INTRODUCTION

Treating communication disorders
There are a number of individuals who present with different communication disorders. Generally, the communication deficits of the client are remediated through the implementation of evidence-based treatment methodologies [1]. The speech-language pathologist (SLP) follows a management regimen, which involves the collection of baseline data to quantify the disorder, implementation of a treatment, and measurement of the treatment to document a change in the communication behavior of interest. During the presentation of the treatment, client responses are elicited, which are subject to some form of accuracy information from the SLP. That is, clients receive external feedback regarding their clinical performance.

The feedback may be in the form of acoustic, physiologic, or perceptual information, but there is a lack of isomorphism among the study methods [2,3]. Perceptual judgement is generally the most frequently used in assessing a client’s articulatory responses over time may result in auditory perceptual confusions. This clinical hypothesis was studied by examining the integrity of judgement accuracy of repeated /r/ stimuli, which varied as a function of correct and incorrect stimuli. Findings showed no statistically significant evidence of auditory perceptual confusions when subjects listened to and evaluated repeated productions of synthesized /r/, /w/ for /r/ substitution, and /r/ distortion embedded in a CV word.

Keywords: Perception, Verbal transformation effect, Phonology, Misarticulations
ever, the use of auditory perceptual judgments can be prone to different perceptual constraints as discussed by different investigators [6].

**Issues in auditory perceptual judgments**

While this measurement modality, which is often considered the “gold standard” of assessment/treatment, can be economical and convenient since it requires no instrumentation, perceptual measures do have limitations in precision due to their subjective nature. Kent [6] wrote that SLPs’ perceptual judgments are grounded in the following assumptions:

(a) SLPs have a common understanding of perceptual labels such as hoarse, nasal, rough, monoloud, excess and equal stress or stutter;

(b) SLPs use essentially the same verbal descriptors and associated scale values to assess a given sample of speech or voice;

(c) SLPs can isolate for judgment one perceptual dimension from several co-occurring dimensions;

(d) SLPs have a uniform reliability in judging the various dimensions that give a complete clinical portrait of speech or voice disorders; and

(e) SLPs can make perceptual judgments for which the interjudge differences are smaller than the differences needed for clinical classification or to discern changes in clinical status. (p. 7).

The issues identified above are important in developing a perspective for understanding the value and limitations of making auditory perceptual judgments. The limitations are what Kent [6] termed potential sources of perceptual inaccuracy. In a review of the literature, he identified nine variables that may adversely affect auditory-perceptual judgments in the research and clinical domains. The sources include phonemic restoration, misperceptions of natural fluent speech, phonemic false evaluation, effects of lexical status on phonetic categorization, equivalence classes in phonetic perception, McGurk Effect, prosodic influences on phonetic classification, talker-listener-utterance interactions, and the verbal transformation effect. The verbal transformation effect is of particular interest, since it involves the repeated listening and evaluation of stimuli and reported change in perceptual judgment by listeners [7].

**Verbal transformation effect**

The verbal transformation effect occurs when an identical auditory stimulus is repeatedly heard but is perceived by the listener to be phonetically altered [7,8]. Initially, a percept matching of the original prime stimulus is perceived, but at some time, erroneous words or nonwords are identified, which signal a change in the perceived stimulus. The transformation effect continues throughout the presentation of the stimuli, leading to perceptual shifts from one presentation to another or a return to the original form. A study reported by Sato et al. [8], found that perceptual alterations of an original stimulus could range from small phonetic changes to semantic confusions.

Different explanations have been provided to explain the verbal transformation effect, and they include both linguistic and perceptual processing explanations. Some hypothesize that the verbal transformation effect occurs when lexical activation of a target word extends to phonologically or semantically related words in the person’s lexicon. The related words then compete with the presented stimuli as potential perceptual responses. An alternative explanation is that the verbal transformation effect occurs when the peripheral or central processing structures that are triggered in response to the target stimuli are fatigued by repeated stimulation [6].

The direct investigation of the verbal transformation effect as it applies to the continuous judgment of children’s misarticulations has not been studied formally, but there are some investigations that have examined judgments of repeated similar misarticulated stimuli. For example, Shriberg [9] assessed the inter-reliability and intra-reliability of articulatory judgments among five experienced SLPs. The SLPs judged audio tapes of four children with nonstandard /r/ and four with deviant /s/. Results suggested that judgements of postvocalic /t/ items showed the most judgement variability. The findings were the following: listeners shifted their judgments when listening to different but repeated misarticulations, and additionally, there was a greater likelihood for reliability to be better when there was a large percentage of the speech stimuli that was perceived as either correct or incorrect.

Additional work relating to perceptual judgements of /t/ in repeated tokens was investigated in a cue-trading task reported by Wolfe et al. [10]. The authors used a /r-w/ cue-trading task to examine the perceptual abilities of graduate students with and without clinical experience. Vowel formants were altered by balancing a temporal-spectral cue on F2 against a spectral cue on F3. Subjects listened to the different stimuli and were required to identify the tokens as containing the /r/ or /w/ speech sounds. The participating speech-language pathology graduate students who had previous clinical
experience exhibited better perceptual sensitivity in determining when a sound was more perceptually similar to a canonical /r/ or /w/ than graduate students who did not have exposure to the clinical setting. The results of this preliminary study indicated that a task utilizing a cue trading paradigm could be useful in evaluating perceptual sensitivity to the acoustic cues representative of misarticulated speech sounds.

A final study conducted by Munson & Brinkman [11] also involved listening to repeated stimuli but the investigators were studying the reliability of perceptual judgments in the administration of an articulation test. That is, they had subjects carry out transcriptions across multiple sampling periods to determine if such an experimental task posed a threat to the reliability of clinical perceptual judgements. The experimental group transcribed words produced by children with speech sound disorders. The listening conditions included a single presentation of the word and seven presentations of the word. Their data consisted of transcriptions of /s/ across various word tokens. Transcribers participated in two separate transcription sessions, with a week between the two sessions. The experimenters found no systematic effect of presentation condition on either accuracy of judgments or intra-rater reliability. In addition, higher interrater reliability was found in the multiple-presentation condition, especially for /s/ tokens that were inaccurate or acoustically intermediate between correct and incorrect words. The data reported imply that multiple presentations of the stimuli had no measurable effect on correct judgments and intra-rater reliability of children’s phonetic accuracy, but they did have a minimal effect on inter-rater reliability.

The studies discussed did not specifically set out to examine the verbal transformation effect but the use of repeated approximate stimuli conditions that could elicit the perceptual illusion are consistent with the study of the effect. It is possible that day to day evaluations of articulatory responses may in fact trigger the verbal transformation effect due to the sheer volume of responses that are assessed by some SLPs. For example, SLPs working in schools have substantial caseloads of children with various misarticulations and repeatedly listen and evaluate the responses of children who present with different errors. This potential issue may pose a threat to the reliability of clinical judgements, which constitute the data for making decisions regarding continuation and/or dismissal from a treatment program. If the verbal transformation effect is a clinical reality, SLPs need to be aware of such a problem and identify ways to minimize the problem. This investigation examined listeners’ ability to make reliable perceptual judgments when presented with repeated speech stimuli.

**METHODS**

**Subjects**

Subjects for this study included 30 graduate students from West Virginia University who were enrolled in the Communication Sciences and Disorders master’s degree program. The subjects consisted of a total of twenty-nine females and one male with a mean age of 23 years, 3 months. To qualify for the investigation subjects must have: (1) completed at least one semester of graduate coursework including a course in speech sound disorders, (2) provided speech treatment to at least one client with a speech sound disorder, and (3) passed screenings of speech, hearing, and auditory perception. The nature of the study was explained to the participants, and a written IRB consent form was obtained.

For each participant, the examiners conducted the preliminary screening procedures (Qualifying Criteria) and administered the experimental stimuli (Listening Task) in one session. The following screening procedures were administered:

1. Pure tone hearing screening of 25 dB at 500 Hz, 1,000 Hz, 2,000 Hz, 4,000 Hz;
2. Articulation screening;
3. Passing a modified variation of the Speech Production Perception Test for /r/ [12].

Subjects were required to pass all screening measures in order to participate in the investigation. Hearing was screened with a Grason-Stadler GSI 17 portable audiometer and speech was assessed through administrations of the Rainbow Passage [13]. The examiners requested that the experimental subjects read the passage to screen articulation skills. A modified version of the Speech Production Perception Test was also administered by the experimenter [12]. This test examined a subject’s perception of /r/ versus substitution of /w/ and a control sound /l/, which shares phonetic features with /r/.

Subjects were required to make perceptual judgments of a target item, which contained /r/. The testing procedure required the experimental subjects to listen and view a picture of the target item (rake). Six random presentations were accompanied by appropriate articulation of /r/ (Is this a rake?), 6 consisted of a /w/ production (Is this a wake?) and 6 consisted of an /l/ production (Is this a lake?). The derhotocized or distorted /r/ was not included in the assessment, because it is not a phoneme of English and would not be consistent with
the test format of assessing a target sound versus its English sound substitutions.

Stimuli
The speech sound stimulus selected for study was the liquid /r/. The rationale for selection was that the speech sound is frequently misarticulated by English-speaking children [14]. In addition, there are two error variants that are described in the literature, which include a substitution of /w/ in place of /r/ and a distortion of /r/ that is classified as a derhoticized version [15-17]. According to Ohde & Sharf [16], /r/ is a speech sound that has rhotic qualities or /r/-coloring, which is the major feature of the auditory percept of /r/. Its lingual production feature is generally described as either bunched or retroflex, although some linguists describe the tongue position as being variable among speakers. The derhoticized or distorted /r/ is a substitution that is devoid of an /r/-quality production due to a deviant lingual articulation [16]. The descriptions of the production vary by author to include excessive lip rounding, a lack of lingual bunching or retroflexion, and lingual positioning that is either too forward or too far posterior in the oral cavity. It has been argued that substitution of /w/ in place of /r/ is not the same production as the /w/ used in an appropriate context (i.e. /ret/—> /wet/) but most literature does identify the error as one of a sound substitution in place of /r/ [14].

Based on the experiments of Sharf et al. [18], synthetic CV stimuli of /r/, /w/, and derhoticized /r/ with the midfront vowel /e/ were created for this current study using the Pratt 6.0.21 articulatory synthesis software. The development of the stimuli was carried out by a linguist with expertise in articulatory synthesis and with assistance of one of the examiners. The tokens were then presented to three certified SLPs with a total of 105 years of clinical experience. Each listened individually and were instructed to identify the /r/ stimuli, the /w/ stimuli, and the distorted /r/. A total of 12 stimuli, 4 of each synthetic CV, were played randomly to each of the SLPs. Each identified the stimuli reliably with a total of 3 selection errors among the 36 presentations of the stimuli.

The stimulus tokens were then electronically copied into sets of 30 stimuli at a presentation level of 65 dB SPL using cross-platform software for recording and editing auditory stimuli. There was an inter-stimulus pause of 2.5 seconds between each token. Upon completion of the editing, there were 3 sets of 30 stimuli from each of the synthesized stimuli. That is, there was a set of 30 repeated stimuli of the synthesized stimulus ray (/re/) designated R, 30 repeated stimuli of the synthesized stimulus way (/we/) designated W, and 30 repeated stimuli of derhoticized /r/ (/re/) designated DR. In total, there were 90 tokens, 30 from each of the three synthesized tokens. The repeated stimuli were then downloaded to a MacBook Pro.

Procedure
The 3 stimuli sets (R, W, DR) were presented to all subjects, with order of presentation counterbalanced across three subgroups of subjects, each of which contained 10 randomly assigned subjects from the group of 30 subjects. The rationale for such a design feature was to minimize order effects as a confounding factor, and it allowed the experimenter to control and measure sequencing effects [1]. Participants who were presented with presentation order 1 (Group I) were first presented with 30 repeated tokens of the CV syllable beginning with the derhoticized /r/ ‘DR’, followed by 30 repeated presentations of the CV syllable beginning with the /w/ ‘W’, and then finally 30 repeated presentations of the CV syllable beginning with the accurately produced /r/ ‘R’. The presentation order for Group II was W, R, and DR, and the ordering for Group III was R, DR, and W.

The listening task was administered via Beyerdynamic DT 211 headphones through a MacBook Pro. Subjects were instructed that they were going to listen to some words that contained the /r/ speech sound and that they were going to judge each item as either right or wrong. After listening to each word, they were required to mark a score sheet. The stimuli were presented diotically at 65 dB SPL. Participants were not provided information regarding the number of trials nor their accuracy judgments for individual trials.

RESULTS
Descriptive statistics
The mean percentage values and the standard deviations for the three subgroups across the three listening conditions are presented in Table 1. The scores show that the subjects exhibited a high level of response accuracy for each of the different test tokens. In most cases mean values are at 80% or higher indicating that they were generally correct in their assessment of the different /r/ tokens. The only difference to this trend is Group I when judging the /w/ substitution for /r/ token. Some of the subjects experienced difficulty with the judgment task, which resulted in a mean score of 69%; however, this decre-
A study of the judgment accuracy in performance was not observed in the other groups for that test token or the other test tokens. Standard deviations show a variable pattern with both limited and substantial dispersion from the mean scores. Those standard deviations as associated with substantial dispersion are reflective of some limited subject outliers across the three listening groups. That is, most subjects displayed a high level of judgment accuracy, but some listeners in each group did not and this is reflected in the large standard deviations for some of the test tokens. For instance, Group III listeners demonstrated a mean score value of 82 and a standard deviation of 36 when they judged the /r/ distortion token (DR). Only 50% of the subjects in Group III were accurate in their judgments, which is reflected in the sizable standard deviation. Again, there was no definite trend for a listening group, but /w/ for /r/ (W) substitution showed the most variability across the three listening subgroups.

Figure 1 is a graphic representation of the results described. Note upon inspection that the tokens beginning with the accurately produced /r/ (R) and /r/ (DR) distortion appear to have the overall highest judged correctness, while the tokens with the /w/ for /r/ (W) substitution appear to have the lowest percent correctness. However, despite the lower values listeners also exhibited a high level of judgment accuracy for the /w/ for /r/ (W) token.

### Table 1. Mean values and standard deviations for Groups across the different test tokens

<table>
<thead>
<tr>
<th>Groups</th>
<th>Group I</th>
<th>Group II</th>
<th>Group III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening Token</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/r/ (R)</td>
<td>$\bar{x} = 100$</td>
<td>$\bar{x} = 81$</td>
<td>$\bar{x} = 100$</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 1$</td>
<td>$\sigma = 33$</td>
<td>$\sigma = 2$</td>
</tr>
<tr>
<td>/r/ → /w/ (W)</td>
<td>$\bar{x} = 69$</td>
<td>$\bar{x} = 89$</td>
<td>$\bar{x} = 87$</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 47$</td>
<td>$\sigma = 27$</td>
<td>$\sigma = 18$</td>
</tr>
<tr>
<td>/r/ Distortion (DR)</td>
<td>$\bar{x} = 98$</td>
<td>$\bar{x} = 99$</td>
<td>$\bar{x} = 82$</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 4$</td>
<td>$\sigma = 2$</td>
<td>$\sigma = 36$</td>
</tr>
</tbody>
</table>

### Summary of results

Most of the subjects exhibited a high percentage of correctness when judging the accuracy of the different test tokens. There were some subjects in the three different subgroups who did not, resulting in a limited number of outliers. Upon inspection of the results respective to the different test tokens, there were no statistically significant differences. In addition, the findings did not show a significant difference between the first half of the presentations of a test token compared to the first fifteen presentations of a token versus the last fifteen presentations of the same token (referred to in this paper as split-half). In addition, there was no significant interaction effect between test token and split-half, since the listeners did not perform differently on the first fifteen presentations compared to the last fifteen presentations of a token as a function of condition, $F(2, 58) = 0.512, p = 0.602, \eta^2 = 0.017$.

### Inferential statistics

Inferential analysis was conducted to determine if there were significant differences among groups with respect to the different listening conditions. The first analysis consisted of the computation of a one-way analysis of variance (ANOVA) using SPSS statistical software to determine whether there were significant differences across the different tokens. The three different tokens or levels of the independent variable were correct /r/ (R), /w/ for /r/ substitution (W), and distorted /r/ (DR), and the dependent measure was the judgment accuracy. There were no statistically significant differences among the different tokens, $F(2, 58) = 1.941, p = 0.153, \eta^2 = 0.063$.

To examine possible accuracy judgments shifts within a listening condition, a $2 \times 3$ repeated measures ANOVA was computed. The analysis was done to test for differences between the first half of the trials within a specific token listening condition compared to the second half of the trials across the three-listening conditions. It was hypothesized that accuracy judgments for a token may change as subjects judged individual presentations of that particular token. The main effect of listening condition was not statistically significant $F(2, 58) = 2.47, p = 0.093, \eta^2 = 0.079$. That is, there was no significant main effect for the first fifteen presentations of a token versus the last fifteen presentations of the same token (referred to in this paper as split-half).
second half. That is, there did not appear to be a perceptual shift in the judgment accuracy of the different tokens with repeated presentations.

**DISCUSSION AND CONCLUSIONS**

**Study findings**
Statistical testing indicated no significant differences among groups when subjects evaluated the different test tokens. Moreover, the majority of subjects performed at a very high level of response accuracy. It was hypothesized that judgment shifts would occur across the stimuli with repeated listening conditions that are similar to an actual clinical situation. This was based on clinical reports of judgment error when engaging in actual clinic treatments that utilized continuous feedback given by an SLP. That is, SLPs have reported that they are unable to discriminate auditorily among correct target sounds and misarticulations in some situations when required to listen and evaluate repeated target stimuli. In order to make the task as difficult as possible, we used /r/ stimuli because the error variants include both phonemic (sound substitution) and nonphonemic (/r/ distortion) misarticulations. Additionally, /r/ is frequently in error and proven to be a difficult remediation target, so we hypothesized that it would be challenging for the study participants [4].

We had proposed that the shifts in judgment that would be identified would be theoretically consistent with the verbal transformation effect, which has consistently demonstrated shifts in judgment when individuals repeatedly listen to various linguistic stimuli [7]. This would suggest that judgments of correct and incorrect stimuli would show the same results as Warren’s subjects who listened to stimuli that did not include speech sound errors. However, our findings are not consistent with previous clinical reports and certainly do not support performance explained by the verbal transformation effect. However, the current findings do not negate the clinical hypothesis, but rather suggest the need for further study.

**Extension of the current research**
It is to be noted that this study is the first that we are aware of that empirically examined the experimental question in the theoretical context of the verbal transformation effect. The findings are preliminary in nature and the trends toward statistical significance indicate the need for further study. The early investigation that was reported by Shriberg [9] did report judgment shifts in perceptual accuracy, but that was a clinical study which included 5 experienced judges. His findings indicated that judgements of postvocalic /r/ items had the most varied judgements of accuracy. He also found that listeners shifted their judgments when listening to different but repeated misarticulations. These variables should be considered in an extension of the current study. For instance, an alteration in sound position such as presenting stimuli in postvocalic position might be implemented along with an increase in the number of stimuli presented. Each of the current stimuli involved 30 presentations and it may be that an increase of 75 to 90 stimuli of each would trigger accuracy judgment changes. In actual clinical situations the SLP is evaluating a number of different clients with speech sound disorders and judging numeruous target productions. The employ of added stimuli would be consistent with typical clinical practice and may allow for a finer grained assessment of a practitioner’s judgment skills to determine if perceptual shifts are a clinical reality.

**Clinical implications**
There are several clinical implications of the study that are relevant to both students and practitioners. First, the vast majority of our student subjects were accurate in their judgments of the different stimuli. They all had clinical experience but not a vast amount of clinical experience when compared to a practitioner. Their performance certainly indicates that they were capable of completing the listening tasks successfully; however, there was a small subgroup who did experience problems and could probably benefit from additional perceptual training in graduate school. Educational programs in the major might want to include additional listening experiences for student clinicians who demonstrate difficulties in the perceptual judgment of correct and incorrect articulatory responses [19]. This could also be extended to practitioners in a variety of ways that might include continuing education credit for participating in a refresher activity to re-calibrate with colleagues. Such experiences can assist those SLPs who must work in such venues as schools and preschool clinics. They are exposed to a variety of speech sound disorders and may vary in their assessment of different clients on certain days or in relation to a specific speech sound error. These are important concerns since their perceptual judgments are the basis for who will continue treatment or be dismissed.

**CONCLUSIONS**
Generally, subjects performed at a high level of accuracy in
the judgment of /r/ (R), /w/ for /r/ substitution (W), and distorted /r/ (DR). There were only six subjects (20%) of the 30 participants who experienced some difficulty with the listening tasks, and there was a minor trend of poorer performance associated with /w/ for /r/ substitution (W). This was a preliminary investigation and further study is needed to verify the clinical hypothesis that listening to repeated stimuli may cause changes in perceptual judgment. Future SLPs and current practitioners need to be cognizant of this potential problem.

REFERENCES