth question and answer verbal interaction) about the participants' interests, while transitioning from one activity to the next. The inclusion of these three most widely used elicitation contexts (play, narrative, and conversation) ensures that the protocol captures a broad range of expressive language abilities even in the most heterogeneous participant samples, which already existing protocols relying on a single elicitation context cannot accomplish.

Each of the eight activities are designed to be interactive, fun and engaging for older children and adolescents regardless of age, sex, or language ability. A list of ELSA-A activities and a brief description of each can be found in Table 3. More detailed descriptions, administration instructions, and a list of required materials for each activity can be found on our website (https://sites.bu.edu/elsa/; https://sites.bu.edu/elsa/elsa-2/manual/). The administration of each activity involves the use of toys and materials, as well as at least two open-ended verbal prompts to be initiated by the examiner.

All ELSA-As were audio recorded using a voice recording app (Voice Recorder HD) on a smartphone worn by the examiner in an armband. The total length of ELSA-A was extracted from the audio files. The first speech utterance by the examiner was considered the start of ELSA-A, and the last speech utterance by examiner or participant before putting away the ELSA-A materials was considered the end of the sample.

ELSA-A administration fidelity was evaluated and operationalized as the number of ELSA-A activities attempted by examiner out of a total of 8. An attempt is defined as engaging the participant with the materials and using a verbal prompt. The choice of administration fidelity measure was motivated by the goal of ELSA-A to be fun and engaging for our participants and to allow for the collection of a language sample long enough to derive measures of language and communication. That is, although very general, this measure captures how easy it is to engage participants with the activities. Prior to administering ELSA-A, each examiner extensively reviewed the instructional manual and watched the instructional video. Only after practicing ELSA-A with adults and receiving feedback, examiners were allowed to collect ELSA-A from the minimally to low-verbal participants.

Table 3. Descriptions of All ELSA-A and ELSA-T Activities

ELSA-A	ELSA-T
Activity Name: description	*Activity Name*: description
Leaf Falling: a joint gross motor activity that gives the participant an opportunity to interact with the examiner by labeling various parts of a tree, putting the leaves on the tree, and talking about different seasons	*Apple Falling*: a joint gross motor activity that gives the participant an opportunity to interact with the examiner by labeling various parts of a tree and picking up apples
Planting an Acorn: a pretend play activity that gives participants the opportunity to pretend to plant an acorn using a shovel	*Picnic Adventure*: a pretend play activity that gives participants the opportunity to pretend to go on a picnic adventure and involves talking about different fruits and vegetables
Discovering Animals: intended to elicit descriptions of animals that most children are already familiar with (e.g., bird, squirrel, racoon) hidden around the room	*Hide and Seek Animals*: intended to elicit descriptions of animals that most younger children are already familiar with (e.g., dogs, cats, etc.) hidden around the room
Helping Animals: a loosely structured pretend play activity in which the examiner and participant can express their creativity and figure out how to help toy animals (e.g., lion, tiger, giraffe) who are hungry, thirsty and/or hurt	*Bath Time*: a loosely structured pretend play activity in which the examiner and participant can express their creativity and give different toy animals a bath
S'mores: a reinforcing activity which contains many opportunities for requesting different snacks, while making a S'more, and an opportunity for the examiner to have conversations with the participant about their interests	*Snack*: a reinforcing activity which contains many opportunities for requesting different snacks (e.g., fruit snacks, goldfish, etc.)
Crafts: a creative activity that gives participants the opportunity to express their preferences with materials of their choosing (pencils vs. crayons; drawing vs. coloring)	*Arts and Crafts*: a creative activity that gives participants the opportunity to express their preferences with materials of their choosing (play-doh, different shapes for play-doh modeling)
Bean Bag Toss: another gross motor activity that allows for opportunities for requesting and turn-taking as the participant is aiming at different animals to earn a different number of points	*Turtle Bean Bag Toss*: another gross motor activity that allows for opportunities for requesting and turn-taking
Pixar Movie Shorts: a narrative activity which includes a discussion of the plot and characters of a movie short after watching it on a tablet	*Storybook Time*: a narrative activity, which includes labeling characters and guessing the plot of a storybook after reading it with the examiner

Abbreviations: ELSA-A, eliciting language samples for analysis-adolescents; ELSA-T, eliciting language samples for analysis-toddlers.

Coding. All ELSA-A audio files were coded for *speech utterances* following a novel, real-time coding approach—coding on the first pass as one is listening to the audio recording of the language sample. Speech utterances were defined as vocalizations that have a syllable structure and consist of vowels and consonants. For the vocalization to be considered speech, it had to consist of at least one consonant-vowel combination, which did not need to

approximate a word. The speech utterance did not necessarily have to be directed at a conversational partner to be coded, and imitations of spoken utterances were coded, as well. Non-speech vocalizations (e.g., sighing, squealing, sneezing) were not coded. Speech utterances that consisted of phrases or full sentences were segmented into communication units defined as independent clauses [Loban, 1976].

Coding was carried out in ELAN Linguistic Annotator software, which is freely available [Lausberg & Sloetjes, 2009]. Trained coders time-stamped the beginning and end of each speech utterance produced by the participant and the examiner, while listening to the audio recorded ELSA-A file. This allows the coding of a 20-min file to be completed in approximately 25 min. Prior to coding ELSA-As for analysis, each coder went through training. Coding training included coding three to five training files and receiving extensive feedback after each completed file. Afterward, each coder coded a minimum of 10 practice files used to assess intercoder reliability. If the coder had achieved high intercoder reliability, defined as an intraclass correlation coefficient larger than 0.9, they went on to code files for analysis. All coders were able to obtain this level of reliability. The coding training, on average, took around 20–30 hr. All coders were naïve, that is, they did not have any prior training in linguistics or language development.

ELSA-A measures. Once coded in ELAN, the ELSA-A audio files were exported and used to compute two key measures of spoken language and communication: frequency of speech utterances per minute (*coded FreqU*) and number of conversational turns per minute (*coded CT*) for both participant and examiner. Coded FreqU was computed by dividing the total number of segmented speaker utterances by ELSA-A length in minutes. Coded CT were computed by exporting the coded ELAN file, which contained the start and end time of each speaker utterance, into an excel document and applying a formula to count CT. Thus, the computation of coded CT did not require additional coding other than what was already done to obtain coded FreqU. A conversational turn was defined as one or more consecutive utterance(s) produced by the same speaker, as defined by the SALT software [Miller et al., 2011].

ELSA-A transcription. In order to validate the real-time coding in ELAN, all ELSA-A samples were transcribed using the SALT-12 software [Miller & Iglesias, 2012]. Each ELSA-A was transcribed by a trained transcriber and checked by a second transcriber. Any disagreements were resolved through consensus. We ensured that no coder both coded and transcribed the same ELSA-A sample. Frequency of utterances per minute (*transcribed FreqU*) and number of conversational turns per minute (*transcribed CT*) were taken from the SALT output files for each participant and examiner.

Results

In the following analyses, we used nonparametric tests for variables that did not have a normal distribution. In particular, for nonparametric correlations, we computed Spearman's correlations and for group comparisons, we used the Mann–Whitney *U* tests.

Evaluation of ELSA-A

Fidelity. On average, examiners administered 7.35 out of 8 (SD = 1.10) ELSA-A activities. The average length of ELSA-A was 20.22 min (SD = 4.23). Participants' coded FreqU during ELSA-A was 3.78 (SD = 3.73) and coded CT was 2.61 (SD = 2.34). Note that coded FreqU and coded CT were significantly correlated for our sample ($r_s[44] = 0.983$, $P < 0.01$), which is not surprising since they were derived from the same segmented utterances.

Effects of sex and age. Independent-samples Mann–Whitney *U* tests comparing male versus female participants on ELSA-A length (male: $M = 20.15$, SD = 4.01, Med = 20.81; female: $M = 20.44$, SD = 4.99, Med = 20.88; $U = 210.00$, $P = 0.881$), coded FreqU (male: $M = 3.85$, SD = 3.68, Med = 3.19; female: $M = 3.57$, SD = 4.01, Med = 2.12; $U = 184.00$, $P = 0.617$), and coded CT (male: $M = 2.73$, SD = 2.42, Med = 2.39; female: $M = 2.25$, SD = 2.18, Med = 1.91; $U = 181.00$, $P = 0.565$) showed no sex differences. To check for effects of age, we ran simple linear regressions, regressing each of the measures onto participants' age in months. No regression reached statistical significance (ELSA length: $R^2 = 0.000$, F [1,44] = 0.001, $P = 0.978$; coded FreqU: $R^2 = 0.033$, F [1,44] = 1.494, $P = 0.228$; coded CT: $R^2 = 0.029$, F [1,44] = 1.325, $P = 0.256$).

Evaluation of Measures

Concurrent validity. To validate the measures derived from the coding of ELSA-A against their corresponding transcript-derived measures, we computed intraclass correlation coefficients (ICC) for both frequency of utterances and conversational turns. Coded FreqU ($M = 3.78$, SD = 3.73) was positively correlated with transcribed FreqU ($M = 3.22$, SD = 3.26; ICC = 0.944, $P < 0.01$). Similarly, a high positive intraclass correlation coefficient (ICC = 0.979, $P < 0.01$) was found for coded CT ($M = 2.61$, SD = 2.34) and transcribed CT ($M = 2.42$, SD = 2.21).

Construct validity. To assess the construct validity of the measures derived from the coding of the language samples, we computed correlations between each measure and SCQ communication domain score (computed by combining the responses from questions 2, 3, 4, 5, 6, 20, 21, 22, 23, 24, 25, 34, and 35; coded FreqU: $r_s(39) = 0.686$, $P < 0.01$; coded CT: $r_s(39) = 0.692$, $P < 0.01$), and VABS

Communication Standard Score (coded FreqU: $r_s(44) = 0.613$, $P < 0.01$; coded CT: $r_s(44) = 0.604$, $P < 0.01$).

Test–retest reliability. Using data from the 10 test–retest participants, we computed ICCs between length of ELSA-A, as well as frequency of utterances and conversational turns at test and at retest. ELSA-A length at test ($M = 18.31$, SD = 6.00) and at retest ($M = 18.31$, SD = 6.48) were significantly positively correlated (ICC = 0.853, $P < 0.01$). Similar results were found for coded FreqU (test: $M = 2.97$, SD = 2.49; retest: $M = 4.26$, SD = 3.97; ICC = 0.774, $P < 0.01$) and for coded CT (test: $M = 2.43$, SD = 1.96; retest: $M = 3.62$, SD = 3.23; ICC = 0.798, $P < 0.01$).

Brief Discussion

On average, ELSA-A samples lasted approximately 20 min indicating that the protocol activities successfully engaged our participants for a relatively long period of time. They produced some utterances (3.71 per minute) and took turns (2.61 per minute) during ELSA-A showing that the protocol is effective in eliciting speech even from older children and adolescents who are minimally to low verbal. Furthermore, the protocol is not biased toward a specific age group or participants' sex as evidenced by the lack of age and sex effects on its length and measures. The high administration fidelity suggests that ELSA-A is easy to administer and has good test–retest reliability.

Not only is ELSA-A a protocol appropriate for participants across a wide range of ages and language abilities, but the coding measures derived from it also have good psychometric properties. The very high concurrent validity of the measures shows that the kind of information obtained from coding is very similar, if not identical, to the information obtained from transcription. Therefore, our coding approach can substitute laborious transcription without sacrificing the quality of the data to derive these specific measures.

The measures of language and communication, frequency of utterances per minute and conversational turns per minute, have good construct validity as demonstrated by their significant correlations with participants' SCQ and VABS communication scores. Although the two coding-derived measures we evaluated are highly correlated in our sample, we included both because they hold the potential to capture distinct abilities in other participant populations. For example, in more verbal speakers, the inverse or weaker correlation between frequency of utterances and conversational turns could indicate high verbal ability accompanied with difficulties in the pragmatic domain.

Study 2: ELSA-T

Many studies enroll younger participants with ASD, particularly those focused on measuring change over time as a result of intervention [e.g., Casenhiser et al., 2015; Kaiser & Roberts, 2013; Paul et al., 2013]. Most of these interventions target expressive language, in particular, because of its predictive role in long-term outcomes [Howlin et al., 2004; Venter et al., 1992]. As a result, most of the participants enrolled in these interventions are younger with limited language and/or still in the process of acquiring it. NLS have often been used to obtain language outcome measures [for review, see Barokova & Tager-Flusberg, 2018] but as with research on older participants, different protocols have been used across studies [e.g., Casenhiser et al., 2015; Deitchman, Reeve, Reeve, & Progar, 2010; Kaiser & Roberts, 2013]. Furthermore, the majority of studies collecting NLS from younger children with ASD rely on parents collecting the sample in the context of free play [e.g., Bang & Nadig, 2015; Fusaroli et al., 2019]. Even though free play with developmentally appropriate toys is a naturalistic context, there is less control over how parents might guide the play with their child, which would limit the ability to compare ability and/or change in ability across participants. Therefore, there is a need for a consistent, semi-structured elicitation protocol appropriate for younger children with ASD, who are likely to be enrolled in language treatment and intervention studies. To address this, we adapted ELSA-A, primarily, by changing the materials and activities to be more appropriate for younger children across language ability, while keeping the same general protocol structure.

Methods

Participants and Procedures

IRB approval was obtained prior to enrolling participants. The sample included 19 (5 girls) preschoolers with ASD between 2;8 and 4;10 years old (see Table 4). Participants were administered a battery of standardized assessments, including the ADOS-2 [Lord et al., 2012] or ADOS—Toddler Module [Luyster et al., 2009] and Mullen Scales of Early Learning (MSEL) [Mullen, 1995], and their parents completed the VABS-3 questionnaire [Sparrow, Cicchetti, & Saulnier, 2016]. Trained examiners collected a natural language sample from each participant following the ELSA-T protocol, which was recorded using video and audio formats.

Eliciting Language Samples for Analysis-Toddler (ELSA-T)

ELSA-T was designed specifically for children between the ages of 1.5 and 5 years across a wide range of language ability. As with ELSA-A, it consists of eight activities across two elicitation contexts: free play and narrative, and the examiner should engage the child in conversation between activities. A list of ELSA-T activities and a brief description of each can be found in Table 3. More

Table 4. Demographic Information and Standardized Assessment Scores for 19 ELSA-T Participants (Study 2)

Characteristic/assessment	*N*				
Sex	19	14 Male 5 Female			
Race	19	13 White 1 Black, African American 3 Asian 1 Hispanic 1 Multiple Races			
Ethnicity	19	18 Non-Hispanic 1 Hispanic			
			M	SD	Range
Age (in months)	19		34.37	11.42	20:58
ADOS-2	18	9 Toddler module 5 Mod 1 4 Mod 2 Calibrated severity score (only for Mod 1 and Mod 2)	6	1.73	3:8
MSEL	17	Expressive language raw score	23.00	13.49	4:49
VABS-3	8	Communication standard score	82.38	18.78	57:111

Abbreviations: ADOS-2, Autism Diagnostic Observation Schedule-2; ELSA-T, eliciting language samples for analysis-toddlers; MSEL, Mullen Scales of Early Learning; VABS-3, Vineland Adaptive Behavior Scale-3.

detailed descriptions, administration instructions, and a list of required materials for each ELSA-T activity can be found on our website (https://sites.bu.edu/elsa/; https://sites.bu.edu/elsa/elsa-t/manual-2/).

Coding and measures. The video recordings were used to code administration fidelity out of eight activities. All ELSA-T audio files were coded and transcribed following the procedures described in Study 1. The same spoken language and communication measures, frequency of utterances per minute and number of conversational turns per minute, were extracted from both the coded files and the transcripts.

Results

Evaluation of ELSA-T

Fidelity. In every ELSA-T collected, a trained examiner administered all eight activities. The average ELSA-T duration was 29.84 min (SD = 5.16; range: 19–39). Participants, on average, produced 4.21 (SD = 3.47) speech utterances and 3.34 (SD = 2.69) conversational turns per minute.

Effects of sex and age. To check for effects of sex, we ran independent-samples *t* tests comparing male versus female participants on ELSA-T length (male: $M = 30.12$, SD = 5.40; female: $M = 29.08$, SD = 4.92; $t[17] = 0.375$, $P = 0.713$), coded FreqU (male: $M = 4.41$, SD = 3.57; female: $M = 3.65$, SD = 3.52; $t[17] = 0.412$, $P = 0.686$), and coded CT (male: $M = 3.48$, SD = 2.78; female: $M = 2.95$, SD = 2.70; $t[17] = 0.369$, $P = 0.717$). None of the comparisons reached statistical significance.

To test for effects of age on ELSA-T length and the coding-derived measures, we ran simple linear regressions. We regressed each variable in turn onto participants' age in months. Regressing ELSA-T length in minutes onto participants' age did not reach statistical significance ($R^2 = 0.017$, $F[1,17] = 0.297$, $P = 0.593$). However, age was a significant predictor of both coded FreqU ($R^2 = 0.322$, $F[1,17] = 8.086$, $P = 0.011$) and of coded CT ($R^2 = 0.311$, $F[1,17] = 7.688$, $P = 0.013$).

Evaluation of Measures

Concurrent validity. To validate the measures derived from coding of ELSA-T against their corresponding transcript-derived measures, we ran ICCs between them. Coded FreqU ($M = 4.21$, SD = 3.47) was positively correlated with transcribed FreqU ($M = 4.32$, SD = 3.08; $ICC = 0.941$, $P < 0.01$). A similarly high, positive ICC (ICC = 0.919, $P < 0.01$) was found for coded CT ($M = 3.34$, SD = 2.69) and transcribed CT ($M = 3.83$, SD = 2.54).

Construct validity. To assess the construct validity of the measures derived from coding, we ran correlations between FreqU and CT and the MSEL Expressive Language Raw Score. Coded FreqU and CT were both positively correlated with MSEL Expressive Language Raw Score (coded FreqU: $r_s(15) = 0.897$, $P < 0.01$; coded CT: $r_s(15) = 0.886$, $P < 0.01$).

Brief Discussion

The findings from ELSA-T replicate what we found for ELSA-A with respect to fidelity of administration, ability

to engage the children, and the validity of the coding measures of utterance frequency and conversational turns. As with ELSA-A, there were no differences between boys and girls on any measure; however, older preschoolers spoke more (higher frequency of utterances) and took more conversational turns than younger preschoolers. This is not surprising since we would expect that this age range reflects a prime period during which language is being acquired, especially by children with ASD enrolled in specialized intervention programs, which was the case for the majority of our participants. The fact that age was a significant predictor suggests that ELSA-T and the measures derived from it are developmentally sensitive and could be used to chart changes over time within children, not just across children differing in age.

Study 3: Comparing ELSA-A and ELSA-T

Many studies of ASD or related neurodevelopmental disorders may want to enroll participants spanning a wide age range, from young toddlers to older adolescents, in a cross-sectional design, or to track developmental change longitudinally from the early years through childhood or beyond [cf. Bal et al., 2019]. Because it would not be appropriate to use the same materials at all ages, we developed two equivalent protocols described in Studies 1 and 2: ELSA-A and ELSA-T. If the two ELSA protocols can be shown to be equivalent, they could be used in studies tracking participants over long periods of time and developmental stages from toddlerhood, when ELSA-T is appropriate, to adolescence, when ELSA-A is appropriate. Considering this advantage, in our final study, we directly compared the two versions of ELSA to evaluate whether they are truly equivalent.

Participants and Procedures

IRB approval was obtained prior to enrolling participants. The sample included 17 (5 girls) young children with ASD between the ages of 4;0 and 6;8 years (see Table 5). This age range was selected because the materials in both versions of ELSA could appeal to the children.

During a single visit to the lab, participants were administered the ADOS-2 [Lord et al., 2012] and Leiter-R [Roid & Miller, 1997], and their parents completed the VABS-3 questionnaire [Sparrow et al., 2016]. In addition, ELSA-A and ELSA-T were administered to the participants by the same examiner. The order of ELSA-A and ELSA-T was counterbalanced across participants, and there was always a break of at least 1 hr between the two protocols.

Coding. Audio recordings of ELSA-A and ELSA-T were coded in ELAN following the same procedures as described in Study 1. Different coders coded the ELSA-A and ELSA-T file for each participant. Length of the language sample in minutes was extracted. The coded files were used to derive the same spoken language and communication measures, coded FreqU and coded CT, for the whole language sample and for each of the eight activities.

Results

General Comparison

First, we ran paired-samples *t* tests to compare the general equivalence of the ELSA-A and ELSA-T protocols in terms

Table 5. Demographic Information and Standardized Assessment Scores for 17 ELSA-A–ELSA-T Participants (Study 3)

Characteristic/assessment	*N*				
Sex	17	12 Male 5 Female			
Race	16	10 White 2 Black, African American 3 Asian 0 Hispanic, Latino, or Spanish Origin 1 Multiple races			
Ethnicity	16	16 Non-Hispanic 0 Hispanic 0 Prefer not to respond			
			M	SD	Range
Age (in months)	17		62.12	9.61	48:80
ADOS-2	17	7 Mod 1 1 Mod 2 9 Mod 3			
		Calibrated severity score	5.94	1.60	4:9
Leiter-R	14	Nonverbal IQ	101.00	8.21	82:115
VABS-3	15	Communication standard score	79.47	22.65	40:132

Abbreviations: ADOS-2, Autism Diagnostic Observation Schedule-2; ELSA-A, eliciting language samples for analysis-adolescents; ELSA-T, eliciting language samples for analysis-toddlers; Leiter-R, Leiter International Performance Scales-Revised; VABS-2, Vineland Adaptive Behavior Scale-2.

of fidelity of administration, length of the samples, coded FreqU and coded CT (see Table 6). ELSA-A and ELSA-T did not significantly differ on any of these characteristics and measures (administration fidelity: $t[16] = -1.461$, $P = 0.163$; length: $t[16] = 1.444$, $P = 0.168$; coded FreqU: $t[16] = -1.700$, $P = 0.108$; coded CT: $t[16] = -1.969$, $P = 0.067$).

Activity by Activity Comparison

Next, we compared the equivalence of the specific matched activities across protocols (e.g., Leaf Falling in ELSA-A vs. Apple Falling in ELSA-T) in terms of coded FreqU (see Table 7) and coded CT produced by participants by running paired-samples t-tests. No differences were found for any of the matched activities.

Brief Discussion

This study confirmed that the two language elicitation protocols were easy to administer and were similar in duration. The protocols elicited a comparable number of speech utterances per minute and conversational turns per minute from 4- to 7-year-old children with ASD. Our activity-by-activity analyses showed that all pairs of activities elicited comparable frequency of utterances and conversational turns. Thus, the activity that elicited the fewest speech utterances per minute in ELSA-T, the Storybook activity, corresponded to the ELSA-A activity that elicited fewest utterances, the Pixar movie shorts activity. A similar pattern was observed for the activities that elicited highest and lowest conversational turns per minute—they corresponded across the ELSA protocols. Overall, the two ELSA protocols were found to be equivalent in terms of administration and in terms of the language that they elicit from children with ASD between the ages of 4 and 7 years.

Table 6. ELSA-A–ELSA-T Comparison on 17 Participants (Study 3)

	ELSA-A, *M* (SD)	ELSA-T, *M* (SD)
Fidelity of administration	7.88 (0.33)	8 (0)
Length in minutes	22.81 (4.18)	21.27 (4.47)
Frequency of utterances per minute	5.74 (3.00)	6.61 (3.32)
Conversational turns per minute	4.94 (2.53)	5.58 (2.52)

Abbreviations: ELSA-A, eliciting language samples for analysis-adolescents; ELSA-T, eliciting language samples for analysis-toddlers.

General Discussion

Our studies have two main findings: (a) ELSA is a language elicitation protocol suitable for children and adolescents across a wide range of abilities and ages. Importantly, ELSA-A is the first protocol to be introduced for older minimally verbal children and adolescents, and ELSA-T, which we have shown to be equivalent to ELSA-A, is developmentally sensitive for young toddlers and preschoolers. (b) Real-time coding of the ELSA protocols yields psychometrically sound measures of language and communication. These results offer an opportunity for future research studies focusing on expressive language in the context of measuring change over time, both in natural contexts and in clinical trials, to evaluate children and adolescents with ASD with objective and valid measures.

Our first finding addresses a significant gap in the field of language and communication in ASD, namely, the lack of assessment tools appropriate across a wide range of ages and language abilities, especially for those who are

Table 7. ELSA-A-ELSA-T Comparison by Activity on Frequency of Utterances Per Minute of 17 Participants (Study 3)

	ELSA-A, *M* (SD)	ELSA-T,*M* (SD)	Comparison significance
Leaf falling versus apple falling	6.15 (1.06)	8.15 (1.48)	$t(16) = 1.705$ $P = 0.108$
Planting an acorn versus picnic adventure	6.50 (1.13)	7.33 (1.30)	$t(16) = 0.931$ $P = 0.366$
Discovering animals versus. hide and seek animals	6.71 (0.78)	7.73 (1.27)	$t(16) = 1.028$ $P = 0.319$
Helping animals versus bath time	5.78 (0.84)	6.21 (0.82)	$t(16) = 0.507$ $P = 0.619$
S'more versus snack	6.00 (1.00)	5.75 (0.90)	$t(16) = -0.203$ $P = 0.842$
Crafts versus craft time	6.70 (0.85)	6.26 (0.74)	$t(16) = -0.576$ $P = 0.573$
Bean bag toss versus turtle toss	7.59 (0.93)	7.77 (1.14)	$t(16) = 0.418$ $P = 0.682$
Pixar movie shorts versus story book	4.25 (0.82)	5.20 (1.12)	$t(16) = 0.952$ $P = 0.355$

Note. All analyses were corrected for multiple comparisons.
Abbreviations: ELSA-A, eliciting language samples for analysis-adolescents; ELSA-T, eliciting language samples for analysis-toddlers.

minimally to low-verbal. Our protocol, ELSA-A, is suitable for children and adolescents between the ages of 4 or 5–20 across a wide range of abilities, and ELSA-T can be used in the same way for younger toddlers and children. Both protocols successfully engage children and adolescents for about 20 min or longer. This should be a sufficient amount of time to evaluate expressive language abilities considering that 1-, 3-, and 7-min-long language samples have been shown to yield reliable language measures [Heilmann, Nockerts, & Miller, 2010]. Preliminary results in our studies show that neither version of ELSA is biased toward males or females; the carefully chosen activities are engaging for both sexes. While ELSA was developed and evaluated here on individuals with ASD, it could easily be used with individuals with other developmental disorders with or without language impairments or with individuals with some degree of language impairment.

Because we have shown that ELSA-A and ELSA-T are equivalent with respect to ease of administration, length of language samples, and measures of both frequency of utterances and conversational turns, they can be used within the same study. This is of particular importance in the context of longitudinal studies tracking change over time and in early treatment and intervention studies. For intervention studies that focus on children in the age range of about 4–7, the two protocols could be used interchangeably at preintervention and postintervention. Relying on two different but equivalent protocols circumvents the risk of practice effects.

Our second major finding is of formidable practical importance. Traditionally, once collected, language samples are transcribed, from which key measures can be derived. However, transcription takes a long time and is costly, which can present an obstacle to the widespread implementation of NLS in ASD research, particularly in clinical trials that often enroll a large number of participants. While the hope is that technology will eventually solve the transcription problem, at this time, there are no reliable means to automatically identify and segment two speakers engaged in a naturistic interaction that will work with the broader ASD population [cf. Jones et al., 2019]. The results of our studies provide a different solution to this problem that capitalizes on the ease with which even naïve listeners can swiftly identify different speakers [Perrachione, Furbeck, & Thurston, 2019]. Our coding of the collected language samples, which takes approximately as long as the duration of the sample, yields the same measures of language and communication as those derived from a more complete transcription process. Moreover, the measures coded in real time as one is listening to the audio recording of the sample, frequency of utterances and conversational turns per minute, which have been quite widely used in ASD research [Barokova & Tager-Flusberg, 2018], were shown to be valid against parent report measures of language and communication for both versions of ELSA. Therefore, coding of ELSA and the measures derived from it can be used to quickly and reliably assess and characterize the language and communication of children across a wide range of ages and abilities. One limitation is that we have not established norms for the coded measures. Such norms for typically developing children could be computed using the SALT reference database [Miller et al., 2011] or derived from corpora of transcripts from the CHILDES database [MacWhinney, 2000]. Such norms would allow for a comparison of language ability between children with ASD and typically developing children. Nevertheless, the lack of norms does not preclude the use of our coding-derived measures for assessing within-child changes over time or in intervention studies [cf. Abbeduto et al., 2020].

There are other advantages of using ELSA to evaluate expressive language abilities, especially in the context of a clinical trial. It is relatively easy to become reliable in its administration and naïve coders can be trained to complete the segmentation of speech utterances fairly quickly. Compared to other assessment tools, the costs associated with ELSA are quite low. Importantly, it can be used to provide objective, blinded assessments of outcomes. Neither the examiners collecting the protocol nor the coders need to be involved in the actual intervention (behavioral or medical). In particular, the coders can be blind not only to which arm of a trial the participant is enrolled in, but also to when the ELSA sample was collected, preintervention or postintervention.

Overall, the ELSA protocols not only try to address assessment gap in the field of ASD but also have the potential to address another gap: the dearth of cross-cultural and cross-linguistic studies. Even though difficulties in social communication are one of the defining criteria for ASD, very little is known about how these may be characterized across different cultural traditions and linguistic practices. The vast majority of ASD research is conducted with English-speaking participants, and most assessment tools are designed for and normed on this same population, which limits their use. In contrast, both ELSA protocols have the potential to be easily tailored to the cultural context of participants. The key to such adaptations is to keep the same high number of activities because the variety of toys and materials are what keeps participants engaged for such a long time. It is also crucial to try and find *equivalent* activities for the specific cultural context. For example, instead of making a S'more or a peanut butter and jelly sandwich the examiner should substitute these snacks with the preferred snack for children and adolescents in the specific region.

In any case, it is essential to monitor the fidelity of administration of the protocols. For the protocols to successfully elicit speech from the most heterogeneous participant samples, it is important to guide administrators

in engaging participants through all the activities and in using open-ended prompts and conversational bids as an avenue to elicit language.

In addition to adapting the ELSA protocols, the measures that can be derived from them can also be tailored to the goals of the research. The language samples can be transcribed and coded for more detailed linguistic analyses (including other traditional measures such as number of different words used), they can be coded for global measures of language and communication as one is listening to a recording of the sample, or they can be coded for a specific measure of interest, for example, social communicative utterances, intelligibility, or stereotyped speech. In addition, if video recordings of the ELSA samples are available, they can be used to code other aspects of communication and behavior such as the use of communicative gestures, use of signs or sign language. Regardless of the measures of choice, future studies should examine their psychometric properties before implementing them.

Limitations and Directions for Future Research

Although very promising, our studies possess a number of limitations that need to be acknowledged and could be used to inform future research. One limitation pertains to our participant samples in terms of both their sizes and their composition. Our conclusions about the characteristics of the ELSA protocols (e.g., length, administration fidelity) and the psychometric properties of the expressive language measures derived from them (e.g., good reliability and validity), should be interpreted within the characteristics of our participants. That is, in Study 1 we have shown that frequency of utterances per minute and conversational turns per minute derived from ELSA-A have good test–retest reliability and concurrent and construct validity for participants, who are minimally to low-verbal and span within a very wide range of ages. Our relatively small sample sizes in Studies 2 and 3, in particular, also preclude more sophisticated analyses to determine whether these psychometric properties vary as a function of participant characteristics that we have measured, like IQ and ASD symptom severity. Our small sample sizes make our conclusions about the lack of sex effects of the protocol rather preliminary. Thus, more research is needed to determine whether the characteristics of the ELSA protocols and the psychometric properties of the measures derived from them will hold with speakers with very different characteristics, for example, more verbally fluent adolescents.

Another area for future research is related to the choice of measures derived from ELSA samples. Indeed, in our studies we have focused on basic measures of expressive language and communication: frequency of utterances and conversational turns per minute. These measures are suitable for studies examining the speech of speakers across a very wide range of ability, including those who are minimally to low verbal. However, other measures of language and communication, like MLU for syntax, NDW for semantics or pragmatic coding of utterance function might be more appropriate when sampling speech from more verbally fluent speakers, and yet other measures like classification of gestures and coding of eye gaze might be more appropriate for nonverbal individuals. Future studies should (a) determine whether such measures can be obtained using our ELSA coding approach, and (b) test the psychometric properties of these measures as derived from ELSA. The choice of measures should ultimately be informed by the goals of the study. Furthermore, even if the measures cannot be coded using our approach, ELSA could still be an appropriate context to collect the language sample.

In our studies, by choosing frequency of utterances per minute and conversational turns per minute we wanted to capture participants' general expressive language ability, as well as their back-and-forth engagement with the examiner. Considering our participants' very limited verbal abilities, it is no surprise that these two measures were very highly correlated. Nevertheless, we have included both because they hold the potential to distinguish between participants who engage more in back-and-forth conversation and those who vocalize more without regard for their conversational partner's utterances. Future studies should examine whether these two measures are sensitive enough to aid in such a distinction.

Another direction for future research is related to examining our measures' sensitivity to change. Across our three studies, we have shown that frequency of utterances and conversational turns per minute have good test–retest reliability, concurrent validity and construct validity. These findings lay the foundation for future studies to examine whether these measures are sensitive to change both as a result of development and as a result of treatment and intervention.

In summary, the ELSA protocols are appropriate for use with children and adolescents across a wide range of abilities, and the broad measures we derived from them have sound psychometric properties. The two protocols, ELSA-A and ELSA-T, are equivalent at eliciting speech from younger children and can thus potentially be used to track change in ability in longitudinal studies. The ELSA protocols can be easily adapted to fit the cultural context of participants, and the measures derived from them can be tailored to the goals of researchers, thus allowing for great flexibility in their use in future research.

Acknowledgments

We thank Dr. Catherine Lord for her wise advice during the development of this project. We are also grateful to our

many coders and transcribers, including Erin Anderson, Phanirekha Chandra, Naomi Chinama, Megan Chung, Laura Doherty, Jared Goldberg, Kloe Hidri, Emily Jackson, Aushy Kaul, Carly Levine, Julia Masterson, Lacey Raymond, Jing Lian Shu, Lucy Stowe, Jessica Tharaud, Vanessa Torrice, Maya Vasishth, Danielle Vitorelo, Rachel Watson, and Ellen Wilkinson. This research was supported by grants from the Simons Foundation (SFARI#383660 and #641201) and from NIH (PO1 HD 64653 and P50 DC 13027) to HTF.

Conflict of interest

The authors declare that they have no conflict of interest. Information about ELSA can be found on our website, https://sites.bu.edu/elsa/, and is not a commercially available product.

References

Abbeduto, L., Berry-Kravis, E., Sterling, A., Sherman, S., Edgin, J. O., McDuffie, A., … 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, 10. https://doi.org/10.1186/s11689-020-09313-6

Bal, V. H., Katz, T., Bishop, S. L., & Krasileva, K. (2016). Understanding definitions of minimally verbal across instruments: Evidence for subgroups within minimally verbal children and adolescents with autism spectrum disorder. Journal of Child Psychology and Psychiatry, 57(12), 1424–1433. https://doi.org/10.1111/jcpp.12609

Bal, V. H., Maye, M., Salzman, E., Huerta, M., Pepa, L., Risi, S., & Lord, C. (2019). The adapted ADOS: A new module set for the assessment of minimally verbal adolescents and adults. Journal of Autism and Developmental Disorders, 50, 719–729. https://doi.org/10.1007/s10803-019-04302-8

Bang, J., & Nadig, A. (2015). Learning language in autism: Maternal linguistic input contributes to later vocabulary. Autism Research, 8, 214–223. https://doi.org/10.1002/aur.1440

Barokova, M., & Tager-Flusberg, H. (2018). Commentary: Measuring language change through natural language samples. Journal of Autism and Developmental Disorders, 50, 1–20. https://doi.org/10.1007/s10803-018-3628-4

Berry-Kravis, E., Des Portes, V., Hagerman, R., Jacquemont, S., Charles, P., Visootsak, J., … Von Raison, F. (2016). Mavoglurant in fragile X syndrome: Results of two randomized, double-blind, placebo-controlled trials. Science Translational Medicine, 8(321), 321ra5. https://doi.org/10.1126/scittranslmed.aab4109

Brown, R. (1973). A first language. Cambridge, MA: Harvard University Press.

Casenhiser, D. M., Binns, A., McGill, F., Morderer, O., & Shanker, S. G. (2015). Measuring and supporting language function for children with autism: Evidence from a randomized control trial of a social-interaction-based therapy. Journal of Autism and Developmental Disorders, 45(3), 846–857. https://doi.org/10.1007/s10803-014-2242-3

Chiang, H.-M. (2009). Differences between spontaneous and elicited expressive communication in children with autism. Research in Autism Spectrum Disorders, 3(1), 214–222. https://doi.org/10.1016/j.rasd.2008.06.002

Colle, L., Baron-Cohen, S., Wheelwright, S., & Van Der Lely, H. K. J. (2008). Narrative discourse in adults with high-functioning autism or Asperger syndrome. Journal of Autism and Developmental Disorders, 38(1), 28–40. https://doi.org/10.1007/s10803-007-0357-5

Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. American Journal of Speech-Language Pathology, 12(3), 349–358. https://doi.org/10.1044/1058-0360(2003/080)

Deitchman, C., Reeve, S. A., Reeve, K. F., & Progar, P. R. (2010). Incorporating video feedback into self-management training to promote generalization of social initiations by children with autism. Education and Treatment of Children, 33(3), 475–488.

Drew, A., Baird, G., Taylor, E., Milne, E., & Charman, T. (2007). The social communication assessment for toddlers with autism (SCATA): An instrument to measure the frequency, form and function of communication in toddlers with autism spectrum disorder. Journal of Autism and Developmental Disorders, 37, 648–666. https://doi.org/10.1007/s10803-006-0224-9

Friedman, L., Sterling, A., Dawalt, L., & Mailick, M. (2019). Conversational language is a predictor of vocational independence and friendships in adults with ASD. Journal of Autism and Developmental Disorders, 49(10), 4294–4305. https://doi.org/10.1007/s10803-019-04147-1

Fusaroli, R., Weed, E., Fein, D., & Naigles, L. (2019). Hearing me hearing you: Reciprocal effects between child and parent language in autism and typical development. Cognition, 183, 1–18.

Guastella, A., Gray, K., Rinehart, N., Alvares, G., Tonge, B., Hickie, I., & Einfeld, S. (2015). The effects of a course of intranasal oxytocin on social behaviors in youth diagnosed with autism spectrum disorders: A randomized controlled trial. Journal of Child Psychology and Psychiatry, 56(4), 444–452. https://doi.org/10.1111/jcpp.12305

Heilmann, J., Nockerts, A., & Miller, J. (2010). Language sampling: Does the length of the transcript matter? Language, Speech & Hearing Services in Schools (Online), 41(4), 393-404A, 404.

Hogan-Brown, A. L., Losh, M., Martin, G. E., & Mueffelmann, D. J. (2013). An investigation of narrative ability in boys with autism and fragile X syndrome. American Journal on Intellectual and Developmental Disabilities, 118(2), 77–94. https://doi.org/10.1352/1944-7558-118.2.77

Howlin, P., Goode, S., Hutton, J., & Rutter, M. (2004). Adult outcome for children with autism. Journal of Child Psychology and Psychiatry, 45(2), 212–229. https://doi.org/10.1111/j.1469-7610.2004.00215.x

Jones, R., Skwerer, D., Pawar, R., Hamo, A., Carberry, C., Ajodan, E., … Tager-Flusberg, H. (2019). How effective is LENA in detecting speech vocalizations and language produced by children and adolescents with ASD in different contexts? Autism Research, 12(4), 628–635. https://doi.org/10.1002/aur.2071

Kaiser, A., & Roberts, M. (2013). Parent-implemented enhanced milieu teaching with preschool children with intellectual

disabilities. Journal of Speech, Language, and Hearing Research, 56(1), 295–309.

Kasari, C., Gulsrud, A., Wong, C., Kwon, S., & Locke, J. (2010). Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism. Journal of Autism and Developmental Disorders, 40, 1045–1056. https://doi.org/10.1007/s10803-010-0955-5

Kasari, C., Kaiser, A., Goods, K., Nietfield, J., Mathy, P., Landa, R., … Almirall, D. (2014). Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. Journal of the American Academy of Child and Adolescent Psychiatry, 53(6), 635–646. https://doi.org/10.1016/j.jaac.2014.01.019

Kover, S., Davidson, M., Sindberg, H., & 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. https://doi.org/10.1044/2014_JSLHR-L-13-0330

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. https://doi.org/10.1111/j.1365-2788.2010.01255.x

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods, 41(3), 841–849. https://doi.org/10.3758/BRM.41.3.841

Loban, W. (1976). Language development: Kindergarten through grade twelve. Urbana, IL: National Council of Teachers of English.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). Autism diagnostic observation schedule (2nd ed.). Torrance, CA: Western Psychological Services.

Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., … Lord, C. (2009). The autism diagnostic observation schedule-toddler module: A new module of a standardized diagnostic measure for autism spectrum disorders. Journal of Autism and Developmental Disorders, 39(9), 1305–1320.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk: Transcription format and programs (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

MacWhinney, B. (2007). The TalkBank project. In Beal, J., Corrigan, K. & Moisl, L. Creating and digitizing language corpora: Synchronic databases (Vol. 1). Houndmills, Basingstoke, Hampshire, Palgrave-Macmillan.

Masi, A., Lampit, A., Glozier, N., Hickie, I. B., & Guastella, A. J. (2015). Predictors of placebo response in pharmacological and dietary supplement treatment trials in pediatric autism spectrum disorder: A meta-analysis. Translational Psychiatry, 5(9), E640. https://doi.org/10.1038/tp.2015.143

Miller, J., Andriacchi, K., & Nockerts, A. (2011). Assessing language production using SALT software: A clinician's guide to language sample analysis. Madison, WI: SALT Software.

Miller, J., & Iglesias, A. (2012). Systematic analysis of language transcripts (SALT), research version 2012 [Computer Software]. Middleton, WI: SALT Software, LLC.

Morley, E., Roark, B., van Santen, J. (2013) The utility of manual and automatic linguistic error codes for identifying neurodevelopmental disorders. Paper presented at the Proceedings of the Eigth Workshop on Innovative Use of Natural Language Processing for Building Educational Applications; Atlanta, Georgia. 2013.

Mullen, E. M. (1995). Mullen scales of early learning (pp. 58–64). Circle Pines, MN: AGS.

Mundy, P., Delgado, C., Block, J., Venezia, M., Hogan, A., & Seibert, J. (2003). A Manual for the Abridged Early Social Communication Scales (ESCS). Coral Gables, FL: University of Miami. Unpublished manuscript.

Park, C. J., Yelland, G. W., Taffe, J. R., & Gray, K. M. (2012). Morphological and syntactic skills in langauge samples of preschool aged children with autism: Atypical development? International Journal of Speech-Language Pathology, 14(2), 95, 22390743–108.

Pasco, G., Gordon, R. K., Howlin, P., & Charman, T. (2008). The classroom observation schedule to measure intentional communication (COSMIC): An observational measure of the intentional communication of children with autism in an unstructured classroom setting. Journal of Autism and Developmental Disorders, 38(10), 1807–1818. https://doi.org/10.1007/s10803-008-0569-3

Paul, R., Campbell, D., Gilbert, K., & Tsiouri, I. (2013). Comparing spoken language treatments for minimally verbal preschoolers with autism spectrum disorders. Journal of Autism and Developmental Disorders, 43(2), 418–431. https://doi.org/10.1007/s10803-012-1583-z

Perrachione, T., Furbeck, K., & Thurston, E. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. The Journal of the Acoustical Society of America, 146(5), 3384–3399. https://doi.org/10.1121/1.5126697

Roid, G. H., & Miller, L. J. (1997). Leiter International Performance Scale-Revised. Wood Dale, IL: Stoelting.

Roos, E. M., McDuffie, A. S., Weismer, S. E., & Gernsbacher, M. A. (2008). A comparison of contexts for assessing joint attention in toddlers on the autism spectrum. Autism, 12(3), 275–291. https://doi.org/10.1177/1362361307089521

Rutter, M., Bailey, A., & Lord, C. (2003). Social communication questionnaire. Los Angeles, CA: Western Psychological Services.

Schoen, E., Paul, R., & Chawarska, K. (2011). Phonology and vocal behavior in toddlers with autism spectrum disorders. Autism Research, 4(3), 177–188.

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). Vineland Adaptive Behavior Scales (3rd ed.) (Vineland-3). San Antonio, TX: Pearson.

Sparrow, S. S., Cicchetti, V. D., & Balla, A. D. (2005). Vineland Adaptive Behavior Scales (2nd ed.). Circle Pines, MN: American Guidance Service.

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. https://doi.org/10.1007/s10803-014-2042-9

Tager-Flusberg, H., & Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. Autism Research, 6(6), 468–478. https://doi.org/10.1002/aur.1329

Tager-Flusberg, H., Rogers, S., Cooper, J., Landa, R., Lord, C., Paul, R., … 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. https://doi.org/10.1044/1092-4388(2009/08-0136)

Venter, A., Lord, C., & Schopler, E. (1992). A follow-up study of high-functioning autistic children. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 33(3), 489–507. https://doi.org/10.1111/j.1469-7610.1992.tb00887.x

Wetherby, A. M., & Prizant, B. M. (2002). CSBS DP manual: Communication and Symbolic Behavior Scales: Developmental profile. Baltimore, MD: Paul H. Brookes Pub.

Woynaroski, T., Oller, D. K., Keceli-Kaysili, B., Xu, D., Richards, J. A., Gilkerson, J., … Yoder, P. (2017). The stability and validity of automated vocal analysis in preverbal preschoolers with autism spectrum disorder. Autism Research, 10(3), 508–519. https://doi.org/10.1002/aur.1667

Yoder, P. J., Oller, D. K., Richards, J. A., Gray, S., & Gilkerson, J. (2013). Stability and validity of an automated measure of vocal development from day-long samples in children with and without autism spectrum disorder. Autism Research, 6(2), 103–107. https://doi.org/10.1002/aur.1271