The Influence of Multiple Presentations on Judgments of Children's Phonetic Accuracy

Benjamin Munson
Kayla N. Brinkman*
University of Minnesota, Minneapolis

Two experiments examined whether listening to multiple presentations of recorded speech stimuli influences the reliability and accuracy of judgments of children's speech production accuracy. In Experiment 1, 10 listeners phonetically transcribed words produced by children with phonological impairments after a single presentation and after the word was played 7 times. Inter- and intratranscriber reliability in the single- and multiple-presentation conditions did not differ significantly. In Experiment 2, 18 listeners provided binary correct/incorrect judgments of /s/ accuracy in single- and multiple-presentation conditions. There was no systematic effect of presentation condition on either accuracy or intrarater reliability. However, greater interrater reliability was noted in the multiple-presentation condition, particularly for tokens of /s/ that were incorrect or acoustically intermediate between an incorrect and a correct /s/. Taken together, the results suggest that multiple presentations have no measurable effect on the accuracy and intrarater reliability of judgments of children's phonetic accuracy, but that they do have a small effect on interrater reliability. Clinical implications are discussed.

Key Words: phonological and articulation disorders, assessment, speech perception

Children with articulation or phonological impairments (henceforth PI) of an unknown origin constitute a large proportion of the caseloads of school speech-language pathologists (Leske, 1981; Whitmire, Karr, & Mullen, 2000). Children with PI produce many addition, deletion, substitution, and distortion errors relative to their peers with typical development. These children may be highly unintelligible and may require speech-language therapy to achieve intelligible speech. Assessments of PI typically consist of a battery of standardized and nonstandardized measures of speech, language, and hearing. A typical component of the assessment battery for PI is a single-word naming test. Many standardized, norm-referenced single-word naming tests for children with PI exist, including the Arizona Articulation Proficiency Scale–Third Edition (AAPS-3; Fudala, 2001), the Bankson-Bernthal Test of Phonology (BBTOP; Bankson & Bernthal, 1990), the Goldman-Fristoe Test of Articulation–Second Edition (GFTA-2; Goldman & Fristoe, 2000), and the Smit-Hand Articulation and Phonology Examination (SHAPE; Smit & Hand, 1997), among others. Most of these tests require examiners to report the accuracy of children's responses in a picture-naming task. Standard scores are calculated by comparing the number of errors made by the child to those made by children in the normative sample.

In many settings, the results from these tests are required to qualify children for services. Moreover, clinicians use the results of these tests to conduct detail analyses of error types and to select initial therapy goals. It is critical, then, that the results of these tests be both reliable and accurate. That is, clinicians must accurately perceive, remember, and record the speech of children with PI during a single-word naming task. However, it is well documented that a number of cognitive factors potentially compromise the reliability and accuracy of people's perception, memory, and reporting of speech sounds. This article explores whether limitations in one of these processes, speech perception, may impact the accuracy and reliability of judgments of speech produced by children with PI.

Many factors bias the process of speech perception, and these biases may affect clinical assessments. In speech perception, individuals must relate a variable acoustic signal to phonological and lexical categories in long-term memory. In this process, people must ignore a great deal of acoustic variability when making the association between an evanescent speech signal and its representation in

*Currently affiliated with the University of Oregon, Eugene.

long-term memory (e.g., Kluender, 1994). This is illustrated by classic categorical perception phenomena in adults. In forced-choice identification tasks, adults may classify acoustically distinct stimuli as members of a single perceptual category. For example, English-speaking adults classify stimuli varying in voice-onset time (VOT) as members of either a voiced or voiceless category; two stimuli with acoustically distinct VOTs (i.e., +10 ms and –30 ms) are typically classified as members of the same category (i.e., /g/, /b/, or /d/) when they fall in the range associated with that perceptual category. In normal speech perception, categorical perception allows people to ignore irrelevant variability. However, this same variability might not be irrelevant in a child's phonological system. Indeed, categorical perception may *compromise* a clinician's assessment of a child with PI who is able to produce an acoustic difference between two target phonemes that falls within one of the clinician's perceptual categories. A number of studies have shown that children with apparent sound-neutralization errors may produce acoustic differences between target sounds that are imperceptible to naïve listeners, because both sounds fall within one of the listener's perceptual categories (Gierut & Dinnsen, 1986; Scobbie, Gibbon, Hardcastle, & Fletcher, 2000).

Categorical perception illustrates only one of the many normal perceptual processes that may compromise assessments of speech-production accuracy. Another limitation may arise due to differences in listeners' familiarity with the child whose speech is being reported. Flipsen (1995) showed that children with PI were more intelligible to their primary caregivers than to unfamiliar people. Simple expectations or linguistic context may also influence speech perception (Ingrisano, Klee, & Binger, 1996; Oller & Eilers, 1975). For example, a listener who is confronted with a sound intermediate between /s/ and /θ/ may be more likely to report a percept of /θ/ if they know that the speaker was attempting to produce the word *think* rather than *sink*. The same sound might be perceived as /s/ in the word *south*, as the word *thouth* is not a real word of English. Speech perception can be biased by social expectations. Listeners are more likely to perceive a sound intermediate between /s/ and /ʃ/ as /s/ when paired with a man's face, and as /ʃ/ when paired with a woman's face, presumably due to their expectation that women produce both fricatives with higher-frequency energy than men (Strand, 1999).

Another confounding factor in speech perception concerns the influence of multiple presentations. When listeners are exposed to multiple presentations of a word, they may report that the word they hear changes over the course of the repetitions (e.g., Kaminska, Pool, & Mayer, 2000; MacKay, Wulf, Ying, & Abrams, 1993; Shoaf & Pitt, 2002). That is, individuals who listen to multiple presentations of a recorded token of the word *sink* may report that it begins to sound like *think* after a number of repetitions. This effect has been termed the *verbal transformation effect* (MacKay et al., 1993). Various explanations have been offered for the verbal transformation effect. While these explanations are not the focus of this investigation, they generally posit that the verbal transformation effect arises when lexical activation for a target word spreads to phonologically or semantically related words in the lexicon. These words then compete with the actual word as perceptual responses. An alternative explanation proposes that the verbal transformation effect occurs when the peripheral or central structures that are activated in response to the target word become fatigued by repeated activation.

The verbal transformation effect could potentially affect clinical transcription. One natural response when attempting to transcribe an uncertain word is to play it multiple times, under the assumption that subsequent presentations will decrease uncertainty and increase the accuracy and reliability of the transcription (Shriberg, Kwiatkowski, & Hoffman, 1984). There is some support from research in nonspeech auditory perception for the efficacy of this strategy. Research in hearing science suggests that general auditory perception might be facilitated by listening to multiple presentations of a stimulus (e.g., Viemeister & Wakefield, 1991). This may extend to low-level speech discrimination. Holt (2003) found that discrimination between the acoustically similar phonemes /s/ and /ʃ/ is facilitated in a multiple-presentation condition relative to a single-presentation condition. If these findings were to translate to judgments of speech production accuracy, then multiple presentations might facilitate perception of the speech of children with PI. However, the bulk of prior research on the verbal transformation effect (e.g., Kaminska et al., 2000; MacKay et al., 1993; Shoaf & Pitt, 2002) would suggest that multiple presentations should inhibit perception. Consequently, people who listen to multiple presentations may experience false percepts of what a child said.

A recent review article (Kent, 1996) hypothesized that multiple presentations may influence assessments of the speech of children with PI and, in particular, may decrease the accuracy and reliability of clinical transcription. Individual examiners who experience incorrect percepts of children's productions due to the verbal transformation effect might transcribe a child's production differently on two different occasions (i.e., the effect might lead to poor *intrarater* reliability). Moreover, this effect might make it difficult for different transcribers to arrive at a consensus (i.e., the effect might lead to less agreement among different coders, which we refer to in this paper as *interrater* reliability). This decreased reliability could potentially affect the diagnosis of PI, if individuals were to perceive the children with PI as producing speech sounds correctly when they are actually producing them incorrectly. Perhaps more importantly, it could influence clinicians' analyses of error types and phonological patterns in children's speech. That is, this perceptual bias could influence speech-language pathologists' determination of the type of errors that a child makes, which could potentially lead to their choosing inappropriate sound teaching strategies.

There has been relatively little research examining factors that influence the reliability of judgments of children's speech production accuracy. Much of this research has dealt with methodological issues. For

example, research has argued that the reliability of phonetic transcriptions cannot be measured using summary statistics designed to assess the presence or absence of a behavior (as can be used to measure reliability of assessments of disfluencies; Lewis, 1994). Thus, a great deal of previous research in this area has focused on developing meaningful measures to assess reliability of phonetic transcriptions (e.g., Cucchiarini, 1996). The few studies that have examined factors influencing transcription have generally focused on the relationship between transcription detail and reliability. Shriberg and Lof (1991) found that reliability was poorer for transcription that utilizes diacritics than for more broad transcription.

The purpose of this study is to examine experimentally the extent to which multiple presentations either compromise or facilitate the accuracy and reliability of judgments of children's phonetic accuracy. This research has two implications. First, it provides empirical verification of the extent to which the verbal transformation effect might influence the transcription of the speech of children with PI. Previous research on this topic has only speculated that this might be a potential confound, based on studies of adults' speech perception (Kent, 1996). Second, the results have the potential to provide clinical transcribers with recommendations regarding the optimal conditions to obtain high intra- and intertranscriber reliability.

This investigation contains two experiments. The first experiment examines the influence of multiple presentations on the reliability of phonetic transcriptions of children's speech. The second experiment investigates the influence of multiple presentations on binary correct/incorrect judgments of children's production of a single sound /s/. This is an exploratory study with no a priori hypotheses: some prior research would suggest that multiple presentations should facilitate phonetic transcription (i.e., increase its accuracy and reliability) because of its facilitative effect on low-level speech discrimination (Holt, 2003). Other research suggests that it should inhibit phonetic transcription because of the effect that it has on inducing verbal transformations (Shoaf & Pitt, 2002). The purpose of this investigation is to examine what effect, if any, multiple presentations have on judgments of children's speech production accuracy.

Experiment 1: Multiple Presentations and Phonetic Transcription

The purpose of Experiment 1 was to examine the extent to which multiple presentations might affect inter- and intrarater reliability (here called *inter-* and *intratranscriber reliability*) of phonetic transcriptions of the speech of children with PI. Phonetic transcription is not a required component of most standardized measures of articulation. For example, standard scores and percentile ranks on the GFTA-2 test are calculated based on the number of errors that the child produces on the target sounds of that test. In contrast, the SHAPE test allows the administrator to select among likely occurring alternative pronunciations of target sounds, with the option to phonetically transcribe a production that is not among the options. However, for tests like the GFTA-2, phonetic transcription can provide more diagnostic information than simple tallies of errors. The full word on the GFTA-2 must be transcribed for the examiner to score a companion assessment, the Kahn-Lewis Phonological Assessment–Second Edition (KLPA-2; Kahn & Lewis, 2002). Indeed, the examiner's manual for the GFTA-2 strongly recommends that some version of phonetic transcription (either transcription of the target sounds or transcription of entire words) be used for children with more severe articulation and phonological deficits (Goldman & Fristoe, 2000, pp. 24–25).

Experiment 1 consisted of two tasks. The primary task of interest was one in which the participants phonetically transcribed the speech of children. This task was administered to each participant twice, in sessions that were separated by at least 1 week. In each session, words were presented in two conditions: one in which they were played only once, and one in which they were played multiple times. Analyses focused on whether there was more consistent transcription across the two sessions for the single versus the multiple presentation condition, and whether there was greater consensus across the transcribers in the multiple- versus the single-presentation condition. The results of this task allow us to examine whether multiple presentations systematically influence the reliability of phonetic transcriptions. It is important to note that this experiment measures transcription reliability only; it does not measure transcription accuracy. The stimuli used in this experiment were natural productions by children with PI. Consequently, there was no unquestionably "accurate" baseline transcription of these words that could be used to measure the participants' transcription accuracy. However, we could compare the participants' transcriptions to themselves and to each other to measure reliability.

In the second task, we measured the participants' recognition memory for words spoken by adults with typical speech and language abilities. This task consisted of a prime phase and a test phase. The prime phase was administered during the first experimental session and involved the participants passively listening to words produced by adults. The test phase was administered in the second session. In this phase, the participants listened to words produced by adults and judged whether they were identical to words that they had heard during the prime phase. This task served two purposes. First, it served as a distracter during the transcription task. Within each session, the participants transcribed the same set of words twice. The prime or test phase of the recognition memory task occurred between the two phonetic-transcription tasks, to minimize the chances that the participants' second transcriptions of the words would be influenced by their memory of what they had transcribed earlier in the session. Second, the results of the recognition memory experiment served as baseline data on the participants' ability to remember specific stimuli from session to session. The participants who demonstrated good recognition memory were predicted to show greater intratranscriber reliability on the transcription task than those who did not.

Methods

Participants

The transcribers were 10 people who were either advanced graduate students in speech-language pathology or beginning speech-language pathologists. All of the participants had received formal training in phonetic transcription and in the assessment of phonological disorders in children, as gauged by self-report. All were native speakers of English, and none reported a history of speech, language, or hearing disorder. The transcribers had experience working in clinical jobs or graduate clinical placements with children with PI. The participants received $5.00 for their participation. They were debriefed about the nature of the experiment after it was completed.

Stimuli

Transcription reliability. The stimuli for this experiment consisted of 45 words produced by 9 different children. These children had been recruited to participate in another research project examining lexical and phonological influences on speech production in children with PI and typically developing age-matched peers (Munson, Swenson, & Manthei, in press). Demographic data on these children are provided in Table 1. As this table shows, both boys and girls produced words used in this experiment. The children all had age-appropriate expressive vocabularies, as indicated by their scoring no lower than 1 *SD* below the mean on the Expressive Vocabulary Test (Williams, 1997). One exception was participant S4, for whom standardized measures were not available. This participant's expressive vocabulary was judged to be age-appropriate by examiners and by his speech-language pathologist. Standard scores on the GFTA-2 show that most of the participants scored within 1 *SD* of the normative sample's mean. Many of the children in the study were enrolled in speech therapy, and their higher-than-expected scores on the GFTA-2 likely reflected progress in speech therapy between the time of initial diagnosis and their participation in the experiment. Nonetheless, the children did make a variety of speech-sound errors and are reflective of the population of better-performing children with PI.

The stimulus words were taken from the GFTA-2 and were collected as part of the regular administration of that test. All of the words were produced in response to a picture stimulus; the examiner provided no auditory prompts for the 45 tokens used in this experiment. Children wore an AKG C420 head-mounted microphone, placed approximately 6 cm from their mouths. Words were recorded directly on to the hard drive of a Roland VS890 Digital Workstation, at a 44.1-kHz sampling rate, with 16-bit quantization. The stimuli were normalized such that the peak amplitudes of all of the stimuli were identical.

TABLE 1. Demographic data for children who provided the stimuli for Experiment 1.

ID	Age	Sex	GFTA-2[a] standard score	EVT[b] standard score
S1	3;8	M	83	102
S2	3;8	M	99	105
S3	4;3	M	86	117
S4	4;4	M	80	
S5	5;1	F	96	99
S6	5;4	M	72	113
S7	5;6	M	89	120
S8	6;7	M	108	118
S9	7;11	F	94	107

Note. The empty cell indicates data were not available.

[a]Goldman-Fristoe Test of Articulation-2 (Goldman & Fristoe, 2000).

[b]Expressive Vocabulary Test (Williams, 1997).

The procedure for choosing the 45 target words was as follows. First, the entire GFTA-2 was transcribed for each child. Each child's production of each word on that test was coded as correct, incorrect with a common error, or incorrect with an uncommon error. Errors were coded as "common" if they were listed as phonological processes on the KLPA-2. From this, it was determined that 18% of the words had been produced correctly, 12% included productions with uncommon error patterns, and 70% with common phonological processes. Eight correctly produced words, five words with uncommon error patterns, and 32 words with common error patterns were chosen quasi-randomly to be used as the stimuli in the experiment. This resulted in a stimulus set in which the percentage of error types (correct, incorrect with an uncommon error, incorrect with a common error) was similar to that in the larger set of words. The selection process was quasi-random in that two additional constraints were imposed on the selection of the stimuli from the larger set of words. First, equal numbers were selected from the 9 children. Second, only those stimuli that did not contain any extraneous noise were chosen.

The first author's phonetic transcription of these 45 words is given in Table 2. As mentioned above, these transcriptions are not meant to serve as the "correct" transcription against which adult participants' transcriptions are evaluated; these transcriptions are no less subject to the biases described in the introduction than those of the research the participants. Rather, these are provided to give the reader a sense of the range of error patterns that were evidenced in the experimental stimuli. As this table shows, the children produced a variety of addition, deletion, substitution, and distortion errors. These ranged from errors illustrating relatively common phonological processes (i.e., the fronting pattern noted in child S5's production of *cup* as [tʊp]; the cluster-simplification process noted in child S8's pronunciation of *spoon* as [pun]) to uncommon or idiosyncratic patterns (i.e., child S1's pronunciation of *blue* as ['θiju]; child S4's pronunciation of *drum* as ['tʊmɪ]).

Recognition memory. Sixty stimuli were used in the recognition memory task. These stimuli consisted of three tokens each of 20 different words. The stimuli were culled from a larger set of recordings made by the first author for use in a computerized version of the Word Intelligibility by Picture Identification test (WIPI; Ross & Lerman, 1979). In the larger set of recordings, 10 different adult talkers (5 men, 5 women) produced the full set of words that appears on the WIPI. Each stimulus was recorded in a sound-treated room using an AKG C420 head-mounted microphone connected to a Roland VS890 digital workstation. A

TABLE 2. The stimuli used in the transcription task, including the first author's transcription.

ID	Word		ID	Word		ID	Word	
S1	blue	ˈθ̲ijuª	S4	drum	ˈt̲ʌmɪ	S7	plane	p̲weɪn
S1	green	d̲in	S4	green	g̲in	S7	shovel	ˈʃʌbə
S1	knife	naɪf	S4	plane	peĩ	S7	thumb	f̲ʌm
S1	tree	t̲sĩ	S4	slide	t̲laɪd	S7	tree	k̲ri
S1	watches	ˈwɔt̲ɪts	S4	watch	jɑt̲ʃ	S7	zipper	ˈzɪpə˞
S2	ball	bɔl	S5	bath	bæ	S8	chair	tʃer
S2	ring	v̲ɪŋ	S5	cup	t̲ʌp	S8	drums	drʌmz
S2	slide	laɪ	S5	light	waɪt	S8	pencils	ˈpɛnsl̩z
S2	window	v̲ɪnoʊ	S5	monkey	ˈmʌki	S8	shovel	ˈʃʌv̩l
S2	zipper	ˈdɪpə˞	S5	slide	s̲waɪd	S8	spoon	pun
S3	brush	b̲ʌs	S6	fishing	fɪt̲sɪn	S9	crying	ˈk̲waɪjɪŋ
S3	drum	t̲wʌm	S6	five	faɪːf̲	S9	frog	f̲wɔg
S3	rabbit	ˈw̲æbɪt	S6	swimming	ˈθ̲ɪmɪn	S9	green	g̲win
S3	shovel	ˈsʌb̩l	S6	thumb	θ̲ʌm	S9	thumb	θʌm
S3	wagon	ˈwægɪt̲	S6	wagon	ˈweɪgɪn	S9	yellow	ˈjɛl̲oʊ

Note. Sounds used in the intertranscriber reliability analysis are underlined.

44.1-kHz sampling rate and 16-bit quantization were used. These stimuli were normalized for peak amplitude. The subset used for this task consisted of three tokens from each of the talkers. The full set of stimuli consisted of three different talkers' productions of each of 20 words, which are presented in the Appendix.

To ensure that the stimuli were uniformly intelligible, a group of six adult listeners was presented with the 60 stimuli at 65 dB SPL in the sound-field and asked to identify the word. The tokens that were chosen were correctly identified by at least five of the listeners, and most were correctly identified by all six listeners.

Procedures

Transcription reliability. Prior to data collection, the participants were told that they would be transcribing children with speech-sound impairments. They were told to use the symbols and the level of transcription detail that they would use in a typical clinical assessment. Each session contained two transcription tasks: one single-presentation task and one multiple-presentation task. The experiment took place in a soundproof booth containing a 17-in. computer monitor. On each trial, the word *LISTEN* was shown in 36-point Courier font in the center of the monitor for 1 s. This was followed by the presentation of a stimulus, concurrent with an orthographic display of the stimulus on the computer screen. In the single-presentation condition, words were presented only once. In the multiple-presentation condition, the participants heard the same token seven consecutive times (separated by pauses of 250 ms) before they transcribed it. The stimuli were presented at a level of approximately 65 dB SLP in stereo, through speakers located at 60° and 300° azimuth from the speaker's head. The experiment was self-paced; the participants pressed a button on a button-box to advance through items.

The order of the single- and multiple-presentation experiments was randomized, both within individual sessions and across the two sessions. Approximately equal numbers of participants participated in the four different experimental orders. Within each session, the two transcription tasks were separated by a recognition memory task.

Recognition memory. There were two portions of the recognition memory task. The *prime* task occurred during the first experimental session, between the two transcription tasks. In this task, the participants were told that they would be hearing a list of words spoken by 10 different adults. They were told that they should listen to the words as carefully as possible, as they would be listening to a similar word list during their second session and judging whether individual words had been presented to them during the first session. The participants listened to words in a sound-treated booth. Words were presented at a level of approximately 65 dB SPL through speakers located at 60° and 300° azimuth from the speaker's head. Each word was presented seven times in succession, with individual repetitions separated by 250 ms. Forty tokens (20 different words, two presentations each by two different talkers) were presented. An orthographic display of the word was presented on a computer screen concurrent with its audio presentation. The presentation rate was self-selected; the participants pressed a button when they were ready to hear the next item. The participants were not allowed to make notes during this task.

The *recognition* task occurred during the second experimental session between the two transcription tasks. In this task, the participants were told that they would be listening to a list of words and judging whether each individual word had been presented during the first experimental session. The participants were told that the 20 words would be the same as those used in the prime phase and that they would hear each word twice. They were also told that one token of each word would be identical to a token heard in the prime phase, and the other token would be spoken by a different talker. A total of 40 tokens (20 words, two presentations each by two different talkers) were presented. Twenty of the tokens presented in the recognition task consisted of a word–talker combination that had been presented in the prime phase. The remaining 20 tokens consisted of a word–talker combination that had not been presented in the prime phase. In both the prime and the recognition phase, all 10 talkers were presented.

Again, the participants were seated in a sound-treated booth. The stimuli were output in stereo from two speakers at a level of approximately 65 dB SPL. The speakers were located at 60° and 300° azimuth from the speaker's head. In this experiment, each word was presented only once. An orthographic display of the word was presented on a computer screen concurrent with its audio presentation. Following each word, the participants judged whether or not the word had been presented before by pressing a button on a button-box. Responses were recorded automatically.

Analysis

Transcription reliability. Two measures of transcriber reliability were calculated. Intrarater reliability for individual participants was calculated by taking the number of phonemes common to the two transcriptions, dividing them by the length of the longer transcription, and expressing the quotient as a percentage. For example, consider participant S101's two transcriptions of the target word *slide* produced by child S2 in the single-presentation condition. The first transcription was [laɪ]. The second transcription was [laɪdə]. In this example, the two-phoneme sequence [laɪ] is common to the two transcriptions, and the longer of the two transcriptions contains four phonemes. Percentage agreement was (2/4)*100 = 50%.

A second example is illustrated by participant S103's transcription of the target word *drum* produced by child S3 in the multiple presentation condition. Transcription 1 was [svʊm] and transcription 2 was [fʌm]. The phoneme [m] is common to the two transcriptions; the longest transcription contains four phonemes. Percentage agreement was (1/4) *100 = 25%. For each participant, mean intratranscriber reliability was calculated separately for the single- and multiple-presentation conditions.

Intertranscriber reliability was calculated separately for the single- and multiple-presentation conditions. Intertranscriber reliability was calculated by examining consensus on transcription of the 23 sounds that are underlined in Table 2. The sounds that were used in the analysis of intertranscriber reliability were selected as follows. Only sounds that the first author transcribed as incorrectly produced were selected as candidates for this analysis. After the participants S101 and S102 had completed the experiment, their responses were compared to the transcriptions of the first author. The intention was to select 8 sounds that had been transcribed the same way by all three transcribers, 8 that had been transcribed differently by all three transcribers, and 8 that had been transcribed the same by two of the transcribers and differently by the third. By selecting some sounds that were consistently transcribed and others that were not, we hoped to minimize the possibility that the data for this analysis would be restricted in range. Only 7 sounds had been transcribed differently by all three transcribers; these were chosen as target sounds for this analysis. The other 16 were chosen by first classifying all of the errors as either consistently transcribed or inconsistently transcribed, then choosing 8 target sounds from each of these categories quasi-randomly. The selection was quasi-random because the constraint was imposed that approximately equal numbers of stimuli were selected from the different children. The data from the first experimental session were used to calculate intertranscriber reliability. Intertranscriber reliability among the 10 participants was calculated separately for the single- and multiple-presentation conditions.

The calculation used to measure intertranscriber reliability is illustrated with the target /tʃ/ in S4's production of *watch*. In the single-presentation condition, 7 of the 10 listeners agreed that it was produced as /tʃ/; intertranscriber reliability was 70% in this condition. In the multiple-presentation condition, 5 of the 10 listeners agreed that it was produced as /tʃ/; intertranscriber reliability was 50% in this condition. A second example is given by the target /bl/ in S1's production of *blue*. In the single-presentation condition, 3 of the 10 transcribers agreed that it was produced as /d/; intertranscriber reliability was 30% in this condition. In the multiple-presentation condition, 5 of the 10 transcribers agreed that it was produced as /θ/; intertranscriber reliability was 50% in this condition.

Recognition memory. To assess recognition memory, *d*-prime (d') statistics were calculated (MacMillan & Creelman, 1991). This measure of signal detection is based both on hits (words in the recognition memory test phase correctly identified as having occurred in the prime phase) and on false alarms (words in the recognition memory test phase incorrectly identified as having occurred in the prime phase). As is conventional (MacMillan & Creelman), d' values over 1.0 were presumed to reflect greater-than-chance performance. Those below 1 indicate chance performance.

Results

Recognition Memory

Table 3 shows the d' statistics that represent performance on the recognition memory task. One of the 10 transcribers (S101) achieved a d' greater than 1, indicating that he could detect the difference between the stimuli that had been presented in the first session and those that had not; the other 9 participants did not detect this difference. This suggests that the participants were not able to recognize which of the stimuli had been presented in the earlier session and which had not been presented at greater-than-chance levels. Importantly, this suggests that the intratranscriber reliability measures were not spuriously inflated by the participants remembering details of the children's productions during Session 2 that they had heard during Session 1.

Transcription Reliability

Intertranscriber reliability. Intertranscriber reliability for the 23 individual items is shown in Table 3. Across the items, mean intertranscriber reliability in the single-presentation condition was 58.8% (*SD* = 20.5%). Mean intertranscriber reliability in the multiple-presentation condition was 60.4% (*SD* = 19.7%). Kolmogorov-Smirnoff tests of normality indicated that these data did not meet the normality assumption required for the use of parametric statistics. Consequently, a nonparametric Wilcoxon signed-ranks test was used to determine statistical significance. The small difference in intertranscriber reliability across the 23 items did not achieve significance ($z = -.339$, $p > .05$). As Table 3 shows, consensus sounds were different in

TABLE 3. Mean intertranscriber reliability for individual items.

ID	Word	Transcription	Single	Sound[a]	Multiple	Sound[a]
S1	blue	ˈθ̲iju	30%	/d/	50%	/θ/
S1	green	d̲in	80%	/d/	100%	/d/
S1	tree	t̲sɪ	60%	/t/	50%	/t/
S1	watches	ˈwɔt̲ɪts	50%	/t/	50%	/t/
S2	ring	v̲ɪŋ	40%	/b/	60%	/v/
S2	window	v̲ɪnoʊ	70%	/v/	80%	/v/
S3	brush	b̲ʌs	70%	/b/	80%	/b/
S3	drum	t̲wʌm	40%	/tw/	40%	/tw/
S3	rabbit	ˈw̲æbɪt	50%	/w/	30%	/w/
S3	wagon	ˈwægɪt̲	40%	omission	40%	omission
S4	drum	ˈt̲ʌmɪ	60%	/t/	70%	/t/
S4	slide	t̲laɪd	20%	/s/, /d/	30%	/s/
S4	watch	jɑt̲ʃ	70%	/tʃ/	50%	/tʃ/
S5	cup	t̲ʌp	70%	/t/	80%	/t/
S6	fishing	fɪt̲sɪn	50%	/θ/	70%	/θ/
S6	swimming	ˈθ̲ɪmɪn	70%	/θ/	70%	/θ/
S6	thumb	θ̲ʌm	60%	/θ/	70%	/θ/
S7	plane	p̲weɪn	40%	/pl/	40%	/pw/
S7	thumb	f̲ʌm	80%	/f/	60%	/f/
S7	tree	k̲ri	20%	/tr/	40%	/kr/
S9	crying	ˈk̲waɪjɪŋ	70%	/kw/	90%	/kw/
S9	frog	f̲wɔg	100%	/fw/	80%	/fw/
S9	green	g̲win	90%	/gw/	80%	/gw/
Mean (*SD*)			58% (21%)		61% (20%)	

[a]The sound that was transcribed most often in a given condition (see text for details).

single- and multiple-presentation conditions for 5 of the 23 items and identical for the other 18 items.

Intratranscriber reliability. Intratranscriber reliability was calculated separately for single- and multiple-presentation conditions. Values for individual transcribers are presented in Table 4. Mean intratranscriber reliability in the single-presentation condition was 83.6% (*SD* = 8.6%, range = 66–92.1%). Mean intratranscriber reliability in the multiple-presentation condition was 84.4% (*SD* = 7.5%, range = 68.8–90.6%). Data did not meet the normality assumptions needed to calculate parametric statistics. Again, a Wilcoxon signed-ranks test was used to measure statistical significance. This difference did not achieve significance ($z = -.866$, $p > .05$). In addition, the 1 participant whose recognition memory was greater than 1.0 (S101) did not show higher intrarater reliability than the other 9 participants.

TABLE 4. Intratranscriber reliability in the single- and multiple-presentation condition and recognition memory *d′* values for each transcriber in Experiment 1.

Transcriber ID	Intratranscriber reliability: Single	Intratranscriber reliability: Multiple	Recognition memory *d′*
S101	66.0%	68.8%	0.910
S102	68.4%	72.3%	1.049
S103	90.3%	88.3%	–0.128
S104	86.8%	87.7%	0.385
S105	92.1%	85.5%	0.524
S106	80.6%	87.0%	0.000
S107	81.1%	80.4%	0.000
S108	82.2%	79.2%	–0.910
S109	86.2%	90.6%	0.000
S110	84.6%	88.7%	0.507
Mean (*SD*)	81.83% (8.6%)	82.85% (7.5%)	0.234 0.568

Discussion

This experiment examined whether the number of presentations systematically affects the reliability of phonetic transcriptions. Previous research (Kent, 1996) speculated that multiple presentations might decrease the reliability of phonetic transcriptions, given the impact that they have on normal perceptual processes. However, no support was found for this hypothesis. In this experiment, the use of single- or multiple-presentation conditions did not systematically affect either inter- or intratranscriber reliability of phonetic transcriptions of children's speech. The clinical implication is that listening to multiple presentations will neither enhance nor detract from transcription reliability.

Perhaps the most surprising finding of this study is not reflected in the results section. The original study was designed to include a larger number of the participants than eventually participated. During the recruitment phase, many people indicated that they were hesitant to participate because they indicated that they did not use phonetic transcription regularly in their assessments of speech-sound disorders in children. Instead, they indicated that they use binary correct/incorrect judgments of target phonemes on whatever test they are using. (Recall that this is not the method recommended by the manual for the GFTA-2, which recommends that some form of phonetic transcription be used for children with more severe

speech-sound disorders.) If this observation were to extend to the population of practicing speech-language pathologists more generally, then the ecological and social validity of Experiment 1 would be considerably weakened: if phonetic transcription were not generally used in assessments, then information about factors that influence its reliability would have little practical application. In response to this finding, a second experiment was constructed examining the influence of multiple presentations on binary correct/incorrect judgments of speech-production accuracy.

Experiment 2: Multiple Presentations and Binary Accuracy Judgments

Experiment 2 examined whether multiple presentations systematically affect the reliability of binary correct/incorrect judgments of children's phonetic accuracy and the rate with which people judged sounds to be correctly produced. This experiment differed from Experiment 1 in four key ways. First, the responses in Experiment 2 were binary correct/incorrect judgments, rather than phonetic transcriptions. Second, the multiple-presentation condition in Experiment 2 contained only three presentations of the stimuli, rather than seven. Informal feedback received from the participants in Experiment 1 during their debriefing indicated that they believed that they would use no more than three repetitions when encountering an unfamiliar word in a real-world clinical assessment. Third, Experiment 2 examined correct/incorrect judgments for only a single sound, /s/. The sound /s/ was chosen because it is a commonly misarticulated sound (Smit, Freilinger, Bernthal, Hand, & Bird, 1990). Consequently, speech-language pathologists should have extensive experience hearing correct and incorrect productions of /s/, and any clinical recommendations resulting from this experiment would have broad applicability. The articulatory and acoustic characteristics of children's productions of /s/ vary greatly as a function of phonetic context. Moreover, they show considerably more intraspeaker variability than is found in productions by adults (Munson, 2004). We predict that practicing speech-language pathologists would have considerable experience hearing a variety of pronunciations of /s/.

Finally, this experiment is different from Experiment 1 in that the stimuli were created by digitally manipulating natural-speech tokens of words containing /s/. One weakness of Experiment 1 was that the lack of control over the stimuli prevented us from assessing the accuracy of the participants' transcriptions. By using digitally manipulated stimuli, we knew a priori which stimuli contained tokens of /s/ whose acoustic characteristics mimicked those of correctly produced /s/ and which did not.

Methods

Participants

Eighteen people participated in Experiment 2. As in Experiment 1, this group contained a mix of advanced graduate students in speech-language pathology and practicing speech-language pathologists. All of the participants had received formal training in phonetic transcription and in the assessment of phonological disorders in children, as gauged by self-report. None reported a history of speech, language, or hearing disorder. The participants were recruited in the University of Minnesota community and in the greater Twin Cities metropolitan area. The participants received $7.00 for participating in Experiment 2. The participants were debriefed about the nature of the experiment after it was completed.

Stimuli

The stimuli were constructed from recordings of seven words containing /s/. These are listed in Table 5. Four of these words contained /s/ in initial position, and three contained /s/ in final position. These words were chosen because they are frequent and familiar, and are likely to be known and used by children.

The stimuli were produced by a typically developing 7-year-old girl who was participating in an unrelated study. As part of her participation in that study, it was determined that she had age-appropriate speech, language, and hearing skills, as evidenced by performance within normal limits on a large battery of standardized tests. Naturally produced tokens of the eight words containing /s/ (seven of which were used as experimental stimuli, and one of which was used as a practice item) and eight distracter items containing /ʃ/ were elicited. Multiple tokens of each word were recorded. Recordings were made with an AKG C420 head-mounted microphone, placed approximately 6 cm from the speaker's mouth. Words were recorded directly onto the hard drive of a Roland VS890 Digital Workstation at a 44.1-kHz sampling rate, with 16-bit quantization.

The duration of all of the recorded tokens was measured using the *Praat* signal-processing program (Boersma & Weenink, 2004). One token of each target word containing /s/ was chosen such that the group of eight tokens matched in total stimulus duration as closely as possible. These tokens were played to a group of five graduate students in speech-language pathology to assess their intelligibility. All of the tokens were correctly identified by the five listeners. The interval of aperiodic energy associated with the /s/ was then digitally removed. The remaining portions of the eight words were then normalized for duration using the PSOLA algorithm in *Praat*, so that the rime portions of the stimuli all were 300 ms long. This involved minimal modification, as the rime portion of the original the stimuli varied from 283 ms to 321 ms. They were then normalized for peak amplitude.

Each stimulus base (i.e., the natural word tokens with /s/ removed) was concatenated with three different synthetic tokens of /s/, for a total of 24 experimental stimuli. These were created using the Klatt synthesizer (Klatt, 1980;

TABLE 5. The stimuli used in Experiment 2.

Initial position	Final position
sign	less
song	race
south	yes
sun	

interface by Qi & Johnson, 1987). The parameter files used in the Klatt synthesizer were taken from a study examining the influence of gender normalization on fricative perception (Strand, 1999). The three tokens of /s/ that were chosen were identified as /s/ 100% of the time in that study (henceforth *correct* /*s*/), 75% of the time (henceforth *intermediate* /*s*/), and 50% of the time (henceforth *incorrect* /*s*/). All of the /s/ tokens were 150 ms long. The correct /s/ had a center of gravity (i.e., first spectral moment; Forrest, Weismer, Milenkovic, & Dougall, 1988) of 8572 Hz and a skewness (i.e., the third spectral moment) of –1.46. The intermediate /s/ had a center of gravity of 6201 Hz and a skewness of –2.06. The incorrect /s/ had a center of gravity of 4546 Hz and a skewness of –2.15. The relative amplitude of the fricative and vowel portions was consistent across the stimuli, as this variable has been shown to influence fricative perception (e.g., Hedrick & Carney, 1997).

Two short pre-experiment tests were given to assess the intelligibility and distinctness of the stimuli. In the first, a group of three graduate students in speech-language pathology was played triplets of the target words containing correct, intermediate, and incorrect /s/. These listeners uniformly judged the tokens with correct /s/ to be better examples of the target words than those containing intermediate or incorrect /s/. In a second pretest, pairs of target words containing all possible combinations of the stimuli (including identical pairs) were played. All three students could reliably discriminate among words containing the three different manipulations of /s/. These listeners' subjective impressions were closely in line with the authors' expectations of how the stimuli should sound. They indicated that tokens of correct /s/ sounded clearly correct, albeit clearly digitally manipulated. Tokens of incorrect /s/ were said to sound clearly incorrect. Subjective judgments about words containing intermediate /s/ were less clear.

Procedures

Prior to participating in the experiment, the participants were told that they would be listening to words spoken by a young girl. They were told that some of the sounds in these words had been manipulated digitally. The experiment took place in a soundproof booth containing a 17-in. monitor. All stimuli were presented over a single loudspeaker located at 0° azimuth from the speaker's head, at a level of approximately 65 dB SPL. On each trial, the word LISTEN was shown in 36-point Courier font in the center of the monitor for 1 s. This was followed by the presentation of a stimulus, concurrent with an orthographic display of the stimulus on the computer screen. After the word was played, the participants were told to press one button if the /s/ (or, in the distracter trials, the /ʃ/) was produced correctly, and a different button if it was produced incorrectly. The distracter stimuli contained various quality tokens of /ʃ/, so that listeners would not always judge the words containing /ʃ/ as correct. In one-half of the trials, the target words were played only once. In the other half of the trials, words were played three times; each stimulus was separated by a pause of 250 ms. The experiment was preceded by a practice block containing a word not used in the experiment. Each stimulus/presentation condition combination was played twice, so that we could assess intrarater reliability. We reasoned that the use of a single talker with multiple manipulations per word would reduce the influence of participants' memory of prior responses on their performance. Hence, we did not use a distracter task (like the recognition memory task in Experiment 1) in this experiment. There were a total of 84 experimental stimuli (7 words × 2 presentation conditions × 3 fricative manipulations × 2 repetitions) and 96 distracter items. Stimulus order was randomized across participants. Experiment 2 took place in a single session lasting approximately 30 min.

Analysis

Correct judgments. The first summary measure in this experiment was the percentage of tokens that the listeners judged to have been produced correctly. These values were calculated separately for words with initial and final /s/, with correct, intermediate, or incorrect /s/, in single- and multiple-presentation conditions, for a total of 12 data points per participant.

Intrarater reliability. The second summary measure in Experiment 2 was mean intrarater reliability. For this measure, we examined the rate at which listeners provided the same judgment for the two presentations of the same stimulus/presentation condition combination. In this analysis, participants scored a 1 if they provided identical judgments for the two different times they heard a particular stimulus in a particular condition, and a 0 otherwise. For example, a participant who rated the token of correct /s/ in the word *south* in the single-presentation condition as correctly produced both times she/he heard it would receive a 1. A participant who rated this sound "correct" one time and "incorrect" the other time would receive a 0. The mean of these was calculated separately for words with initial and final /s/, with correct, intermediate, or incorrect /s/, in single- and multiple-presentation conditions, for a total of 12 data points per participant.

Interrater reliability. The final measure taken from this experiment was a measure of interrater reliability. This measure was the proportion of times a word was rated the same by all of the 18 listeners in the study, averaged across the two conditions in which the word was presented. For example, if 14 of 18 listeners rated the intermediate /s/ in *less* in the single-presentation condition as correct the first time it was played, and 16 rated it as correct the second time it was played, then mean interrater reliability would be 83.3% (i.e., 100*(((14/18) + (16/18))/2)). This was measured separately for each stimulus, with each /s/ manipulation, in each repetition condition, for a total of 42 (7 × 3 × 2) data points.

Results

Each summary statistic was submitted to a three-factor within-subjects ANOVA in order to examine the influence of word position (initial vs. final), condition (single vs. multiple) and fricative manipulation (correct vs. incorrect vs. intermediate). Partial η^2 measures of effect size are reported for all significant main effects and interactions. This is a measure of the variance in the summary statistic

that is accounted for by word position, presentation condition, or fricative manipulation.

Mean Correct Judgments

Kolmogorov-Smirnoff tests confirmed that the data collected in this experiment met the normality assumptions required to use fully factorial parametric ANOVA. A strong, significant main effect of manipulation was found, $F(2, 34) = 188.6$, $p < .001$, partial $\eta^2 = 0.92$. Words containing correct /s/ were rated as having been correctly produced 97% of the time. In contrast, words containing intermediate /s/ were rated as correct 76% of the time, and those containing incorrect /s/ were rated as correct 15% of the time. Post hoc Scheffe tests showed that all pairwise differences were significant at the .01 level. There was no effect of word position: the rate with which /s/ was judged to have been correctly produced was similar for word-initial and word-final tokens. Importantly, there was no significant main effect of condition. The rate at which words were judged as correctly produced was very similar in single- and multiple-presentation conditions ($M = 62.7\%$, $SD = 9.8\%$ for the single-presentation condition; $M = 62.9\%$, $SD = 10.4\%$ for the multiple-presentation condition).

Figure 1 shows the percentage of correct judgments in the single- and multiple-presentation conditions for words containing /s/ in word-initial position (top) and word-final position (bottom). As this figure shows, the rate with which word-final /s/ was judged to be correctly produced was nearly identical for single- and multiple-presentation conditions. In word-initial position, small (approximately 5%) differences were noted between single- and multiple-presentation conditions for the words containing incorrect and intermediate /s/. In both of these cases, higher accuracy rates were noted in the multiple-presentation condition. However, neither of these differences achieved statistical significance.

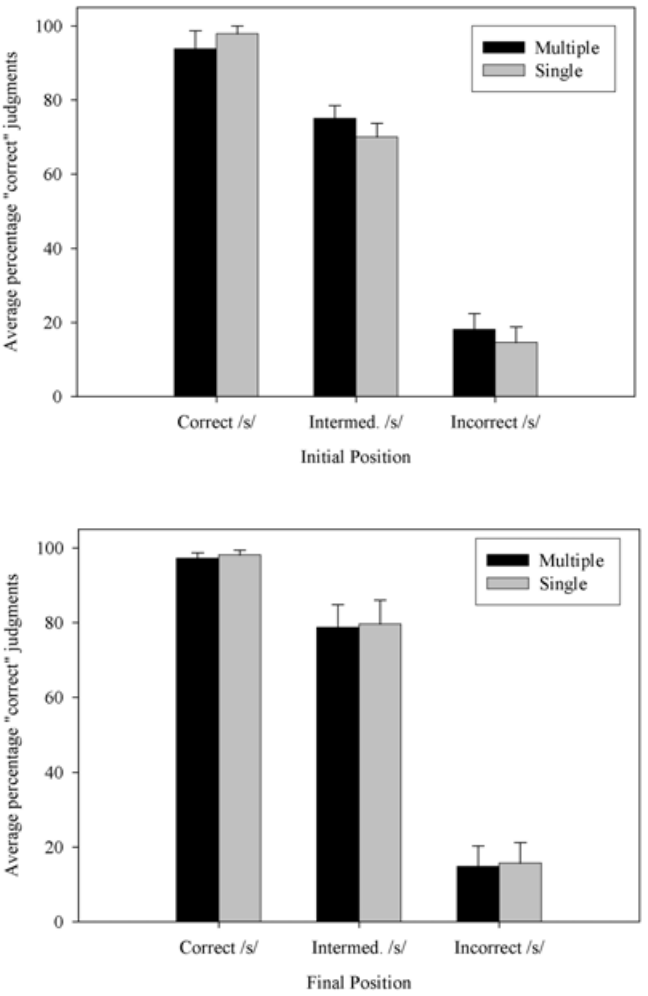
Intrarater Reliability

In this analysis, a significant main effect of manipulation was found, $F(2, 34) = 5.2$, $p = .01$, partial $\eta^2 = 0.23$. The mean reliability was 94% for words containing correct /s/, 84% for words containing intermediate /s/, and 79% for words containing incorrect /s/. Post hoc Scheffe tests indicated that all pairwise differences were significant at the .05 level. A significant main effect of word position was also found, $F(1, 17) = 4.0$, $p = .04$, partial $\eta^2 = 0.17$. Greater consistency was noted in judgments of the accuracy of /s/ in word-initial position ($M = 88\%$) than word-final position ($M = 83\%$). As in the analysis of accuracy judgments, no significant effect of presentation condition was found. Intrarater reliability was nearly identical in the single-presentation ($M = 86\%$) and multiple-presentation ($M = 85\%$) conditions. Figure 2 shows the mean intrarater reliability for correct, intermediate, and incorrect /s/ in word-initial position (top) and word-final position (bottom), in the two presentation conditions. Readers should note that the variance accounted for by these ANOVAs (as measured by partial η^2) was considerably lower than the variance accounted for in the ANOVAs on the rate with which sounds were judged to have been produced accurately.

Interrater Reliability

A significant main effect of manipulation was found,

FIGURE 1. Mean percentage of tokens judged to be correctly produced for /s/ in word-initial position (top) and word-final position (bottom) in single- and multiple-repetition conditions. Error bars represent 1 *SEM*.



$F(2, 30) = 10.6$, $p < .001$, partial $\eta^2 = 0.41$. The effect of fricative manipulation interacted with position, $F(2, 30) = 3.3$, $p = .049$, partial $\eta^2 = 0.18$. This interaction occurred because there was a significant effect of manipulation on interrater reliability in final position, $F(1, 15) = 8.6$, $p = .003$, but not initial position, $F(1, 21) = 1.6$, $p > .05$. There was also a significant main effect of condition, $F(1, 30) = 7.5$, $p = .01$, partial $\eta^2 = 0.20$. This factor interacted with manipulation, $F(2, 30) = 7.5$, $p = .002$, partial $\eta^2 = 0.33$. Post hoc tests of significant main effects showed that this interaction was due to there being a significant influence of multiple presentations on interrater reliability of incorrect /s/ ($F(1, 12) = 25$, $p < .001$) and intermediate /s/, $F(1, 12) = 5.8$, $p = .03$, but not correct /s/, $F(1, 12) = 2.4$, $p > .05$.

The variance in interrater reliability accounted for by the experimental variables (as measured by partial η^2) was slightly higher than the variance in intrarater reliability that was accounted for by the same variables. Data on interrater reliability measures are shown in Figure 3, which shows

FIGURE 2. Mean intrarater reliability (measured as the percentage of tokens judged similarly both times they were presented) for /s/ in word-initial position (top) and word-final position (bottom) in single- and multiple-presentation conditions. Error bars represent 1 *SEM*.

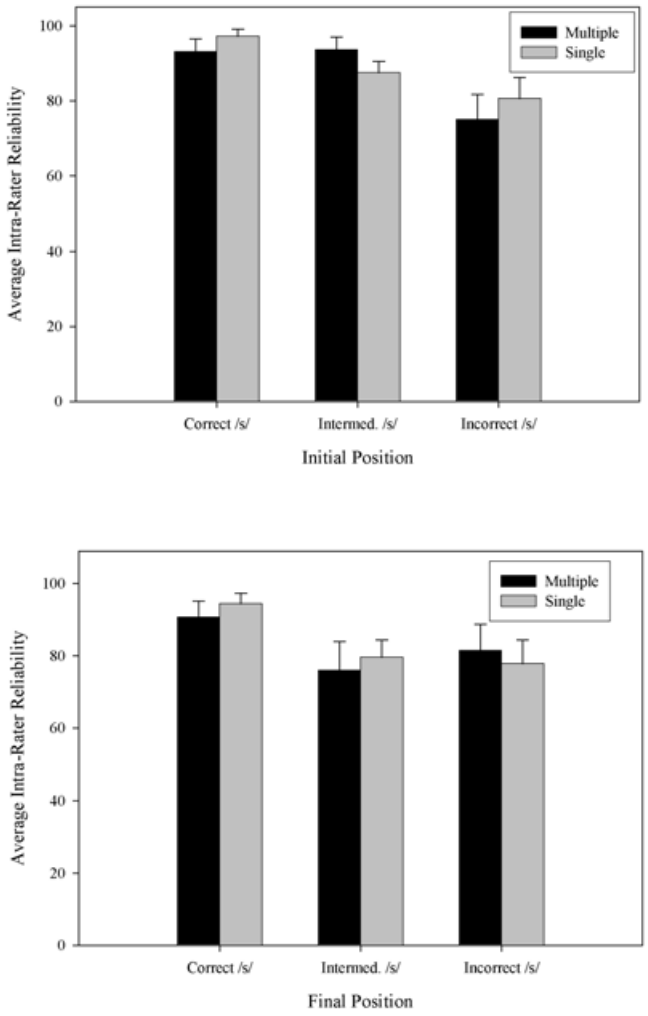
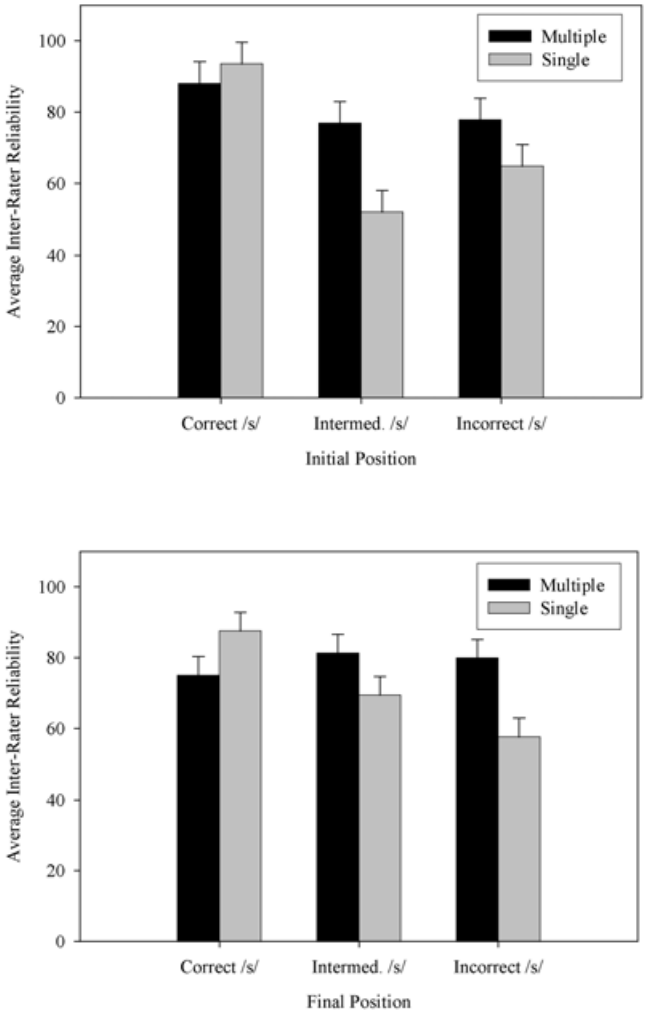


FIGURE 3. Mean interrater reliability (measured as the percentage of the 18 listeners who judged the sound similarly) for /s/ in word-initial position (top) and word-final position (bottom) in single- and multiple-presentation conditions. Error bars represent 1 *SEM*.



the mean interrater reliability for correct /s/, intermediate /s/, and incorrect /s/ in the single- and multiple-presentation conditions for /s/ in word-initial (top) and word-final (bottom) position.

Discussion

This experiment failed to find an effect of multiple presentations either on the rate with which /s/ was judged to be correctly produced or on intrarater reliability of binary correct/incorrect judgments of children's production of /s/. The use of the stimuli containing synthetic versions of /s/ allowed us to know a priori which of the stimuli contained correctly produced /s/, incorrectly produced /s/, or a variant of /s/ intermediate between the two. As expected, people rated the three types of synthetic /s/ differently. More interestingly, however, was the influence of the type of /s/ on intrarater reliability. The participants were more consistent in their ratings of correct /s/ than in their ratings of intermediate and incorrect /s/. If this were to translate to clinical practice, then we would anticipate that there would be greater reliability in judgments of the accuracy of children's speech production relative to judgments of its inaccuracy. This would suggest that the problems introduced by poor intrarater reliability might lead to more false-negative diagnoses of PI (due to the relative instability of the judgments of inaccurate speech sounds) rather than false-positive diagnoses (due to the relative stability of judgments of children's speech production accuracy).

Unlike Experiment 1, Experiment 2 found an effect of multiple presentations on interrater reliability. Greater consensus across the 18 participants was found in the multiple-presentation condition than in the single-presentation condition. The mean consensus was 70.8% in the

single-presentation condition and 79.8% in the multiple-presentation condition. We regard this 9% improvement to be clinically significant and to argue persuasively for the use of multiple presentations when more than one person is assessing phonetic accuracy of children's speech. This interacted with the type of /s/ that was being rated. Tokens of correct /s/ were uniformly rated as accurate across the two presentation conditions. There was more agreement in the accuracy of tokens of /s/ that were either incorrect or intermediate between correct and incorrect productions in the multiple-presentation condition than in the single-presentation condition. Recall that these were the two types of /s/ that showed the lowest intrarater reliability. Taken together, the results suggest that judgments of the accuracy of incorrect and intermediate /s/ are the least stable within individuals, but that greater consensus across individuals can be achieved when multiple presentations are used.

General Discussion

Summary

Two experiments investigated the influence of multiple presentations on the reliability of judgments of children's speech production and the rate with which sounds were judged to have been produced accurately. Experiment 1 used phonetic transcription and found no systematic influence of multiple presentations on intratranscriber reliability. Intertranscriber reliability was measured for a subset of the stimuli. Again, there was no systematic influence of multiple presentations on intertranscriber reliability. Experiment 2 utilized binary correct/incorrect judgments of the accuracy of /s/ production. There was no systematic influence of multiple presentations on rate of correct judgments or on intrarater reliability. There was, however, an effect of this variable on interrater reliability. Greater consensus on the judgments of the accuracy of incorrect /s/ and intermediate /s/ was achieved when multiple presentations were used.

One salient finding of the study was that the mean intrarater reliability, while not unprecedented (e.g., Shriberg & Lof, 1991), was undesirably low in both experiments. Mean intratranscriber reliability in Experiment 1 was 82.3%. Collapsed across the conditions, values ranged from 67.4% to 89.3%. In Experiment 2, the mean was only slightly higher, at 85.5%. Individual values ranged from 73.7% to 100%, and 14 of the 18 participants had mean intrarater reliability below 90%. Two factors may have affected these values. One factor that may have inflated them was the use of a concurrent orthographic display during the transcription/rating tasks. Previous research (Oller & Eilers, 1975) showed that expectations about how words are produced affect phonetic transcription. The orthographic display may have biased the listeners to transcribe or rate things as correct when they might not have in a task without such a display. Our choice to use the concurrent orthographic display was driven by our desire to maintain ecological validity: In real-world transcriptions of single-word naming tasks, examiners generally know the target words. The second factor may have decreased reliability. McNutt, Wicki, and Paulsen (1991) showed that transcriptions are less reliable when done in audio-only conditions than when done with concurrent audio and visual displays. The use of an audio-only display in this experiment may have decreased reliability.

One factor that qualifies the results of Experiments 1 and 2 concerns the use of a 250-ms delay interval. One reasonable response to the result that multiple presentations did not have an effect on intrarater reliability might be to assert that the 250-ms delay interval between presentations was not long enough for listeners to establish a percept of the word. This is unlikely, given research showing that listeners typically establish percepts of words in advance of their acoustic offsets (e.g., Collison, Munson, & Carney, 2004). A more potent criticism of the use of a 250-ms delay interval is that it limits the ecological validity of the study. Clinicians who rate children's speech from tape-recorded samples must stop the tape, pause, rewind, and play between presentations; this process rarely can be completed in 250 ms. However, as discussed below, this criticism would not apply to situations in which children are rated by people using digital audio players and digitized speech samples.

Implications and Future Research

Across the two experiments, no effect of multiple presentations was found on measures of intrarater reliability of measures of children's phonetic accuracy. In addition, Experiment 2 found no effect of multiple presentations on the rate with which people judged /s/ to be accurate or inaccurate. In contrast, Experiment 2 found that interrater reliability was facilitated in the multiple-presentation condition. The clear recommendation that emerges from this study is that clinicians who strive to have greater interrater reliability should use multiple presentations, particularly when transcribing or judging phonemes whose accuracy is unclear. Although this recommendation could not be implemented in live-voice administration of picture-naming assessments of phonology, it would be easy to implement if assessments were recorded. The hardware needed to record assessments digitally is relatively inexpensive; a minimum setup would include a microphone attached to a computer equipped with a sound card. There are many free, downloadable software packages that can be used to listen to and edit digital speech files (e.g., *Praat*). The use of digitized speech would allow raters to listen to multiple presentations of speech tokens rapidly. This would obviate the criticism that the 250-ms delay interval used in Experiments 1 and 2 was too short to have ecological validity.

Another recommendation resulting from this study concerns the use of phonetic transcription in research. For example, this finding has implications for research in normal phonological development. People working in that area should use multiple presentations to achieve greater interrater consensus. This recommendation would also apply to people working in the development of large, phonetically transcribed corpora of children's speech. Recently, a number of large corpora of spoken language

have been created to assist in the development of speech technology applications, such as automatic speech recognition. In addition to having industry applications, large corpora have been shown to be potentially valuable tools to study phonological patterns in naturally occurring language (e.g., Bell et al., 2003; Greenberg, Hollenback, & Ellis, 1996). The phonological characteristics of the words in these corpora show patterns of reduction and variation that are characteristic of words spoken in conversation and that are very unlike citation forms of the same words. Not surprisingly, the transcription of such databases has proven to be a challenge, given the variability in pronunciation that occurs in conversational speech. Future research in this area should utilize multiple repetitions to achieve higher intratranscriber reliability.

The original goal of this research was to determine whether multiple presentations influence the reliability of judgments of children's speech-production accuracy. The study sought to determine whether multiple presentations influence the reliability of clinical assessments of speech-sound disorders in children and to gauge whether clinical recommendations can be made to improve the reliability of transcriptions. Ultimately, research that examines methods for maximizing the reliability of assessments should include two parallel lines of inquiry. The first line of inquiry is illustrated by studies like this one that examine reliability in a controlled, experimental setting. The second line of inquiry should examine how assessments are done in real-world settings. There exists a large body of research showing that problem-solving behaviors (such as those that occur during real-world assessments of speech and language) are very sensitive to the resources that exist in different environments (e.g., see Clark, 1997, chap. 3 for a review). An individual who adopts one problem-solving strategy in a given context (i.e., such as using multiple repetitions to achieve high interrater reliability in a laboratory study such as this one) might use a different strategy to maintain the same level of reliability in a setting in which that resource is not available (i.e., a real-world clinical assessment utilizing live voice only). Research on factors that affect transcription accuracy and reliability in naturalistic settings could provide a powerful complementary line of research to the one illustrated in this study. Together, both lines of inquiry will provide a clearer picture of methods for maximizing the reliability and accuracy of assessments of speech and language.

Acknowledgments

This research was supported by a Faculty Summer Research Fellowship from the University of Minnesota Graduate School to the first author to conduct Experiment 1, and by a University of Minnesota Undergraduate Research Opportunity Program grant to the second author to conduct Experiment 2. We thank Keith Johnson and Liz Strand for sharing the Klatt Synthesizer parameter files that were used to create the stimuli for Experiment 2. The results of Experiment 1 were presented at the 2003 Symposium for Research on Child Language Disorders in Madison, WI. We thank audiences at that conference for useful input. We are especially grateful to Rebecca Herman for providing valuable insight into the design of Experiment 2, particularly regarding the number of repetitions in our multiple-presentation condition. We also thank the children who produced the stimuli, and the adults who participated in the experiments.

References

Bankson, N., & Bernthal, J. (1990). *Bankson-Bernthal Test of Phonology*. Austin, TX: Pro-Ed.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America, 113*, 1001–1024.

Boersma, P., & Weenink, D. (2004). Praat v. 4.1.7 (Computer software). Amsterdam: Institute of Phonetic Sciences.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Collison, E. A., Munson, B., & Carney, A. E. (2004). Relations among linguistic and cognitive skill and spoken word recognition in adults with cochlear implants. *Journal of Speech, Language, and Hearing Research, 47,* 496–508.

Cucchiarini, C. (1996). Assessing transcription agreement: Methodological aspects. *Clinical Linguistics and Phonetics, 10*, 131–155.

Flipsen, P. (1995). Speaker-listener familiarity: Parents as judges of delayed speech intelligibility. *Journal of Communication Disorders, 28,* 3–19.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, P. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America, 84*, 115–123.

Fudala, J. (2001). *The Arizona Articulation Proficiency Scale–Third Edition*. Los Angeles: Western Psychological Services

Gierut, J., & Dinnsen, D. (1986). On word-initial voicing: Converging sources of evidence in phonologically disordered speech. *Language and Speech, 29*, 29–114

Goldman, R., & Fristoe, M. (2000). *The Goldman-Fristoe Test of Articulation–Second Edition*. Circle Pines, MN: American Guidance Service.

Greenberg, S., Hollenback, J., & Ellis, D. (1996). Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In *Proceedings of the International Conference on Spoken Language Processing '96*. Philadelphia: Author.

Hedrick, M., & Carney, A. (1997). Effect of relative amplitude and formant transitions on perception of place of articulation by adult listeners with cochlear implants. *Journal of Speech, Language, and Hearing Research, 40*, 1445–1457.

Holt, R. A. F. (2003). *Non-sensory factors in children's speech perception*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Ingrisano, D., Klee, T., & Binger, C. (1996). Linguistic context effects on transcription. In T. W. Powell (Ed.), *Pathologies of speech and language: Contributions of clinical phonetics and linguistics* (pp. 101–108). New Orleans, LA: International Clinical Linguistics and Phonetics Association.

Kahn, L., & Lewis, N. (2002). *Kahn-Lewis Phonological Analysis–Second Edition*. Circle Pines, MN: American Guidance Service.

Kaminska, Z., Pool, M., & Mayer, P. (2000). Verbal transformation: Habituation or spreading activation*? Brain and Language, 71*, 285–298.

Kent, R. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*, 7–23.

Klatt, D. (1980). Software for a cascade/parallel formant

synthesizer. *Journal of the Acoustical Society of America, 67*, 971–995.

Kluender, K. (1994). Speech perception as a tractable problem in cognitive science. In M. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 173–217). San Diego, CA: Academic Press.

Leske, M. (1981). Prevalence estimate of communicative disorders in the U.S.: Speech disorders. *Asha, 23*, 217–225.

Lewis, K. (1994). Reporting observer agreement on stuttering event judgments: A survey and evaluation of current practice. *Journal of Fluency Disorders, 19*, 269–284.

MacKay, D., Wulf, G., Ying, C., & Abrams, L. (1993). Relations between word perception and production: New theory and data on the verbal transformation effect. *Journal of Memory and Language, 32*, 624–646.

MacMillan, N., & Creelman, C. (1991). *Detection theory: A user's guide*. Cambridge, England: Cambridge University Press.

McNutt, J., Wicki, L., & Paulsen, J. (1991). Judgments of phoneme errors under four modes of audio-visual presentation. *Journal of Speech-Language Pathology and Audiology, 15*, 37–42.

Munson, B. (2004). Variability in children's and adults' productions of /s/: Evidence from dynamic measures of spectral mean. *Journal of Speech, Language, and Hearing Research, 47*, 69–80.

Munson, B., Swenson, C., & Manthei, S. (in press). Lexical and phonological organization in children: Evidence from real-word and nonword repetition. *Journal of Speech, Language, and Hearing Research*.

Oller, D., & Eilers, R. (1975). Phonetic expectation and transcription validity. *Phonetica, 31,* 288–304

Qi, Y., & Johnson, K. (1987). *KLSYN: A formant synthesis program*. Unpublished manuscript, Ohio State University, Columbus.

Ross, M., & Lerman, J. (1979). A picture identification test for hearing impaired children. *Journal of Speech and Hearing Research, 13*, 44–53.

Scobbie, J., Gibbon, F., Hardcastle, W., & Fletcher, P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V* (pp. 194–207). Cambridge, England: Cambridge University Press.

Shoaf, L., & Pitt, M. (2002). Does node stability underlie the verbal transformation effect? A test of node structure theory. *Perception & Psychophysics, 64*, 795–803.

Shriberg, L., Kwiatkowski, J., & Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research, 27*, 456–465.

Shriberg, L., & Lof, G. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics, 5*, 225–279.

Smit, A. B., Freilinger, J. J., Bernthal, J. E., Hand, L., & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55*, 779–798.

Smit, A. B., & Hand, L. (1997). *Smit-Hand Articulation and Phonology Examination*. Los Angeles: Western Psychological Services.

Strand, E. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology, 18*, 86–99.

Viemeister, N. V., & Wakefield, G. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America, 90*, 858–865.

Whitmire, K., Karr, S., & Mullen, R. (2000). Action: School services. *Language, Speech, and Hearing Services in Schools, 31*, 402–406.

Williams, K. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Received May 4, 2004
Accepted August 7, 2004
DOI: 10.1044/1058-0360(2004/034)

Contact author: Benjamin Munson, Department of Speech-Language-Hearing Sciences, University of Minnesota, 115 Shevlin Hall, 164 Pillsbury Drive SE, Minneapolis, MN 55455. E-mail: munso005@umn.edu

Appendix
Words Used in the Recognition Memory Task

Bad	Cup	Girl	Name
Ball	Dad	Hat	Pool
Barn	Fish	Knee	Race
Bear	Fly	Mad	Ship
Corn	Foot	Mouse	Shirt