Original Article

http://e-cacd.org/ eISSN: 2508-5948

Clinical Archives of Communication Disorders / Vol. 4, No. 3:163-176 / December 2019 Open Access

https://doi.org/10.21849/cacd.2019.00143

Connected speech of two-year-olds: Test-retest reliability for assessment of phonetic inventory and word shape analysis

Shari Leigh DeVeney, Lucia Scheffel

University of Nebraska at Omaha, Omaha, Nebraska, USA

Purpose: Clinicians and researchers depend in part on informal measures, those that are not standardized or norm referenced, to reliably represent young children's speech sound productions. However, few investigations have explored the reliable use of such measures with young children. Consequently, little is known about the consistency of extrapolated findings from informal measures. The present study aimed to address this issue by testing the short-term (one week) reliability of informal measures used for independent analyses of speech sound productions, analyses that describe productions without comparison to an adult standard, with two-year-old children.

Methods: Participants were eleven two-year-old monolingual American English-speaking toddlers without communication delays. The two informal independent measures studied were phonetic inventory and word shape analysis. The researchers compared the analysis outcomes of two 20-minute parent-child, play-based connected speech samples collected one week apart under near-identical circumstances.

Results: Findings indicated measures of phonetic inventory for consonants across word positions (initial, medial, and final) and in clusters were stable over a short-term (one-week) time frame. Although the word shape analysis studied did not reach the level of consideration for short-term reliability, this could be an artifact of procedural differences in conducting this type of analysis.

Conclusions: The present study findings offered partial support for the continued use of informal independent measures for clinicians and researchers working with young children. Support was noted for the reliable use of phonetic inventory; however, administration of word shape analyses may result in unreliable representations of young children's speech sound repertoires.

Keywords: Two-year-olds, Test-retest reliability, Independent analysis, Speech, Phonology

Received: October 17, 2019
Revision: December 12, 2019
Accepted: December 13, 2019

Correspondence:
Shari Leigh DeVeney

University of Nebraska at Omaha,
6001 Dodge Street, Omaha, Nebraska,
USA
Tel: +4025542993
Fax: +4025543572
E-mail: sdeveney@unomaha.edu

INTRODUCTION

To comprehensively evaluate children's speech and language development, speech-language pathologists (SLPs) typically rely on a combination of formal and informal assessment tools. When integrated, these tools are useful in determining eligibility for intervention services and providing descriptive information to establish baseline performance for speech-language progress monitoring [1]. In contrast to formal measures such as standardized tests (for an overview of standardized assessment tools appropriate for young children, see Claessen et al. [2]), informal assessment tools are not standardized or norm-referenced, but often add descriptive and naturalistic data to the assessment process across a variety of communication disorders and individuals [1]. Types of informal assessment measures include clinical observation, parent report, and analysis of a sample of the child's speech obtained from pre-determined word list, observation, or conversation [1]. For young children, informal assessment customarily involves collecting and analyzing a spontaneous conversational speech sample [1,3]. When the purpose of analysis is to ascertain the child's speech sound development and functional use of sounds in words, descriptive analyses of the speech sample include two major types, relational and independent analyses [4].

Both relational and independent analyses offer a holistic framework for representing a child's phonological system and allow for description of both typical and disordered productions [4], but differ in important respects. Relational analyses are used to compare a child's speech production to an idealized adult standard of production [4,5] for the purpose of describing and categorizing patterns evident in the speech production of young children that do not conform to developmental and/or adult standards [4]. However, relational analyses are limited by their comparative nature due to underlying differences in how sounds are organized between adults and young children, with adults relying on a phoneme-level representation of contrastive sound units and young children utilizing a less refined representation of contrast at the syllable or even the word level [6]. This difference in representation can complicate comparison of young children's productions to adult-based phonemes as it leads to variable intraword productions, meaning a child may produce a single word several different ways [4]. As noted by Sosa and Stoel-Gammon [7], the transition from a holistic child-like representation to the phonemic, adult-like representation does not occur quickly and is often a gradual process, limiting the utility of comparative analyses.

For independent analyses, a child's production is described without comparison to adult productions. Because of this, independent measures are able to provide useful diagnostic information about speech sound productions of children who present with many sounds in error, those who have few speech sounds, as well as those who are dual language learners [3,5,8]. For two well-known independent analyses available to SLPs that can be used with conversational speech samples, *phonetic inventory* and *word shape analysis*, reliability of use with young children has been called into question [8-10].

Phonetic inventory provides a detailed description and distribution of speech sounds (phones) in terms of their articulatory, acoustic, and psychoacoustic features [4]. A phonetic inventory is an index of the different sounds and sound sequences children use across word positions, even when the phones are not produced in words according to the idealized adult standard [4,5]. For example, if a child said [dæt] for *b̲at*, the [d] production in the initial word position and [t] for final position would be recorded in the phonetic inventory even though the child's attempt of the adult word form *bat* was not accurate. Stoel-Gammon and Dunn [4] noted that a phonetic inventory allowed for grouping of phones according to word position distribution (e.g., initial, medial, final) as well as manner of production (e.g., stop-plosive, fricative, affricate, etc.). According to Stoel-Gammon and Dunn [4], a phone production can further be distinguished as "marginal" when only produced once or twice in a given word position.

Another independent measure free of adult standard comparison is a syllable or word shape analysis. This type of analysis is comprised of a list of sound sequences identified as consonant (C) and/or vowel or diphthong (V) structures that children use in syllables and words they produce [5,4]. For example, the words *two* and *bee* produced as [tu] and [bi] are examples of CV (consonant-vowel) word shapes. The word *balloon* produced as [bun] is an example of a CVC word shape. A child's word shape may not have a 1:1 correspondence with the target word; therefore, word shape analyses include immature forms of words children produce that have meaning, but are not necessarily conventional adult productions. These shapes are recorded along with their number of occurrences within the sampled speech and analyzed for frequency of occurrence and patterns [4]. Different procedures for word shape analyses have been used and studied. For example, some researchers [8,10] used word shape analysis procedures in which a finite number of different word shapes (i.e. V, CV, CVCV, VC, CVC, CCVC, CVCC, and CVCVC) were specifically analyzed. Both sets of researchers noted a ceiling effect because most of their young participants produced at least two different words in each of the target word shapes. Stoel-Gammon and Dunn [4] described the use of an "open-ended" tally procedure for syllable and word shape analyses that offered a less constrained technique than a "closed-set" methodology. Use of an "open-ended" procedure could decrease the likelihood of a ceiling effect being noted.

Irrespective of the type of analysis used, when examining assessment results, SLPs need to know that they are relying on measurements that are consistent and accurate as these tools "serve as gateways to services" (3, p342). However, work with young children presents unique challenges to determining the consistent and reliable use of phonetic analyses due to variable word productions observed. The presence of normal articulatory variability complicates evaluation of speech sound production for young children (for an in-depth discussion see [11-13]). As noted above, Ferguson and Farwell [6], suggested that young children do not use phonemes as the minimal unit of lexical representation; rather, they organize with the syllable or word level as the minimal unit. This "whole word" representation, as noted by Sosa [14] means that "young children may be operating with holistic rather than segmental phonological representations, resulting in high rates of intraword variability" (14, p26). Sosa and Stoel-Gammon [7] defined intraword variability as "multiple tokens of the same word produced differently at the same point in time (same chronological age, recording session, etc.)" (7, p32). Researchers have found a considerable amount of intraword variability even for children with typical development at two- and three-years of age [7,13-15]. This variability, while still observed to be present in the speech of 3-year-old children, tends to gradually decrease with age [14] and is associated with specific word productions and phonetic complexity such as later-developing consonant and complex syllable structures like consonant clusters [16]. However, despite the presence of intraword variability, researchers have observed that atypical speech patterns can be reliably identified in children as young as two years of age [17].

A number of investigators interested in preschool and toddler populations have used phonetic inventory [18-23] and word shape analyses [24,25] to describe and compare speech sound productions across typical and disordered populations. Some have even used these measures for studies in which a child's performance on a measure was repeatedly compared over time (e.g. [21,22]) or compared to other participants (e.g. [19,24]), drawing conclusions from use of these independent measures. For example, Robb and Bleile [21] and Stoel-Gammon [22] found that as children age, their phonetic repertoires expand and productions tend to become more stable. Carson et al. [18] found that toddlers with expressive language delay showed significant differences across speech production measures, including phonetic inventory, compared with peers who were typically developing. However, even as these measures are used with assumed reliability, few researchers have explicitly examined their temporal reliability in order to determine if such comparisons represent true differences across samples or rather represent "an artifact of an unstable measure" (8, p46).

Although attempts have been made to evaluate the effectiveness and reliability of informal measures (see Limbrick et al. [26]), only a few have focused on independent analyses. Morris [8] and Wittler and DeVeney [10] evaluated the reliable use of independent analyses with young children who were typically developing. DeVeney and Sheridan [9] evaluated the reliable use of these measures with a small sample of children identified as late talkers and Heilmann et al. [27] focused on school-age children at risk for speech-language delays. Morris [8] conducted a short-term test-retest evaluation with play-based spontaneous conversational samples of ten 18-to-22-month-old children with typical language development. She evaluated the reliable use of phonetic inventory and word shape analysis among other informal independent analyses. In the study, mother-child dyads participated in two 20-minute play sessions one week apart, after which the children's speech was analyzed from videotape. Results indicated the test-retest reliability of some aspects of the analyses were unstable and did not necessarily represent the same number or range of speech sounds produced from session to session. In particular, Morris [8] found the range and number of initial consonant productions were the least stable measure over time. Although final consonant productions and word shape analyses were found to be moderately stable, neither were deemed to be significantly reliable. According to Morris [8], if the informal measures were not shown to be significantly reliable, even over such a short time frame, perceived improvements attributed to speech-language therapy may not truly represent progress.

Similar findings were noted in an exploratory study conducted by Wittler and DeVeney [10] with three participants 25- to 33-months of age using procedures aligned with those of Morris [8]. Findings indicated two of the three participants obtained inconsistent phonetic inventories for word-initial sound productions and for all three participants word-final productions were relatively consistent. Consonant cluster productions (i.e. production of two adjacent consonant sounds such as [sn] in *sn*ake or [pl] in *pl*ate) while not addressed in the Morris [8] study; were inconsistent for two of the three participants. Morris [8] noted that word shape analysis was moderately but not significantly consistent across participants and in the feasibility data all three participants exhibited consistency across the target word shapes analyzed.

DeVeney and Sheridan [9] analyzed the spontaneous speech samples of three toddlers (24- to 31-months-of-age) identified as late talkers. They noted that one participant demonstrated inconsistent outcomes for phonetic inventory in initial and final position consonants and two participants produced inconsistent results for the word shape analysis. Heilmann et al. [27] collected two communication samples one week apart for 20 kindergartners at risk for speech-language delays using a structured interview procedure. In contrast to the studies reviewed above regarding toddler populations, their findings indicated strong test-retest reliability for informal relational analyses calculated from structured, rather than spontaneous, communication samples.

Given the clinical relevance of independent measure use with young children and previous findings of inconsistencies in the test-retest reliability of such measures, the aim of the current study was established. The current study aimed to test the reliable use of two independent analyses, phonetic inventory and word shape analysis, to assess two-year-olds. This target population was slightly older than the sample (18- to 22-months) used in the Morris study [8]. The 2-year-old age range was the focus because many toddlers who receive speech-language services at this age do so to increase expressive vocabulary size and transition from single- to multi-word utterances [30]. Lack of practice with word productions potentially puts these children at risk for delays in acquiring and mastering speech sounds and, consequently, establishes a need for reliable measures to accurately record phonetic skills for service verification decisions, progress monitoring, and intervention planning. Noting the earliest ages at which SLPs can reliably use independent phonological analyses to measure speech sound production skills is important for evidence-based clinical practice. Determining at what age use of independent analyses indicate consistent short-term results is also important for reliable interpretation of research findings involving these measures. In order to establish the age at which measures can be reliably used, information is needed for both typically developing and delayed populations. The current study was not a longitudinal study, but an effort to address this information for a typical population of young children: Children who were not presenting with communication delays.

In the present study, the researchers aimed to determine if inconsistencies in reliability noted for a sample of very young children were also present with 2-year-olds given the same time period, one week. The following questions were addressed:

1. For 2-year-old children without communication delays, what is the short-term (within one week) test-retest reliability of phonetic inventory when calculated using a 20-minute spontaneous conversation sample?
   Based on previous findings from existing literature and high occurrence of variable word productions characteristic of this young population, the researcher hypothesized that short-term test-retest for phonetic inventories would be unreliable across word positions.
2. For 2-year-old children without communication delays, what is the short-term (within one week) test-retest reliability of word shape analysis when calculated using a 20-minute spontaneous conversation sample?

For this question, the researcher hypothesized that the short-term test-retests for word shapes would also be unreliable given the high occurrence of intraword variability with young children in this age range as well as differences in word shape analysis procedures planned for this study (see below) compared with previous studies.

METHODS

Participants

Participants were eleven (four male, seven female) monolingual American English-speaking toddlers between 25- and 33-months of age (M=27.70, SD=3.16) recruited from educational settings (e.g. childcare centers) and clinical agencies (e.g. pediatricians' offices) in a United States Midwestern metropolitan area. Participant recruitment, interactions, and project procedures were conducted in accordance with the ethical standards of the authors' university institutional review board and the study was approved by this governing body prior to the beginning of data collection. Participants were identified as having no known delays in communication development through completion and scoring of three screening measures and parent-reported observations. Screenings were administered during the first of two sessions prior to the collection of the play-based conversational samples and included the following:

(1) *MacArthur Bates Communicative Development Inventory* (*CDI*): *Words and Sentences* - 2nd edition [31] and *MacArthur Bates Communicative Development Inventory-III* (*CDI*) [31] for participants over 30 months of

age, standardized, norm-referenced assessment tools used to measure expressive vocabulary.

(2) *Preschool Language Scale* - 5th edition (PLS-5) [32], a standardized, norm-referenced assessment instrument used to evaluate the receptive and expressive language skills of young children.

(3) *Ages & Stages Questionnaire* [33], a criterion-referenced development screener providing information about general developmental across five domains: communication, gross motor, fine motor, problem solving, and personal-social.

To be included in the study as a child without communication delays, participants needed to meet the following criteria: (1) CDI: scores at or above the 20th percentile using gender-specific norms (M=40.45; SD=20.79) to designate expected performance in this area as children scoring at or below the 10th percentile on the CDI are typically noted as having a language delay, (2) PLS-5: total language standard score of 85 or higher (M=104.91; SD=7.15) to indicate performance within one standard deviation of the mean, and (3) ASQ-3: passing scores to signify consistency with typical developmental progress across a range of skill domains. Although 13 children were initially assessed for potential participation, two did not meet the eligibility criteria for inclusion in the study. See Tables 1 and 2 for descriptive information about the participants and obtained conversational samples.

Additional intake information was collected through parent interviews regarding birth and developmental history, potential presence of sensory deficits, and unusual family circumstances that may influence their child's performance. All parents reported typical birth and developmental histories and indicated no concerns for speech-language development or hearing and vision abilities. All parent participants were mothers and reported ethnicity as Caucasian. The highest level of

Table 1. Participant descriptive data and standardized measures

Participant	Age[a]	G[b]	CDI/CDI-III[c]	PLS-5 Expressive[d]		PLS-5 Auditory[e]		PLS-5 Total lang[f]		Utterances included[g]	
			Percentile	SS	Percentile	SS	Percentile	SS	Percentile	Session 1	2
A	31	F	25	103	58	98	45	100	50	100	100
B	25	M	70	119	90	118	88	120	91	100	100
C	31	M	20	103	58	104	61	104	61	100	100
D	30	F	25	111	77	98	45	105	63	94	100
E	25	M	20	100	50	86	18	92	30	100	100
F	33	F	55	106	66	112	79	110	75	100	100
G	26	F	25	103	58	106	66	105	63	100	100
H	31	F	75	103	58	107	68	105	63	100	100
I	26	F	40	106	66	106	66	106	66	100	100
J	25	M	30	91	27	106	66	98	45	48	73
K	25	F	60	116	86	100	50	109	73	98	100

[a]age in months; [b]gender; [c]MacArthur Bates Communicative Development Inventory – Words and Sentences (CDI)/MacArthur Bates Communicative Development Inventory Extension (CDI III) administered to children over 30 months of age. Five participants were 30 months of age or older so these participants have a reported percentile rank from the CDI III; [d]Preschool Language Scales – 5th edition, Expressive Communication subtest; [e]Preschool Language Scales – 5th edition, Auditory Comprehension subtest; [f]Preschool Language Scales – 5th edition, Total Language Score; [g]Number of intelligible utterances included in analysis corpus by session (maximum: 100).

Table 2. Descriptive language sampling data from sessions 1 and 2

Language sample measures	Session 1		Session 2		Mean difference
	M	SD	M	SD	
Mean length of utterance (MLU)	2.25	0.87	2.30	0.93	+0.05
# of different words used	75	28.07	78	34.6	+3.00
Total # of words used	222	84.03	225	89.3	+3.00
Type-token ratio (TTR)[a]	0.34	0.07	0.35	0.06	+0.01

[a]TTR was calculated by dividing the # of different words by the total # of words and multiplying the result by 100.

maternal education ranged from "some college" (n=2) to post-graduate work (n=5) with one parent declining to report.

Setting and procedures

Following completion of the screening measures and parent interviews, the first of two spontaneous conversation samples was obtained. The second sample was collected one week later. All screening and data collection was conducted in a clinic room designed for individual sessions in a university speech-language pathology clinic. Using procedures consistent with Morris [8] and Wittler and DeVeney [10], two 20-minute spontaneous conversation samples were collected from each child one week apart under identical circumstances. The samples were collected during play sessions between the child and the child's parent as they interacted together with researcher-provided, age-appropriate toy sets (e.g. farm set, cars/garage set, grocery cart/groceries set, blocks/construction set). All parents were provided the same standard instructions prior to each 20-minute play session, "I want to see what kind of activities [your child] enjoys. I'd like to see how [your child] communicates when s/he enjoys what s/he is doing. So, play and have fun. Help [your child] enjoy what s/he's doing". During the play sessions, the child and caregiver could utilize multiple toy sets simultaneously or one at a time based on the child's interest.

All play sessions were video recorded for later review using a Cannon HD R500 camcorder with a mounted external microphone. Consistent with procedures utilized by Binger Ragsdale, and Bustos [34], each camera and tripod were moved as needed during the sample to provide maximal view of the child's face without interrupting the play interactions. Because the sound quality from the camera with external microphone was sufficient for transcription based on the judgement of the first author after its tested use and confirmed by student transcriptionists, additional microphones were not employed during the data collection. Although the child participants did move around the room, the clinic rooms were small, limited spaces and the tripod with camera and microphone were adjusted and moved as needed to capture audio and video recordings.

After data collection, intelligible utterances were transcribed using the International Phonetic Alphabet (IPA) by three trained independent coders who were not involved in collecting the samples. The coders were upper-level undergraduate students in speech-language pathology who had completed formal training in phonetic transcription as part of a university course they successfully finished the previous semester. In addition, they received further training by the first author on IPA transcription procedures using videos from participants who did not meet criteria to continue in the present study. The first author is a licensed speech-language pathologist with formal training in phonetic transcription and analysis who teaches university coursework on pediatric speech sound disorders which involves consistent use and demonstration of IPA transcription techniques. For coding purposes, 'utterance' was defined according to linguistic conventions as a unit of speech bounded by breaths/pauses and 'words' included both words that met adult standards of production as well as immature forms of words children produce that have meaning.

During training, when instances of disagreement occurred on particular phones, student coders and the author reviewed the child's production of the disagreed-upon utterance and reached an agreement. Initial inter-rater reliability with the author and the other coders ranged from 80-86% (20-14% disagreement); however, after reviewing all utterances with disagreements, the transcribers and author resolved 100% of disagreements by discussing the videos together. All students were trained to note the session time of each transcribed utterance to ease utterance matching across transcription comparisons. Each of the 22 sessions (11 participants x 2 sessions) was transcribed by two coders and the resulting transcriptions were compared. Although all vocalizations were transcribed and glossed, only vocalizations interpreted as productions of the same word by both coders were eligible for inclusion in the analysis corpus. Secondly, among these agreed-upon words, only those in which the coders agreed upon the phonetic transcription of consonant sounds were included in the analysis corpus. Using these criteria, initial agreement was 72.30% (27.7% disagreement), closely aligned with the reliability noted by Morris [8], 73% agreement (27% disagreement). When instances of disagreement occurred between coders, together the two reviewed the child's production of the disagreed upon utterance and reached an agreement. If an agreement could not be reached, the utterance was not used in the final analysis.

Low inter-rater transcription reliability is a documented concern within the field [35,36], particularly when phonetically transcribing productions of infants and young children due to a variety of issues including their limited phonological development, lack of well-formed speech sounds, transcribers confidence and experience, and lack of valid and reliable methods to accomplish this task in research and clinical practice [36-38]. However, given the high social validity of language and communication sampling [39,40], this type of measurement continues to be a valued, though clinically underutilized (see [41] tool for assessing phonological productions.

Following transcription and establishment of the analysis corpus, the first 100 intelligible utterances of each resulting corpus were further analyzed for the present investigation. As shown in Table 1, most participant corpus data met this established threshold. Because their samples did not meet the threshold, all intelligible utterances were included in the analysis for Participants D, J, and K. The author trained an upper-level undergraduate student majoring in speech-language pathology who was not involved in data collection or transcription to conduct the data analyses which consisted of mean length of utterance (MLU), type-token ratio (TTR), and the two targeted independent analyses: phonetic inventory and word shape analysis.

At the end of training, the inter-rater reliability between the student and author was 100% for MLU calculations, 99% for TTR, 97% for phonetic inventory, and 98% for word shape analysis. During data analysis, the author re-calculated MLU, TTR, phonetic inventory, and word shape analysis for 20% of the sample, resulting in inter-rater reliability of 94-97% for MLU calculation, 95-98% for TTR calculation, 94-100% for phonetic inventory, and 95-98% for word shapes.

Phonetic Inventory

Data documented for the phonetic inventory analysis included word position of the phone (i.e. initial, medial, final word positions) to provide a descriptive account of phone production progression within and across word positions and presence of consonant clusters in the sample. This word position categorization was used to align with Stoel-Gammon and Dunn's description [4] as well as guidelines for clinical use provided by Watson, Murthy, and Wadhaw [42]. Consequently, for the present study, the use of "medial" was selected for a word position construct to maintain consistency with [4,42].

All phones produced were included in the phonetic inventory, including those produced as substitutions. In an effort to provide further descriptive information regarding phone production stability, phones produced at least twice in two different words at a given position were additionally categorized as "productive" according to procedures described by Watson et al. [42]. Phones that did not meet this criterion were additionally categorized as "emerging" [42]. For example, if a child produced [k] in cat and car, the [k] was classified as productive in word-initial position. If the child produced [k] in only cat, the sound was classified as emerging in word-initial position. This categorization aligned with Stoel-Gammon and Dunn's identification of "consistent" versus "marginal" phone use [4] as well as Watson et al.'s [42] more lenient categorical criteria. Stoel-Gammon and Dunn [4] advocated for a phone to be deemed "marginal" when produced less than three times in a given word position, but Watson et al. [42] used a less stringent criteria of less than two times. The researcher opted to adhere to criteria provided by Watson et al. [42] as these guidelines were written explicitly for clinical practice. Once categorized, productive and emerging sound were then analyzed when combined together for a total number of phones produced across word positions and clusters.

Word shape

A word shape analysis categorizes productions in terms of consonant-vowel sequences. For example, the word *foot* [fʊt] has a CVC structure while *football* [fʊtbɔl] has a more complex CVCCVC structure. In an independent analysis, a production of [fʊ] for *foot*, would be classified as CV regardless of the intended target word shape. There are a number of different conventions for measuring syllable and/or word shape complexity with young children such as the *Mean Babble Level* [43], *Syllable Structure Level* [19] and the *Index of Phonetic Complexity* [44]. Although other measurement options were considered for this study, the researcher decided to use a simple count and tally system in an attempt to replicate what is likely typical clinical practice for assessing word shape variability with young children. Only word shapes with at least one vowel were included since vowels serve as a syllable's nucleus, consistent with syllable descriptions from Stoel-Gammon [45] and following specific procedures outlined by Watson et al. [42]. Thus, a prolonged [ʃ] for a gas sound with the cars/garage set was not included in analysis. Word shapes that included one vowel or diphthong were considered monosyllabic, and word shapes that included more than one vowel/diphthong and could be segmented at the level of the syllable were considered multisyllabic [5]. An inventory of all word shapes present in each sample was recorded and the count tallied when used multiple times.

Language sample descriptors (Table 2) and independent measures (Table 3) were reported for each data collection session. Additionally, Wilcoxon's sighned-ranks comparisons

Table 3. Descriptive information from the independent measures for sessions 1 and 2

Independent measures	Session 1		Session 2		Mean difference
	M	SD	M	SD	
Productive initial consonants[a]	12.18	3.74	12.36	3.33	+0.18
Emerging initial consonants[b]	2.64	1.43	3.00	2.37	+0.36
Productive medial consonants	4.27	2.57	4.36	2.54	+0.09
Emerging medial consonants	3.09	1.38	3.91	1.51	+0.82
Productive final consonants	7.36	3.08	7.27	3.41	-0.09
Emerging final consonants	2.36	1.75	3.09	1.58	+0.73
Productive consonant clusters	1.73	1.56	2.73	3.61	+1.00
Emerging consonant clusters	7.09	4.91	9.18	6.31	+2.09
# of different word shapes[c]	14.82	4.79	16.91	7.02	+2.09
# of monosyllabic word shapes[d]	7.00	2.41	7.73	2.15	+0.73
# of multisyllabic word shapes[e]	7.82	2.68	9.18	5.56	+1.36

[a]"Productive" consonants and consonant clusters were those produced at least twice in two different words at a given word position; [b]"Emerging" consonants and consonant clusters were those produced in a given word position only once during the sample; [c]# of different word shapes comprised a count of all the differing word shapes used in the sample that included at least one vowel; [d]Monosyllabic word shapes were those word shapes that included one vowel or diphthong (e.g. CV, CVC); [e]Multisyllabic word shapes are those word shapes that included more than one vowel or diphthong (e.g. CVCV).

Table 4. Inferential statistics for selected language sampling and independent measures for sessions 1 and 2

Independent measures	Session 1		Session 2		Wilcoxon z	Significance	Spearman ρ	Significance
	M	SD	M	SD				
Mean length of utterance (MLU)	2.25	0.87	2.30	0.93	0.800	0.424	0.856	0.001*
Total[a] initial consonants	14.82	5.61	15.36	5.56	0.256	0.798	0.854	0.001*
Total medial consonants	7.36	2.10	8.27	2.05	1.132	0.258	0.751	0.008
Total final consonants	9.73	3.54	10.36	3.36	0.716	0.474	0.751	0.008
Total consonant clusters	8.82	4.49	11.91	6.00	1.339	0.181	0.825	0.002*
# of different word shapes	14.82	4.79	16.91	7.02	1.177	0.239	0.624	0.040

[a]Totals include both productive and emerging consonant productions.
*Statistically significant result.

and Spearman's correlations were conducted to assess changes in measures between data collection sessions (Table 4). Nonparametric inferential analyses were conducted due to the study's relatively small sample size and potential for heterogeneity within the group data as participants' assessment scores indicated a wide range of non-disordered language skills. Because of these limitations, the study data could not be determined to meet normal distribution assumptions and, thus, nonparametric analyses were employed. The statistical analyses, Wilcoxon sighned-ranks comparison and Spearman's correlation, were both performed using the SPSS version 24.0 statistical software program [46]. A Bonferroni correction was applied for six comparisons (five independent phonological measures and one language sample measure) by dividing the predetermined critical *p*-value by the number of comparisons being made (i.e. 0.05 divided by six) to protect against experiment-wise error. Thus, the acceptable significance value was ≤0.008. Consistent with conditions in the Morris [8] study, for a measure to be considered reliable, it had to meet two criteria: (a) no significant differences between data collection sessions as noted by Wilcoxon's sighned-ranks comparisons and (b) high test-retest correlations as noted by Spearman's correlations.

RESULTS

Temporal stability of language samples

In order to support meaningful comparisons regarding the reliable use of informal phonological measures across sessions, temporal stability of the language samples needed to be established. Time length was held constant at 20 minutes, representing duration stability across the two sessions. Mean MLUs across the two sessions ($M=2.25$, $SD=0.87$, *range*= 1.21-3.69 in Session 1, $M=2.3$, $SD=0.93$, *range*=1.08-4.14 in Session 2) showed no significant differences ($z=0.93$, $p=0.800$) and were significantly positively correlated ($\rho=0.856$, $p=0.001$), indicating stable mean utterance lengths across the two sessions. Further, the samples were descriptively, but not inferentially analyzed for total number of words used (*range*= 74-361 for Session 1, 101-394 for Session 2) and total number of different words (NDW) used (*range*=23-127 and 35-153, respectively), which were employed to calculate TTR. The mean NDW for session 1, 75.36 ($SD=26.76$), and session 2, 78.27 ($SD=33.00$) differed by only +2.91, indicating consistency across the two sessions although a broad range of different words used were noted across participants, as signified by standard deviation measurements. The mean TTR for session 1, 0.34 (*range*= 0.24-0.42, $SD=0.06$), and session 2, 0.35 (*range*= 0.25-0.44, $SD=0.06$) differed by only +0.01, indicating consistent measurements of word types used across the two sessions.

As noted previously, most participants met the 100-utterance threshold across both sessions and of the three who did not, only participant J exhibited a substantial difference in total utterances across the two sessions, 48 to 73. However, even this participant demonstrated relatively stable measures of MLU (1.21 and 1.08, respectively) and TTR (0.31 and 0.35, respectively) across the two sessions. Consistent with the language samples obtained by Morris [8], the participants of the present study who produced higher total numbers of words likewise produced higher total numbers of different words across both language samples. All together, these indicators demonstrated relative temporal stability across the language samples obtained during the two data collection session.

Test-retest reliability of phonetic inventory

There were no significant differences between sessions 1 and 2 for the number of consonants in the phonetic inventory (initial, medial, final, consonant clusters). Exact consonant productions in each session differed somewhat, but a general pattern emerged across sound classes regarding manner of production in that stop-plosive and nasal phones occurred the most frequently across participants and samples with fricatives and glides occurring less frequently and affricates and liquids occurring the least often. Further, this production pattern was noted across word positions with far more consonant sounds produced in initial position - and, of these, of particular consistency were stop-plosives and nasals - than those present in medial and final positions. Regarding place of production stability, bilabials such as [p], [b], [m], and [w] occurred the most frequently and consistently across participants and samples, followed by sounds with lingua-alveolar placements such as [t], [d], [s], [z], and [n]. Interdental placements were present least often. Exact consonant cluster productions differed more dramatically across participants and samples than did singleton consonant productions; however, among clusters those containing [s] and [r]/[w] (with [w] a frequent substitution of [r] in clusters) represented the majority of initial word position clusters. There were fewer clusters noted and much more variability across productions in medial word position with very little continuity across exact cluster productions. However, although a variety of consonant clusters were present in final word position, the clusters [ts], [nt], and [ŋk] occurred more frequently across participants' samples.

For the purposes of this study, number, not type of consonants and clusters, was used for correlations to facilitate comparisons with previous studies conducted in this area (e.g. Morris, [8]). No significant differences were found for initial consonants productions measured in session 1 and session 2 ($z=0.256$, $p=0.798$), medial consonant productions ($z=1.132$, $p=0.258$), final consonant productions ($z=0.716$, $p=0.474$), and consonant cluster productions ($z=0.181$, $p=0.825$).

There were statistically significant strong positive correlations between session 1 and session 2 for initial consonant productions ($\rho=0.856$, $p=0.001$) and consonant cluster productions ($\rho=0.825$, $p=0.002$). Further, there were strong positive correlations for medial ($\rho=0.751$, $p=0.008$) and final consonant productions ($\rho=0.751$, $p=0.008$) that achieved significance with the Bonferroni correction. Given the criteria established above for reliability, it had to meet two criteria: (a) no significant differences between data collection sessions as noted by Wilcoxon's sighned-ranks comparisons and (b) high test-retest correlations as noted by Spearman's correlations.

Test-retest reliability of word shape analysis

Although there were no significant differences between sessions for word shape analysis results ($z=1.177$, $p=0.239$), there was not a significant test-retest correlation for this measure ($\rho=0.624$, $p=0.040$). Rather, a non-significant positive test-retest correlation was noted. Given the criteria established above for reliability, the word shape analysis did not meet both criteria. Specifically, there was no significant differences between data collection sessions as noted by Wilcoxon's sighned-ranks comparisons, but significant test-retest correlations as noted by Spearman's correlations was not present.

Differences in means across the two sessions ranged from +0.73 for monosyllabic word shapes to +1.36 for multisyllabic word shapes indicating more variability across the two sessions in multisyllabic word shape productions. The most stable monosyllabic word shapes across participants and samples were CVC shapes, followed by shapes consisting of CV and V. There were much fewer and more varied multisyllabic shapes produced; however, the most stable across participants and samples were CVCV, followed by CVCVC, and then VCVC.

DISCUSSION

Phonetic inventory

This study addressed research questions regarding the short-term (within one week) test-retest reliability of phonetic inventory and word shape analysis for 2-year-old children without communication delays when calculated from 20-minute conversational samples. The present study findings indicated four stable measures of phonetic inventory: consonants in initial word position and those in clusters and to a lesser extent, word medial and final consonants. Each will be discussed in the following paragraphs.

Consonants in initial word position

The present study outcome regarding initial consonant consistency partially supported the results of Wittler and DeVeney [10] who observed one of three typical 2-year-old participants obtained consistent phonetic inventories for word-initial sound productions. The finding differed from that of Morris [8] who reported that word-initial consonants were the least stable measure of phonetic inventory for 18-24-month-old children. A possible explanation for this discrepancy in the literature is the nature of variability in speech sound development and mastery. As children age, their phonetic and phonemic repertoires expand and productions tend to become more stable [21,22,24]. This process of expansion and stabilization is typically demonstrated first in initial position of syllables and words, and substantive differences are noted even between a few months of age among toddlers [21,24,25]. This process was noted in the data collected for the present study as well. Far more consonant sounds were present in initial word position compared to medial and final positions. However, it should be noted that researchers have reported on a lack of articulatory stability across children's speech sound productions and found high variability for children within this study's age group (see McLeod & Hewett [13]). While variability of productions may decrease with age (Holm et al., [11]), it does not entirely dissipate and may continue to limit the reliable use of independent measures across populations of young children as well as populations presenting with speech sound disorders [12].

Consonant clusters

Present findings indicated measures of phonetic inventory for consonant clusters were stable over a short-term (one-week) time frame. This finding offered partial support for results noted by Wittler and DeVeney [10] in which one of three participants had consistent consonant cluster productions. Further comparison to existing literature is limited by the few investigations of phonetic inventory with toddlers in which consonant clusters were included. Most investigators only reported findings for initial and final word/syllable positions (e.g. [8,18,21-24]. However, Watson and Scukanec [25] examined consonant cluster productions with 2-year-old children and noted inconsistent consonant cluster productions across older 2-year-olds (30- to 33-month-olds) and a lack of consonant cluster productions in younger 2-year-olds (24- to 29-month-olds). The data from the present study were not analyzed in terms of age range differences across participants and with only a cursory observation of type across word positions. Given previous findings of variability in consonant cluster productions for 2-year-olds, the present study findings of reliable consonant cluster profiles should be interpreted with cautious optimism until additional study outcomes are reported to support or refute.

Medial consonants

Statistical analyses from the present study indicated that two additional measures of phonetic inventory, medial and final consonant productions, were also significantly reliable and stable over the one-week time frame. More research is needed to corroborate or refute the present study findings regarding consonants in medial position. Claessen et al. [2] noted the dearth of comparable study findings related to 2-year-old speech which they attributed to a number of factors such as limited number of participants, narrow age range restrictions, and limited ranges of speech behaviors described, indicating a continued need for further investigation in this area.

Final consonants

The finding regarding final consonants provides support for results noted by Wittler and DeVeney [10], in which one of three participants had consistent final consonant productions, and Morris [8], who noted moderate but not significant reliability for word-final phonetic inventory. As noted by McIntosh and Dodd [17], many 2-year-olds consistently use the phonological error pattern of final consonant deletion. Additionally, most 2-year-olds, when they do produce word-final consonants, show a restricted production variability compared with initial consonants. At this age, most of final consonant productions include stop-plosives (e.g. [p], [k]), but some may include the presence of a nasal, fricative, and/or liquid production [24]. The descriptive data derived from this study suggested a similar pattern of heavy utilization of stop-plosives followed by nasals compared with fricatives and glides, which occurred less often, and affricates and liquids, which occurred least frequently. Due to infrequent use of final consonants in general and restrictions in the manner of productions when used, it is not entirely surprising to note that this aspect of phonetic inventory showed short-term reliability.

Word shape analysis

In the present study, consistent with the researcher hypothesis, the measure of word shape did not indicate short-term reliability across data collection sessions. This finding is in contrast to Wittler and DeVeney [10] who noted that the word shape analysis studied was relatively stable over time. However, as noted in the introduction, different procedures for word shape analyses have been used and studied. Those who used a “closed-set” procedure in which specific word shapes were targeted [8,10], noted a ceiling effect across participants. To avoid a similar issue and reflect variability in clinical practice use, an “open-ended” tally procedure was utilized in the present study. Watson and Scukanec [25] noted substantial changes in word shapes present in the speech of 24- to 36-month old children characterized by decreased productions of CV shape over time and increased productions of CVC, CCVC, and CVCC words as children begin to produce more word-final consonants and consonant clusters. These changes in word shape production across the second year of life may render open-ended tallies of word shape productions an unreliable measure given the great deal of variability noted in the population. However, a closed-set analysis indicated more reliability, but a ceiling effect for 2-year-olds may be noted depending on shapes targeted.

Clinical significance

For 2-year-olds without communication delays, the measure of phonetic inventory calculated from play-based conversational speech samples was significantly reliable given a short timeframe (one week) across word positions (initial, medial, and final) and consonant clusters. This finding is encouraging news for researchers and practitioners alike who are assessing the speech sound productions of this important age group. Indications that short-term stability is present for this informal measure is helpful for clinicians who rely on this measure for accurate diagnostic information and researchers who use this measure to compare within and across participants to determine appropriate speech sound acquisition.

To the contrary, the word shape measure assessed did not indicate short-term reliability for this young population. Findings indicated that word shape analysis procedures following an open-ended tally method should be used with caution in both clinical and research environments until more information is available to support or refute the present study results.

Taken together, the present study findings regarding phonetic inventory and word shape analysis showed partial support for continued use of independent analyses along with standardized assessment tools when conducting a comprehensive assessment of speech-language skills with young children. Informal measures should be used with caution when assessing and comparing the speech sound production skills of 2-year-old children with atypical communication skills because little is known about the reliability of these measures with clinically-relevant populations [9]. However, present study findings regarding young children without communication delays assist in establishing the evidence-base for the earliest ages at which SLPs can reliably use informal measures to assess speech sound production skills. For this, information is needed for populations presenting with and without communication delays. Until evidence of reliability is provided regarding clinically-relevant populations (e.g., toddlers who are late talkers, those presenting with communication disorders such as language and/or speech delay/disorder), clinicians should continue to rely on a variety of data sources during diagnostic evaluations invoking both formal and informal measures when available such as standardized assessments, parental report, and observational data [1]. SLPs should not determine the presence of a speech-language deficit based solely on the results of informal assessment tools.

Limitations and future directions

Although the homogeneous small sample represented in the present study was consistent with previous research in this area, a larger more heterogeneous sampling would benefit inferential data analyses and provide a more reflective exemplification of the larger population. Additionally, two of the 11 participants were high performers and scored higher than typical range on screening measures, which could potentially present a confounding variable regarding the target population's authentic representation in the study. Although parents were asked to report the highest level of maternal education, this information was not used in the study analysis to determine the impact of high maternal education level on participant outcomes. Additionally, even though data was collected regarding the language composite of each participant sample, this data was not analyzed beyond its relative stability for conducting the speech sound analyses. This type of comparative analysis was not the focus of the present paper; however, further research regarding the interconnection of language and speech is warranted and important for clinical practice implications.

There are inherent risks for use of statistical analyses with small studies for which sampling procedures could influence statistical outcomes. Confidence in the generalizability of the study results would be enhanced by replication of the present study with a larger sample of diverse participants including those from under-represented geographic, ethnic, and/or socioeconomic groups. The screening measures included in the present study did not include a standardized measure of speech sound/phonology production. Although parents reported no concerns about speech or language development for all participants, a standard measure of this content would add value to individual participant descriptions and data analyses. Similarly, a report of parent responses to the representativeness of their child's speech across collected samples (e.g., "Was your child's speech typical today"?) would offer information regarding the ecological validity of the work.

Another limitation involved differences in measurement and reporting procedures across similar studies. Differences occurred in measurement and reporting procedures of the targeted informal independent phonological analyses between this and previous yet similar studies. Systematic investigations of reliability for these analyses would benefit from standardized reporting conventions and procedures utilized (e.g., reporting all productions across all word positions for phonetic inventories and use of standardized procedures for word shape analyses). There are also reliability limitations to measures of lexical diversity such as TTR that would benefit from researchers consistently obtaining a larger spontaneous speech sample from participating children. To further expand on issues of limited consistency across studies, consistent use of terminology in reporting and analysis regarding word positions would also benefit from standardization. For example, rather than using the term "medial" to describe word position, researchers would add more precision to discrete categorical descriptions by adhering to terminology introduced by Grunwell in 1985: Syllable initial within word (SIWW) and syllable final within word (SFWW).

The present study - and most studies involving toddlers - conducted analyses based on spontaneous speech samples, which present inherent limitations in content stability despite context controls. An analysis of phones produced is dependent on the number of items attempted within that sample for each phone in each word position. For example, the present study included one child who had only 48 utterances in session 1, which was less than half compared to the majority of participants. However, Heilmann et al. [27] noted stability for informal measurements based on structured conversations with kindergarten-aged students. Determining the utility of this type of conversational sampling technique with toddler and/or preschool populations would be a worthy investigation that could further inform effective evidence-based practices. Finally, the present study represents reliability of measures based on phonological productions of young children who do not present with communication delays. Clinicians need this same type of information pertaining to clinically-significant populations of young children (e.g., late talkers, those with speech sound disorders and/or childhood apraxia of speech) in order to make well-informed clinical decisions regarding assessment and treatment.

Conclusion

The purpose of the present study was to determine the short-term temporal reliability of phonetic inventory and a word shape analysis for 2-year-old children without communication delays when calculated based on spontaneous conversational samples obtained during parent-child play-based interactions. Results of the study indicated measures of phonetic inventory for consonants across word positions (initial, medial, and final) and in clusters were stable over a short-term (one-week) time frame. However, the word shape analysis used was not determined to have short-term reliability, but this could be a function of the nature of the open-ended tally procedure used. These findings indicated partial support for the continued use of independent analyses for clinicians and researchers working with this young population in that support was implicated for the reliable use of phonetic inventory, but not for the word shape analysis studied. Until more is known regarding the reliability of these measures for use with young, clinically-relevant populations, SLPs should continue to rely on a variety of data sources during diagnostic evaluations including cautious but optimistic use of phonetic inventory for reliable determination of speech sound productions present in the speech of young children.

ACKNOWLEDGMENTS

This research was supported by the American Speech-Language-Hearing Association's Advancing Academic Research Career (AARC) funding mechanism and a University of Nebraska at Omaha Research and Creative Activity faculty grant both awarded to the first author.

REFERENCES

1. Paul R. Introduction to clinical methods in communication disorders. 3rd ed. Baltimore, MD: Brookes; 2014.
2. Claessen M, Beattie T, Roberts R, Leitao S, Whitworth A, Dodd B. Is two too early? Assessing toddlers' phonology. Speech, Language, and Hearing. 2017;20:1-11.
3. Crais ER. Testing and beyond: Strategies and tools for evaluating and assessing infants and toddlers. Language, Speech, and Hearing Services in Schools. 2011;42:341-364.
4. Stoel-Gammon C, Dunn C. Normal and disordered phonology in children. Baltimore, MD: University Park Press, 1985.
5. Bernthal J, Bankson N, Flipsen P. Articulation and phonological disorders: Speech sound disorders in children. 8th ed. Boston: Pearson; 2017.
6. Ferguson CA, Farwell CB. Words and sounds in early language acquisition. Language. 1975;51:419-439.
7. Sosa AV, Stoel-Gammon, C. Patterns of intra-word phonological variability during the second year of life. Journal of Child Language. 2006;33:31-50.
8. Morris SR. Test-retest reliability of independent measures of phonology in the assessment of toddlers' speech. Language, Speech, and Hearing Services in Schools. 2009;40:46-52.
9. DeVeney S, Sheridan K. Short-term stability of phonological measures in a sample of two-year-old late talkers. Clinical Archives of Communication Disorders. 2017;2:227-237.
10. Wittler K, DeVeney S. Test-retest reliability of independent phonological measures of 2-year-old speech: A pilot study. Journal of Special Education and Rehabilitation. 2016;17:71-88.
11. Holm A, Crosbie S, Dodd B. Differentiating normal variability from inconsistency in children's speech: Normative data. International Journal of Language and Communication Disorders. 2007; 42:467-486.
12. Kim HY, Ha S. Articulatory variability in 24- to 36-month-old typically developing children. Communication Sciences and Disorders. 2016;21:333-342.
13. McLeod S, Hewett SR. Variability in the production of words containing consonant clusters by typical 2- and 3-year-old children. International Journal of Phoniatrics, Speech Therapy and Communication Pathology. 2008;60:163-174.
14. Sosa AV. Intraword variability in typical speech development. American Journal of Speech-Language Pathology. 2015;24:24-35.
15. Macrae T. Lexical and child-related factors in word variability and accuracy in infants. Clinical Linguistics & Phonetics. 2013;27:497-507.
16. Sosa AV, Stoel-Gammon C. Lexical and phonological effects in early word production. Journal of Speech, Language, and Hearing Research. 2012;55:596-608.
17. McIntosh B, Dodd, BJ. Two-year-olds' phonological acquisition: Normative data. International Journal of Speech-language Pathology. 2008;10:460-469.
18. Carson CP, Klee T, Carson DK, Hime, LK. Phonological profiles of 2-year-olds with delayed language development: Predicting clinical outcomes at age 3. American Journal Speech-Language Pathology. 2003;12:28-39.
19. Paul R, Jennings P. Phonological behavior in toddlers with slow expressive language development. Journal of Speech and Hearing Research. 1992;35:99-107.
20. Rescorla L, Ratner NB. Phonetic profiles of toddlers with expressive language impairment (SLI-E). Journal of Speech and Hearing Research. 1996;39:153-165.
21. Robb MP, Bleile KM. Consonant inventories of young children from 8 to 2 months. Clinical Linguistics and Phonetics. 1994;8:295-320.
22. Stoel-Gammon C. Phonetic inventories, 15-24 months: A longitudinal study. Journal of Speech and Hearing Research. 1985;28:505-512.
23. Van Severen L, Van Den Berg R, Molemans I, Gillis S. Consonant inventories in the spontaneous speech of young children: A bootstrapping procedure. Clinical Linguistics and Phonetics. 2012;26: 164-187.
24. Stoel-Gammon C. Phonological profiles of 2-year-olds. Language, Speech, and Hearing Services in Schools. 1987;18:323-329.
25. Watson MM, Scukanec GP. Profiling the phonological abilities of 2-year-olds: longitudinal investigation. Child Language Teaching and Therapy. 1997;13:3-14.
26. Limbrick N, McCormack J, McLeod S. Designs and decisions: The creation of informal measures for assessing speech production in

children. International Journal of Speech-Language Pathology. 2013;15:296-311.

27. Heilmann J, DeBrock L, Riley-Tillman TC. Stability of measures from children's interviews: The effects of time, sample length, and topic. American Journal of Speech-Language Pathology. 2013;22: 463-475.

28. Gershkoff-Stowe L, Smith LB. A curvilinear trend in naming errors as a function of early vocabulary growth. Cognitive Psychology.1997;34:37-71.

29. Owens RE. Language development: An introduction (8th ed.). Boston, MA: Allyn & Bacon; 2012.

30. Hedge MN. Treatment in speech-language pathology (4th ed.). San Diego, CA: Plural Publishing; 2018.

31. Fenson L, Marchman VA, Thal DJ, Dale PS, Reznick, JS, Bates E. MacArthur communicative development inventories - 2nd edition. Baltimore, MD: Brookes; 2007.

32. Zimmerman IL, Steiner VG, PondRE. Preschool Language Scale - Fifth Edition. San Antonio, TX: Psychcorp; 2011.

33. Bricker D, Squires J. Ages & Stages Questionnaires. Baltimore: Brookes; 2003.

34. Binger C, Ragsdale J, Bustos A. Language sampling for preschoolers with severe speech impairments. American Journal of Speech-Language Pathology, [Advance online publication]. 2016;1-15.

35. Oller DK, Ramsdell HL. A weighted reliability measure for phonetic transcription. Journal of Speech, Language, and Hearing Research. 2006;49:1391-1411.

36. Preston JL, Ramsdell HL, Oller KD, Edwards ML, Tobin, SJ. Developing a weighted measure of speech sound accuracy. Journal of Speech, Language, Hearing Research. 2011;54:1-18.

37. Munson B, Johnson JM, Edwards J. The role of experience in the perception of phonetic detail in children's speech: A comparison between speech-language pathologists and clinically untrained listeners. American Journal of Speech-language Pathology. 2012; 21:124-139.

38. Ramsdell HL, Oller, DK, Ethington CA. Predicting phonetic transcription agreement: Insights from research in infant vocalizations. Clinical Linguistics & Phonetics. 2007;21:793-831.

39. Flipsen P. Measuring the intelligibility of conversational speech in children. Clinical Linguistics and Phonetics. 2006;20:303-312.

40. Kwiatkowski J, Shriberg LD. Intelligibility assessment in developmental phonological disorders: Accuracy of caregiver gloss. Journal of Speech and Hearing Research. 1992;35:1095-1104.

41. Pavelko SL, Owens Jr, RE, Ireland M, Hahs-Vaughn DL. Use of language sample analysis by school-based SLPs: Results of a nationwide survey. Language, Speech, and Hearing Services in Schools. 2016;47:246-258.

42. Watson M, Murthy SN, Wadhaw N. Phonological Analysis Practice. Eau Claire, WI: Thinking Publications; 2003.

43. Stoel-Gammon C. Language Production Scale. In Olswang, L., Stoel-Gammon, C., Coggins, T., & Carpenter, R. (Eds.), Assessing prelinguistic and early linguistic behaviors in developmentally young children (pp. 120-150). Seattle, WA: University of Washington Press; 1987.

44. Jakielski KJ, Maytasse R, Doyle E. Acquisition of phonetic complexity in children 12-36 months of age. Poster session presented at the annual convention of the American Speech- Language-Hearing Association, Miami, FL, 2006.

45. Stoel-Gammon C. Prespeech and early speech development of two late talkers. First Language. 1989;26:207-223.

46. IBM Corporation. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp; 2016.