Journal of Child Language (2024), 1–14
doi:10.1017/S0305000924000643

CAMBRIDGE UNIVERSITY PRESS

BRIEF RESEARCH REPORT

Remote collection of language samples from three-year-olds

Jinyoung Jo and Megha Sundara

UCLA Department of Linguistics, 3125 Campbell Hall, Los Angeles, CA 90095-1543, USA
Corresponding author: Jinyoung Jo; Email: jinyoungjo@ucla.edu

(Received 09 February 2024; revised 04 November 2024; accepted 05 November 2024)

Abstract

We characterised language samples collected remotely from typically developing three-year-olds by comparing them against independent language samples collected in person from age-matched peers with and without language delays. Forty-eight typically developing, English-learning three-year-olds were administered a picture description task via Zoom. The in-person comparison groups were two sets of independent language samples from age-matched typically developing as well as language-delayed children available on the Child Language Data Exchange System. The findings show that although language samples collected remotely from three-year-olds yield numerically dissimilar lexical and grammatical measures compared to samples collected in person, they still consistently distinguish toddlers with and without language delays.

Keywords: remote language assessment; language sample analysis; language delays

Introduction

There is increasing interest in remote language assessment, particularly in light of the COVID-19 pandemic and the rise of telehealth. Remote language assessment offers numerous advantages, one of them being the significant decrease in travel time for patients and families alongside increased accessibility of clinical services. Additionally, remote assessments can lower the costs associated with in-person sessions, including the maintenance of physical lab spaces.

However, conducting language assessments remotely comes with several challenges. First, participants must have access to the necessary technology, which can be a barrier to participation. Any method of language assessment typically requires a stable internet connection and high quality of audio recordings to preserve the intelligibility of children's speech. Additionally, young children may struggle with tasks on a digital screen, and there are more likely to be distractions in the home environment, for example, toys and family members, making it challenging to accurately evaluate their language abilities. Despite their challenges, remote language assessments are a tool available to clinicians, and therefore it is imperative to understand the extent to which such measures are similar

to or different from traditional in-person assessments to determine their clinical utility in identifying children at risk for language impairment.

Over the last several years, research comparing in-person and online language assessment is emerging. Such comparisons are now available for children around four years and older, both typically developing (Dam & Pham, 2023; Manning et al., 2020; McElwain et al., 2022; Pratt et al., 2022) and those at risk for language impairment (Magimairaj et al., 2022; Sutherland et al., 2017; Waite et al., 2010). The focus in these comparisons has typically been on evaluating the reliability and validity of standardised assessment tools (for an overview, see special issue edited by Peña and Sutherland (2022)). For instance, based on these studies, we know that remote assessments of vocabulary, morphosyntax, narrative, and nonverbal IQ are highly correlated with in-person assessments for children with and without language impairment (Schmitt et al., 2022; Pratt et al., 2022; Castilla-Earls et al., 2022; Magimairaj et al., 2022). The high correlation between assessments administered in person and remotely is perhaps unsurprising, given that assessments are highly structured, with only limited response types. Additionally, remote assessments of receptive vocabulary, phonological awareness, and conceptual print knowledge across time show similar growth slopes in typically hearing as well as deaf children with hearing aids and cochlear implants (Lund & Werfel, 2022). Therefore, such assessments are reliable for assessing vocabulary in children with language impairment and comparable at identifying language and literacy disorders (Nelson & Plante, 2022). These are all necessary first steps to inform clinicians about the potential as well as the limitations of remote assessments conducted using Zoom, Qualtrics, or PowerPoint.

In the clinical setting, however, language sample analyses are widely used to supplement norm-referenced standardised tests for identifying language disorders, as they can provide a more comprehensive view of the patient's language profile to aid in accurate diagnosis. Many children with a small vocabulary during their early years tend to catch up with their peers by age 3 (Leonard, 1998; Rescorla & Lee, 2000), and those who do not face a higher probability of experiencing long-term language impairment (Rescorla & Lee, 2000; Rescorla & Schwartz, 1990). As a result, language samples elicited from three-year-olds, at least in person, are very informative for assessing the risk of language impairment. However, language samples, whether elicited during free play, picture description, narrative retelling, or questions and answers, are fundamentally less structured and more open-ended than the assessments discussed above. Therefore, it is not clear whether language samples elicited remotely have the potential to provide supplementary information to aid in accurate diagnosis. To address this gap, we present a qualitative and quantitative comparison of language samples elicited from three-year-olds remotely and in person.

There is a long tradition of using quantitative measures from language sample analyses to identify infants at risk for language impairment (Barokova & Tager-Flusberg, 2020; Hux et al., 1993; Kemp & Klee, 1997; Loeb et al., 2000). Quantitative measures calculated from language samples include measures of lexical diversity, such as type-token ratio and the number of different words (Finestack & Satterlund, 2018; Watkins et al., 1995), mean length of utterance (MLU; e.g., Eisenberg et al., 2001; Loeb et al., 2000), and the Index of Productive Syntax (IPSyn; Scarborough, 1990), a measure of grammatical complexity based on scoring of 60 syntactic forms across four categories, namely, noun phrases, verb phrases, questions/negations, and sentence structures.

To be most useful for risk assessment, ideally, a language outcome measure should exhibit improvement as children grow older, especially in typically developing children. It should also consistently differentiate between typically developing children and those

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

with language impairments. Further, outcomes should evaluate a specific aspect of development (e.g., lexical) without being influenced by other factors (e.g., grammar; Yang, Rosvold et al., 2022). Therefore, both lexical and grammatical outcome measures are typically derived from language samples to provide a composite view of language development in three-year-olds.

In the present study, our primary goal was to assess how language samples obtained through remote sessions compared to those collected in person. This comparison is particularly relevant due to the growing popularity of telehealth practices. Additionally, as a secondary goal, we sought to identify measures that are most useful in detecting differences between typically developing children and those with language delays. For this purpose, we elicited language samples from typically developing three-year-olds using a picture description task administered via a 20-minute session on Zoom. We compared several outcomes from this set of remote recordings to those from independent samples. These independent samples were collected in person and are available on the Child Language Data Exchange System (CHILDES) database. We included all available samples from three-year-olds, both typically developing and language-delayed toddlers, recorded in a session lasting from 10 to 30 minutes. Our comparison thus involved three groups: our sample elicited remotely from Typically Developing toddlers, CHILDES Typically Developing toddlers, and CHILDES Language Delayed toddlers.

From the language sample, we calculated the number of different words in the first 100 words (NDW; Watkins et al., 1995) as a proxy for lexical diversity, as well as two measures of grammatical development, MLU and IPSyn. We decided not to include the Type Token Ratio (TTR), which is commonly used as a measure of lexical diversity, because of concerns raised against using TTR for clinical purposes (Charest & Skoczylas, 2019; Charest et al., 2020; Yang, Rosvold et al., 2022). TTR is highly influenced by the size of the language sample and does not consistently increase with age, even in typically developing children (Templin, 1957; Yang, Rosvold et al., 2022). Importantly, TTR does not effectively distinguish between typically developing children and those with language impairment (Watkins et al., 1995; Yang, Rosvold et al., 2022).

In sum, assessing children as young as three years old, who may not be familiar with remote testing setups, presents unique challenges. Thus, it is not certain whether language samples can be reliably collected from such young children in a remote session. The goal of the present study was to (a) characterize language samples obtained from remote sessions, (b) determine how they compare to language samples collected in person in order to assess the feasibility and clinical utility of remote elicitations from three-year-olds, and (c) identify measures that are most useful in detecting differences between children with and without language delays.

Methods

Subjects

During the COVID-19 pandemic, we recruited 48 monolingual English-learning three-year-olds (female = 24; age mean = 36.6, range = 34.7–38.2; percent English exposure mean = 98.8, range = 90–100) between Feb. 2021 and June 2022 to evaluate various measures of language outcomes. All assessments were conducted remotely. None of the children had any reported hearing, speech, or language impairments.

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

Comparison groups

We compared our language samples, elicited remotely, against two sets of independent language samples obtained in person. The in-person samples included all language samples from all 35 to 37-month-olds available on CHILDES, typically developing as well as children with language delays (Ambrose, 2016; Bang & Nadig, 2015; Bliss, 1988; Conti-Ramsden & Dykins, 1991; Conti-Ramsden et al., 1995; Conti-Ramsden & Jones, 1997; Eisenberg & Guo, 2013; Feldman et al., 1989; Hargrove et al., 1986; Keefe et al., 1989; Nicholas & Geers, 1997; Rescorla et al., 2000; Rollins, 1999). This dataset includes transcripts from 74 typically developing toddlers and 102 toddlers with language delays. The most common elicitation method in this independent sample was free play with parents using a set of toys, with some instances of book reading and a few instances of elicited picture descriptions by clinicians. We only included recordings where the duration of the sessions fell within the 10 to 30-minute range, roughly comparable to the recording time in the present study.

Procedures

We collected language samples using a picture description task on Zoom, administered by research assistants. We followed the procedure outlined in Eisenberg and Guo (2013, 2015). For this task, each child was presented with 7 coloured line drawings randomly selected from a set of 15, available on the project OSF page (https://osf.io/dh9za/?view_only=2057dafea8ca46a38f5985e72da6abac). The pictures were unrelated to each other. The order of picture presentation was randomised. To initiate each session, each child sat in front of a laptop, facing the webcam. The experimenter used prompts provided in Eisenberg and Guo, which are also available on the OSF page, to elicit the children's utterances. In cases where the child was hesitant to interact with the experimenter, the caregiver was allowed to prompt the child. Following the child's description of each picture, the experimenter verbally complimented them and added stars on the screen as a reward. Each session lasted between 20 and 25 minutes and was video-recorded on Zoom for transcription purposes.

The recorded Zoom sessions were transcribed by undergraduate research assistants, following guidelines developed by the first author. We used a subset of the CHAT conventions (MacWhinney, 2000) that were essential for capturing the information required for our analysis (the set of symbols used is available on the OSF page). The same author conducted a review of all the transcripts to resolve any inconsistencies across transcribers.

Subsequently, we used the *mor* function in Computerized Language ANalysis (CLAN) to identify morphological boundaries and syntactic categories for each morpheme. We then used the *kideval* function in CLAN to calculate NDW, MLU, and IPSyn scores.

Analyses

We employed Bayesian models to compare the total number of utterances, NDW, MLU, and IPSyn scores obtained from our remotely collected sample with those calculated from independent language samples of toddlers with and without language disorders, obtained from CHILDES. We used default priors for intercepts and a Normal (0,1) prior for the coefficients; the raw data and code for model testing as well as sensitivity analyses can be found on the OSF page. The model was fitted using the *brms* package (Bürkner 2017),

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

which is based on the Stan programming language (Stan Development Team, 2024) in the R programming environment (R Core Team, 2022). We used a No U-Turn Sampler to draw 10,000 samples in each of four chains from the posterior distribution over parameter values, discarding the first 1,000 for warm-up.

We report the median values for the coefficient associated with the predictors of interest, as well as their corresponding 95% Credible Intervals. If the credible interval includes 0, we also provide the probability of an effect in the direction of the coefficient sign (referred to as p-direction), irrespective of its magnitude. This is determined by examining the proportion of samples from the posterior distribution over coefficient values that fall on one side of zero. This measure ranges from 0.5 (50%), indicating equal evidence for an effect in either direction, to 1 (100%), indicating strong evidence for a directional effect when all posterior samples align on one side of zero. In this paper, we consider p-direction values greater than 0.95 (95%) as indicative of a scientifically meaningful level of evidence. With Bayesian analysis, however, one could also evaluate the effect of a predictor more gradiently using the p-direction, which we provide. We refer readers who are unfamiliar with the Bayesian framework to Nicenboim and Vasishth (2016) and Vasishth et al. (2018) for more details on the interpretation of Bayesian models.

Results

We first compare language outcomes obtained from the remote sample with those obtained in person from both typically developing and language-delayed children. We then compare the relationship among language outcomes in the remote sample with that in the in-person samples.

Comparison of language outcomes obtained remotely and those obtained in person

In Figure 1, we present a violin plot of the number of utterances produced by three-year-olds in each of the three groups. Following an anonymous reviewer's suggestion, we used a negative binomial regression instead of a Poisson regression, as the data were over-dispersed. Toddlers produced fewer utterances over Zoom (current study mean = 118, SD = 52, n = 48) than the CHILDES Typically Developing cohort (mean = 169, SD = 76, n = 74; $\beta = -0.36$ [−0.54, −0.18]), and did not credibly differ from the CHILDES Language Delayed cohort (mean = 131, SD = 73, n = 102; $\beta = -0.10$ [−0.30, 0.10]; p-direction: 84.9%). The Typically Developing children in the CHILDES cohort produced more utterances than their Language Delayed peers ($\beta = 0.26$, [0.08, 0.43]). We discuss the possible reasons for the differences in number of utterances produced by typically developing children in our sample compared to the children in the CHILDES cohort in the Discussion.

Next, we compared the quality of the language sample collected over Zoom to that obtained from CHILDES. In Figure 2, we present a violin plot comparing the NDW across all three groups. Because NDW is calculated over the first 100 words, only toddlers who produced at least 100 words were included in this analysis. The NDW from the current sample (mean = 45, SD = 7.3, n = 46) was credibly lower than the NDW obtained from the CHILDES Typically Developing cohort (mean = 48, SD = 7.2, n = 73; $\beta = -0.06$, [−0.11, −0.005]), although the difference was very small. However, it was not credibly different from the NDW of the CHILDES Language

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

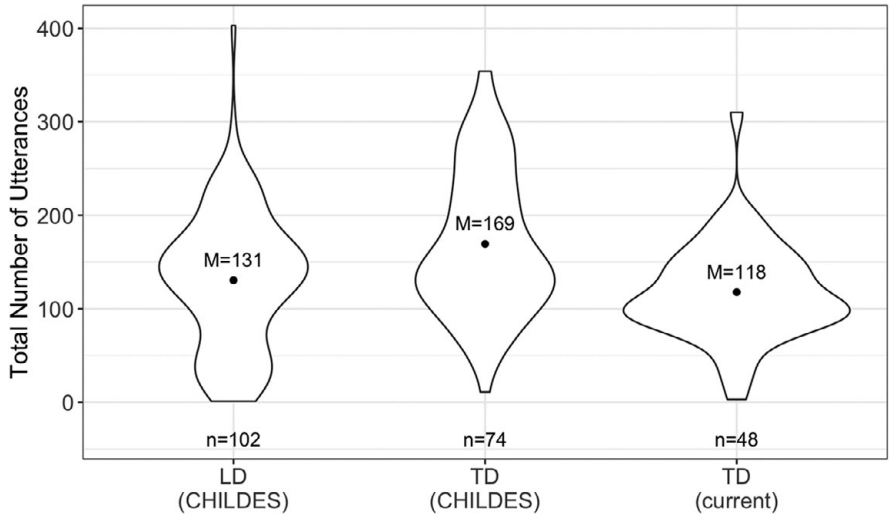

Figure 1. Total number of utterances in the current sample and the Compiled CHILDES Corpus. LD = Language Delayed, TD = Typically Developing.

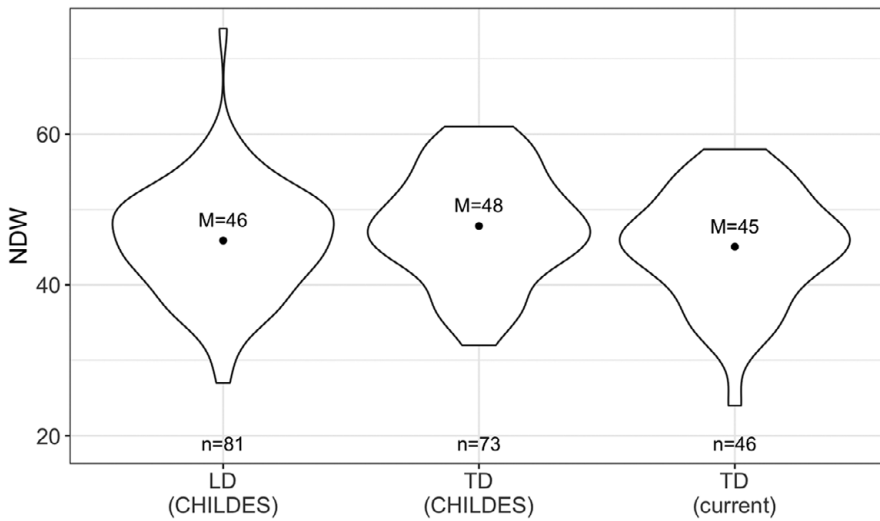

Figure 2. NDW in the current sample and the Compiled CHILDES Corpus. LD = Language Delayed, TD = Typically Developing.

Delayed cohort (mean = 46, SD = 7.5, n = 81; β = −0.02, [−0.07, 0.04]; p-direction: 74.0%). There was strong evidence showing that the CHILDES Language Delayed cohort had a lower NDW than the CHILDES Typically Developing cohort (β = −0.04, [−0.09, 0.006]; p-direction 96.0%). Thus, at three years, a lexical measure like NDW can be used to detect a difference between groups of children with and without language delay when both groups are evaluated in person. However, NDW obtained from the remote sessions with

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

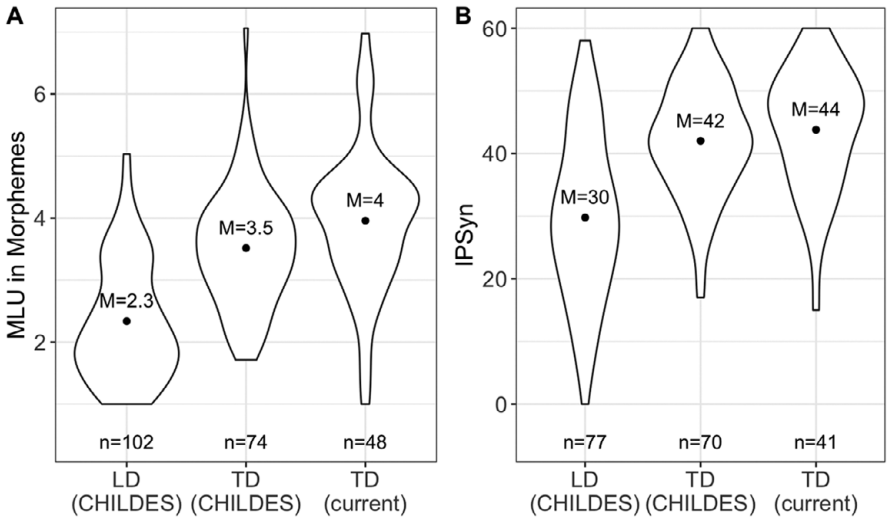

Figure 3. MLU and IPSyn scores in the current sample and the Compiled CHILDES Corpus; LD = Language Delayed, TD = Typically Developing.

Typically Developing children was systematically lower than from in-person sessions and not reliably different from the NDW measures obtained from toddlers with language impairment evaluated in person.

In Figure 3A, we present the MLU in morphemes across the three groups. We found that the MLU of the current sample (mean = 4.0, SD = 1.2, n = 48) was higher than that obtained from the Typically Developing children in the CHILDES cohort (mean = 3.5, SD = 1.0, n = 74; β = 0.12, [0.01, 0.23]). Additionally, the MLU of children with language delay in the CHILDES cohort (mean = 2.3, SD = 1.0, n = 102) was credibly lower than that of Typically Developing peers in the CHILDES cohort (β = −0.41, [−0.51, −0.31]) and in the current study (β = −0.53, [−0.65, −0.41]). Thus, a grammatical measure like MLU can be used to detect a group difference between Language Delayed three-year-olds and Typically Developing peers whether the sample is collected in person or remotely. Further, MLU was higher in the remote sample compared to the in-person sample from Typically Developing toddlers.

In Figure 3B, we present a violin plot comparing the IPsyn measure across all three groups. Because IPSyn measures are calculated only when there are at least 50 IPSyn-eligible utterances (see Yang, MacWhinney et al., 2022, for details), only a subset of toddlers was included in this analysis. Again, following the anonymous reviewer's suggestion, we used a negative binomial regression instead of a Poisson regression, as it provides a better fit to the data. IPSyn scores obtained from the current study (mean = 44, SD = 9.8, n = 41) did not credibly differ from those of the Typically Developing toddlers in the CHILDES cohort (mean = 42, SD = 9.0, n = 70; β = 0.04, [−0.05, 0.13]; p-direction: 91.5%), although there was a trend towards higher IPSyn measured in the current study. IPSyn scores of the Language Delayed children in the CHILDES cohort (mean = 30, SD = 13.0, n = 77) were credibly lower than those of both groups of Typically Developing children, ones in the CHILDES cohort (β = −0.34, [−0.46, −0.22]) and ones in the current study (β = −0.38, [−0.51, −0.25]). Thus, a grammatical measure like IPSyn can be

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

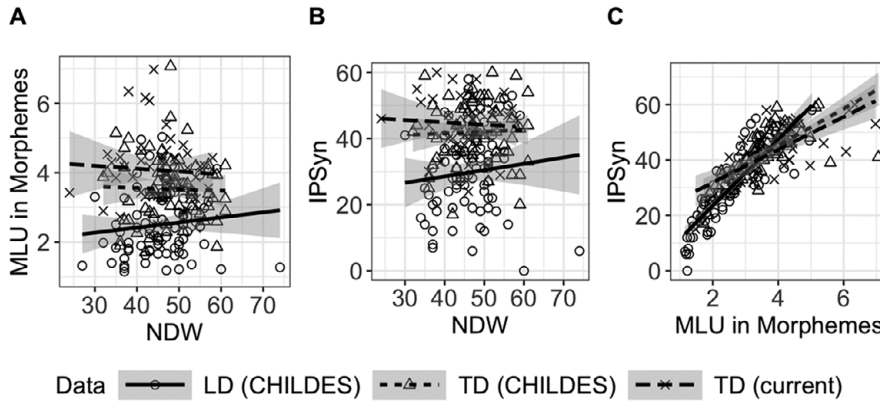

Figure 4. Bivariate correlations between NDW, MLU, and IPSyn in the current sample and the Compiled CHILDES Corpus; LD = Language Delayed, TD = Typically Developing.

used to detect a group difference between three-year-olds with and without language delays, whether the samples are collected remotely or in person, and the IPSyn measures obtained from in-person and remote sessions are comparable.

Relationships among language outcomes

Next, we compared the relationship between NDW, a lexical measure, and MLU, a grammatical measure, in the current study and the CHILDES cohorts with and without language delay by calculating bivariate Pearson correlations (Figure 4A). There was no evidence of a relationship between NDW and MLU in any of the three groups (Typically Developing toddlers in the present study $r = -0.06$, $p = 0.70$; Typically Developing cohort in the CHILDES $r = -0.03$, $p = 0.80$; Language Delayed cohort in the CHILDES $r = 0.12$, $p = 0.29$).

We also compared the relationship between NDW and IPSyn in the current study, and the CHILDES cohorts with and without language delay (Figure 4B). As was the case with MLU, there was no evidence of a relationship between NDW and IPSyn for Typically Developing three-year-olds in this study ($r = -0.06$, $p = 0.71$) or in the CHILDES cohort ($r = 0.05$, $p = 0.71$), or in the Language Delayed toddlers in the CHILDES cohort ($r = 0.10$, $p = 0.37$).

Finally, we compared the relationship between two measures of grammatical development – MLU and IPSyn – in the current study and the CHILDES cohorts with and without language delay (Figure 4C). In all three groups, MLU and IPSyn were positively correlated (current study $r = 0.69$, $p < 0.001$; CHILDES Typically Developing cohort: $r = 0.70$, $p< 0.001$; CHILDES Language Delayed cohort $r = 0.86$, $p < 0.001$).

Discussion

In the present study, we compared three-year-olds' language outcomes obtained from samples collected remotely and those obtained from samples collected in person. We first confirmed previous findings that in three-year-olds tested in person, NDW, a measure of lexical development, and both measures of grammatical development, MLU and IPSyn

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

scores, can reliably detect group differences between toddlers with and without language delay. Crucially, we also found group differences in the grammatical measures - MLU and IPSyn between typically developing children based on outcomes obtained from remote language samples and toddlers with language delays with in-person language samples.

In the following discussion, we highlight the similarities and differences in the language sample elicited remotely from typically developing toddlers using a picture description task and samples elicited in person, and then the relationship between lexical and grammatical measures, ending with recommendations for eliciting language samples remotely.

Overall, language samples elicited remotely and in person were quantitatively and qualitatively dissimilar. It is perhaps not too surprising that when language samples were elicited in a remote session, toddlers produced fewer utterances than the typically developing cohort tested in person. Likely as a result, NDW, a proxy for lexical diversity, was also lower in the remote sample when compared to the in-person sample from typically developing toddlers and surprisingly as low as that of children with language delays. Despite producing fewer utterances, typically developing toddlers had a higher MLU in the sample elicited remotely compared to the in-person sample.

However, before we attribute the differences between outcome measures from language samples elicited in this study and the CHILDES typically developing cohort to the modality of elicitation, we need to consider several alternate explanations. The language samples obtained from the CHILDES cohort were typically obtained during free play or book reading with parents. In contrast, in the remote sessions, we elicited language samples using a picture description task conducted by an unfamiliar research assistant. It is, of course, quite likely that children produce fewer utterances with strangers than with parents. Further, we know from previous research that although children produce more utterances in free play, their MLUs are typically lower compared to speech elicited from other activities (Sealey & Gilmore, 2008; Southwood & Russell, 2004). In contrast, utterances obtained from narrative speech or storytelling tend to be longer and more complex compared to those elicited from free play or conversation (Mirsaleh et al., 2011; Southwood & Russell, 2004; Stalnaker & Creaghead, 1982; Wagner et al., 2000). Consistent with this literature, in our study with a picture description task as well, toddlers produced more structured, longer sentences with higher MLUs and consequently lower NDW; longer utterances contain more word repetitions and a greater number of function words, both of which lower lexical diversity. Although plausible, we were unable to confirm that the differences documented here could be attributed to task differences alone. Language samples elicited using picture descriptions are available on CHILDES for 17 children with Specific Language Impairment and their 17 age-matched peers (Eisenberg & Guo, 2013). Recall that we adapted a picture description task for our remote language assessment originally used by Eisenberg and Guo. Thus, our sample and Eisenberg and Guo's sample share the same task type but differ only in the mode of test administration. Unfortunately, children in Eisenberg and Guo's sample were older (41.6 months for typically developing children and 41.3 months for children with SLI) than those in our sample (36.6 months). Because lexical and grammatical measures are expected to systematically increase with age, we chose not to compare the two cohorts directly.

Overall, though, there were several other differences between our sample and the CHILDES typically developing cohort. All children in our cohort were monolingual, recorded in the presence of just one parent, and from middle- to high-income families.

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

This information is unavailable for the children in the CHILDES cohort but is likely to be more varied given the many different labs and cities where these data were collected. It should also be noted that some of the CHILDES recordings date back to the 1980s, so we cannot rule out systematic differences in children's speech patterns across time. We ourselves were unable to test the same children in person after testing them remotely because these data were collected during the height of the COVID pandemic, and all labs were closed. Clearly, future research is needed to delineate the effects of task differences from that of remote assessment of language samples.

Despite the differences in both the elicitation activity and modality of elicitation between the current sample and the CHILDES cohorts, it was still possible to reliably detect group differences between typically developing children and language-delayed children, based on the two grammatical measures MLU and IPSyn. The finding underscores the credibility of remote language assessments as a means to effectively identify three-year-olds at risk for language impairment.

A secondary goal of this study was to investigate the relationship among the various measures of language outcomes. We found no significant correlation between the lexical and grammatical measures in any of the three samples (aligning with Yang, Rosvold et al., 2022, but contrasting with the results of Klee (1992), and Watkins et al. (1995)), underscoring the fact that measures like NDW provide information that is complementary to that provided by MLU and IPSyn. The two grammatical measures, MLU and IPSyn, however, were positively correlated across all three groups. The replication of the relationship among language outcomes observed in the in-person samples within the remote samples attests to the validity of remote testing.

It should be noted that all the language outcomes we used to detect differences between English-learning typically developing and language-delayed toddlers – total number of utterances, NDW, MLU, and IPSyn – were calculated automatically using CLAN. Coupled with the benefit of remotely collecting language samples, this automated process offers better test-retest reliability as well as quicker, clinically relevant assessment.

Finally, we highlight challenges encountered in remotely eliciting language samples from three-year-olds through a picture description task and provide recommendations for overcoming these obstacles. First, three-year-olds are typically not familiar with interacting with experimenters through a screen, especially when the picture is presented full screen and the experimenter's thumbnail is small. In this situation, caregivers should be asked to assist by helping children locate the experimenter and explaining that they are participating in a storytelling activity with the experimenter. Additionally, children often had difficulty locating specific parts of the picture when the experimenter used the cursor to point to it. In such cases, caregivers can point to the indicated part of the picture and prompt the child to talk about it. Further, children in home environments were prone to distractions, such as toys and family members other than the caregiver participating in the experiment. To mitigate this, we recommend that experimenters request that caregivers eliminate potential distractors before beginning the experimental session. Lastly, maintaining engagement and contingent reinforcement during online testing is challenging for three-year-olds. To enhance focus and reward engagement, we used the Annotate function within Zoom to draw "stars" for children, motivating them to talk.

The present study also highlights the growing necessity for equipping practitioners with training in remote language assessment. Previous research has shown that practitioners' confidence in delivering telehealth services was positively influenced by prior experience in telehealth services and exposure to more diverse training programs

https://doi.org/10.1017/S0305000924000643 Published online by Cambridge University Press

(Biggs et al., 2022). Hence, it is crucial for speech-language pathologists to receive training in remote language assessment services to effectively address the unique challenges it presents.

In conclusion, our findings show that language samples collected remotely from three-year-olds are sufficient to detect differences between groups of children with and without language impairment. Because remote collection of language samples offers the potential for enhancing the accessibility of language assessment, we believe that continued research to ensure its validity, coupled with training for professionals, is necessary to maximize its benefits.

Acknowledgements. This study was funded by NSF BCS-2028034 awarded to Megha Sundara & Bruce Hayes.

Competing interest. The author(s) declare none.

References

Ambrose, S. E. (2016). Gesture use in 14-month-old toddlers with hearing loss and their mothers' responses. *American Journal of Speech-Language Pathology*, **25**, 519–531. https://doi.org/10.1044/2016_AJSLP-15-0098.

Bang, J., & **Nadig, A.** (2015). Language learning in autism: Maternal linguistic input contributes to later vocabulary. *Autism Research*, **8**(2), 214–233. https://doi.org/10.1002/aur.1440.

Barokova, M., & **Tager-Flusberg, H.** (2020). Commentary: Measuring language change through natural language samples. *Journal of Autism and Developmental Disorders*, **50**(7), 2287–2306. https://doi.org/10.1007/s10803-018-3628-4.

Biggs, E. E., **Rossi, E. B.**, **Douglas, S. N.**, **Therrien, M. C. S.**, & **Snodgrass, M. R.** (2022). Preparedness, training, and support for augmentative and alternative communication telepractice during the COVID-19 pandemic. *Language, Speech, and Hearing Services in Schools*, **53**(2), 335–359. https://doi.org/10.1044/2021_LSHSS-21-00159.

Bliss, L. (1988). Modal usage by preschool children. *The Journal of Applied Developmental Psychology*, **9**, 253–261. https://doi.org/10.1016/0193-3973(88)90028-7.

Bürkner, P. (2017). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, **80** (1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Castilla-Earls, A, **Ronderos, J.**, **McIlraith, A.**, & **Martinez, D.** (2022). Is bilingual receptive vocabulary assessment via telepractice comparable to face-to-face? *Language, Speech, and Hearing Services in Schools*, **53**(2), 454–465. https://doi.org/10.1044/2021_LSHSS-21-00054.

Charest, M., & **Skoczylas, M. J.** (2019). Lexical diversity versus lexical error in the language transcripts of children with developmental language disorder: Different conclusions about lexical ability. *American Journal of Speech-Language Pathology*, **28**, 1275–1282. https://doi.org/10.1044/2019_AJSLP-18-0143.

Charest, M., **Skoczylas, M. J.**, & **Schneider, P.** (2020). Properties of lexical diversity in the narratives of children with typical language development and developmental language disorder. *American Journal of Speech-Language Pathology*, **29**, 1866–1882. https://doi.org/10.1044/2020_AJSLP-19-00176.

Conti-Ramsden, G., & **Dykins, J.** (1991). Mother–child interactions with language-impaired children and their siblings. *British Journal of Disorders of Communication*, **26**, 337–354. https://doi.org/10.3109/13682829109012019.

Conti-Ramsden, G., **Hutcheson, G. D.**, & **Grove, J.** (1995). Contingency and breakdown: Specific language impaired children's conversations with their mothers and fathers. *Journal of Speech and Hearing Research*, **38**(6), 1290–1302. https://doi.org/10.1044/jshr.3806.1290.

Conti-Ramsden, G., & **Jones, M.** (1997). Verb use in specific language impairment. *Journal of Speech and Hearing Research*, **40**, 1298–1313. https://doi.org/10.1044/jslhr.4006.1298.

Dam, Q. D., & **Pham, G. T.** (2023). Remote first-language assessment: Feasibility study with Vietnamese bilingual children and their caregivers. *Language, Speech, and Hearing Services in Schools*, **54** (2), 618–635. https://doi.org/10.1044/2023_LSHSS-22-00123.

Eisenberg, S. L., **Fersko, T. M.**, & **Lundgren, C.** (2001). The use of MLU for identifying language impairment in preschool children: A review. *American Journal of Speech-Language Pathology*, **10**, 323–342. https://doi.org/10.1044/1058-0360(2001/028).

Eisenberg, S. L., & **Guo, L.-Y.** (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools*, **44** (1), 20–31. https://doi.org/10.1044/0161-1461(2012/11-0089).

Eisenberg, S. L., & **Guo, L.-Y.** (2015). Sample size for measuring grammaticality in preschool children from picture-elicited language samples. *Language, Speech, and Hearing Services in Schools*, **46** (2), 81–93. https://doi.org/10.1044/2015_LSHSS-14-0049.

Feldman, H., **Keefe, K.**, & **Holland, A.** (1989). Language abilities after left hemisphere brain injury: A case study of twins. *Topics in Special Education*, **9**, 32–47. https://doi.org/10.1177/027112148900900104.

Finestack, L. H., & **Satterlund, K. E.** (2018). Current practice of child grammar intervention: A survey of speech-language pathologists. *American Journal of Speech-Language Pathology*, **27** (4), 1329–1351. https://doi.org/10.1044/2018_AJSLP-17-0168.

Hargrove, P. M., **Holmberg, C.**, & **Zeigler, M.** (1986). Changes in spontaneous speech associated with therapy hiatus: A retrospective study. *Children Language Teaching and Therapy*, **2**, 266–280. https://doi.org/10.1177/026565908600200302.

Hux, K., **Morris-Friehe, M.**, & **Sanger, D. D.** (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools*, **24**, 84–91. https://doi.org/10.1044/0161-1461.2402.84.

Keefe, K., **Feldman, H.**, & **Holland, A.** (1989). Lexical learning and language abilities in preschoolers with perinatal brain damage. *Journal of Speech and Hearing Disorders*, **54**, 395–402. https://doi.org/10.1044/jshd.5403.395.

Kemp, K., & **Klee, T.** (1997). Clinical speech and language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy*, **13**, 161–176. https://doi.org/10.1177/026565909701300204.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders*, **12**(2), 28–41. https://doi.org/10.1097/00011363-199202000-00005.

Leonard, L. B. (1998). *Children with specific language impairment*. Boston: MIT Press.

Loeb, D. F., **Kinsler, K.**, & **Bookbinder, L.** (2000). *Current language sampling practices in preschools* [Poster presentation]. , Washington, D.C: Annual Convention of the American Speech-Language-Hearing Association.

Lund, E., & **Werfel, K. L.** (2022). The effects of virtual assessment on capturing skill growth in children with hearing loss. *Language, Speech, and Hearing Services in Schools*, **53** (2), 391–403. https://doi.org/10.1044/2021_LSHSS-21-00074.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Magimairaj, B. M., **Capin, P.**, **Gillam, S. L.**, **Vaughn, S.**, **Roberts, G.**, **Fall, A.-M.**, & **Gillam, R. B.** (2022). Online administration of the test of narrative language-second edition: Psychometrics and considerations for remote assessment. *Language, Speech, and Hearing Services in Schools*, **53** (2), 404–416. https://doi.org/10.1044/2021_LSHSS-21-00129.

Manning, B. L., **Harpole, A.**, **Harriott, E. M.**, **Postolowicz, K.**, & **Norton, E. S.** (2020). Taking language samples home: Feasibility, reliability, and validity of child language samples conducted remotely with video chat versus in-person. *Journal of Speech, Language, and Hearing Research*, **63** (12), 3982–3990. https://doi.org/10.1044/2020_JSLHR-20-00202.

McElwain, N. L., **Hu, Y.**, **Li, X.**, **Fisher, M. C.**, **Baldwin, J. C.**, & **Bodway, J. M.** (2022). Zoom, Zoom, baby! Assessing mother-infant interaction during the still face paradigm and infant language development via a virtual visit procedure. *Frontiers in Psychology*, **12**, 734492. https://doi.org/10.3389/fpsyg.2021.734492.

Mirsaleh, Y. R., **Abdi, K.**, **Rezai, H.**, & **Kashani, P. A.** (2011). A comparison between three methods of language sampling: Freeplay, narrative speech and conversation. *Iranian Rehabilitation Journal*, **9** (14), 4–9.

Nelson, N. W., & **Plante, E.** (2022). Evaluating the equivalence of telepractice and traditional administration of the test of integrated language and literacy skills. *Language, Speech, and Hearing Services in Schools*, **53** (2), 376–390. https://doi.org/10.1044/2022_LSHSS-21-00056.

Nicenboim, B., & **Vasishth, S.** (2016). Statistical methods for linguistic research: Foundational ideas—Part II. *Language and Linguistics Compass*, **10**, 591–613. https://doi.org/10.1111/lnc3.12207.

Nicholas, J. G., & **Geers, A. E.** (1997). Communication of oral deaf and normally hearing children at 36 months of age. *Journal of Speech, Language, and Hearing Research*, **40**, 1314–1327. https://doi.org/10.1044/jslhr.4006.1314.

Peña, E. D., & **Sutherland, R.** (2022). Can you see my screen? Virtual assessment in speech and language. *Language, Speech, and Hearing Services in Schools*, **53** (2), 329–334. https://doi.org/10.1044/2022_LSHSS-22-00007.

Pratt, A. S., **Anaya, J. B.**, **Ramos, M. N.**, **Pham, G.**, **Muñoz, M.**, **Bedore, L. M.**, & **Peña, E. D.** (2022). From a distance: Comparison of in-person and virtual assessments with adult–child dyads from linguistically diverse backgrounds. *Language, Speech, and Hearing Services in Schools*, **53** (2), 360–375. https://doi.org/10.1044/2021_LSHSS-21-00070.

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rescorla, L., **Dahlsgaard, K.**, & **Roberts, J.** (2000). Late-talking toddlers: MLU and IPSyn outcomes at 3;0 and 4;0. *Journal of Child Language*, **27** (3), 643–664. https://doi.org/10.1017/s0305000900004232.

Rescorla, L., & **Lee, E.** (2000). Language impairment in young children. In T. Layton, E. Crais, & L. Watson (Eds.), *Handbook of early language impairment in children: Nature* (pp. 1–55). Albany, NY: Delmar.

Rescorla, L., & **Schwartz, E.** (1990). Outcome of toddlers with expressive language delay. *Applied Psycholinguistics*, **11**, 393–407. https://doi.org/10.1017/S0142716400009644.

Rollins, P. R. (1999). Pragmatic accomplishments and vocabulary development in pre-school children with autism. *American Journal of Speech-Language Pathology: A Journal of Clinical Practice*, **8**, 181–190. https://doi.org/10.1044/1058-0360.0802.181.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, **11** (1), 1–22. https://doi.org/10.1017/S0142716400008262.

Schmitt, M. B., **Tambyraja, S.**, **Thibodeaux, M.**, & **Filipkowski, J.** (2022). Feasibility of assessing expressive and receptive vocabulary via telepractice for early elementary-age children with language impairment. *Language, Speech, and Hearing Services in Schools*, **53** (2), 445–453. https://doi.org/10.1044/2021_LSHSS-21-00057.

Sealey, L. R., & **Gilmore, S. E.** (2008). Effects of sampling context on the finite verb production of children with and without delayed language development. *Journal of Communication Disorders*, **41**(3), 223–258.

Southwood, F., & **Russell, A. F.** (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research*, **47** (2), 366–376. https://doi.org/10.1044/1092-4388(2004/030.

Stalnaker, L. D., & **Creaghead, N. A.** (1982). An examination of language samples obtained under three experimental conditions. *Language, Speech, and Hearing Services in Schools*, **13**, 121–128. https://doi.org/10.1044/0161-1461.1302.121.

Stan Development Team. 2024. Stan Modeling Language Users Guide and Reference Manual, Version 2.35. https://mc-stan.org

Sutherland, R., **Trembath, D.**, **Hodge, A.**, **Drevensek, S.**, **Lee, S.**, **Silove, N.**, & **Roberts, J.** (2017). Telehealth language assessments using consumer grade equipment in rural and urban settings: Feasible, reliable and well tolerated. *Journal of Telemedicine and Telecare*, **23** (1), 106–115. https://doi.org/10.1177/1357633X15623921.

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships* (Vol. **10**). Minneapolis, MN: University of Minnesota Press.

Vasishth, S., **Nicenboim, B.**, **Beckman, M. E.**, **Li, F.**, & **Kong, E. J.** (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of phonetics*, **71**, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008.

Wagner, C. R., **Nettelbladt, U.**, **Sahlén, B.**, & **Nilholm, C.** (2000). Conversation versus narration in pre-school children with language impairment. *International Journal of Language & Communication Disorders*, **35** (1), 83–93. https://doi.org/10.1080/136828200247269.

Waite, M. C., **Theodoros, D. G.**, **Russell, T. G.**, & **Cahill, L. M.** (2010). Internet-based telehealth assessment of language using the CELF-4. *Language, Speech, and Hearing Services in Schools*, **41** (4), 445–458. https://doi.org/10.1044/0161-1461(2009/08-0131.

Watkins, R. V., **Kelly, D. J.**, **Harbers, H. M.**, & **Hollis, W.** (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech, Language, and Hearing Research*, **38** (6), 1349–1355. https://doi.org/10.1044/jshr.3806.1349.

Yang, J. S., **MacWhinney, B.**, & **Bernstein Ratner, N.** (2022). The index of productive syntax: Psychometric properties and suggested modifications. *American Journal of Speech-Language Pathology*, **31**, 239–256. https://doi.org/10.1044/2021_AJSLP-21-00084.

Yang, J. S., **Rosvold, C.**, & **Bernstein Ratner, N.** (2022). Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations. *Frontiers in Psychology*, **13**, 905789. 10.3389/fpsyg.2022.905789.

Cite this article: Jo, J., & Sundara, M. (2024). Remote collection of language samples from three-year-olds. *Journal of Child Language* 1–14, https://doi.org/10.1017/S0305000924000643