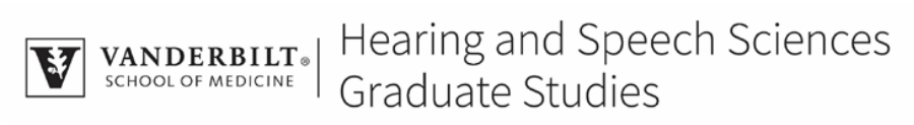


# An Investigation of a Sentence Imitation Task as a Measure of Speech Sound Accuracy

**Kaitlyn J. Hamers, B.A.**

Vanderbilt University School of Medicine

Department of Hearing and Speech Sciences

Nashville, TN

**Master's Thesis**

completed in fulfillment of the requirements of the Thesis Specialty Track

within the Master of Science in Speech-Language Pathology Program

Defense Date:

May 12, 2022

C. Melanie Schuele, Ph.D., Committee Chair

Stephen Camarata, Ph.D.

Tiffany Woynaroski, Ph.D.

Antje Mefferd, Ph.D.

Jacob Feldman, Ph.D.

Ian Morton, Ph.D.

**Abstract**

**Purpose:** The long-term goal of this line of inquiry is to develop an ecologically valid and efficient measure of speech sound accuracy in connected speech that is suitable for use in clinical practice for children aged 3;6-4;11. In this study, we explored a novel sentence imitation task, the Story-Sentence Imitation Task for Speech (SSITS), as a measure of speech sound accuracy. We investigated whether the measure (1) leads children to repeat a sufficient proportion of target consonants, (2) demonstrates inter-rater, intra-rater, and test-retest reliability and scoring stability, and (3) demonstrates convergent validity with published measures of speech sound accuracy and intelligibility, and (4) whether a scoring training leads speech-language pathologists (SLPs) to follow the intended protocols for deriving the final score on the SSITS.

**Methods:** Eleven typically developing children with normal receptive language skills, between the ages of 36 and 59 months, participated. Speech sound production abilities freely varied so that participants' speech production skills encompassed a wide range of accuracy. Children completed a battery of assessments including a single-word intelligibility measure, a word-level speech accuracy measure, and the SSITS. Nine children returned for a second study visit, in which they completed only the SSITS. The SSITS was presented as a story script accompanied by a wordless picture book, where the children repeated each utterance after the examiner. Target sounds were selected for scoring based on phoneme frequency in child conversational speech. Five SLPs scored all child videos at least once after completing a self-guided PowerPoint training on the SSITS.

**Results:** The SSITS demonstrated high feasibility, good inter-rater reliability, excellent intra-rater reliability and test-retest reliability, high scoring stability, and strong convergent validity with a measure of single-word speech sound accuracy and a measure of intelligibility.

**Conclusion:** This study adds to the evidence base surrounding the assessment of children's speech sound production by demonstrating that a sentence imitation task presented as a narrative illustrated in a wordless picture book can have strong psychometric properties.

*Keywords:* speech accuracy, speech sound disorder, intelligibility, assessment

## An Investigation of a Sentence Imitation Task as a Measure of Speech Sound Accuracy

Speech sound assessments are essential in the field of speech-language pathology. They help clinicians to identify a child's speech errors and error patterns to determine the need for services, generate treatment goals, and plan the course of intervention. Anecdotally, the assessments that are used most frequently and perhaps almost exclusively to assess speech sound accuracy capture production only at the single-word level. However, the primary interests of speech sound assessment and intervention is speech sound production in connected speech, such as in conversation, and intelligibility. Intelligibility is generally defined as the degree to which a child's spoken message can be understood by a listener. However, direct assessments of intelligibility are time-consuming and resource-intensive, making their use challenging in routine clinical practice. Therefore, an assessment of speech sound accuracy in connected speech that is efficient would be ideal to capture initial status and measure change in speech sound accuracy over the course of intervention. Such an assessment could be more ecologically valid than a single-word speech accuracy measure, as well as perhaps provide an overall impression of intelligibility. Although speech sound accuracy is not equivalent to speech intelligibility, an investigation is warranted to determine how closely an assessment of speech sound accuracy in connected speech could approximate intelligibility. The purpose of the present study was to validate a sentence imitation (i.e., connected speech) measure of speech sound accuracy that could be used to assess severity and to track growth.

**Measuring Speech Sound Accuracy**

In a survey of practice patterns among speech-language pathologists (SLPs), Skahan et al. (2007) stated that 74% of participants self-reported that they “always” use single-word tests to determine percentile rank and standard score when evaluating children with suspected speech sound disorder. Two commonly used measures of speech sound accuracy are the Goldman-Fristoe Test of Articulation (GFTA; Goldman & Fristoe, 2015) and the Arizona Articulation and Phonology Scale (Arizona; Fudala & Stegall, 2017; Skahan et al., 2007). The word level subtests on these measures require the examinee (the child) to name pictures (i.e., speech targets). If the child does not appropriately name the picture, the examiner provides the target word and the child imitates it. The Arizona evaluates selected vowels, single consonants, and consonant clusters, whereas the most recent edition of the GFTA, the GFTA-3, evaluates all consonants and vocalic /r/s in the stimulus words. On the Arizona, scored speech sounds are weighted to derive a raw score, whereas on the GFTA-3, each speech sound contributes one point to the raw score. The Arizona and GFTA-3 also include a sentence-level speech production subtest. Skahan et al. did not appear to evaluate the extent to which SLPs use these sentence tasks. Anecdotally, however, on both tests, only the word-level subtest is routinely administered in clinical practice. Moreover, the sentence-level subtests on the Arizona and the GFTA are heavily loaded with “Late-8” sounds (Shriberg, 1993). The reliance on Late-8 sounds in the sentences likely makes the tasks inappropriate for capturing a global impression of speech sound accuracy, particularly for children with errors on earlier-developing sounds.

The Arizona-4 and GFTA-3 often are used to identify the speech sound errors a child makes with the end goal of making a clinical judgment as to the developmental appropriateness of the errors and the child's eligibility or need for speech intervention services. To make initial eligibility determinations, an assessment that tests speech sound production at the word level is often sufficient. However, speech accuracy at the single-word level typically precedes speech accuracy in connected speech (Glaspey et al., 2022). Moreover, emerging sounds may be produced more accurately on single-word assessments than in connected speech (Morrison & Shriberg, 1992). Therefore, single-word assessments may not be sensitive to errors that children make in connected speech. Particularly when a child is enrolled in speech therapy that focuses on drill practice of single words, the child may become adept at producing speech sounds at the single-word level but may not yet generalize that skill to connected-speech contexts. Therefore, re-assessing a child who has received speech sound intervention using a single-word speech sound assessment may not yield results that reflect the child's speech sound accuracy outside of the therapy environment (e.g., classroom, home) or in conversational speech. Thus, an assessment that measures speech sound accuracy in a functional, connected-speech context may provide a more valid picture of the child's speech sound performance, critical for measuring speech outcomes as intervention progresses.

**Measuring Intelligibility**

Experts in speech sound disorder routinely argue the necessity of evaluating speech intelligibility as a component of a comprehensive assessment for children with suspected speech sound disorder (Skahan et al., 2007). Skahan et al. stated that 75%

of their participants reported “always” estimating intelligibility, but they did not report the methods by which the survey respondents capture intelligibility. Moreover, intelligibility is influenced by many factors. Certainly, speech sound accuracy has a significant impact on intelligibility, but many other factors also play a role, including pragmatic, suprasegmental, and linguistic characteristics of utterances, context, clarity of the visual and auditory signals, and the listener’s familiarity with the speaker (Weston & Shriberg, 1992; Kent et al., 1994).

Speech intelligibility can be assessed broadly in two ways: scaling and item identification (Kent et al., 1989). Scaling methods involve a listener assigning a rating from a closed set forming a continuum of intelligibility. Such measures are easily implemented because they are time-efficient, accessible, and require only a single listener and a single speech sample (Schiavetti, 1992). However, scaling methods may not differentiate effectively between children whose intelligibility falls within the middle points on the scale. Therefore, scaling methods may not be helpful for monitoring treatment progress (Samar & Metz, 1988; Ertmer, 2010). Additionally, wide variability can be present among even experienced raters when assessing intelligibility via scaling methods (Schiavetti, 1992).

Although scaling methods have value in some situations, item identification measures can provide a more objective and sensitive measure of speech intelligibility (Ertmer, 2010). Item identification involves a listener listening to a child’s speech and attempting to identify the words produced (Schiavetti, 1992). Such tasks can either be open-set, where the listener freely generates their identification of the words in the utterance, or closed-set, where the listener is provided with possible interpretations in a

multiple-choice format (Ertmer, 2010). In both cases, the number of words understood correctly by the listener is divided by the number of total words attempted to provide a percentage or proportional measure of intelligibility (Schiavetti, 1992).

Open-set item identification has the most ecological validity in that it mimics a typical scenario of attempting to understand a child's speech without foreknowledge of what the child might be saying. However, it is very difficult to calculate percent intelligibility based on an open-set identification task because (1) the child's intended utterances are largely unknown or unverified, and (2) the number of words in a sample may be difficult to ascertain, particularly for children who are highly unintelligible. Several methods have been proposed for measuring intelligibility using open-set item identification (e.g., Ertmer, 2010; Flipsen, 2006; Yoder et al., 2016). Regardless of the method, scoring intelligibility based on connected conversational speech is demanding and time-consuming on the part of the clinician because it requires transcription of a language sample with unknown targets.

Closed-set item identification, on the other hand, does not parallel a listener's experiences in conversation. In conversation, a listener does not have access to the target utterance. Furthermore, it can be difficult to map performance on a closed-set identification task to true intelligibility. Morris et al. (1995) discussed a single-word closed-set intelligibility assessment, the Preschool Speech Intelligibility Measure (PSIM), which they concluded had concurrent validity with the first edition of the GFTA and with teacher ratings of intelligibility.

In sum, scaling methods of speech intelligibility are easy to administer but may not differentiate effectively between children whose intelligibility falls within the middle

categories. Item identification methods, in contrast, may be more objective and sensitive but are more costly. Currently, intelligibility is very difficult to quantify while maximizing efficiency and validity.

**Percentage of Consonants Correct**

Percentage of consonants correct (PCC) as a measure of speech sound accuracy was initially presented by Shriberg and Kwiatkowski (1982). PCC is defined as the number of consonants articulated correctly in a connected speech sample divided by the total number of total consonants in the glossed speech sample (i.e., adult targets). Shriberg and Kwiatkowski envisioned PCC as a metric that could reflect the severity of a child's speech sound disorder. They provided detailed instructions for counting errors when calculating PCC. Their protocol involved counting deletions, substitutions, partial voicing of initial sounds, distortions, and additions as errors. Additionally, they specified that allophones should be scored as correct, but that judgment is left up to the examiner. They did not provide information on coarticulation effects and how to score sound productions affected by coarticulation. They also did not specify an ideal length of sample, but they used one-minute speech samples to validate the measure.

In Shriberg and Kwiatkowski's (1982) initial implementation of PCC, they selected existing stimulus tapes which included continuous speech of children in conversation with an adult. In these tapes, the adult conversation partner glossed each utterance online, meaning that after every child utterance, the adult conversation partner immediately repeated what the adult thought the child said. Shriberg and Kwiatkowski were interested in three primary variables that are relevant here. One is the

metric they refer to as “severity of involvement,” referring to the severity of a child’s speech sound disorder. They assessed severity of involvement by providing SLPs with a one-minute audio recording of child speech, slightly edited to remove long pauses. The clinicians assigned a severity rating to each child on a scale from 3 to 7, allowing numerical answers ending in “.5” to produce a 9-point interval scale. They listened to each sample only one time before assigning a severity rating.

The second variable that Shriberg and Kwiatkowski (1982) were interested in was intelligibility, quantified by percent of intelligible words. They calculated this by providing listeners (a different population than those who judged severity of involvement) the same one-minute audio recordings and asking them to gloss the words that they could identify in each utterance. The tape was paused immediately after each child utterance (before the adult conversation partner’s gloss) to allow the listener time to transcribe. After the listener finished transcribing an utterance, the tape was unpaused.

The third variable that Shriberg and Kwiatkowski (1982) investigated was Percentage of Consonants Correct (PCC). Kwiatkowski, a researcher experienced with transcription of children’s speech, calculated PCC for all samples based on the same 1-minute samples based on the described rules. Shriberg and Kwiatkowski found that PCC was a strong predictor of severity of involvement, but age and suprasegmental characteristics of speech also played a role. Therefore, children with a lower PCC tended to have more severe speech sound disorders than children with a higher PCC. PCC was also moderately correlated with the intelligibility metric in their study ($p$ = .42).

**Story Sentence Imitation Task for Speech**

The assessment evaluated in the study reported here, which we call the *Story Sentence Imitation Task for Speech* (SSITS), is a speech accuracy assessment that was designed to replicate speech sound production in conversational connected speech. It builds on preliminary work by Taddeo et al. (2018) and a sentence imitation speech accuracy assessment developed by Johnson et al. (2004).

Johnson et al. (2004) developed a set of sentences telling the story illustrated in a commercially available wordless picture book. The examiner told the story sentence-by-sentence to the child and the child repeated each sentence. Following administration of the task, the examiner reviewed an audio recording and narrowly transcribed the child's sentence repetitions. The task stimuli included 36 sentences with 273 consonants. In an advantage over assessment of speech sound production based on purely conversational speech, all targets are known. In an advantage over single-word assessments, the child's speech production is evaluated in a connected-speech context, which is more ecologically valid and more directly related to the goals of speech sound intervention than single-word speech accuracy. Moreover, storytelling is a familiar task for preschool-age children. Thus, embedding speech sound assessment into storytelling should result in a more valid measure of a child's true performance in everyday communication interactions.

Although the Johnson et al. (2004) assessment had strong potential, several limitations are noted. First, the designated procedure requires narrow transcription of the entire sample post facto based on a recording, a quite time-consuming endeavor. Second, the design of this assessment did not consider the effects of coarticulation. The

target transcription is citation format for each word in each sentence. For instance, in the sentence ‘He got cold,’ the final consonant in ‘got’ is scored as a /t/. However, in conversational speech, many Standard American English speakers produce a /ʔ/ in place of /t/. According to the procedure, production of a glottal stop here would be marked as an error. Similarly, in the sentence ‘Mom says sit down,’ the final /z/ in ‘says’ and the initial /s/ in ‘sit’ are scored. In Standard American English speakers’ conversational speech, the production would likely be [sɛz̬sɪtda͡ʊn] or [sɛsːɪda͡ʊn], not [sɛz̬sɪtda͡ʊn] or [sɛz̬sɪda͡ʊn]. In other words, the final /z/ in ‘says’ would likely be devoiced and therefore indistinguishable from the initial /s/ in ‘sit.’ The described procedure does not account for such variations in production. By scoring children’s productions based on the citation form of words and not adequately considering the effects of coarticulation, the assessment may yield a PCC that underestimates the child’s true speech accuracy. Moreover, although the Johnson et al. assessment is associated with a wordless picture book, in the opinion of this author, the narrative of the story is not effectively conveyed via the sentences. Thus, it is not clear that a child would construe that she/he was retelling a story or simply repeating sentences. Our hypothesis is that when sentences to be imitated convey a narrative, the child’s accuracy approximates connected speech more so than unrelated sentences.

The SSITS, like the assessment developed by Johnson et al. (2004), entails the imitation of utterances forming a story illustrated in a wordless picture book. The wordless picture book associated with the SSITS, *A Day at the* Zoo, was developed by Taddeo et al. (2018). In contrast to Johnson et al. (2004), the SSITS does not require transcription; judgments are made only as to whether a sound is produced correctly or

incorrectly. Moreover, only some speech sounds in the SSITS utterances are scored for accuracy, rather than all sounds, which is hypothesized to allow for reliable live scoring rather than requiring the examiner to score from an audio recording. In the SSITS, like in the Johnson et al. assessment, the examiner models target utterances while displaying pages from the wordless picture book and the child is asked to imitate the utterances one at a time. The utterances together form a story that is illustrated in the wordless picture book. In the opinion of this author, the SSITS sentences form a more engaging story than the Johnson et al. sentences. To score the SSITS, the examiner scores online on a hard copy by marking target sounds produced in error (see form in Appendix A).

The SSITS includes 36 utterances. The utterances range from two to seven words (*M* = 4.86*, SD* = 1.22) and three to seven morphemes (*M* = 5.22*, SD* = 1.29). Rice et al. (2010) reported that the mean MLU in spontaneous speech for children 3;6-3;11 is 3.36, for children 4;0-4;5 it is 3.64, and for children 4;6-4;11 it is 3.95. The mean MLU in SSITS utterances is slightly higher than these age-indexed MLUs. Each utterance is scored for 0 to 6 target sounds (*M* = 3.72*, SD* = 1.45). In total, there are 134 target sounds distributed across the 36 utterances.

One goal when developing the SSITS was for it to mirror connected speech. If this goal is achieved, the SSITS may be able to serve as a proxy for speech intelligibility. Speech intelligibility is difficult to capture clinically, as detailed previously. Therefore, a goal of the SSITS design was for overall accuracy on the SSITS to provide a representation of intelligibility.

To approach this goal, the number of times that each consonant is represented as a target sound in the SSITS corresponds directly with the frequency of that consonant in child spoken English, based on frequencies reported by Mader (1954; see Table 1). For example, Mader reported that /n/ made up 13.14% of all consonants produced by his participants in conversational speech. In the SSITS, a similar proportion of target sounds are /n/, at 12.69%. Data on relative frequency of the occurrence of sounds in children 3;6-4;11, the target age range of the SSITS, was not available; frequency data based on conversational speech of first through third grade children was used instead (see Table 1; Mader, 1954). Errors on more frequently occurring sounds are expected to have a greater impact on a child's speech intelligibility. Therefore, proportional selection of target sounds may allow scores on the SSITS to correlate with speech intelligibility.

In the SSITS, each consonant sound in the American English phonemic inventory[1] is scored in multiple different contexts and words. Each sound is scored in no more than two instances of the same token (e.g., /ð/ is scored in two separate instances of the word 'the').

The final score derived from the SSITS is referred to as the SSITS PCC. This score is derived by dividing the number of target sounds produced correctly divided by the number of target sounds that the child attempted. Target sounds can be marked as (a) correct, (b) error, or (c) not attempted. Omissions and substitutions are considered errors, but distortions are not, based on the Percentage of Consonants Correct—

[1] /ʒ/ was excluded due to its relative rarity.

Revised methods (Shriberg et al., 1997). Target sounds that are not attempted are those in which the child did not repeat the word containing the target sound.

Speech sound accuracy in connected speech is a primary goal of speech sound intervention. Therefore, the SSITS may be useful in clinical practice regardless of whether it truly correlates with intelligibility or serves solely as a speech accuracy assessment.

Another goal was for the SSITS to be neutral to a child's language skills, to the extent possible, and to their dialect of spoken English. The target sounds in the SSITS were selected intentionally to maximize the likelihood that a child's errors are due to true speech sound errors rather than linguistic factors. Firstly, vocabulary in the SSITS was intentionally chosen such that all words are expected to be in the lexicon of children in the target age range. Furthermore, in a sentence imitation task, omission of inflectional morphemes could result from speech deficits, language deficits, or language differences. As such, in the SSITS, no target sounds are scored within inflectional morphemes (e.g., /ŋ/ is never scored within an *-ing* suffix). Furthermore, sentences were developed and target sounds were selected with the goal of not scoring sounds in words that might be omitted or altered based on the child's variety of English. For example, some speakers of African American English might use a zero copula (Craig & Washington, 2001), so no target sounds were selected within the copula. Additionally, the effects of coarticulation and casual speech were considered such that all target sounds are expected to be produced in the target manner despite those effects. For example, word-final plosives at the ends of utterances are not scored because in casual speech they may be unreleased or replaced with glottal stops.

In a clinical setting, the SSITS would likely have the most utility as a global measure of speech production. It does not provide significant information about the nature of children's errors, but rather quantifies the number of errors across a sample of speech that approximates the distribution sound in child spoken American English.

### Purpose

The purpose of the present study was to assess the psychometric properties of the SSITS, leading to eight research questions: (1) Do children 3;6-4;11 repeat a sufficient proportion of target sounds when participating in the SSITS? (2) Does the SSITS have adequate inter-rater reliability? (3) Does the SSITS have adequate intra-rater reliability? (4) Does the SSITS have adequate test-retest reliability? (5) Does the SSITS have adequate overall scoring stability? (6) Does the SSITS have convergent validity with an existing single-word measure of speech sound accuracy? (7) Does the SSITS have convergent validity with a measure of speech intelligibility? (8) Do the SSITS training materials lead SLPs to follow the intended protocols for deriving the SSITS PCC?

## Methods

Study procedures were approved by the Vanderbilt University Institutional Review Board. The study tasks were completed by the author or the faculty thesis advisor.

### Participants

#### *Children*

Study participants included 12 children between the ages of 3;10 and 5;9 at the time of Study Visit 1 (see Table 2 for demographic information). At this stage of

assessment development and validation, we chose four inclusionary criteria: (a) between the ages of 3;6 and 5;11, (b) monolingual English speaking, (c) receptive language skills within normal limits and (d) normal hearing. The age range of 3;6-5;11 was chosen because the SSITS was designed with this population in mind. Inclusion of only monolingual English speakers meant we did not have to consider the influence of learning more than one language. Receptive language skills within normal limits was selected as a criterion because the assessment of interest requires imitation of developmentally-appropriate sentences. If a participant was unable to imitate the sentences, we wanted to rule out receptive language deficits as a possible source. To evaluate receptive language, the Quick Interactive Language Screener (QUILS, Golinkoff et al., 2017) was administered. A child was excluded from data analysis if scores on the QUILS fell below the 25$^{th}$ percentile; Golinkoff et al. (2017) recommended no further screening if a child scores at or above the 25$^{th}$ percentile. One participant was excluded on this basis. The performance of the remaining 11 children fell between the percentiles of 44.10 and 99.70 (*M* = 78.15 percentile, *SD* = 17.00). Finally, we included only children with normal hearing because we predicted that hearing impairment may negatively affect children's ability to successfully repeat the target sentences, which was not a factor that we wanted to consider at this time. To eliminate other factors that could be confounds in addressing the research questions, exclusionary criteria included diagnosis of autism spectrum disorder, hearing loss, intellectual disability, neurological impairment (e.g., cerebral palsy), and/or uncorrected visual impairment. Children were screened prior to study consent with the Test of Articulation Performance – Screen

(TAP-S; Bryant & Bryant, 1983) to assure that we included participants with a range of speech production accuracy.

Child participants were recruited via emails to a listserv of Vanderbilt University Medical Center employees who have agreed to be contacted about research studies at Vanderbilt University, the Department of Hearing and Speech Sciences email listserv (i.e., staff, faculty), Vanderbilt Kennedy Center Study Finder, and flyer distribution to the Vanderbilt Acorn School. In addition, the study flyer was posted around the Vanderbilt University Medical Center. There was no randomization or assignment to condition.

***SLPs***

Study participants also included five SLPs (see Table 3 for demographic information). SLP participants were recruited via the Schuele lab listserv that includes SLPs who have attended professional development sponsored by the lab. There was no randomization or assignment to condition. SLPs who completed an eligibility survey were contacted based on the chronological order in which they completed the survey. The first five eligible SLPs who replied and who lived or worked within a 25-mile radius of Vanderbilt University Medical Center were enrolled in the study. This radius was imposed due to our desire to hand deliver study materials to participants. Inclusionary criteria were (a) at least five years of full-time work experience as an SLP, (b) at least five years of experience working with children with speech sound disorders in preschool and/or elementary school, (c) currently working with at least one child with a speech sound disorder who was between 3;6 and 5;11, (d) licensed by either the Tennessee Department of Education or Department of Health, (e) hold the Certificate of Clinical Competence (CCC) in Speech-Language Pathology from the American Speech-

Language-Hearing Association, (f) self-reported normal hearing acuity, and (g) achieve at least 92% scoring agreement with a master key for a simulated administration following completion of the SSITS training program. If the SLP failed to meet this criterion on the first simulation, one additional chance was afforded. Failure to meet the criterion would eliminate the SLP from further study participation. All five SLPs met the criterion.

## Measures

### *Children*

Child participants completed an assessment battery that included a screening measure, two eligibility measures, and multiple dependent measures.

**Screening Measure.** The TAP-S (Bryant & Bryant, 1983) requires children to verbally label pictures. If a child does not produce the target word, semantic prompts are provided; if the child still does not label the picture with the target word, the word is elicited in imitation. All word productions are transcribed phonetically. At completion, the examiner compares transcription of the child's speech to transcription of the adult target pronunciation. For the purposes of this study, the TAP-S was used descriptively; a numerical score was not generated.

**Eligibility Measures.** Two eligibility measures were administered. Children participated in a 30 decibel (dB) HL pure-tone hearing screening at 1000 Hz, 2000 Hz, and 4000 Hz. Though the American Speech-Language and Hearing Association recommends screening at 20 dB HL, an increased level was used for this screening due to background noise in the room where screenings were conducted (American Speech-

Language and Hearing Association, n.d.). To pass, the child was required to respond via conditioned play to two of four tone presentations at each dB level in each ear.

The QUILS (Golinkoff et al., 2017) is a receptive language screener with automated administration and scoring. The child is seated in front of a computer or tablet with an adult monitoring the child's attention to the task, guiding compliance as needed. For each item, picture stimuli (a test stimulus and answer options) are presented on the screen accompanied by a pre-recorded verbal prompt. The child responds to one of the answer option pictures by touching the picture on the screen. The screen advances to the next item immediately after the child responds. The QUILS program records child responses. Scoring is automated; after all test items have been administered, percentile ranks are generated and displayed on the screen and stored automatically in the program. The program provides sub-scores for Vocabulary, Syntax, and Language Learning Process as well as an Overall receptive language score.

**Dependent Measures.** Three dependent measures were administered. To recap, the SSITS is a sentence imitation task where the utterances form a narrative that is illustrated in an accompanying wordless picture book. Across the 36 utterances of the assessment, there are 134 target sounds that are scored by the examiner. Target sounds were selected intentionally with the goals of (1) proportional consonant frequency among target sounds similar to proportional consonant frequency in children's conversational speech and (2) low likelihood of omission or alteration due to coarticulation, a child's language skills, or a child's dialect of spoken English. Target sounds are recorded by the examiner as correct, error, or not attempted. A target sound is considered an error when there is a substitution or omission. A target sound is

considered “not attempted” when the child does not attempt repetition of the word containing that target sound. The outcome variable is PCC derived from the child’s attempted target sounds.[2]

The Preschool Speech Intelligibility Measure (PSIM; Morris et al., 1995), as described earlier, is a measure of children’s speech intelligibility at the single-word level. In this task, a child imitates 50 single words spoken by the examiner. Later, based on an audio recording, a naïve listener listens to the child’s production of each word and makes a closed-set selection of the word they believe the child to have said, from a field of 12 phonetically similar words. The outcome variable is the percentage of words identified accurately.

The Arizona-4 Word Articulation subtest requires children to label pictures to produce target words (Fudala & Stegall, 2017). Across all words, select speech sound productions are scored as correct or incorrect. Scored sounds are weighted, with some sounds having a higher point value than others. The Word Articulation score is calculated as a sum of the weights for correctly produced sounds. From the Word Articulation score, a standard score and a percentile rank can be derived; normative data is presented in the manual based on the child sex and age. For the purposes of this study, Word Articulation raw scores out of 100 was the outcome variable, rather than standard score or percentile rank, because SSITS PCC and PSIM percent identified are not adjusted for age.

[2] Following the work of Shriberg and colleagues, the mathematical calculation for the final PCC score is as follows: $\frac{number\ of\ correct\ target\ sounds}{134\ -\ number\ of\ target\ sounds\ not\ attempted}$

### *SLPs*

SLP participants participated in a self-guided PowerPoint training module and then scored the SSITS for all child participants based on video recordings. The PowerPoint training module consisted of directions for scoring the SSITS and opportunities for participants to practice scoring utterances from simulated administrations of the SSITS. The training module culminated in a criterion test, where SLPs scored full-length simulated administrations of the SSITS and compared their responses to a provided key.

## Procedures

### *Children*

An initial eligibility screening interview took place over the phone or in-person with the parent or guardian ("caregiver") following verbal assent. The caregiver was asked to answer questions regarding the child's age, language skills, and other exclusionary criteria as detailed above. If the child was deemed eligible based on this interview, the TAP-S was administered following verbal assent from the caregiver and the child. This administration took place in-person or virtually via videoconferencing. If the child's performance on the TAP-S aligned with eligibility, the caregiver was verbally provided with information about the study procedures and a study visit (Study Visit 1) was scheduled.

Upon arrival for Study Visit 1, the caregiver was asked to sign the written study participant consent form and complete the demographics form. Verbal assent for study participation was obtained from the child. Then, the study tasks were administered in the following order: QUILS, PSIM, SSITS, Arizona-4, hearing screening.

The Arizona-4 and the QUILS were administered according to the manualized instructions (Fudala & Stegall, 2017; Golinkoff et al., 2017). The Arizona-4 was transcribed online by the examiner and video recorded for later evaluation of intra-rater reliability. Because the QUILS is a fully computer-based and automated assessment, it was not audio or video recorded and reliability of response recording is not necessary.

The PSIM was administered according to the instructions detailed by Morris et al. (1995). The administration was audio recorded and later, audio clips that contained only the child's productions were isolated and used for scoring of the PSIM. Two naïve listeners who were unfamiliar with the child participants completed the scoring of the PSIM. One listener was a 17-year-old male high school student and the other was a 67-year-old male with a doctorate in developmental psychology; both have self-reported normal hearing acuity. The audio recording was played for each listener who chose a word from the field of 12 to match the listener's perception of the word spoken by the child. The listener could ask for the recording to be paused while a selection was made, but generally this was not necessary.

The hearing screening took place in a quiet room and was implemented via conditioned play. Each child was trained to drop a toy block into a tin when a sound was heard. Each child practiced this routine at least two times using sounds at suprathreshold levels. Then, the child wore supra-aural headphones and tones were presented at 1 kHz, 2 kHz, and 4 kHz at 30 dB HL. Each tone was presented no more than four times. The child was considered to pass the screening if the child responded to each stimulus upon at least two out of four presentations in both ears. This measure was not audio or video recorded.

The procedure for the SSITS was developed for this study (see Appendix A); the SSITS was audio and video recorded. The examiner began by saying, “I am going to show you some pictures and tell you a story. I want you to tell the story back to me. You’ll say just what I say. Let’s practice.” Then, the examiner showed the child the cover page of the storybook. The examiner said, “Say: The sky is blue” and pointed to the sky. If the child imitated the sentence appropriately, the examiner continued with two more practice sentences. If the child did not imitate the sentence appropriately, the examiner reminded the child to “Say just what I say. Say, ‘The sky is blue.’” If a child was unable to imitate the three practice sentences, we planned to discontinue administration. In the present study, all participants imitated the practice sentences successfully; therefore, no participants were disqualified based on this criterion. Verbal praise was provided. Next, the administrator read the utterances (one at a time) from the Sentence Repetition Form while pointing to relevant areas of the pictures in the storybook. The administrator indicated on the Sentence Repetition Form whether each target sound in an utterance was produced correctly or with an error before proceeding to the next utterance. If the child imitated an utterance incorrectly but all words with target sounds were produced, the examiner proceeded to the next utterance. If the child imitated an utterance incorrectly due to omission or substitution of a word with a target sound, the examiner scored as many target sounds as possible based on the child’s imitation. Then, the examiner said, “Remember, you say just what I say” and repeated the target utterance. If the child again omitted or replaced a word containing a target sound on the second attempt, the omitted or replaced word was marked with a large ‘X’ and the examiner proceeded to the next sentence. On the child’s second production of the utterance, the

examiner scored only target sounds in words that were not attempted in the child's initial production of the utterance. All target sounds in a word that was not repeated by the child (and therefore marked with an 'X') were considered "not attempted."

The examiner circled the target sound that was produced in error (i.e., substitution or omission). Distortions and insertions were not counted as errors, and thus were not recorded. The categorization of only substitutions and omissions as errors is in line with the Percentage of Consonants Correct-Revised (PCC-R; Shriberg et al., 1997). After the measure was administered, the examiner tallied the number of errors in each utterance on the Sentence Repetition Form and wrote that number in the blank on the right-hand side of each utterance (see Appendix A). Then, on each page, the examiner tallied the number of not attempted sounds on that page, summed the number of errors on that page, and wrote the resultant numbers in the appropriate blanks at the bottom of the page. Finally, the examiner transferred the number of errors and number of not attempted sounds from the bottom of each page of the Sentence Repetition Form to the Scoring Summary Form (see Appendix B). On the Scoring Summary Form (Appendix B), the examiner summed the number of error sounds across the entire assessment to obtain the number of Total Errors on Target Sounds. Then, the examiner summed the number of "not attempted" sounds across the entire assessment to obtain the Total Target Sounds Not Attempted. The examiner then subtracted the Total Target Sounds Not Attempted from 134 to calculate the Total Target Sounds Attempted. The Total Errors on Target Sounds was subtracted from the Total Target Sounds Attempted to obtain the Total Target Sounds Correct. Finally, the examiner divided the Total Target

Sounds Correct by the Total Target Sounds Attempted and multiplied the quotient by 100 to obtain the SSITS PCC final score.

At the conclusion of Study Visit 1, the caregiver was invited to have the child participate in Study Visit 2, which could take place at the lab or at another location convenient to the family (e.g., home, public library) and could take place in person or via videoconferencing. The caregiver was informed that we would like all participants to participate in Study Visit 2, if possible. If the parent indicated interest, Study Visit 2 was scheduled (weeks between study visits, Range = 1 – 9 weeks, $M$ = 3.73, $SD$ = 2.42). At Study Visit 2, the child completed the SSITS only, according to the same procedure as in Study Visit 1. Child participants were compensated $40 for participating in Study Visit 1 and $20 for participating in Study Visit 2.

***SLPs***

SLP participants completed a consent form and a demographics form via REDCap, a secure web platform for building and managing online databases and surveys. For all study activities, SLP participants used AudioTechnica ATH-M20x headphones. Each SLP participant used their own personal laptop or desktop computer running either Mac OS 10.5.8 or newer or Windows 7 or later with a screen size of at least 11 inches, measured diagonally.

The SLP participants were provided with research materials – AudioTechnica ATH-M20x headphones and four folders. One folder, labeled “Training,” contained four blank copies of the Sentence Repetition Form and four blank copies of the Procedures & Scoring Summary Form. The second folder, labeled “Keys,” contained two copies of the Sentence Repetition Form and two copies of the Procedures & Scoring Summary

Form, with scoring already completed by the author.[3] Another folder, labeled “Task 1,” contained 20 blank copies of the Sentence Repetition Form and 20 blank copies of the Procedures & Scoring Summary Form. The fourth folder, labeled “Task 2,” contained five blank copies of the Sentence Repetition Form and 5 blank copies of the Procedures & Scoring Summary Form.

The SLP participants were provided with a training protocol in Microsoft PowerPoint slideshow format via email. Each SLP participant independently completed this self-guided training protocol, which was developed to familiarize the participants with the SSITS and teach the scoring procedures.

The slideshow began with a “Familiarize” section. This section contained informational slides orienting the participant to the Sentence Repetition Form. This section also contained a video clip of a simulated SSITS administration in which an SLP graduate student played the role of a child. In this clip, the SLP graduate student did not make speech errors. Next, the slideshow continued into the “Learn” section. In this section, slides were displayed with information for the SLP participant to read regarding scoring basics for the SSITS, written answers to questions that the authors anticipated regarding SSITS scoring, and directions (including examples) for calculating the SSITS PCC. In the “Practice and Check” section, participants watched embedded video clips, practiced scoring those clips on provided copies of the Sentence Repetition Form and the Procedures & Scoring Summary Form from the provided “Training” folder, and compared their responses to answers in the slides. In these video clips, an SLP graduate student played the role of a child and intentionally made speech errors. The

[3] After distribution of physical materials, the author identified mistakes on the distributed keys. The author then distributed a corrected version of the key via email in PDF format.

first video clips consisted of one utterance and the video clips progressively lengthened throughout the “Practice and Check” section, culminating in a video clip of a full SSITS administration. During the “Practice and Check” section, participants were permitted to review earlier slides and play video clips multiple times.

Finally, at the end of the training, participants completed the "Criterion Test.” Participants scored a video clip of a full SSITS administration, again with an SLP graduate student playing the role of a child. For the criterion test, SLP participants were not permitted to rewind or replay any portion of the video, so as to approximate an administration with a live scoring. However, they were permitted to pause the video between utterances, because in live scoring a clinician would pause to complete scoring an utterance before modeling the next utterance for the child to repeat. Upon completion of the criterion test, participants compared their responses to a provided key. For each target sound, the participant marked whether her judgment of the sound as correct, error, or not attempted was the same as the judgment in the provided key. If 92% or more of the target sounds were judged identically between the participant’s scoring and the key, the participant passed the criterion test. If less than 92% of the target sounds were judged identically between the participant’s scoring and the key, the participant was directed to review the training slides, score a new video of a full SSITS administration, and again compare their responses to a provided key. If the participant again achieved less than 92% agreement with the key, the participant would be discontinued from the study. All participants achieved greater than 92% agreement with the provided key on the first criterion test video.

The research protocol required each SLP participant to score multiple SSITS videos. First, each SLP scored 11 Study Visit 1 videos and 9 Study Visit 2 videos (20 videos in total). The order of the videos was randomized for each SLP participant. To score these videos, SLP participants used Sentence Repetition Forms and Procedures & Scoring Summary Forms from the “Task 1” folder. Videos were accessed via password-protected Vimeo links, which were distributed to each SLP participant individually via email. SLP participants were allowed to pause the video between utterances but were not allowed to rewind or replay any portions of the video. Second, each SLP participant scored for a second time five randomly assigned Study Visit 1 videos using Sentence Repetition Forms and Procedures & Scoring Summary Forms from the “Task 2” folder. Again, videos were accessed via password-protected Vimeo links, which were distributed to each participant individually in random order via email. After completing all study procedures, the SLP returned all forms but kept the headphones. Each SLP was compensated $100 for completing the study procedures.

**Data Preparation**

After the SLP participants returned the completed Sentence Repetition Forms and Procedures & Scoring Summary Forms, the author prepared the data. The author reviewed each score form, identified instances in which the SLP participant made mathematical errors or did not adhere to the directions (“SLP mistakes”), and made corrections where appropriate. SLP mistakes fell into 9 categories which will be referred to as Type A-I errors; see Table 4. To summarize, Type A-E errors are instances in which SLPs did not follow the intended SSITS protocols. Type F-I errors are errors that the SLPs made in counting/tallying or in mathematical calculations.

The author corrected Type A, B, C, and D errors because the marking of each individual target sound as an error, not an error, or not attempted remained clear. For each Type A error, the author adjusted the total number of not attempted target sounds such that each target sound in the word marked with an 'X' was included in the sum of not attempted target sounds. For each Type B error, the author adjusted the sums such that not attempted target sounds were only counted in the sum of not attempted target sounds and were not counted in the sum of errors on target sounds. For each Type C error, the author adjusted the total number of not attempted target sounds such that only target sounds and not any other sounds in the SSITS were included in the sum. For each Type D error, the author adjusted the total number of errors on target sounds such that only target sounds and not any other sounds in the SSITS were included in the sum. Type E errors were not corrected because the author was unsure of the SLP participant's intentions. Lastly, the author corrected Type F, G, H, and I errors by performing the appropriate calculations using a calculator.

## Results

A summary of children's performance on all dependent measures can be found in Table 5. On the Arizona-4 Word Articulation subtest, the mean score was 77.95 (*SD* = 16.08, range = 56.50 to 100.00). On the PSIM, the mean number of words correctly identified by the naïve listeners was 51% (*SD* = 16%, range = 26% to 74%). For Study Visit 1, the mean SSITS PCC was 82.00% (*SD* = 14.29%, range = 60.67% to 100.00%; see Table 6). For Study Visit 2, the mean SSITS PCC was 83.75% (*SD* = 12.97%, range = 65.97% to 100.00%; see Table 7).

**Research Question 1: Do children 3;6-4;11 repeat a sufficient proportion of target sounds when participating in the SSITS?**

For the present study, a child who "completed" the SSITS must have attempted at least 95% of the target consonants (128 out of 134). A priori, we set the criterion for feasibility as at least 90% of the participants completing the SSITS. By the author's judgment following video review of all Study Visit 1 SSITS administrations, 100% of children attempted at least 128 of the 134 consonants ($M$ = 133.3, $SD$ = 1.8, range = 128 to 134); therefore, the criterion was met.

**Research Question 2: Does the SSITS have adequate inter-rater reliability?**

To establish inter-rater reliability, SLP participants scored the Study Visit 1 SSITS of all child participants (Table 6). The SSITS PCCs that each rater obtained were compared and reliability was computed through intra-class correlation coefficients (ICCs), which reflect the degree of correlation and agreement between measures (Koo & Li, 2016). Koo and Li (2016) suggested an interpretation of ICCs <.50 as "poor agreement", between .50 and .75 as "moderate agreement", between .75 and .90 as "good agreement," and greater than .90 as "excellent agreement." The two-way random effects single measures absolute ICC was .883—that is, good agreement—with a 95% confidence interval of .708 to .964, $F(10,40) = 67.44$, $p < .001$.

We can also gain more insight into this inter-rater reliability of the SSITS by examining agreement between clinicians on a point-by-point basis. In other words, how often did SLP participants agree as to whether a given child's production of a target sound was correct, error, or not attempted? Overall, among all one-on-onepairs of SLP participants and across all initial administrations of the SSITS, each one-to-one pair of

SLPs agreed with each other on a mean of 88.9% of target sounds. Data on individual pairwise comparisons can be found in Table 8. All five of the SLPs collectively agreed as to whether a given individual target sound was correct, error, or not attempted on 75.71% of sounds.

Yet another element of inter-rater reliability could consider whether, despite small differences in scoring, the rank ordering of children's performance remained consistent across SLPs. See Table 9 for full rankings for each SLP participant.

**Research Question 3: Does the SSITS have adequate intra-rater reliability?**

To establish intra-rater reliability, after scoring the 11 Study Visit 1 and 9 Study Visit 2 SSITS videos (Task 1), SLP participants scored 5 randomly assigned Study Visit 1 videos a second time (Task 2). A single measures, two-way random effects, absolute ICC was calculated, using the same interpretations described previously. This value was .987—that is, excellent agreement—with a 95% confidence interval of .968 to .994, $F(24,24) = 161.441$, $p < .001$.

Intra-rater agreement can also be examined on a point-by-point basis by quantifying frequencies of disagreement between Task 1 and Task 2 within the same video and the same rater. Each SLP scored five of the videos twice and there were five SLP participants; thus, for each of the 134 target sounds, there are 25 data points indicating whether the SLP scored an individual child's production of an individual sound in the same way or a different way (correct, error, not attempted) on both iterations of rating the same video (Task 1 and Task 2). Across the 134 target sounds, raters made the same judgements during Task 1 and Task 2 on average 24.00 out of a possible 25 times (*SD* = 1.49; range = 19 to 25).

**Research Question 4: Does the SSITS have adequate test-retest reliability?**

To evaluate the test-retest reliability of the SSITS, a single measures, two-way random effects, absolute ICC was calculated between SSITS PCCs at Study Visit 1 and Study Visit 2 averaged across raters (Task 1 judgments; see Table 10). This value was .935—that is, excellent reliability—with a 95% confidence interval of .750 to .985, $F(8,8) = 33.015$, $p < .001$. The mean change in SSITS PCC between Study Visit 1 and Study Visit 2 was 2.26% (*SD* = 4.35%, range = -4.82% to 7.71%).

When assessing test-retest reliability, we can also examine whether, despite slight changes in performance and/or scoring between two time points, children maintain the same rank order between those time points. See Table 11 for a rank order of children at Study Visit 1 and Study Visit 2 based on average SSITS PCC across raters.

A Pearson correlation coefficient was calculated between change in SSITS PCC between Study Visit 1 and Study Visit 2 and the number of days that passed between Study Visit 1 and Study Visit 2. Guidelines for interpretation of Pearson correlation coefficients set forth by Cohen (1988) were followed; that is, a correlation between .10 and .30 was considered weak, between .30 and .50 was considered moderate, and above .50 was considered strong. The calculation revealed a Pearson correlation coefficient of -0.18—that is, a weak negative correlation—between change in SSITS PCC score between Study Visit 1 and Study Visit 2 and the number of days that passed between Study Visit 1 and Study Visit 2; however, this correlation was not statistically significant at the $p = .05$ level.

**Research Question 5: Does the SSITS have adequate scoring stability overall?**

The next research question pertained to the overall stability of the SSITS. A *G* study was conducted to evaluate the overall contribution of each contributor to variance in SSITS PCC. Children who completed Study Visit 1 and Study Visit 2 were included in this analysis. This analysis revealed that individual differences between child participants accounted for 82.67% of variance in SSITS PCCs. Day (Study Visit 1 versus Study Visit 2) accounted for only 0.66% of the variance. SLP participant accounted for 5.12% of the variance. The interaction between child and day accounted for 3.99% of variance, the interaction between subject and rater accounted for 5.17% of the variance, and the interaction between day and rater accounted for 0.14% of the variance. The interaction between subject, day, and rater accounted for 2.25% of the variance. The overall absolute G coefficient was .95.

**Research Question 6: Does the SSITS have adequate convergent validity with an existing measure of speech sound accuracy?**

To evaluate the convergent validity of the SSITS with an existing single-word speech accuracy measure, SSITS PCC was correlated the Arizona-4 Word Articulation Score (Fudala & Stegall, 2017). Normal distribution of each variable was verified; for SSITS PCC, skewness was -0.15 and kurtosis -1.704 and for Arizona-4 scores, skewness was .117 and kurtosis was -1.720. Guidelines for interpretation of Pearson correlation coefficients set forth by Cohen (1988) were followed; that is, a correlation between .10 and .30 was considered weak, between .30 and .50 was considered moderate, and above .50 was considered strong. Study Visit 1 SSITS PCC (averaged

across raters) and Arizona-4 Word Articulation Score were strongly positively correlated, $r(11) = .937$, $p < .001$.

**Research Question 7: Does the SSITS have adequate convergent validity with a measure of intelligibility?**

To evaluate the convergent validity of the SSITS with an existing intelligibility measure, SSITS PCC was correlated with PSIM percent words identified (Morris et al., 1995). Normal distribution of each variable was verified; for SSITS PCC, skewness was -0.15 and kurtosis -1.704 and for PSIM percent words identified, skewness was .341 and kurtosis was -1.089. Study Visit 1 SSITS PCC (averaged across raters) and Arizona-4 scores were found to be strongly correlated, $r(11) = .914$, $p < .001$.

**Research Question 8: Do the SSITS training materials lead SLPs to follow the intended protocols for deriving the SSITS PCC?**

To evaluate whether the training materials were sufficient to lead SLP participants to follow the intended protocols for deriving the SSITS PCC, all SLP errors were documented. Table 12 details the frequency of each error type in the Task 1 scoring forms provided by each SLP (see Table 4 for descriptions of error types).

To evaluate the impact of errors on SSITS PCCs, SSITS PCCs calculated by SLPs (“uncorrected SSITS PCCs”) were compared with SSITS PCCs calculated by the author after making necessary corrections (“corrected SSITS PCCs”). Corrected SSITS PCCs differed from uncorrected SSITS PCCs by a mean of -0.02 percentage points (*SD* = 1.22, range = -7.46 to 7.62). Comparisons of corrected and uncorrected SSITS PCCs for each SLP participant can be found in Table 13.

## Discussion

The present study was undertaken to evaluate a sentence imitation task, the SSITS, as a measure of speech sound accuracy in children aged 3;6-4;11. The major findings were that the SSITS has high feasibility and reliability, as well as high convergent validity with scores on an existing single-word speech sound assessment and an intelligibility assessment.

### Research Question 1: Do children 3;6-4;11 repeat a sufficient proportion of target sounds when participating in the SSITS?

All child participants attempted at least 95% of the target sounds in Study Visit 1 SSITS administration. An assessment tool is most useful clinically if an SLP selecting it as part of an assessment battery can trust that the child will complete the task. The feasibility of completion of the SSITS is also important because the target sounds are carefully selected to reflect the frequency of sounds in children's speech. When too many words are not attempted by the child, those alignments of frequencies can become mismatched. Therefore, a task such as this has the most utility if children consistently attempt the majority of target sounds. In this study, children did consistently attempt the majority of target sounds; all children in the sample met our criterion for feasibility. Furthermore, multiple child participants were observed to initiate conversation and ask questions about the story and the book associated with the SSITS. These interactions suggest that children see the SSITS as reading a story with an adult, rather than taking a test.

Though feasibility was generally strong, some words and utterances in the assessment could be considered less feasible than others. In other words, some target

sounds were more frequently not attempted than others. Out of the 134 target sounds, 110 were never marked as not attempted during Study Visit 1; three were marked as not attempted for two or more children.

The target sound most frequently marked as not attempted was the sentence-initial /ʃ/ in “She’s sliding through the bush,” which was marked by at least one SLP as not attempted for three children during Study Visit 1. These three children repeated the sentence as “Her sliding through the bush.” This word substitution emphasizes the importance of creating utterances that are neutral to language skill and to dialect. In hindsight, we see that this sentence subject should not have been included in the assessment as the substitution of “her” for “she” was foreseeable. This word substitution could occur due to a dialectical difference (Craig & Washington, 2002) or delayed language development (Loeb & Leonard, 1991).

Another two children were scored by at least one SLP as not attempting the word ‘please’ in ‘Can you please help me?’ during Study Visit 1. Nonrepetition of this word could be related to the increased linguistic demand of interrogative sentences. Additionally, this utterance is still grammatical if ‘please’ is omitted, which could have further contributed to nonrepetition of ‘please.’ One child responded to the examiner’s presentation of this utterance and another interrogative utterance in the assessment (‘Where is she going?’) by answering the question rather than imitating the sentence. Though in this study only that one child experienced noticeable difficulty repeating these utterances, interrogative stimulus sentences pose an unnecessary potential challenge. Therefore, in future iterations of the SSITS, feasibility could be strengthened by inclusion of only declarative utterances.

Additionally, two children were scored by at least one of the five SLP participants as not attempting the word ‘say’ in ‘Mom and dad say time for lunch’ during Study Visit 1. It is unclear why this word in particular was not attempted multiple times, but perhaps the syntactic complexity of the embedded clausal object of the verb ‘say’ posed too high of a linguistic load for some children.

**Research Question 2: Does the SSITS have adequate inter-rater reliability?**

An intra-class correlation coefficient of .883 suggests that the SSITS has good inter-rater reliability overall. Inter-rater reliability is a crucial characteristic of any assessment. When an SLP working clinically selects an assessment to administer to a child, they should be able to trust that another SLP administering the same assessment to the same child would arrive at a similar final score. A child’s overall score should not depend on the administrator of the test. For comparison, Fudala and Stegall (2017) evaluated inter-rater reliability between two raters on the Arizona-4. They reported an ICC of .90 for the Word Articulation subtest and .85 for the Sentence Articulation subtest. The ICC calculated for the SSITS is in line with those values.

In the development of another sentence imitation task with different sentences and methods, Johnson et al. (2004) found 93% point-by-point agreement between two raters for imitative sentences. Our point-by-point agreement between each pair of clinicians, at 88.9%, was slightly lower than that reported by Johnson et al., but a difference that could be explained by our differences in protocol. Though Johnson et al. do not specify, we assume that their raters were allowed to pause and replay the tapes; ours were allowed to pause but not to replay. The Johnson et al. raters scored every

consonant in the assessment, which is a contrast from ours where SLPs only scored select target sounds.

All SLPs rated the same child as the child with the highest score on the SSITS. Four of the SLPs rated the same children as #2 and #3, respectively; the fifth SLP had these two children's positions reversed. Ranks #4, #5, and #6 were consistent across all raters. Each of the SLPs included the same group of children in ranks #7 to 11, but more variability between participants in the specific placement of these children was noted than among the children in higher ranks. This suggests that there may be more agreement among SLPs as to the performance of children with fewer speech errors than children with more speech errors.

**Research Question 3: Does the SSITS have adequate intra-rater reliability?**

An ICC calculated between SLPs' first and second times scoring the same video (during Task 1 and Task 2) was .987, suggesting excellent intra-rater reliability. Intra-rater reliability is a crucial characteristic of any assessment because SLPs need to be consistent with themselves in how they are evaluating a child's skill. If an assessment has low intra-rater reliability, that could indicate that the directions for scoring are not clear or that SLPs are inconsistent in their judgments when scoring the assessment.

Intra-rater reliability can also be examined on a point-by-point basis. Overall, the SLPs scored 94.7% of the target sounds the same on their first time and second time scoring the same child's SSITS (during Task 1 and Task 2). This is marginally higher than the point-by-point intrajudge agreement found by Johnson et al. (2004) which was 93% following a different scoring protocol.

On a point-by-point basis, some target sounds were associated with higher frequency of disagreement between Task 1 and Task 2 when a single SLP scored the same child's SSITS. Three target sounds were identified as especially problematic, as each was scored the same in only 19/25 instances, the lowest frequency of point-by-point intra-rater agreement among the 134 target sounds in the SSITS. These three target sounds were: (a) the first /r/ in 'Everybody, look there!' (b) the sentence-initial /ð/ in 'There's the snake!' and (c) the /ð/ in 'I really like this new hat.' In contrast, 53 target sounds were always scored the same (i.e., 25/25). Word position appeared to have no influence on intra-rater agreement. However, scoring productions of /ð/ appeared challenging for SLPs. For the seven instances of /ð/, five were agreed upon 22/25 times or fewer. One explanation for this lack of agreement may be that stopping of /ð/ may not significantly impact intelligibility and therefore may be more difficult for SLPs to identify.

**Research Question 4: Does the SSITS have adequate test-retest reliability?**

An ICC calculated between children's scores at Study Visit 1 and Study Visit 2 was .935, indicating excellent test-retest reliability. Test-retest reliability is important for an assessment to have, particularly one like the SSITS which could be used as a progress-monitoring tool. If a child is administered a given assessment at the start of a school year and again at the end of the school year, the SLP must be confident that an improvement in the child's score reflects improvement in the child's speech production skills, rather than instability in the assessment.

Of the nine children who participated in both study visits, when PCC scores were averaged across all raters' Task 1 judgments, two scored lower on Study Visit 2, five scored higher on Study Visit 2, and two scored within one percentage point of their first-

visit score. It would be preferable for there to be more of an even split between those who scored higher and lower on the Study Visit 2; however, the overall average improvement in scores between the first and study visit was only 2%. There were a few methodological challenges that could have contributed to improvement in scores other than those inherent to the assessment. Eight of the nine second-round visits were conducted via video chat, whereas all first-round visits were conducted in person. It is possible that deterioration in the sound quality and inconsistent ability to visualize children's mouths during the video chat recordings could have altered SLPs perceptions of sounds. An empirical study could compare performance based on live and video recordings to assess whether SLPs' accuracy in judgment of children's speech sound accuracy differs between mode of test administration.

The time that elapsed between the two study visits was inconsistent between participants, ranging from 10 to 61 days with a mean of 26 days. There was concern that participants whose Study Visit 2 was close chronologically to Study Visit 1 could have experienced practice effects contributing to improvement in scores. Alternatively, those whose Study Visit 2 was distant chronologically from Study Visit 1 could have experienced true improvement in speech sound production. The Pearson correlation coefficient of -0.18 suggests a weak negative correlation between time elapsed and change in SSITS PCC, supporting the suggestion that fsome practice effects may have contributed to the performance of children whose Study Visit 2 came close chronologically to Study Visit 1; however, these practice effects, if present, were minimal.

Another aspect of test-retest reliability to consider is whether, despite small changes in children's performance between Study Visit 1 and Study Visit 2, the children maintained the same rank order. As demonstrated in Table 11, the six highest-scoring children maintained the same rank ordering between Study Visit 1 and Study Visit 2. The children ranked #7, #8, and #9 switched positions between the two study visits. These shifts in position suggest that children with lower speech sound accuracy may demonstrate more variability in performance on the SSITS from day-to-day than children with higher speech sound accuracy.

**Research Question 5: Does the SSITS have adequate overall scoring stability?**

An analysis of variance was conducted to evaluate the role and interaction effects of different components as contributors to children's scores. The factors considered were the individual child, the date of evaluation (Study Visit 1 or Study Visit 2), and the SLP rater. This study revealed that 82.7% of the variability in scores was attributable to individual differences between children. Moreover, the absolute G coefficient was 0.95, indicating high generalizability of these results. This stability, like measures of reliability, is crucial for an SLP using this assessment clinically. Stability allows clinicians to be confident that differences between scores are true differences between children or changes in speech sound production, rather than variability inherent to the measure.

**Research Question 6: Does the SSITS have convergent validity with an existing measure of speech sound accuracy?**

Scores on the SSITS from Study Visit 1 were correlated with scores on the Arizona-4 and a Pearson correlation coefficient of .937 was obtained, indicating a strong

correlation between these two measures. Thus, the SSITS appears to have convergent validity with the Arizona-4. Convergent validity is important for any new assessment because it evaluates how closely related two assessments that purport to measure the same construct—in this case, speech accuracy—are. The Arizona-4 is a well-established assessment of speech sound accuracy, so a strong correlation between the SSITS and the Arizona-4 suggests that the SSITS also has validity as an assessment of speech sound accuracy.

Another important element when considering the alignment between a new speech accuracy assessment and an existing one is whether the two assessments place children in the same order in terms of performance. When considering SSITS scores at Time 1 averaged across raters and Arizona Word Articulation scores, these two assessments placed the 7 highest-scoring children in the same order, whereas the 4 lowest-scoring children were scored in different orders between the two assessments. These rankings suggests that there may be more discrepancies between the two assessments when assessing children with lower speech sound accuracy.

Importantly, the SSITS and the Arizona-4 do not measure the exact same construct. The Arizona-4 Word Articulation Test, the subtest administered for the present study, evaluates speech sound production at the single word level, whereas the SSITS evaluates speech sound production in connected speech. Therefore, it is expected for children to demonstrate differences in performance between the two assessments.

**Research Question 7: Does the SSITS have convergent validity with a measure of intelligibility?**

Scores on the SSITS from Study Visit 1 were correlated with scores on the PSIM and a Pearson correlation coefficient of .914 was obtained, indicating a strong correlation between these two factors. It is known in the literature that speech sound accuracy and speech intelligibility are not the same; there are many factors other than speech sound accuracy that contribute to intelligibility (Kent et al. 1994). Largely, speech sound accuracy has been measured in the literature using single-word speech accuracy measures, which may not replicate the errors that children make in connected or conversational speech. Therefore, this question was posed here to investigate whether by measuring speech accuracy within a connected speech context, intelligibility can be more closely approximated—in other words, whether speech accuracy at the sentence level can serve as a proxy for intelligibility.

The correlation found here between SSITS scores and PSIM scores suggests, as expected, that there is some relation between speech sound accuracy and intelligibility. However, due to the nature of the PSIM, the relation between SSITS scores and overall intelligibility remains unclear. The PSIM is a single-word intelligibility measure, which raises concerns about its ability to approximate conversational intelligibility—not to mention the wide variability amongst professionals as to the definition of intelligibility. Therefore, it may be wise to assess intelligibility at the conversation level prior to making any judgments about the relation between intelligibility and speech sound accuracy on a task such as the SSITS.

**Research Question 8: Do the SSITS training materials lead SLPs to follow the intended protocols for deriving the SSITS PCC?**

For a measure's psychometric properties to be maintained and maximized, persons who score and administer that assessment must be able to do so according to a specified set of directions. With that in mind, the adherence of SLP participants to the SSITS scoring protocols was explored. After an SLP perceived and judged a target sound for accuracy, did that SLP (a) record it appropriately, (b) follow the directions for counting errors on target sounds and target sounds not attempted, and (c) perform calculations correctly? Some variation was observed in the degree of adherence to scoring protocols (see Table 12).

The most common types of errors observed were mathematical errors (Type F, G, H, I; see Table 4). The SSITS scoring forms (Sentence Repetition Form and Procedures & Scoring Summary Form; Appendix A; Appendix B) entail several steps of calculation. The scorer must sum the number of errors in each utterance, the number of errors on each page, the number of not attempted target sounds on each page, transfer those numbers to the Scoring Summary Form, and perform more calculations on that page. Mathematical errors were by far the most commonly observed at the first step—SLP participants frequently circled a number of target sounds in an utterance (to mark them as errors) but wrote a different number in the blank intended for the total number of errors in that utterance (Type F error). One explanation for the high frequency of this type of error could be that SLPs may have tallied the number of errors in each sentence at the same time as they scored the sentence. It was not specified in the directions, but the authors assumed that the SLPs would tally the number of circled target sounds

(errors) in each sentence following the full administration of the assessment, or that they would pause the video in between utterances to complete this tallying. In future development of the SSITS it could be beneficial to specify that the SLP should tally errors after the completion of the assessment. Mathematical errors of other types (G, H, I) were rare.

Another group of errors involved adherence to directions for what sounds to count in what categories (error, not attempted, correct). Type A errors were the most common. According to the directions, if a word with multiple target sounds was omitted by the child, each target sound in that word should be counted separately as "not attempted." For example, imagine that a child said 'What will we see?' instead of 'What will we see today?' 'Today' includes two target sounds, /t/ and /d/. In this case, the SLP should tally two sounds as not attempted. There were 15 instances across four SLP participants in which an SLP counted scenarios like this as only one sound not attempted (Type A error).

Another error type within this group was counting not attempted sounds as simultaneously not attempted and an error (Type B error). Each target sound is counted in one category—correct, error, or not attempted. However, SLP 3[4] counted a sound as both an error and "not attempted" in 13 instances. No other participants made this error.

Some errors were also made in identifying what sounds should be counted toward the "not attempted" sum and the "error" sum. A single SLP participant, SLP 3, in 12 instances counted a non-target sound toward the "not attempted" sum (Type C error). In other words, a child omitted a word that did not have any target sounds in it

[4] To distinguish between SLP participants in this discussion, SLP participants will be referred to as SLP 1, SLP 2, SLP 3, SLP 4, and SLP 5.

and yet SLP 3 factored that omission into the “not attempted” sum. A different SLP participant, SLP 4, once counted a non-target sound (presumably produced in error) toward the sum of errors for that sentence (type D error).

One SLP participant, SLP 4, marked one target sound in a word as “not attempted,” but other target sounds in the same word as correct or errors (a Type E error) in five instances. The marking of “not attempted” was intended by the authors to be used only when the child did not successfully repeat a word that included one or more target sound(s). We hypothesize that perhaps SLP 4 used the “not attempted” marking to indicate omissions at the sound level, which according to the manualized instructions are considered errors and should be marked as errors, not as “not attempted.”

When first reviewing SLP 4’s scoring forms, the author questioned whether this SLP participant may have been trying to account for syllable deletion. In review of the videos, however, the author does not believe syllable deletion to have been occurring at the times where SLP 4 made this Type E error. It is possible that the SLP may have become confused about how to deal with omissions of target sounds, which are errors in the SSITS, and instead marked omissions of target sounds as “not attempted.”

Although syllable deletion was not an issue in this sample of participants, in some cases, a child might omit one syllable of a multisyllabic word. It is debatable whether target sounds in that omitted syllable should be considered errors, because syllable deletions may occur due to speech or language processes. In future development of this assessment, it would be beneficial to create specific instructions about how to deal with this situation, such as the opportunity to mark sounds within an

omitted syllable as an error. The current instructions state that the “not attempted” mark should be used when an entire word is not repeated by the child, but do not specify how to proceed with scoring when a syllable is deleted.

It is clear from this information that the directions for scoring “not attempted” sounds and the training that the SLPs received in this area were insufficient. In future research on the SSITS, the training that SLPs receive should be more explicit about when a sound is considered “not attempted.”

Overall, the number of errors was highly variable between participants—SLP 1 only evidenced one scoring error, whereas SLP 3 evidenced 38. Because different types of errors impact the calculation of SSITS PCC in different ways, it is difficult to set a criterion for an acceptable frequency of error amongst SLP raters. Even 38 errors, however, averages to less than two errors per child administration, which could be considered an acceptable level—but for this SLP, 22 of those errors were evidenced on a single video. On that video, all her errors were Type A, B, and C errors, suggesting a significant lack of understanding of how to deal with not attempted target sounds.

In sum, it appears that the training materials can lead SLPs to follow the intended protocols for deriving the SSITS PCC, as evidenced by the fact that SLP 1 only made one error. However, the high number of errors for some SLPs suggests that the materials do not consistently lead SLPs to follow the intended protocols for deriving the SSITS PCC. Clarity regarding when the SLPs should tally error sounds is needed, as well as more thorough instruction on dealing with target sounds that are not attempted.

## Conclusion

This study adds to the evidence base surrounding the assessment of children's speech sound production by demonstrating that sentence imitation forming a narrative illustrated in a wordless picture book can have strong psychometric properties. The SSITS was found to be feasible for children and SLPs, have adequate reliability and stability, and have convergent validity with measures of both speech sound accuracy and of intelligibility. However, the training materials for SLPs to learn how to score the SSITS need further development.

### Clinical Implications

This study suggests that sentence imitation can be an effective method for measuring children's speech sound accuracy. It may be useful in progress monitoring, particularly when a child is performing well on drill-based single-word articulation tasks and the SLP wants to know how well this speech production ability has generalized to connected speech.

### Future Directions for Research

As discussed throughout the discussion section, there are several changes that should be made to the SSITS as development of this assessment moves into the next stage. A non-inclusive list of some suggested revisions to the SSITS utterances can be found in Table 14. The sentence stimuli should be modified to ensure that target sounds are not within words that children frequently omit or alter when repeating. Target sounds with lower intra-rater reliability should be eliminated. Interrogative sentences should be rephrased as declarative sentences. Importantly, any changes to the utterances or the target sounds could affect the frequency distribution of consonants. If changes are

made, other changes would likely need to be made in response to maintain a frequency distribution of consonants similar to that in children's conversational speech.

The SSITS training materials also need further development. SLP participants frequently made mathematical mistakes and were evidently unsure how to deal with not attempted target sounds. Future iterations of these training materials should include more examples of not attempted target sounds with accompanying video clips. Additionally, protocols should be developed for how to score syllable deletions.

One reason for using sentence imitation to assess speech sound production (rather than single words) is that we hypothesized that children's speech production in sentence imitation would more closely approximate their speech sound production in conversational speech than a single-word assessment would. An important step, then, in validating this assessment, would be to transcribe children's conversational speech and analyze the relation between children's speech sound accuracy in that environment and SSITS PCC. Analysis using conversational speech could also further elucidate the relationship between SSITS PCC and intelligibility.

Relatedly, further analysis could be performed comparing children's speech accuracy in the SSITS and the Arizona-4. This information, in combination with conversational speech data, could be useful in establishing whether sentence imitation truly provides a more accurate picture of a child's conversational speech accuracy than do single-word speech accuracy assessments.

Another characteristic of the SSITS that has not fully been explored is its ability to differentiate between children. Based on the present study, the SSITS may have utility as a global measure of speech sound accuracy that can generally capture the

overall magnitude of a child's speech errors and impairment in intelligibility. However, it is yet unknown how well this assessment can differentiate between children with more similar error profiles or how well it can identify whether a child has a speech sound disorder.

One of the goals while developing utterances and selecting target sounds for the SSITS was for the assessment to be neutral to the variety of English that the child speaks. In particular, we thought of Standard American English and African American English. Our sample of child participants was not reflective of the diversity in varieties of English found in the United States. To make claims regarding this assessment's neutrality toward a child's variety of English, it would be necessary to obtain a larger sample with a higher number of children who vary in dialect and investigate whether these groups of children differ in their scores on the SSITS.

Expanding the target audience of the SSITS further, further research could examine whether the measures is feasible with children with receptive and/or expressive language disorders. Roughly 11-15% of six-year-olds with speech sound disorders have specific language impairment (Shriberg et al., 1999). An additional subset of children with speech sound disorders have nonspecific language impairments; prevalence data in this area was unavailable. For this assessment to be the most clinically useful, it should be feasible for children with and without language impairments; however, the task may be too linguistically and/or cognitively taxing to be useful in assessing children with specific and nonspecific language impairments. It may be useful to adapt the SSITS to create different versions for children with varying cognitive and linguistic abilities.

## References

American Speech-Language-Hearing Association. (n.d.). *Childhood hearing screening*. https://www.asha.org/practice-portal/professional-issues/childhood-hearing-screening/

Bryant, B., & Bryant, D. (1983). *Test of Articulation Performance–Screen*. Austin, TX: Pro-Ed.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Routledge.

Craig, H. K., & Washington, J. A. (2002). Oral language expectations for African American preschoolers and kindergartners. *American Journal of Speech-Language Pathology*, *11*(1), 59–70. https://doi.org/10.1044/1058-0360(2002/007)

Ertmer, D. J. (2010). Relationships between speech intelligibility and word articulation scores in children with hearing loss. *Journal of Speech, Language, and Hearing Research*, *53*(5), 1075–1086. https://doi.org/10.1044/1092-4388(2010/09-0250)

Flipsen, P. (2006). Measuring the intelligibility of conversational speech in children. *Clinical Linguistics & Phonetics*, *20*(4), 303–312. https://doi.org/10.1080/02699200400024863

Fudala, J. B., & Stegall, S. (2017). *Arizona Articulation and Phonology Scale, Fourth Revision (Arizona-4)*. Western Psychological Services.

Glaspey, A. M., Wilson, J. J., Reeder, J. D., Tseng, W., & MacLeod, A. A. N. (2022). Moving beyond single word acquisition of speech sounds to connected speech development with dynamic assessment. *Journal of Speech, Language, and*

*Hearing Research, 65*(2), 508-524. https://doi.org/10.1044/2021_JSLHR-21-00188

Goldman, R. & Fristoe, M. (2015). *Goldman-Fristoe Test of Articulation 3 (GFTA-3)*. Pearson.

Golinkoff, R. M., De Villiers, J., Hirsh-Pasek, K., Iglesias, A., & Wilson, M. S. (2017). *Quick Interactive Language Screener*. Brookes.

Johnson, C. A., Weston, A. D., & Bain, B. A. (2004). An objective and time-efficient method for determining severity of childhood speech delay. *American Journal of Speech-Language Pathology*, *13*(1), 55–65. https://doi.org/10.1044/1058-0360(2004/007)

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech. *American Journal of Speech-Language Pathology*, *3*(2), 81–95. https://doi.org/10.1044/1058-0360.0302.81

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, *54*(4), 482–499. https://doi.org/10.1044/jshd.5404.482

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Loeb, D. F., & Leonard, L. B. (1991). Subject case marking and verb morphology in normally developing and specifically language-impaired children. *Journal of Speech, Language, and Hearing Research*, *34*(2), 340–346. https://doi.org/10.1044/jshr.3402.340

Mader, J. B. (1954). The relative frequency of occurrence of English consonant sounds in words in the speech of children in grades one, two, and three. *Communications Monographs, 21*(4), 194-300. https://doi.org/10.1080/03637755409375122

Morris, S. R., Wilcox, K. A., & Schooling, T. L. (1995). The Preschool Speech Intelligibility Measure. *American Journal of Speech-Language Pathology*, *4*(4), 22–28. https://doi.org/10.1044/1058-0360.0404.22

Morrison, J. A. & Shriberg, L. D. (1992). Articulation testing versus conversational speech sampling. *Journal of Speech and Hearing Research, 35*(2), 259-273. https://doi.org/10.1044/jshr.3502.259

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, *53*(2), 333–349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Samar, V. J., & Metz, D. E. (1988). Criterion validity of speech intelligibility rating-scale procedures for the hearing-impaired population. *Journal of Speech, Language, and Hearing Research*, *31*(3), 307–316. https://doi.org/10.1044/jshr.3103.307

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders: Theory, measurement, and management* (pp. 11-34). John Benjamins.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The Percentage of Consonants Correct (PCC) metric: Extensions and reliability data.

*Journal of Speech, Language, and Hearing Research*, *40*(4), 708–722. https://doi.org/10.1044/jslhr.4004.708

Shriberg, L. D., & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders*, *47*(3), 256–270. https://doi.org/10.1044/jshd.4703.256

Shriberg, L. D., Tomblin, J. B., & McSweeny, J. L. (1999). Prevalence of speech delay in 6-year-old children and comorbidity with language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(6), 1461–1481. https://doi.org/10.1044/jslhr.4206.1461

Skahan, S. M., Watson, M., & Lof, G. L. (2007). Speech-language pathologists' assessment practices for children with suspected speech sound disorders: Results of a national survey. *American Journal of Speech-Language Pathology, 16*(3), 246–259. https://doi.org/10.1044/1058-0360(2007/029)

Taddeo, T., White, A. Z., Redfern, A., & Schuele, C. M. (2018, November). *The Childhood Intelligibility Assessment: A translational clinical research project* [Seminar]. Annual Convention of the American Speech-Language-Hearing Association, Boston, MA.

Weston, A. D., & Shriberg, L. D. (1992). Contextual and linguistic correlates of intelligibility in children with developmental phonological disorders. *Journal of Speech, Language, and Hearing Research*, *35*(6), 1316–1332. https://doi.org/10.1044/jshr.3506.1316

Yoder, P. J., Woynaroski, T., & Camarata, S. (2016). Measuring speech comprehensibility in students with Down syndrome. *Journal of Speech,*

*Language, and Hearing Research*, *59*(3), 460–467.

https://doi.org/10.1044/2015_JSLHR-S-15-0149

**Table 1**

*Proportional Frequency of Consonants*

| Sound | Expected percentage[a] | SSITS percentage[b] |
|---|---|---|
| n | 13.14% | 12.69% |
| t | 11.74% | 11.19% |
| d | 10.25% | 9.70% |
| r | 7.83% | 7.46% |
| s | 6.50% | 5.97% |
| ð | 6.40% | 5.97% |
| l | 5.55% | 5.22% |
| w | 5.33% | 5.22% |
| m | 4.63% | 4.48% |
| k | 4.25% | 3.73% |
| z | 3.70% | 3.73% |
| h | 3.33% | 2.99% |
| b | 2.97% | 2.99% |
| p | 2.73% | 2.24% |
| g | 2.38% | 2.24% |
| v | 1.91% | 2.24% |
| f | 1.83% | 2.24% |
| ŋ | 1.61% | 2.24% |
| θ | 0.93% | 1.49% |
| ʃ | 0.84% | 1.49% |
| j | 0.77% | 1.49% |
| ʤ | 0.69% | 1.49% |
| ʧ | 0.55% | 1.49% |
| ʒ | 0.01% | 0.00% |

*Note.* The expected frequency of each American English consonant sound compared with the actual frequency among target sounds in the Story-Sentence Imitation Task for Speech (SSITS)

[a] Frequency percentage of each consonant sound in the conversational speech of 1st-3rd grade children (Mader, 1954). Note that Mader's 'hw' and 'w' sounds have been combined here.

[b] Frequency percentage of each consonant sound among target sounds in the Story-Sentence Imitation Task for Speech (SSITS).

**Table 2**

*Demographics of Child Participants*

| Characteristic | Number | Percentage |
|---|---|---|
| Age at Study Visit 1 | | |
| 3;6-3;11 | 1 | 9% |
| 4:0-4:5 | 7 | 64% |
| 4;6-4;11 | 0 | 0% |
| 5;0-5;5 | 2 | 18% |
| 5;6-5;11 | 1 | 9% |
| Age at Study Visit 2 | | |
| 3;6-3;11 | 1 | 11% |
| 4:0-4:5 | 6 | 67% |
| 4;6-4;11 | 0 | 0% |
| 5;0-5;5 | 1 | 11% |
| 5;6-5;11 | 1 | 11% |
| Currently receiving speech/language therapy | | |
| Yes | 1 | 9% |
| No | 10 | 91% |
| Previously received speech/language therapy | | |
| Yes | 1 | 9% |
| No | 10 | 91% |
| Race | | |
| American Indian or Alaska Native | 0 | 0% |
| Asian | 1 | 9% |
| Black or African American | 1 | 9% |
| Native Hawaiian or Other Pacific Islander | 0 | 0% |
| White | 9 | 82% |
| Highest education level attained by Parent 1 | | |
| Less than high school | 0 | 0% |
| High school or GED | 0 | 0% |

| | | |
|---|---|---|
| Post-secondary vocational school or training | 0 | 0% |
| Some college (four-year or two-year college) | 0 | 0% |
| Associate degree | 0 | 0% |
| Bachelor's degree | 3 | 27% |
| Graduate degree | 8 | 73% |
| Highest education level attained by Parent 2 | | |
| Less than high school | 0 | 0% |
| High school or GED | 1 | 9% |
| Post-secondary vocational school or training | 0 | 0% |
| Some college (four-year or two-year college) | 1 | 9% |
| Associate degree | 0 | 0% |
| Bachelor's degree | 6 | 55% |
| Graduate degree | 3 | 27% |
| Biological Parent 1 history of speech/language therapy | | |
| Yes | 1 | 9% |
| No | 9 | 82% |
| Unknown | 1 | 9% |
| Biological Parent 2 history of speech/language therapy | | |
| Yes | 1 | 9% |
| No | 9 | 82% |
| Unknown | 1 | 9% |
| Biological sibling with current or past speech/language therapy | | |
| Yes | 2 | 18% |
| No | 8 | 73% |
| Unknown | 1 | 9% |

*Note.* All demographic data reported by parent/guardian.

**Table 3**

*Demographics of Speech-Language Pathologist Participants*

| Characteristic | Number | Percentage |
|---|---|---|
| Year obtained master's degree in speech-language pathology | | |
| 1990-1999 | 1 | 20% |
| 2000-2009 | 1 | 20% |
| 2010-2019 | 3 | 60% |
| Years of employment as a speech-language pathologist | | |
| 5-9 | 2 | 40% |
| 10-14 | 1 | 20% |
| 15-19 | 1 | 20% |
| 20-24 | 0 | 0% |
| 25+ | 1 | 20% |
| Current primary employment facility | | |
| School | 4 | 80% |
| Healthcare facility | 1 | 20% |
| Gender | | |
| Female | 5 | 100% |
| Male | 0 | 0% |
| Age | | |
| 20-29 | 0 | 0% |
| 30-39 | 2 | 40% |
| 40-49 | 1 | 20% |
| 50+ | 2 | 40% |
| Race | | |
| American Indian or Alaska Native | 0 | 0% |
| Asian | 0 | 0% |
| Black or African American | 0 | 0% |
| Native Hawaiian or Other Pacific Islander | 0 | 0% |

| | | |
|---|---|---|
| White | 5 | 100% |
| Hispanic or Latino/a | | |
| Yes | 0 | 0% |
| No | 5 | 100% |

*Note.* All data self-reported.

**Table 4**

*Types of Errors Made by SLPs[a] while Scoring the SSITS[b]*

| SLP error type | Description of SLP error |
|---|---|
| A | When there are multiple not attempted target sounds within a single not attempted word, counting only one not attempted target sound |
| B | Double-counting a not attempted target sound as a not attempted target sound and an error |
| C | Counting a non-target sound toward the sum of not attempted target sounds |
| D | Counting a non-target sound toward the sum of errors |
| E | Marking one target sound in a word as not attempted and another target sound in the same word as correct or an error |
| F | Miscounting the number of errors in a single utterance |
| G | Error in summing the total number of errors on a page |
| H | Miscounting the number of not attempted target sounds on a page |
| I | Error in the final SSITS PCC[c] calculation |

[a] Speech-language pathologists
[b] Story-Sentence Imitation Task for Speech
[c] Percentage of consonants correct

**Table 5**

*Dependent Measure Assessment Results by Child*

| Child | Arizona Word Articulation total score[a] | PSIM[b] | SSITS[c] PCC[d] Study Visit 1[e] | SSITS PCC Study Visit 2[f] |
|---|---|---|---|---|
| 1 | 60.5 | 39% | 70.79% | 65.97% |
| 2 | 71.5 | 42% | 72.31% | 79.52% |
| 3 | 72.5 | 43% | 82.69% | 81.31% |
| 4 | 97 | 65% | 97.91% | |
| 5 | 95.5 | 74% | 98.51% | 98.21% |
| 6 | 100 | 65% | 100.00% | 100.00% |
| 7 | 61 | 42% | 70.63% | |
| 8 | 68 | 26% | 60.67% | 67.62% |
| 9 | 91 | 74% | 94.18% | 96.41% |
| 10 | 84 | 46% | 87.00% | 89.70% |
| 11 | 56.5 | 43% | 67.33% | 75.04% |

*Note.* Each child's results on each dependent measure. Blank cells indicate that the participant did not complete the study visit.
[a] Word Articulation total scores from the Arizona Articulation and Phonology Scale—4th Revision (Fudala & Stegall, 2017)
[b] Preschool Speech Intelligibility Measure (Morris et al., 1995); mean of scores obtained by two naïve listeners
[c] Story-Sentence Imitation Task for Speech
[d] Percentage of Consonants Correct
[e] Mean score across five SLP raters
[f] Mean score across five SLP raters

**Table 6**

*Study Visit 1 SSITS[a] PCC[b] by SLP[c] Participant*

| Child participant | SLP 1 | SLP 2 | SLP 3 | SLP 4 | SLP 5 | Mean[e] |
|---|---|---|---|---|---|---|
| 1 | 78.95% | 64.66% | 65.41% | 72.73% | 72.18% | 70.79% |
| | 81.34%[d] | — | — | 72.18% | — | — |
| 2 | 80.60% | 67.91% | 70.15% | 74.22% | 68.66% | 72.31% |
| | — | 63.43% | — | — | 70.68% | — |
| 3 | 89.55% | 76.87% | 88.06% | 79.10% | 79.85% | 82.69% |
| | 90.30% | — | — | 74.63% | — | — |
| 4 | 100.00% | 95.52% | 98.51% | 97.76% | 97.76% | 97.91% |
| | — | — | 98.51% | 97.76% | 97.01% | — |
| 5 | 98.51% | 97.01% | 100.00% | 99.25% | 97.76% | 98.51% |
| | — | 97.01% | — | — | 96.27% | — |
| 6 | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| | — | 99.25% | — | 100.00% | — | — |
| 7 | 79.70% | 66.17% | 69.17% | 67.42% | 70.68% | 70.63% |
| | — | 64.66% | 69.92% | — | — | — |
| 8 | 79.23% | 57.69% | 59.85% | 52.76% | 53.85% | 60.67% |
| | 73.85% | — | — | 51.18% | 60.47% | — |
| 9 | 94.78% | 91.79% | 97.01% | 91.79% | 95.52% | 94.18% |
| | — | 92.54% | 97.01% | — | — | — |
| 10 | 92.54% | 82.84% | 88.06% | 88.72% | 82.84% | 87.00% |
| | 90.30% | — | 87.31% | — | 85.07% | — |
| 11 | 74.05% | 62.60% | 64.18% | 71.97% | 63.85% | 67.33% |
| | 72.93% | — | 63.64% | — | — | — |

*Note.* SSITS PCC on Study Visit 1, as scored by each SLP participant, and means

[a] Story-Sentence Imitation Task for Speech

[b] Percentage of Consonants Correct

[c] Speech-language pathologist

[d] Where there are two numbers in a cell, that SLP rated the same video of the same child twice for intra-rater reliability. The top value is the PCC score that that SLP obtained during their first rating of the video; the bottom value is the PCC score that that SLP obtained during their second rating of the video.

[e] Calculation of the mean of each SLP's first rating of each video (during Task 1).

**Table 7**

*Study Visit 2 SSITS[a] PCC[b] by SLP[c] Participant*

| Child participant | SLP 1 | SLP 2 | SLP 3 | SLP 4 | SLP 5 | Mean |
|---|---|---|---|---|---|---|
| 1 | 77.61% | 59.70% | 58.21% | 67.91% | 66.42% | 65.97% |
| 2 | 85.07% | 75.37% | 81.34% | 79.85% | 75.94% | 79.52% |
| 3 | 89.55% | 77.61% | 84.33% | 75.97% | 79.10% | 81.31% |
| 4 | — | — | — | — | — | — |
| 5 | 99.25% | 97.01% | 97.76% | 100.00% | 97.01% | 98.21% |
| 6 | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| 7 | — | — | — | | — | — |
| 8 | 78.79% | 62.40% | 65.60% | 60.77% | 70.54% | 67.62% |
| 9 | 97.76% | 94.03% | 98.51% | 94.74% | 97.01% | 96.41% |
| 10 | 94.03% | 85.82% | 90.30% | 89.55% | 88.81% | 89.70% |
| 11 | 84.73% | 70.99% | 69.23% | 72.31% | 77.10% | 74.87% |

*Note.* SSITS PCC on Study Visit 2, as scored by each SLP participant, and means Participants 4 and 7 did not complete Study Visit 2.
[a] Story-Sentence Imitation Task for Speech
[b] Percentage of Consonants Correct
[c] Speech-language pathologist

**Table 8**

*Point-by-Point Agreement on the SSITS[a] between SLP Participants*

| SLP | SLP 1 | SLP 2 | SLP 3 | SLP 4 | SLP 5 |
|---|---|---|---|---|---|
| SLP 1 | — | 87.11% | 88.87% | 87.38% | 88.06% |
| SLP 2 | | — | 89.96% | 87.79% | 90.23% |
| SLP 3 | | | — | 89.55% | 90.64% |
| SLP 4 | | | | — | 89.48% |
| SLP 5 | | | | | — |

*Note*. Agreement across 11 Study Visit 1 SSITS videos; percentages indicate the proportion of target sounds on which each pair of SLPs agreed as to whether it was an error, correct, or not attempted.
[a] Story-Sentence Imitation Task for Speech
[b] Speech-language pathologist

**Table 9**

*Ranking of Children's SSITS[a] PCCs[b] as Scored by each SLP[c]*

| Ranking | SLP 1 | SLP 2 | SLP 3 | SLP 4 | SLP 5 |
|---|---|---|---|---|---|
| Child with the highest SSITS PCC | 6 (tie) | 6 | 6 | 6 | 6 |
| | 4 (tie) | 5 | 5 | 5 | 5 |
| | 5 | 4 | 4 | 4 | 4 |
| | 9 | 9 | 9 | 9 | 9 |
| | 10 | 10 | 10 | 10 | 10 |
| | 3 | 3 | 3 | 3 | 3 |
| | 2 | 2 | 2 | 2 | 1 |
| | 7 | 7 | 7 | 1 | 7 |
| | 8 | 1 | 1 | 11 | 2 |
| | 1 | 11 | 11 | 7 | 11 |
| Child with the lowest SSITS PCC | 11 | 8 | 8 | 8 | 8 |

*Note*. Rankings generated based on Study Visit 1 scores.
[a] Story-Sentence Imitation Task for Speech
[b] Percentage of Consonants Correct
[c] Speech-language pathologist

**Table 10**

*SSITS[a] PCCs[b] at Study Visit 1 and Study Visit 2*

| Child participant | Study Visit 1 SSITS PCC | Study Visit 2 SSITS PCC | Difference between Study Visit 1 and Study Visit 2 SSITS PCC |
|---|---|---|---|
| 1 | 70.79% | 65.97% | -4.82% |
| 2 | 72.31% | 79.52% | 7.21% |
| 3 | 82.69% | 81.31% | -1.37% |
| 5 | 98.51% | 98.21% | -0.30% |
| 6 | 100.00% | 100.00% | 0 |
| 8 | 60.67% | 67.62% | 6.95% |
| 9 | 94.18% | 96.41% | 2.23% |
| 10 | 87.00% | 89.70% | 2.70% |
| 11 | 67.33% | 75.04% | 7.71% |

*Note.* SSITS PCCs represented here are the mean of the judgments of 5 speech-language pathologist participants.
[a] Story-Sentence Imitation Task for Speech
[b] Percentage of Consonants Correct

**Table 11**

*Ranking of Children's SSITS[a] PCCs[b] at Two Time Points*

| Ranking | Study Visit 1 | Study Visit 2 |
|---|---|---|
| Child with the highest SSITS PCC | 6 | 6 |
| | 5 | 5 |
| | 9 | 9 |
| | 10 | 10 |
| | 3 | 3 |
| | 2 | 2 |
| | 1 | 11 |
| | 11 | 8 |
| Child with the lowest SSITS PCC | 8 | 1 |

*Note.* Rankings compiled based on the mean of 5 speech-language pathologists' scoring of each child.
[a] Story-Sentence Imitation Task for Speech
[b] Percentage of Consonants Correct

**Table 12**

*Frequency of Each Type of Error Made by SLPs[a] while Scoring the SSITS[b]*

| SLP error type[c] | SLP 1 | SLP 2 | SLP 3 | SLP 4 | SLP 5 | Total |
|---|---|---|---|---|---|---|
| A | | 3 | 9 | 1 | 2 | 15 |
| B | | | 13 | | | 13 |
| C | | | 12 | | | 12 |
| D | | | | 1 | | 1 |
| E | | | | 5 | | 5 |
| F | 1 | 1 | 3 | 13 | 4 | 22 |
| G | | | 1 | 1 | 1 | 3 |
| H | | | | 1 | | 1 |
| I | | | | | 1 | 1 |
| Total | 1 | 4 | 38 | 22 | 8 | |

*Note*. Data in this table represents the number of times that each SLP participant made each type of SLP error across 20 videos of child administrations of the SSITS (11 Study Visit 1 videos and 9 Study Visit 2 videos) after receiving training on the scoring of the SSITS.

[a] Speech-language pathologists

[b] Story-Sentence Imitation Task for Speech

[c] See Table 4 for a description of SLP error types

**Table 13**

*Difference between Corrected[a] and Uncorrected SSITS[b] PCCs[c]*

| Statistic | SLP[d] 1 | SLP 2 | SLP 3 | SLP 4 | SLP 5 | Overall |
|---|---|---|---|---|---|---|
| Mean | -0.04% | -0.07% | 0.69% | -0.30% | -0.39% | -0.02% |
| SD | 0.17% | 0.31% | 1.83% | 0.75% | 1.70% | 1.22% |
| Minimum | -0.75% | -1.38% | -0.75% | -1.52% | -7.46% | 7.62% |
| Maximum | 0.00% | 0.00% | 7.62% | 1.52% | 0.75% | -7.46% |

[a] Corrected scores refer to the score calculated by the author after data preparation/correction of SLP errors; uncorrected scores refer to the score calculated by the SLP participant
[b] Story-Sentence Imitation Task for Speech
[c] Percentage of Consonants Correct
[d] Speech-language pathologist

**Table 14**

*Suggested Revisions to SSITS[a] Utterances[b]*

| Utterance number | Utterance | Suggested revision | Rationale |
|---|---|---|---|
| 1 | Welcome to the zoo. | Do not score /ð/ in 'the' | Low intra-rater reliability |
| 11 | Everybody, look there! | Do not score /r/ in 'everybody' | Low intra-rater reliability |
| 21 | There's the snake. | Do not score /ð/ in 'there's' | Low intra-rater reliability |
| 35 | I really like this new hat. | Do not score /ð/ in 'this' | Low intra-rater reliability |
| 14 | Can the pretty bird help? | Rephrase as a declarative sentence | Increased linguistic load of interrogative sentences |
| 20 | Where is she going? | Rephrase as a declarative sentence | Increased linguistic load of interrogative sentences |
| 25 | What about the rhino? | Rephrase as a declarative sentence | Increased linguistic load of interrogative sentences |
| 31 | Can you please help me? | Rephrase as a declarative sentence | Increased linguistic load of interrogative sentences |
| 19 | She's sliding through the bush | Rephrase, do not score /ʃ/ in 'she's' | Frequent nonrepetition of 'she's' |
| 24 | Mom and dad say time for lunch | Rephrase, do not score /s/ in 'say' | Frequent nonrepetition of 'say' |
| 31 | Can you please help me? | Rephrase, do not score /z/ in 'please' | Frequent nonrepetition of 'please' |

*Note*. This list is not meant to be inclusive of all recommended or possible changes to the SSITS. Each change described could affect the frequency distribution of consonants; in response, other changes would likely need to be made to maintain a frequency distribution of consonants similar to that in children's' conversational speech.
[a] Story-Sentence Imitation Task for Speech
[b] All utterances and target sounds can be found on the Sentence Repetition Form (Appendix A)

# Appendix A
Sentence Repetition Form

Video number code: _____ Page 1

**Sentence Repetition Form**

1. Welcome to the zoo. **Errors:** _____

| | | | |
|---|---|---|---|
| /wɛlkəm/ | /tu/ | /ðə/ | /zʊ/ |
| /wɛlkəm/ | /tə/ | /ðə/ | /zʊ |
| w m | t | ð | z |

2. What will we see today? **Errors:** _____

| | | | | |
|---|---|---|---|---|
| /wʌt/ | /wɪl/ | /wi/ | /si/ | /təde/ |
| /wʌtəl/ | | /wi/ | /si/ | /təde/ |
| w | | | s | t d |

3. Watch me. I'm resting! *[point to cheetah in tree]*. **Errors:** _____

| | | | |
|---|---|---|---|
| /watʃ/ | /mi/ | /aɪm/ | /rɛstɪŋ/ |
| /watʃ/ | /mi/ | /aɪm/ | /rɛstɪŋ/ |
| tʃ | m | | r s t |

4. Hey you, jump on in! *[point to alligators]* **Errors:** _____

| | | | | |
|---|---|---|---|---|
| /heɪ/ | /ju/ | /dʒʌmp/ | /an/ | /ɪn/ |
| /heɪ/ | /ju/ | /dʒʌmp/ | /an/ | /ɪn/ |
| h | j | dʒ p | n | |

5. Hush, I need a nap. *[point to bear]* **Errors:** _____

| | | | | |
|---|---|---|---|---|
| /hʌʃ/ | /aɪ/ | /nid/ | /ə/ | /næp/ |
| /hʌʃ/ | /aɪ/ | /nid/ | /ə/ | /næp/ |
| ʃ | | n d | | n |

6. They have long necks. *[point to giraffes]* **Errors:** _____

| | | | |
|---|---|---|---|
| /ðe/ | /hæv/ | /lɔŋ/ | /nɛks/ |
| /ðe/ | /æv/ | /lɔŋ/ | /nɛks/ |
| ð | v | l ŋ | n |

7. Good for picking things up. *[point to mama giraffe]* **Errors:** _____

| | | | | |
|---|---|---|---|---|
| /gʊd/ | /fɚ/ | /pɪkɪŋ/ | /θɪŋz/ | /ʌp/ |
| /gʊd/ | /fɚ/ | /pɪkɪŋ/ | /θɪŋz/ | /ʌp/ |
| g d | | p k | θ ŋ | |

8. You took my hat! *[point to boy and then to the hat]* **Errors:** _____

| | | | |
|---|---|---|---|
| /ju/ | /tʊk/ | /maɪ/ | /hæt/ |
| /ju/ | /tʊk/ | /maɪ/ | /hæt/ |
| | t | m | |

9. Now that bad monkey wants it. *[point to monkey]* **Errors:** _____

| | | | | | |
|---|---|---|---|---|---|
| /naʊ/ | /ðæt/ | /bæd/ | /mʌŋki/ | /wants/ | /ɪt/ |
| /naʊ/ | /ðæt/ | /bæd/ | /mʌŋki/ | /wants/ | /ɪt/ |
| n | ð | d | ŋ | w | |

Total Errors Page 1_____ Total Phonemes Not Attempted Page 1 _____

Video number code: _____ Page 2

10. Throw it over here! *[point to monkey]* **Errors:** _____

| /θro/ | /ɪt/ | /ovə-/ | /hir/ |
|---|---|---|---|
| /θro/ | /ɪt/ | /ovə-/ | /hir/ |
| θ r | t | v | |

11. Everybody look there! *[point to hat in the monkey's hand]* **Errors:** _____

| /ɛvribʌdi/ | /lʊk/ | /ðɛr/ |
|---|---|---|
| /ɛvribədi/ | /lʊk/ | /ðɛr/ |
| v r b d | l | ð |

12. A monkey in a tall tree. *[point to monkey]* **Errors:** _____

| /ə/ | /mʌŋki/ | /ɪn/ | /ə/ | /tal/ | /tri/ |
|---|---|---|---|---|---|
| /ə/ | /mʌŋki/ | /ɪn/ | /ə/ | /tal/ | /tri/ |
| | m k | n | | t | |

13. Let me have my hat! *[point to boy]* **Errors:** _____

| /lɛt/ | /mi/ | /hæv/ | /maɪ/ | /hæt/ |
|---|---|---|---|---|
| /lɛ/ | /mi/ | /hæv/ | /maɪ/ | /hæt/ |
| l | | | | h |

14. Can the pretty bird help? *[point to bird]* **Errors:** _____

| /kæn/ | /ðə/ | /prɪti/ | /bɝd/ | /hɛlp/ |
|---|---|---|---|---|
| /kæn/ | /ðə/ | /priti/ | /bɝd/ | /hɛɬp/ |
| k n | | t | b d | |

15. Watch the hat fall down. *[trace hat falling]* **Errors:** _____

| /watʃ/ | /ðə/ | /hæt/ | /fɔl/ | /daʊn/ |
|---|---|---|---|---|
| /watʃ/ | /ðə/ | /hæt/ | /fɔl/ | /daʊn/ |
| w | | | f | d |

16. Right onto that old snake. *[point to snake's head]* **Errors:** _____

| /raɪt/ | /ɑntʊ/ | /ðæt/ | /old/ | /snek/ |
|---|---|---|---|---|
| /raɪt/ | /ɑntə/ | /ðæt/ | /old/ | /snek/ |
| r t | n t | t | | |

17. Keep it away from the lion. *[point to female lion]* **Errors:** _____

| /kip/ | /ɪt/ | /əweɪ/ | /frʌm/ | /ðə/ | /laɪən/ |
|---|---|---|---|---|---|
| /kip/ | /ɪt/ | /əweɪ/ | /frʌm/ | /ðə/ | /laɪən/ |
| k p | | | f | | l |

18. The snake just got away. *[point to snake]* **Errors:** _____

| /ðə/ | /snek/ | /dʒʌst/ | /gat/ | /əwe/ |
|---|---|---|---|---|
| /ðə/ | /snek | /dʒʌst/ | /gat/ | /əwe/ |
| | | dʒ | t | w |

Total Errors Page 2_____ Total Phonemes Not Attempted Page 2 _____

Video number code: _____ Page 3

19. She's sliding through the bush. *[point to snake]* **Errors:** _____

| /ʃiz/ | /slaɪdɪŋ/ | /θru/ | /ðə/ | /bʊʃ/ |
|---|---|---|---|---|
| /ʃiz/ | /slaɪdɪŋ/ | /θru/ | /ðə/ | /bəʃ/ |
| ʃ | l d | | | b |

20. Where is she going? **Errors:** _____

| /wɛr/ | /ɪz/ | /ʃi/ | /goɪŋ/ |
|---|---|---|---|
| /wɛr/ | /ɪz/ | /ʃi/ | /goɪŋ/ |
| w | | | g |

21. There's the snake. *[point to snake]* **Errors:** _____

| /ðɛrz/ | /ðə/ | /snek/ |
|---|---|---|
| /ðɛrz/ | /ðə/ | /snek/ |
| ð | | s |

22. Mom! Do you see her nose? *[point to mom]* **Errors:** _____

| /mam/ | /du/ | /ju/ | /si/ | /hɝ/ | /noz/ |
|---|---|---|---|---|---|
| /mam/ | /də/ | /jə/ | /si/ | /ɝ/ | /noz/ |
| m | d | | | | n z |

23. Maybe Dad can reach my hat. *[point to dad]* **Errors:** _____

| /mebi/ | /dæd/ | /kæn/ | /ritʃ/ | /maɪ/ | /hæt/ |
|---|---|---|---|---|---|
| /mebi/ | /dæd/ | /kən/ | /ritʃ/ | /maɪ/ | /hæt/ |
| m b | d | | r | | |

24. Mom and dad say time for lunch. *[point to parents]* **Errors:** _____

| /mam/ | /ænd/ | /dæd/ | /se/ | /taɪm/ | /fɝ/ | /lʌntʃ/ |
|---|---|---|---|---|---|---|
| /mam/ | /n/ | /dæd/ | /se/ | /taɪm/ | /fɝ/ | /lʌntʃ/ |
| | n | | s | t | | n |

25. What about the rhino? *[point to rhino]* **Errors:** _____

| /wʌt/ | /əbaʊt/ | /ðə/ | /raɪno/ |
|---|---|---|---|
| /wʌt/ | /əbaʊt/ | /ðə/ | /raɪno/ |
| t | | ð | r n |

26. Now let's find a yummy snack. *[point to parents]* **Errors:** _____

| /naʊ/ | /lɛts/ | /faɪnd/ | /ə/ | /jʌmi/ | /snæk/ |
|---|---|---|---|---|---|
| /naʊ/ | /lɛts/ | /faɪnd/ | /ə/ | /jʌmi/ | /snæk/ |
| | | n d | | j | n |

27. Let's go. *[point to parents]*

| /lɛts/ | /go/ |
|---|---|
| /lɛts/ | /go/ |
| | |

Total Errors Page 3_____ Total Phonemes Not Attempted Page 3 _____

Video number code: _____ Page 4

28. Such a long line. *[point to parents in line]* **Errors:** _____

| /sʌtʃ/ | /ə/ | /lɔŋ/ | /laɪn/ |
|---|---|---|---|
| /sʌtʃ/ | /ə/ | /lɔŋ/ | /laɪn/ |
| s | | | n |

29. I miss that green hat. *[point to boy]* **Errors:** _____

| /aɪ/ | /mɪs/ | /ðæt/ | /grin/ | /hæt/ |
|---|---|---|---|---|
| /aɪ/ | /mɪs/ | /ðæt/ | /grin/ | /hæt/ |
| | s | ð | g n | |

30. Hello Zookeeper! **Errors:** _____

| /hɛlo/ | /zukipɚ/ |
|---|---|
| /hɛlo/ | /zukipɚ/ |
| h | z k |

31. Can you please help me? **Errors:** _____

| /kæn/ | /ju/ | /pliz/ | /hɛlp/ | /mi/ |
|---|---|---|---|---|
| /kæn/ | /ju/ | /pliz/ | /hɛlp/ | /mi/ |
| | | z | h | |

32. I lost my favorite hat. **Errors:** _____

| /aɪ/ | /lɔst/ | /maɪ/ | /fevrɪt/ | /hæt/ |
|---|---|---|---|---|
| /aɪ/ | /lɔs/ | /maɪ/ | /fevrɪt/ | /hæt/ |
| | l s | | f t | |

33. Cheer up, I have a great idea. *[point to zookeeper]* **Errors:** _____

| /tʃir/ | /ʌp/ | /aɪ/ | /hæv/ | /ə/ | /gret/ | /aɪdiə/ |
|---|---|---|---|---|---|---|
| /tʃir/ | /ʌp/ | /aɪ/ | /æv/ | /ə/ | /gret/ | /aɪdiə/ |
| tʃ | | | | | r t | d |

34. You can have my hat!

| /ju/ | /kæn/ | /hæv/ | /maɪ/ | /hæt/ |
|---|---|---|---|---|
| /ju/ | /kən/ | /hæv/ | /maɪ/ | /hæt/ |
| | | | | |

35. I really like this new hat. *[point to boy]* **Errors:** _____

| /aɪ/ | /ɹli/ | /laɪk/ | /ðɪs/ | /nu/ | /hæt/ |
|---|---|---|---|---|---|
| /aɪ/ | /ɹli/ | /laɪk/ | /ðɪs/ | /nu/ | /hæt/ |
| | r l | | ð s | n | |

36. We're ready for the zebra exhibit *[point to family]* **Errors:** _____

| /wɚ/ | /ɹɛdi/ | /fɔr/ | /ðə/ | /zibrə/ | /ɛgzɪbɪt/ |
|---|---|---|---|---|---|
| /wɚ/ | /ɹɛdi/ | /fɚ/ | /ðə/ | /zibrə/ | /əgzɪbət/ |
| w | r d | | | z r | |

And that's the end of our story!

Total Errors Page 4_____ Total Phonemes Not Attempted Page 4 _____

# Appendix B
Procedures & Scoring Summary Form

Video number code: _____ Page 5

## Story Sentence Imitation Task for Speech

Rev. Date: 1.12.22 | Hamers thesis study version

Administration and Scoring Instructions

1. For each sentence, the first line of boxes on the Sentence Repetition Form is citation transcription, the second line of boxes is coarticulated / casual transcription, and the third line of boxes includes phonemes measured in that sentence (i.e., target sounds).
2. Administer the practice items according to the directions below.
3. Present each page of the book and read each sentence, one at a time, for the child to repeat. Only explicitly ask the child to repeat as needed (i.e., if the child is not repeating automatically).
4. If the child omits or replaces a word or multiple words that include target sounds, score those target sounds in the words the child repeated. Then readminister the sentence one time to obtain repetition of omitted or replaced word(s). Score only the previously omitted or replaced word(s) from the second administration. If the child omits the word(s) again, proceed to the next sentence.
5. On the third line of boxes for each target sentence, **circle** the phonetic symbol for any target sound produced **incorrectly**. Deletions and substitutions **are** counted as errors. Distortions and insertions **are not** counted as errors.
6. If a word with a target sound(s) is not **attempted after two administrations of a sentence**, mark a **large "X"** through all rows representing the word (e.g., if the child deletes or replaces an entire word).
7. A horizontal RED line marks where to turn to the next page in the book.
8. After the administration is completed,
    a. Tally the number of errors (circled sounds) in each sentence and write that number in the "errors" blank to the right of the sentence.
    b. Tally the number of errors on each page. Note the sum at the bottom of the page.
    c. Tally the number of target sounds that were not attempted (target sounds in words with a large "X" through them). Note the sum at the bottom of the page.
9. Complete the summary on the reverse of this sheet to calculate the child's **SSITS PCC** (percent consonants correct).

**Definitions** Sentence: A sentence in the SSITS that is repeated by the child.
Target sound: A phoneme that is scored for accuracy

---

**Administration Procedure**

**Examiner to child**: *We are going to read my story book together. I am going to show you some pictures and tell you the story. I want you to tell the story back to me. You'll say just what I say. Let's practice*. Here's the book. [show the child the cover page] Let's read at the title, A Fun Day at the Zoo.

Sentences to practice imitation task. Read the text just as written. Circle YES or NO as to whether the child repeats every word. If NO, then put an X on the words not repeated. As needed, prompt the child to repeat the practice sentences, wait until you are finished before saying the sentence, etc. The goal in the practice sentences is that the child knows by the last practice sentence to repeat the sentence without being told to.

| | | |
|---|---|---|
| YES | NO | *SAY: The sky is blue. NOW YOU SAY IT. GREAT!* |
| YES | NO | *SAY: It's time for a fun day. YOU SAID IT JUST LIKE ME.* |
| YES | NO | *Look at the boy. GOOD WORK.* |

*You did a great job on those practice sentences. Now let's do the rest of the story. I think you will like this story and the pictures!*

A few times while doing the task, provide verbal praise to the child for working hard, saying just what you say, etc.

Video number code: _____ Page 6

## Story Sentence Imitation Task for Speech: Summary

| Calculation of Total Errors on Target Sounds | |
|---|---|
| | # Errors |
| Page 1 | |
| Page 2 | |
| Page 3 | |
| Page 4 | |
| **Total Errors on Target Sounds (A)** | |

| Calculation of Target Sounds Not Attempted | |
|---|---|
| | # Target Sounds Not Attempted |
| Page 1 | |
| Page 2 | |
| Page 3 | |
| Page 4 | |
| **Total Target Sounds Not Attempted (B)** | |

| | | |
|---|---|---|
| | **134** | |
| – | | Total Target Sounds **Not** Attempted (B) |
| = | | Total Target Sounds Attempted (C) |
| – | | Total Errors on Target Sounds (A) |
| = | | Total Target Sounds Correct (D) |

$$\frac{\text{Total Target Sounds Correct (D)}}{\text{Total Target Sounds Attempted (C)}} = \text{ } \times 100 = \text{ \%}$$

**SSITS PCC**
final score