December 2015

Effects of Ultrasound as Visual Feedback of the Tongue on Generalization, Retention, and Acquisition in Speech Therapy for Rhotics

Greta Sjolie
Syracuse University

Abstract

Purpose: The purpose of this study was to provide an initial comparison of exposure to ultrasound visual feedback of the tongue and no exposure to ultrasound in speech therapy for postvocalic rhotics (the /r/ family of sounds). Effects of the two treatments on acquisition, retention, and generalization were explored in participants ages 7-9.

Methods: A single-subject randomized block design replicated across four participants was used. Each week for seven weeks, one session containing high frequency ultrasound use and one session containing no ultrasound use were randomly ordered. A Training Probe List and Generalization Probe List consisting of monosyllabic words, multisyllabic words, phrases, and short sentences were used to measure acquisition within each session as well as retention and generalization between two consecutive sessions. Data analyses included: (a) descriptive statistics to complement visual inspections of single-subject graphs, (b) effect size calculation, and (c) statistical results from a randomization test.

Results: One participant showed a significant advantage for ultrasound sessions over no ultrasound sessions in acquisition scores; however, there were no differences between treatment conditions for any participants in generalization or retention.

Conclusion: For some children, acquisition may be facilitated by ultrasound visual feedback. No evidence suggested that ultrasound visual feedback inhibited retention or generalization in speech tasks. As a whole, treatment was effective for 2 of the 4 participants when comparing pre/post generalization data. Future studies should focus on evaluating the effectiveness of ultrasound visual feedback therapy given a larger dose (i.e., treatment duration) and differing age groups.

Effects of Ultrasound as Visual Feedback of the Tongue on Generalization, Retention, and Acquisition in Speech Therapy for Rhotics

By

Greta M. Sjolie

B. A. Simpson College

MASTER'S THESIS

Submitted in partial fulfillment of the requirements for the degree of

Master of Science in Speech-Language Pathology

Department of Communication Sciences and Disorders

Syracuse University

December, 2015

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to all the families that participated in this research, to those who referred children, and to my superiors and friends who offered their encouragement and academic support. Further, I would like to express my deepest gratitude to my advisor, Dr. Jonathan Preston, for agreeing to undertake this considerable project with me and for his excellent guidance, teaching, and patience along the way for without his continued support, my research would not have been possible. Also, thanks in particular to Megan Leece for your energetic supervision during hours of rhotic therapy and your vital feedback, support, and words of encouragement that have helped shape my clinical skills. Furthermore, I am most grateful for the comments and feedback from my committee members: Dr. Jonathan Preston, Dr. Mary Louise Edwards, Dr. Tanya Eckert, and Dr. Ellyn Riley. Thanks in particular to graduate student Olivia Flaherty for assisting with reliability and fidelity checks. Additionally, special thanks to graduate students Monique Uy and Kerry McNamara for assisting with reliability. This research was supported by NIH grant R03DC012152.

TABLE OF CONTENTS

CHAPTERS:

FIGURES:

TABLES:

APPENDICES:

Introduction

Speech Sound Disorders

A speech sound disorder (SSD) is a disorder in producing the sounds of a language (Bernthal, Bankson, & Flipsen Jr., 2009; Lewis et al., 2006; Shriberg, Tomblin, & McSweeny, 1999). The term SSD encompasses articulation disorders and phonological disorders, though it provides a classification that is free of theoretical biases (Shriberg, 2010). SSDs are one of the most common communication impairments seen in children (Baker & McLeod, 2011; Bernthal et al., 2009; Shriberg et al., 1999) and a large number of speech pathologists in school based settings work with children who demonstrate difficulty in the acquisition of speech sounds (Ruscello, 1995a). Specifically, rhotics (i.e., /r/, the r-family of sounds, including rhotic diphthongs) are among the most common sounds seen in error in preschool and school-age children with SSDs (Bernthal et al., 2009; Hodson & Padden, 1991; Ruscello, 1995a; Secord, Boyce, Donahue, Fox, & Shine, 2007)[1]. Children who have a SSD are at risk for developing future academic and socioemotional difficulties; thus, it is important for speech pathologists to use therapy techniques that have been shown to be effective when treating SSDs (Aram & Hall, 1989; Lewis, Freebairn, & Taylor, 2000; McCormack, McLeod, McAllister, & Harrison, 2009; Nathan, Stackhouse, Goulandris, & Snowling, 2004; Shriberg & Kwiatkowski, 1988).

The term Speech Delay has been used to characterize three to nine year-old children with significant speech sound deletions and substitutions that may normalize with treatment (Shriberg, 2010). Shriberg (1997) found that by age six, 75% of children with a history of a SSD have normalized their speech errors; however, 25% of children with a SSD will still have speech errors past the age of six. Although some variation occurs, it is generally accepted that the

[1] Instead of the IPA symbol /ɹ /, /r/ will be used throughout this paper to encompass all rhotic variants.

production of American English speech sounds should be mastered by children nine years of age (Shriberg, 1994; Smit, 1986; Smit, Hand, Freilinger, Bernthal, & Bird, 1990). Shriberg (2010) has used the term Speech Errors when referencing speakers with speech sound distortion errors that persist; however, these sound errors are not coupled with adverse social and academic risks. Similarly, the term Residual Speech Sound Error (RSSE) has been used for children who exhibit certain speech errors beyond the typical age of acquisition when they reach the age of nine (Preston et al., 2014; Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997). Later developing sounds, such as /r, s, z, l/, are among sounds frequently in error for those with RSSEs (Preston & Edwards, 2007; Preston et al., 2014; Shriberg, 1994). The present study focuses on those children with Speech Delays, whose distortions and speech errors may soon turn into RSSEs.

While other errors also occur in children ages 7 to 9, /r/ errors are commonly observed (Bernthal et al., 2009; Ruscello, 1995a; Secord et al., 2007; Shriberg et al., 1997). The liquid /r/ is one of the last phonemes to be acquired by children when compared to all other English phonemes (Arlt & Goodban, 1976; Smit et al., 1990; Templin, 1957). Templin (1957) found that /r/ reached a 75% level of accuracy at age 4; 0 in males and females. Sax (1972) found that when mastery was defined as 93% accuracy, females mastered /r/ at the end of third grade, while males did not reach mastery of /r/ even by the end of fifth grade. More recently, Smit et al. (1990) found that /r/ reached a 90% level of accuracy at 8; 0 in males and females. Thus, there is a wide range of variation in reported acquisition of /r/. As /r/ is among the last sounds to be acquired in typical speech sound development, failure to achieve correct production by approximately age nine may result in a RSSE.

Lingual Components Necessary for /r/

The articulatory and phonetic variability of rhotics, as well as their complexity, make rhotics one of the most common sound classes in error (Bernthal et al., 2009; Ruscello, 1995a; Secord et al., 2007). The American English rhotics may appear in several positional contexts as either a vowel or consonant. Rhotic vocalic variants appear in the nucleus of a syllable (i.e., /ɝ/ and /ɚ/). Rhotic consonantal variants appear in prevocalic positions (i.e., /r/) and in postvocalic positions which are commonly referred to as "rhotic diphthongs" (i.e., /ar/, /ɛr/, /ɪr/, /ʊr/, and /ɔr/) (Klein, McAllister-Byun, Davidson, & Grigos, 2013; Secord et al, 2007). Although the articulatory constrictions (i.e., narrowing of the vocal tract) are similar regardless of phonetic context, the relative timing of the constrictions may vary depending on word position **(**Gick, Campbell, Oh, & Tamburri-Watt et al., 2006)[2]. For simplicity, /r/ will be used throughout this paper to encompass all rhotic variants.

No matter the positional context of /r/, the following lingual components are typically necessary for the proper production of /r/: a pharyngeal constriction, an oral constriction (achieved by tongue tip retroflex or bunching of the tongue body), tongue midline grooving, and (in most speakers) lateral bracing (Bacsfalvi, 2010; Bernhardt et al., 2008; Bernhardt, Gick, Bacsfalvi, & Alder-Bock, 2005b). The pharyngeal constriction refers to the tongue root retracting into the pharynx and involves a posterior narrowing of the vocal tract. Anterior tongue shapes for /r/ are generally divided into two main categories: retroflex and bunching. For a retroflex /r/, the tongue tip is raised and curled back slightly toward the alveolar ridge; for a bunched /r/, the tip of the tongue lowers, while the dorsum rises toward the hard palate (Alder-Bock, Bernhardt, Gick, & Bacsfalvi, 2007; Bernhardt et al., 2005b; Klein et al., 2013; McAllister-Byun, Hitchcock, & Swartz, 2014). The dorsum of the tongue does not usually

---

[2] Constriction is used throughout this paper to specify a narrowing at any point in the vocal tract.

contact the hard palate; however, lateral bracing is often exhibited wherein the sides of the posterior tongue contact the rear upper molars (Bernhardt et al., 2005b; McAllister-Byun et al., 2014). As the sides of the posterior tongue elevate to contact the rear molars, a midline groove is often seen in the center of the tongue (Bernhardt et al., 2005b; McAllister-Byun et al., 2014; Preston, Brick, & Landi, 2013). The most anterior constriction associated with a correct production of /r/ is a slight degree of lip rounding (Bernhardt et al., 2005b; Klein et al., 2013; McAllister-Byun et al., 2014). The magnetic resonance image (MRI) picture in Figure 1 displays the following constrictions necessary for a correct production of /r/.

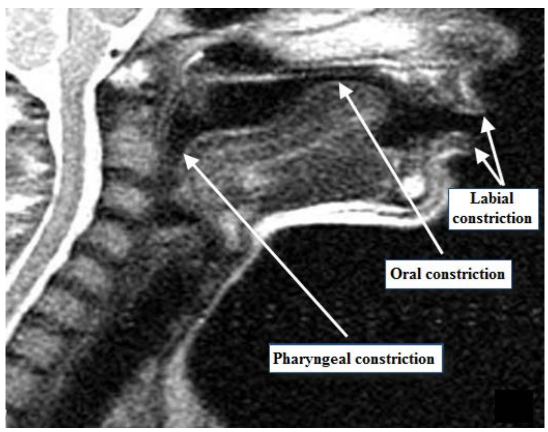


Figure 1: MRI image of a correct /r/ production with labial, oral, and pharyngeal constrictions (Image from Secord et al., 2007).

While these components are necessary for /r/ production, each component will not be uniformly met across speakers, and individual variation exists; for example, MRI midsagittal images reveal a variety of tongue shapes with differing amounts of pharyngeal constriction and midline grooving to produce a perceptually correct post-vocalic /r/ (Secord et al., 2007). Recently, research has revealed that the least accurate /r/ productions are associated with a highly curved, posteriorly located tongue shape which likely results in loss of pharyngeal constriction since the tongue root is not retracted properly into the pharynx (Boyce, Combs, & Rivera-Campos, 2011; Klein et al., 2013). An accurate /r/ cannot be produced with only one component; thus, a speaker must possess the motor abilities to simultaneously produce all lingual components necessary for /r/, adding to the complexity of this phoneme (Gick et al., 2007).

Review of Traditional /r/ Therapy

The variability and complexity of the lingual components of /r/ contribute to persisting /r/ errors in children and brand treatment of /r/ as difficult (Adler Bock et al., 2007; Bernthal et al., 2009; Gruber, 1999; Klein et al., 2013; McAllister-Byun et al., 2014; Preston et al., in press; Ruscello, 1995a; Ruscello, 1995b). Traditional methods for /r/ therapy include imitation, contextual facilitation, phonetic placement and shaping, motokinesthetic training, and awareness training (Alder-Bock et al., 2007; Bernthal et al., 2009; Ruscello, 1995a; Secord et al., 2007).

A variety of tongue shapes may lead to a perceptually correct /r/, including different degrees of retroflexed or bunched tongue shapes; thus, there is not a single way to verbally cue the correct production of /r/. Therefore, numerous verbal cues are used in the clinical setting to elicit /r/; however, these cues alone may not be sufficient when specific information concerning the oral cavity is needed to facilitate a correct /r/ production. Verbal cues encouraging children to curl the tongue tip up or bunch the tongue toward the back of the mouth are extremely abstract and relay minimal information about the oral configuration necessary to produce /r/. Tactile feedback concerning tongue position is limited: lateral bracing is the only tactile feedback available, and this gives no specific information about where to place the front of the tongue in order to achieve the desired constrictions. Additionally, lip rounding, and a relatively small oral cavity opening while producing /r/ restricts clinicians from seeing inside the mouth to properly cue tongue positions that will facilitate /r/. Thus, clinicians are forced to rely on acoustic information and clinical experience to cue the desired tongue shape (Secord et al., 2007).

While a myriad of intervention techniques exist to help facilitate correct production of /r/, a limited number of evidence based methods are available for treating /r/ (Ruscello, 1995b). Shriberg (1975) outlined an eight step approach to evoking the rhotic vowel /ɝ/. This shaping procedure required shaping /l/ into /ɝ/. Of the 65 children with developmental /ɝ/ errors who were administered this approach, approximately 70% produced a correct /ɝ/ within six minutes of therapy, and 10% needed additional training sessions to produce a correct /ɝ/.

Shriberg (1980) also developed a diagnostic teaching procedure to use with children exhibiting persistent rhotic errors. Children were first classified into one of two categories based on the articulation of /r/ in the assessment phase. When the clinician became familiar with the child's error pattern, the treatment phase began, seeking to eliminate any incorrect articulatory movements. Lastly, the bite stick was introduced to assist in facilitating a perceptually correct /r/. A bite stick is a four to six inch long dowel that is approximately three eights of an inch in diameter. Shriberg (1980) noted that variations of this procedure had been conducted on over a dozen children, and while all children dramatically modified their gestures, not all children produced /r/ correctly after one session. While these traditional approaches may be successful for some children, they do not consistently result in correct production of pre-vocalic and vocalic /r/. Additionally, both of these approaches target productions of /r/ in isolation only.

Most /r/ articulation therapy is provided to individuals older than nine years of age (i.e., those with RSSEs) because /r/ is not acquired until a later age. However, providing articulation therapy for /r/ to younger children with Speech Delays may help avoid the progression of /r/ errors into RSSEs. While these traditional approaches to /r/ therapy are still relevant, increasing access to new technologies warrants consideration for enhancing the effectiveness of /r/ therapy as clinical observations reveal that numerous children remain in /r/ therapy for years, or are dismissed from therapy on the grounds that no improvements have been made. Therefore, when traditional approaches to /r/ therapy have not worked for a child, alternative approaches incorporating technology into therapy should be explored (Alder-Bock et al., 2007; Ruscello, 1995a; Ruscello, 1995b).

Review of Biofeedback in /r/ Therapy

Biofeedback refers to the presentation of information that allows individuals to gain greater awareness of physiological functions (Volin, 1998). This is usually achieved by the use of instrumentation that provides real-time information on the desired activity. In regards to speech, biofeedback allows aspects of speech that are usually difficult to perceive (i.e., tongue movement) to be brought under conscious control using real-time information and instrumentation that provides physiological feedback (McAllister-Byun et al., 2014; Volin, 1998). One of the most common applications of biofeedback in speech therapy is the use of a mirror to provide visual feedback on the movement of the lips and jaw during speech articulation. However, the lingual gestures of /r/ cannot be seen in a mirror.

Research on alternative biofeedback methods in speech therapy has been conducted for approximately the past half-century (Fletcher 1982, 1983; Fletcher, Dagenais, & Critz-Crosby, 1991; Fletcher, McCutcheon, & Wolf, 1975). The use of technology in articulation therapy has fostered interest in current literature as these technologies become more refined and the prices decrease, making them more accessible in the clinical setting (Bernhardt et al., 2003; Gick, Bernhardt, & Bacsfalvi, 2004). Although limited, research has evaluated the general effectiveness of spectral biofeedback, electropalatography (EPG), and two-dimensional ultrasound across a variety of clients (Bacsfalvi & Bernhardt, 2011; Bacsfalvi, Bernhardt, &

Gick, 2007; Fletcher et al., 1991; Gibbon & Hardcastle, 1989; McAllister-Byun et al., 2014; Preston et al., 2014; Shuster, Ruscello, & Toth, 1995).

Spectral biofeedback. Spectral biofeedback refers to the use of real-time spectrographs to provide visual feedback about the acoustic signal to help facilitate correct speech sound productions (Schuster et al., 1995). The use of spectral biofeedback was shown to be effective in remediation of individuals who failed to respond to traditional forms of treatment for /r/ (Schuster, Ruscello, & Smith, 1992; Shuster et al., 1995). More recently, McAllister-Byun & Hitchcock (2012) sought to determine if children who failed to respond to traditional forms of treatment for /r/ would benefit from spectral biofeedback intervention. Clinicians provided spectral biofeedback by uploading an appropriate formant height template for a given sound, which allowed the child to match their results with the appropriate template. Results revealed that at the group level no significant changes were evidenced in the accuracy of /r/ productions after a period of traditional intervention, whereas statistical analysis revealed that /r/ productions were significantly more likely to be rated as perceptually correct after spectral biofeedback.

Although spectral feedback is non-invasive and only requires a microphone, the display may not always be clearly understood by clients. Moreover, acoustic templates of /r/ vary to some degree. A child’s /r/ production is not necessarily wrong if the formant pattern does not identically match the clinician’s formant template. Further, matching formant patterns does not provide the child with precise information on the articulatory movements needed to produce these formants, and traditional methods of cueing tongue placement still need to be used (Gick et al., 2004).

Electropalatography (EPG). EPG requires the client to wear a custom-fit pseudopalate that contains electrodes which record the timing and location of tongue-palate contact. This allows for a tongue-palate contact pattern to be displayed, providing visual biofeedback (Bacsfalvi et al., 2007). Prior research has examined the usefulness of EPG as visual biofeedback in speech therapy. Fletcher (1982) found that deaf speakers were able to use biofeedback and EPG training to perceive and produce key elements of speech. Additional studies using EPG displayed mixed results; however, Fletcher et al. (1991) demonstrated that six subjects with profound hearing impairments had improved linguapalatal contact patterns for consonantal sound targets. They also reported overall improved intelligibility from listeners using EPG in therapy. Gibbon and Hardcastle (1987) used EPG to provide detailed information in a therapy program to correct a lateralized /s/. They reported that normal contact patterns for /s/ were achieved and maintained in a six-month follow up. Gibbon, Stewart, Hardcastle, and Crampin (1999) determined that a therapy program using EPG for visual feedback was successful, as a child with a developmental speech disorder who previously had abnormal tongue palate contact showed typical EPG patterns for alveolar stops in speech post-therapy. Schmidt (2007) found an EPG system to be successful in treatment in several case studies. Thirteen clients with articulation disorders resulting from a range of causes, and who had received years of previous speech treatment, consistently produced acoustically acceptable placements for their target sounds (including /r/) using the EPG system.

Previous research has demonstrated the effective use of EPG as visual biofeedback in speech therapy (Bacsfalvi & Bernhardt, 2011; Bacsfalvi et al., 2007; Dent, Gibbon, & Hardcastle, 1995; Fletcher et al., 1991; Gibbon & Hardcastle, 1989; Schmidt, 2007; Schmidt & Beamer, 1998). EPG is useful because it provides a clear, two-dimensional contact pattern between the tongue and the palate, which is especially useful for alveolar and palatal sounds; however, studies have not shown that EPG works better than ultrasound for one class of sounds over another in speech habilitation for deaf clients (Bernhardt, Gick, Bacsfalvi, & Ashdown, 2003). While the use of EPG allows the client to see if lateral bracing is occurring for the production of /r/, it lacks direct visual feedback pertaining to tongue placement for the anterior constriction and the pharyngeal constriction. Additionally, the pseudopalate must be custom made to fit the client and requires time and additional monetary investment from the client. The client must also agree to use the pseudopalate and be cooperative in the invasive process of wearing the pseudopalate (Bernhardt et al., 2005b; Zharkova, 2013).

Ultrasound. The use of ultrasound in speech therapy has been shown to be safe and to have no detrimental side effects (Gick et al., 2004; Preston et al., 2014). When using an ultrasound, gel is applied to the tip of the ultrasound probe to enhance image quality. The probe is placed beneath the speaker's chin and allows for the image to be displayed on a screen as ultrasonic waves are reflected back towards the probe when they encounter air from the oral cavity just above the tongue (Bernhardt et al., 2005b).

Ultrasound is particularly useful in treatment of /r/ because several possible tongue configurations may produce /r/. Ultrasound displays of tongue shapes and movements help the clinician select appropriate cues to facilitate a correct production of /r/. The most anterior constriction of /r/ is labial (i.e., lip rounding). The labial constriction is not visible with ultrasound. The medial oral constriction produced by the tongue tip or blade and the posterior constriction produced as the tongue root retracts toward the pharyngeal wall are both visible with ultrasound. The depth of the midline groove in the tongue is also visible with ultrasound using a coronal view (Bernhardt et al., 2005b; Bernhardt, Bacsfalvi, Gick, Radanov, & Williams, 2005a). The constrictions visible on an ultrasound display are shown in Figure 2 and Figure 3.

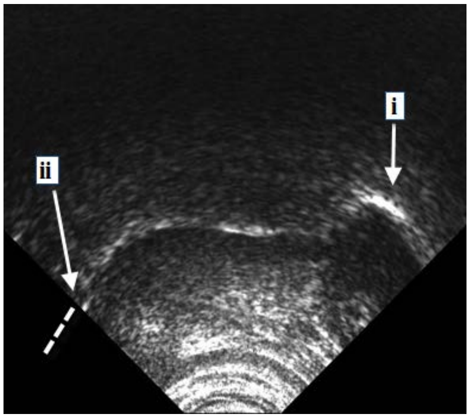


Figure 2: Sagittal ultrasound display where (i) shows the tongue tip/blade creating the oral constriction and (ii) shows the projected tongue shape creating the pharyngeal constriction

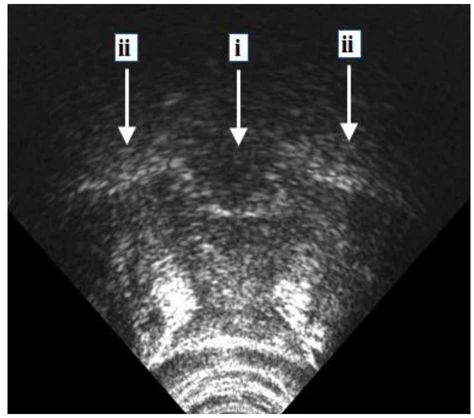


Figure 3: Coronal ultrasound display where (i) shows the midline tongue groove and (ii) shows the raised sides of the tongue

Ultrasound is a less invasive technique than EPG. While tongue-palate contact information is not available with the use of ultrasound, approximate measures of where the palate would be in relation to the tongue can be determined if a clinician desires (Bernhardt et al., 2005b). Ultrasound is especially ideal for younger children as nothing needs to be placed inside the mouth. Furthermore, ultrasound provides a 2-dimensional image of the tongue that can be viewed in either a sagittal or coronal view. Ultrasound allows children to visualize the exact movement of the tongue in real time. Additionally, the direct tongue images may be presented in real time or frozen at certain points in time (Bernhardt et al., 2005b). This enables the clinician to provide specific feedback on tongue placement and movement. Figures 4 and 5 provide examples of how specific tongue placement cues are available with the use of ultrasound.

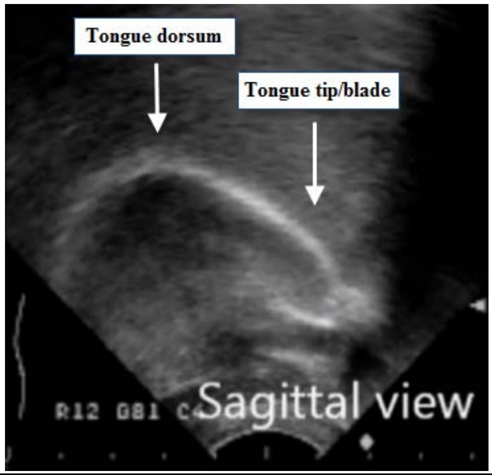


Figure 4: A tongue position that will not elicit a correct /r/ production because the tongue tip/blade is very low and the posterior dorsum of the tongue is very high; thus, the following placement cue would be used: "lower the back of the tongue and raise the front of the tongue" (Image from Haskins Laboratories, 2014).

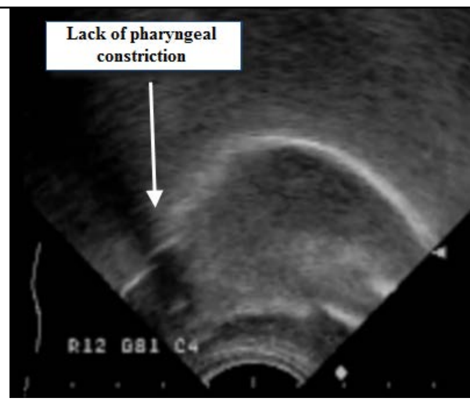


Figure 5: A tongue position that will not elicit a correct /r/ production because the tongue root is not retracted far enough into the pharynx; thus, the following placement cue would be used: "pull the back of the tongue into the throat" (Image from Haskins Laboratories, 2014)

Specific feedback on tongue placement and movement aids a child in understanding the unique tongue shape needed to produce /r/; this information is not available while using EPG or spectral biofeedback. Additionally, the ultrasound is small and portable, allowing clinicians to easily administer therapy in situations most convenient for a client (i.e., clinic or home settings) (Bacsfalvi & Bernhardt, 2011; Bernhardt et al., 2005b; Lee, Wrench, & Sancibrian, 2015). The primary advantages of ultrasound include dynamic and static visual feedback imaging abilities, information concerning the medial and posterior constrictions necessary for /r/, its non-invasive nature, and portability.

Efficacy of Ultrasound in Speech Therapy

Two-dimensional ultrasound as visual feedback has helped facilitate articulatory gains in speech therapy for a variety of clients (Bacsfalvi et al., 2007; Bacsfalvi & Bernhardt, 2011; Bernhardt, 2003; Cleland, Scobbie, & Wrench, 2015; Lee et al., 2015). Bernhardt et al. (2008) demonstrated that ultrasound has positive benefits with less than three hours of actual practice in establishing sounds in speech therapy habilitation when clients possess motivation and have less pervasive residual speech impairments. Preston et al. (2014) found eight participants with RSSEs increased accuracy for at least one treated sound during a treatment program including ultrasound biofeedback. McAllister-Byun et al. (2014) demonstrated that ultrasound biofeedback intervention can be highly effective for children with persistent rhotic errors. Additionally, they reported that therapy is more effective when opportunities to explore different tongue shapes are included; this helps elicit a tongue shape facilitative to a perceptually correct rhotic, when compared to only cueing a bunched tongue shape. Further, Cleland et al. (2015) found that seven children with persistent speech sound disorders who were previously unresponsive to traditional therapy approaches made significant progress evidenced by perceptual measures and tongue shape analysis after intervention that included the use of ultrasound. Preston et al. (2013) found that ultrasound as biofeedback resulted in improved accuracy of treated sound sequences (e.g., /ar, kr/) in children with persisting Childhood Apraxia of Speech. Preliminary results further evidenced that visual feedback using ultrasound could be used to assist in increased performance of /r/ in adults with acquired Apraxia of Speech due to a cerebral vascular accident (Preston & Leaman 2013). The above studies support the use of ultrasound in speech therapy; however, no studies have directly compared ultrasound therapy to no ultrasound therapy in order to observe the specific benefits ultrasound may provide within speech therapy.

Ultrasound has been shown to be an effective tool in establishing a correct production of phonemes and a useful aid to provide biofeedback of the tongue in speech therapy; yet, an optimal therapy plan for incorporating ultrasound into traditional speech therapy has not been defined (Lee et al., 2015). It is reasonable to believe that children with speech sound errors have unsuccessfully established a proper motor plan involving the appropriate shaping of articulators, or they have trouble with accurate timing of the movements they have learned (Preston et al., 2014). Additionally, as Shriberg (1980) noted, children may have learned incorrect articulatory patterns that must be rectified through speech therapy. Therefore, approaching therapy from a motor learning framework may facilitate learning of the movements required for rhotics (Preston et al., 2014).

Review of Motor Learning

Motor learning principles have recently been applied to treatment for children’s speech sound errors (Hitchcock & McAllister-Byun, 2015; Preston et al., 2014). While it is unknown if speech motor control is equally responsive to the same principles of learning as nonspeech motor control, a reasonable and advocated hypothesis is as follows: principles of motor learning govern the motor skills necessary for speech production that are similar to those of nonspeech motor control (Duffy, 2005; Maas et al., 2008). Thus, the following factors that have been shown to facilitate motor skill acquisition and learning in general are considered when developing effective interventions for SSDs: structure of practice, stimulus selection, and nature of feedback.

In order to understand the principles of motor learning, the distinction between motor performance during acquisition of a skill and motor performance during retention and generalization must be made. Motor *performance* may be observed as a temporary change in executing a movement during acquisition. However, accurate movement in untrained contexts (*generalization*) must be observed over time (*retention*) to achieve permanent change in movement resulting in *learning* (Schmidt & Lee, 1999). The goal of speech therapy is to enhance communication in a person's everyday life (Bernhardt et al., 2005a); accordingly, the long-term goal is for clients to learn by demonstrating successful retention and generalization of target sounds in untrained stimuli and untrained contexts.

Therefore, it is crucial to the field of speech-language pathology that our treatment properly distinguishes acquisition, retention, and generalization so the long term goal of learning is achieved in treatment. Acquisition reflects motor performance, while retention and generalization are critical elements that reflect motor learning. Acquisition refers to successful attempts during practice, which aids in the capability of rehearsed movements. Although some degree of skill acquisition is necessary before learning can occur, acquisition does not imply learning. In speech therapy, acquisition occurs as trained speech sounds are repeatedly practiced and shaped into new movements. As acquisition of rehearsed movements increases, performance within a session also increases as target sounds are correctly produced in repeated trials. Several individual factors may affect practice performance within a session, including: warm up, fatigue, and attention. Therefore, it is critical to understand that performance during acquisition is not an indicator of retention and generalization: it does not imply that learning is taking place. Thus, motor performance is reflected through acquisition of rehearsed movements;

however, motor learning must be observed through retention and/or generalization (Maas et al., 2008; Schmidt & Lee, 1999).

Retention refers to performance levels after practice is completed. In speech therapy, retention can be observed by determining the number of trained items containing a target sound practiced during a session that the client correctly produces when asked to produce the target sound at a later time. The improved capability of rehearsed movements should not only be observed during acquisition, but should be retained over time. Generalization refers to how practice on one movement affects similar, but untrained, movements. In speech therapy, generalization may be observed by determining the number of untrained items containing the target sound that are correctly produced after a duration of time. The improved capability of rehearsed movements should not only be retained over time, but should also be generalized (i.e., transfer) to similar, but untrained movements to demonstrate that learning has occurred. Therefore, acquisition must precede learning; however, speech therapy seeks to ultimately maximize learning (retention and generalization), not acquisition during therapy sessions (Maas et al., 2008; Schmidt & Lee, 1999). To date, no study has examined the effects of acquisition and learning in therapy that includes the frequent use of ultrasound compared to no ultrasound.

Feedback. In order to learn a new motor plan for a target sound (e.g., /r/), acquisition of the basic movement pattern must first take place. Feedback is a vital component of establishing a new motor plan and two *types of feedback* are typically distinguished: knowledge of performance (KP) and knowledge of results (KR). KP feedback refers to feedback related to how the movement was produced, while KR feedback refers to feedback related to the results produced in terms of the goal (e.g., correct or incorrect). While the effects of KR feedback and KP feedback have not been thoroughly studied in speech motor learning, KP feedback and KR feedback appear to be equally effective in most speech tasks; however, when a task is novel (e.g., teaching a new speech sound) KP feedback may be more helpful when provided after a practice attempt (Maas et al., 2008; Schmidt & Lee, 1999).

Traditional approaches to /r/ therapy supply KP feedback with verbal instruction (e.g. “elevate your tongue a little more”); however, ultrasound offers an alternative way to offer KP feedback during speech therapy. Thus, in addition to verbal KP feedback offered by the clinician, ultrasound as biofeedback allows clients to receive KP feedback visually. Since KP feedback is especially useful when trying to master an unfamiliar task, providing two differing levels of support for KP feedback should help facilitate motor performance (Maas et al., 2008 & Preston et al., 2014).

It is important to acknowledge the effects increased KP may have on acquisition and generalization. KP feedback has been found to facilitate the rate of acquisition and overall performance level when a non-speech task is not well known (Newell, Carlton, & Antoniou, 1990). However, Hodges and Franks (2001) demonstrated that an increase in KP feedback during non-speech tasks may impede motor learning. The effects of increased KP feedback using ultrasound as visual feedback on acquisition and generalization in speech tasks that are not well known has not yet been established in previous research. The present study seeks to determine how the increase of KP feedback by use of high frequency ultrasound within a session will affect acquisition within a session, and retention and generalization between two consecutive sessions compared to the use of no ultrasound.

Other important considerations regarding feedback in motor learning include *frequency of feedback* whereby reduced frequency of feedback has benefits for motor learning while frequent feedback appears to augment novel learning. Thus, the initial phase of therapy should begin with frequent KP feedback and KR feedback to help aid in novel learning of the target sound. KP feedback and KR feedback should be reduced once the target sound has been established. Whether KP feedback and KR feedback are provided *immediately or delayed* is also an important consideration. The limited evidence available for speech motor learning presents preliminary evidence that delayed feedback may enhance speech motor learning (Austermann Hula, Robin, Maas, Ballard, & Schmidt, 2008; Maas et al., 2008).

Practice. The *complexity of stimuli* practiced in speech therapy is important to consider. The challenge point framework (Hitchcook & McAllister Byun, 2014) considers the amount of information available and interpretable by the child when factors such as functional task difficulty, nominal task difficulty, and the child's skill level are influential in maximizing learning for a child. Thus, learning is optimized when a child is challenged; however, a shortage or excess of challenging material will have the reverse affect (Guadagnoli & Lee, 2004). Therefore, selected stimuli should increase in complexity to a level consistent with a child's motor performance to increase the task difficulty. Instructional steps for production training allow for this increase in stimuli difficulty to occur (Bernthal et al., 2009). For example, training typically begins at the syllable level and progresses to words, starting with monosyllabic words and moving to multisyllabic words, phrases, then full sentences. Lastly, the target sound is practiced at the conversational level using everyday speech **(**Van Riper & Erickson, 1996**)**.

Once stimuli are selected for use in speech therapy, the use of blocked versus random practice must be determined. Blocked practice refers to a practice schedule wherein the client practices a group of the same target movements where the target is predictable before moving on to the next target movement (e.g., AAA, BBB, CCC). Random practice refers to a practice schedule wherein different movements are successively practiced, and the target for the upcoming trial is not predictable to the client (e.g., ABC, CBA, and BCA). There is preliminary support suggesting that random practice enhances motor learning for intact and impaired speech motor systems (Maas et al., 2008; Skelton & Hagopian, 2014). Concerning nonspeech functions, learning may be optimized by first practicing in blocked order and then transitioning to random order (Lai & Shea 1998; Lai, Shea, Wulf, & Wright, 2000). Therefore, clients may benefit from first using a blocked practice schedule and then transitioning to a random practice schedule once the target sound has reached a high level of performance.

When selected stimuli are presented in speech therapy, KP feedback and KR feedback are provided to supply the child with knowledge about his/her attempt of the target sound. Ultrasound as visual feedback provides a higher level of KP feedback than is available with traditional verbal or tactile cues as the child is provided with a visual of the moving articulators. Therefore, it is necessary to examine how ultrasound as visual feedback providing KR feedback and KP feedback will influence a child's motor performance and motor learning, and to evaluate how this compares to treatment that does not have an ultrasound as visual feedback component.

Level of evidence. A multi-phase progression framework may be used to examine treatment research on ultrasound as visual feedback (Fey, Finestack, & Schwartz, 2009; Robey, 2004). Phase I research seeks to establish the practicality of the initial concept and typically includes low levels of experimental controls such as case studies (Robey, 2004). Phase I research has suggested that ultrasound as visual feedback may have beneficial effects for individuals with residual and developmental speech sound errors (Alder-Bock et al., 2007; Modha, Bernhardt, Church, & Bacsfalvi, 2007).

Phase II research targets outlining treatment procedures and demonstrating outcomes when testing the protocol (Robey 2004). Phase II research involves studies pursuing more experimental control (McAllister-Byun et al., 2014; Preston et al., 2014) and demonstrated that ultrasound as visual feedback improves articulatory accuracy in children with Childhood Apraxia of Speech and in children with RSSEs (Preston et al., 2013; Preston, Maas, Whittle, Leece, & McCabe, 2015). While phase II research is underway in examining ultrasound as visual feedback, it is limited. Thus, there is a need for more phase II research to explore the optimal treatment methods and demonstration of effects when using ultrasound as visual feedback in therapy.

Present Study

The present study was guided by the principles of motor learning. This study examined the differences between how two treatments influence acquisition, retention, and generalization in children with Speech Delay characterized by rhotic errors. Though learning (i.e., retention and generalization) is the primary goal of speech therapy, acquisition is a necessary prerequisite to learning. Thus, the present study examined how ultrasound as visual feedback treatment compared to no ultrasound treatment in facilitating performance within a session to trained targets, retention between consecutive sessions to trained targets, and generalization between sessions to untrained targets. Prior to the present study, no studies had directly compared a treatment program with ultrasound as visual feedback to a treatment program with no ultrasound that is similar in target selection, treatment duration, practice schedule, and feedback frequency. This study sought to answer three questions concerning the use of ultrasound as visual feedback in treatment: 1) Does frequent exposure to ultrasound feedback of the tongue facilitate acquisition of rhotic diphthongs better than no ultrasound feedback; 2) Does frequent exposure to ultrasound facilitate retention of rhotic diphthongs better than no ultrasound feedback; and 3) Does frequent exposure to ultrasound facilitate generalization of rhotic diphthongs to other rhotics (e.g., /r/, stressed schwar, etc.) and differing positions (e.g., initial /r/, /r/ in clusters, etc.) better than no ultrasound feedback.

Principles of motor learning were used to further examine the above questions. According to Newell et al. (1990), KP feedback may facilitate the rate of acquisition and overall performance level in non-speech tasks that are not well established. Thus, concerning speech tasks, ultrasound feedback may enhance acquisition of a sound by providing a visual display of the tongue that can be evaluated (i.e., comparing actual tongue shapes to targets drawn over the ultrasound display); this type of KP feedback cannot be provided during traditional therapy. Therefore, ultrasound use was expected to facilitate acquisition of rhotic diphthongs compared to no ultrasound within treatment sessions. The use of ultrasound would enhance the KP feedback participants receive because they would be provided with direct visual KP feedback in addition to verbal KP feedback.

In contrast, Hodges and Franks (2001) demonstrated that motor learning may be impeded when KP feedback is increased during non-speech tasks. Concerning speech tasks, Preston et al. (2015) found that children with persisting speech sound errors associated with CAS increased the accuracy of rhotic productions within treatment sessions that included ultrasound visual feedback of the tongue; however, they did not demonstrate generalization to untreated words. Thus, concerning unfamiliar speech tasks, frequent exposure to ultrasound as visual feedback may impede retention and generalization of rhotic diphthongs as participants may become dependent on the ultrasound in speech therapy (cf. Preston et al., 2015). Consequently, individuals may be unable to generalize trained sounds to untrained contexts without the use of the ultrasound. The use of high frequency ultrasound may only affect the words trained per session where success may be seen within the session, but not between sessions (i.e., retention) and not to other similar contexts (i.e., generalization). As a result, the increase of KP feedback that ultrasound provides may impede retention and generalization.

In sum, the use of ultrasound as KP feedback is expected to facilitate motor performance within a session; however, the use of ultrasound as KP feedback may hinder motor learning. The present study seeks to determine how the increase of KP feedback by use of high frequency ultrasound within a session will affect acquisition within a session, and retention and generalization between two consecutive sessions compared to the use of no ultrasound.

To answer the above questions, a single-subject randomized blocked phase scheme was employed in the present study to address three *hypotheses*, as derived from motor learning theory:

1. Frequent exposure to ultrasound as visual feedback of the tongue will better facilitate acquisition of rhotic diphthongs compared to no ultrasound within treatment sessions.
2. Frequent exposure to ultrasound as visual feedback of the tongue may impede retention of rhotic diphthongs compared to no ultrasound between treatment sessions.
3. Frequent exposure to ultrasound as visual feedback of the tongue may impede generalization of rhotic diphthongs compared to no ultrasound between treatment sessions.

Based on the levels of evidence described by Baker and McLeod (2011) and adopted by ASHA (www.asha.org/members/ebp/assessing.htm), the present study qualifies as level IIb evidence: single case experimental design with replication. The present study was considered Phase II research, as a single-subject design was used to assess the therapeutic effects of ultrasound as visual feedback and to estimate the effect size of the treatment when compared to the use of no ultrasound in therapy. Existing studies had only demonstrated that ultrasound treatment is more effective than no treatment; the present study allowed for the first comparisons to be drawn between treatment using ultrasound as visual feedback and treatment with no ultrasound. This essential step added a valuable component to existing literature, which had not included studies directly comparing the effectiveness of ultrasound as visual feedback during treatment to other treatments. This provided an important step of advancing ultrasound treatment research as the present study built upon existing research to establish new findings in the efficacy of ultrasound to support future, large scale studies (Baker & McLeod, 2011). Results of the present study have informed the development of future Phase III research seeking to establish the efficacy of the protocol through randomized controlled trials (Robey, 2004). Additionally, it may provide guidance for clinicians who are considering treatment options for children with SSDs, including those with Speech Delays.

Methods

Study Design

A single-subject randomized block design replicated across four subjects was used for the present study providing multiple opportunities for the intervention to demonstrate an effect. A randomized block design is similar to an alternating treatment design except that each week the treatment order is randomly determined instead of remaining the same. Rvachew (1988) and McReynolds and Thompson (1986) outlined the advantages of a single-subject design: greater control over subject variability, economy, and the ability to observe unique clients.

Kratochwill and Levin (2014) further explained that the addition of randomization to replication across sessions within a single-subject design improves the credibility of the experiment and reduces threats to internal validity. In addition to graphical-visual analysis from single-subject designs, adding randomization allows for a variety of data-analysis strategies to be computed including effect size and statistical significance (Rvachew, 1988).

In order to strengthen the causal and statistical conclusions of the present study, a within-series randomized phase single-subject design was used. Specifically, a blocked phase randomization scheme was used to test the present study's hypotheses. Each successive week was considered a block. Each week involved two different treatment sessions: one ultrasound (US) and one no ultrasound (NoUS). US represents a session of treatment with ultrasound as visual feedback and NoUS represents a session of treatment with no ultrasound. The order of one US session and one NoUS session was randomly assigned to each week. This type of assignment method guaranteed that conditions of the same type could not appear in more than two consecutive time periods ensuring that a large dose of one treatment did not successively occur (Kratochwill & Levin, 2014). An additional tier of randomization was added to the present study during baseline, as a randomized number of baseline sessions from 3-5 was assigned to each replication across participants. Kratochwill and Levin (2014) explained that, generally, including more replications in the study design concerning number of cases or number of time periods will yield more statistical power during the statistical analysis. Therefore, the presented study completed 14 sessions (7 randomized blocks) replicated across 4 subjects.

Source of Data

Two probe lists were administered to the client during US treatment and NoUS treatment: a Training Probe List comprised of 20 words was administered to observe acquisition and retention; a Generalization Probe List comprised of 25 words was administered to observe generalization. The Generalization Probe List and Training Probe List were combined into one randomized master list containing 45 words total. The master list was administered at the beginning of each session with a different start point and order each time to ensure probes were presented in random order. At the end of the session, only the Training Probe List was administered in random order. Both probe lists were administered using imitation (e.g., clinician states, “say chair”) because the present study included children ages 7-9 who may not have mastered reading.

Acquisition. Acquisition was quantified by administration of the Training Probe List during both the beginning of the session and the end of the session. The percent increase in accuracy from the beginning to the end of each session served as the dependent variable to address Hypothesis 1, and the data provided insight into whether US treatment versus NoUS treatment was more effective at facilitating acquisition on trained items within a session.

Two rhotic contexts (/Vr/ combinations) were trained for each participant. Each rhotic context included 10 initial training words: five to be trained during US sessions and five to be trained during NoUS sessions. Therefore, only 10 words were trained during each session. All training items were designated as an US or NoUS training item before therapy began and any given word was trained under only one of the conditions. The use of differing words for NoUS and US sessions allowed for additional conclusions to be drawn that compared the success of acquisition training words between the two therapy types without contamination/carry-over between the two treatment conditions (please see Appendix A for the Training Probe List Sample).

Twenty Training Probe Items were administered at the beginning of each session. The same Training Probe Items administered at the beginning of the session and determined for training were always the same Training Probe Items administered at the end of the session. The number of Training Probe Items incorrectly produced at the beginning of the session provided the number of probes available for improvement out of 10 (i.e., 10 Training Probe Items for that session). Then, a percentage was calculated by subtracting the number of correctly produced Training Probe Items out of 10 at the beginning of the session from the number of correctly produced Training Probe Items out of 10 at the end of the session. This method determined the percentage increase of correctly produced Training Probe Items from the beginning of the session to the end of the same session.

All Training Probe Items within a session were taken from the Training Probe List. Each Training Probe Item began at the monosyllabic level (please see Appendix B for Training Probe Chains); if a participant correctly produced the monosyllabic Training Probe Item at the beginning of the session, a varied longer phrase (i.e., a phrase or short sentence containing six to seven syllables) including a multisyllabic word form of the original monosyllabic word replaced the original monosyllabic word in the Training Probe List. This helped to ensure that the highest level of each chain was reached. Additionally, in order to avoid continuous training of a varied longer phrase that had been correctly produced during administration of the Training Probe List at the beginning of the training session, a word bank for additional chains containing each target sound was kept (see Appendix C for sample Target Sound Word Bank). Thus, when a participant correctly produced a varied longer phrase at the beginning of the session, a new, varied longer phrase containing a different multisyllabic target word was administered.

A protocol was developed for adapting Training Probe Items during each session in order to ensure a target sound within a chain had been retained or generalized before replacing the word (please see Appendix D for more detail on the procedures used for adapting the Training Probe List). The procedures for adapting the Training Probe List were designed to ensure participants began each session with a chance of acquiring as close to 10 words during each session as possible. That is, if a participant began the session with 3/10 correct on the Training Probe List, only 7 could be acquired during the session. After sampling up to three alternative chains per word during each session to identify words in error, ten items were always trained in the session (which could include items that were correct at the beginning of the session). Thus, if a participant began a session with some Training Probe Items correctly produced, the number of correctly produced Training Probe Items was subtracted from 10 and the remaining items left to be trained for that session were considered equal to 100%. For example, if a participant began a session with 3/10 correct Training Probe Items, 7 items could be acquired and therefore 7 was used as the denominator to calculate acquisition; the numerator was calculated by subtracting the number of correctly produced Training Probe Items at the beginning of the session from the number of correctly produced Training Probe Items at the end of the session. All 10 Training Probe Items were always trained to keep the number of training items consistent among all sessions.[3]

Retention. Administration of the same 20 Training Probe Items at the beginning of the next consecutive session allowed for retention to be quantified; accuracy on these probes served as the dependent variable addressing Hypothesis 2. All 20 Training Probe Items were always administered at the beginning of each session; however, retention was only quantified for the 10 Training Probe Items from the previous session. Retention refers to performance levels after practice is completed; thus, the assumption of this study was that retention was attributed to the prior treatment session. Therefore, the percentage of the 10 Training Probe Items trained in the previous session that were correctly produced by participants at the beginning of the next session

[3] It was possible to get a negative acquisition rate: if a participant began the sessions with 2/10 Training Probe Items correct and then ended the session with 1/10 Training Probe Items correct. Then, the acquisition rate would be -1/8 (-12.5%).

were calculated to reflect retention[4].

The Training Probe Items always allowed for a 100% chance of retention as the same 10 Training Probe Items that were practiced during the previous session were administered at the beginning of the next consecutive session. Therefore, at the beginning of each session retention began at 0% and if 2 Training Probe Items were produced correctly then 20% acquisition was observed. The quantification of retention provided insight into whether ultrasound treatment versus no ultrasound treatment was more effective at facilitating retention of the Training Probe Items between two consecutive sessions.

Generalization. Administration of the 25 Generalization Probe Items at the beginning of each session allowed for generalization to be quantified; accuracy on these probes served as the dependent variable addressing Hypothesis 3. The Generalization Probe List included 25 monosyllabic words, multisyllabic words, phrases, and short sentences each containing one scored rhotic. The Generalization Probe Items included trained sound contexts as well as untrained sound contexts, but they did not contain any Training Probe List words, phrases, or short sentences. The following includes all sounds probed in the Generalization Probe List: /ɚ/, /ar/, /ɛr/, /ɝ/, /ɪr/, /ɔr/, /r/, /gr/, /dr/, /fr/, /pr/, /str/, /θr/, and /spr/ (see Appendix E for the complete Generalization Probe List).

Generalization was observed by comparing the percentage of correctly produced Generalization Probe Items between two successive sessions. The difference observed between two consecutive sessions accounted for the amount of generalization that occurred from the beginning of one session to the beginning of the next session. For example, if a participant

[4] While retention could have been measured immediately after practice was completed, for the purpose of this study, retention was observed by determining the number of trained items containing a target sound practiced during a session that the participant correctly produced at the beginning of the next consecutive session. This demonstrated the effects of practice remained with the participant for a short duration once practice was completed.

produced 16% of Generalization Probe Items correctly in session 3 and 28% of Generalization Probe Items correctly in session 4, this would indicate that a 12% generalization increase was attributed to the treatment that occurred in session 3 as the assumption was that generalization was attributed to the prior treatment session. This generalization data provided insight into whether ultrasound treatment versus no ultrasound treatment is more effective at facilitating generalization on untrained items between sessions.

Data Analysis

In the present study, the treatment was the independent variable which had two levels: treatment with ultrasound and treatment without ultrasound. There were three dependent variables: probe scores reflecting acquisition, retention, and generalization. Data analysis methods included (a) visual inspection of single-subject graphs, (b) effect size calculation (raw and standardized difference between the two treatment conditions by computing Cohen's d), and (c) statistical results from a randomization test (Kratochwill & Levin, 2014).

Descriptive statistics (i.e., means and standard deviations associated with each treatment condition) were used in the present study to complement visual inspections of single-subject graphs. Additionally, an effect size was computed because it represented a standard measure by which all outcomes could be assessed and described the difference in standard deviations between the means of US treatment and NoUS treatment. The effect size was determined in the present study by computing Cohen's d.

While the descriptive term of an effect size relies on research context and differs depending on the source, Cohen (1988) described a common conventional frame of reference when interpreting the magnitude of d used in the behavioral sciences where an effect of .2 corresponds to small, .5 to medium, and .8 to large (Zakzanis, 2001). More recently, Gierut, Morrisette, & Dickinson (2015) evaluated the effect size for single-subject design in treatment of children with functional phonological disorders by computing the Standard Mean Difference (All with Correction for Continuity) to measure the amount of generalization gain that accumulated longitudinally from treatment for each child. The results established benchmarks for interpretation of effect sizes (Standard Mean Difference) in that population as follows: 1.4 corresponds to small, 3.6 corresponds to medium, and 10.1 corresponds to large (Gierut et al., 2015).

Since the interpretations of effect size are subjective and field specific, neither the Zakzanis (2001) or the Gierut et al. (2015) studies examine effect size in a manner that is consistent with the present study. Designated benchmarks are often based on certain applications and the established referents may have little to no efficacy in other contexts; thus, effect sizes should be determined based on specific behaviors of interest and populations (Beeson &Robey, 2006; Gierut et al., 2015). The present study examines an effect size that takes place between consecutive sessions and reflects accuracy on a single phoneme. Since the efficacy of benchmarks used for interpretation for this context has not been established in current research, the present study will utilize interpretations of comparisons within this study only as a guide for assigning qualitative labels to the derived effect sizes.

The null hypothesis states that the two treatment conditions are equal. For single-subject experiments, the null hypothesis is that at each treatment time, the measurement provided by a subject would be the same, even if the alternative treatment were given. Therefore, the p-value may be computed in an attempt to reject the null hypothesis (Edgington, 1987). The test statistic was the mean difference between US sessions and NoUS sessions which was computed for each block (each week of therapy). A randomization test using all possible permutations of data was used to obtain p-values (significance values) conducted using the Single Case Randomization Test in the statistical software R (R Core Team, 2015). The p-value corresponding to the observed mean difference is obtained by ranking the obtained value in the distribution of all assignment possibilities (Bulte & Onghena, 2008). By taking all possible permutations of each data set and ordering each possible outcome, the results from the randomization test evaluated whether the observed advantage for one treatment over the other is likely due to the manipulation of the independent variables, not random chance. Two directional p-values per condition were computed: one p-value testing the assumption that the advantage was for the US session and the other p-value testing the assumption that the advantage was for the NoUS session. Each p-value was then compared with a critical value for significance testing where this study had a critical value of alpha=.05.[5] Given the exploratory nature of this single case study, multiple corrections for the alpha level were not used.

Participants

Four participants with a Speech Delay (Shriberg, 1994, 1997) were selected for the present study. Participants were between the ages of 7;0 to 9;7 at the time initial baseline data was collected. Participants were recruited through Institutional Review Board (IRB) approved flyers. It took approximately one hour to gather and administer the inclusionary information and assessments necessary to determine if a participant was initially eligible for the present study during the first baseline visit. The remaining testing administered to provide additional information on each participant at baseline took approximately an additional two and a half hours

[5] Randomized single-subject designs satisfy the requirement of random assignment; however, they do not satisfy the requirement of random sampling. Therefore, only tests valid in the absence of random sampling can be used when examining data. Randomization tests may be applied to data collected from single-subject experiments as long as randomization has occurred in the experiment. In randomized single-subject designs, the randomization is random assignment of treatment times (Edgington, 1987). Thus, the randomization tests were performed to obtain a test statistic which could be used to derive a significance level.

spread over 2-4 more visits depending on the random number of baseline sessions assigned to the participant.

All participants were native English speakers from mono-lingual English-speaking homes who agreed to be audio recorded as this is essential in tracking changes in speech over time. Parents of participants reported no present concerns of hearing loss and participants passed a hearing screening at 20 dB at 1, 2, and 4 kHz bilaterally using standard audiometric screening. Visual acuity was tested with the Snellen Eye Chart where participants' visual acuity in one eye was at least 20/40. In order to avoid unclear interpretations of the relations between treatment outcomes and cognitive abilities, participants' visual spatial perception and reasoning abilities were evaluated. Participants had to score no lower than -1 SD below the mean on the *Matrix Reasoning Index* score of the *Wechsler Abbreviated Scale of Intelligence* (*WASI*; Wechsler, 1999), indicating nonverbal cognition broadly within normal limits. Additional tasks used to evaluate participants are described below. Descriptive data for each participant can be found below in Table 1.

Table 1: Descriptive Data for Participants

	Participants			
	1003	1004	1008	1010
Age (years)	9;0	8;4	9;7	7;0
Intervention History	18 months-3 years; 2nd grade-present	3 years-present	7 years-present	3 years-present
Articulation				
GFTA-2 Standard Score/Percentile*	70/3	60/<1	73/1	73/7
Sentence Imitation Task percent /r/ correct*	6	0	15	0
Oral Language				
PPVT-4 Standard Score/Percentile*	103/58	96/39	111/77	106/66
CELF-4 Formulated Sentences Scaled Score/Percentile*	11/63	9/37	7/16	7/16
CELF-4 Recalling Sentences Scaled Score/Percentile*	12/75	12/75	8/25	9/37
Phonological Processing				
CTOPP-2 Elision Scaled Score/Percentile	9/37	10/50	5/5	9/37
CTOPP-2 Blending Scaled Score/Percentile	10/50	7/16	2/<1	11/63
Nonword Repetition Percent Phonemes Correct	83	75	82	79
SAILS Number Correct (of 100)	86	84	70	72
SAILS /r/ correct (of 20)	17	19	14	14
Motor Speech				
Multisyllabic Word Repetition PCC	83	75	71	69
Multisyllabic Word Repetition Percent Correct Lexical Stress	70	60	40	70
Emphatic Stress Score (Max=24)	24	10	15	20
Inconsistency Task Number of Different Tokens	1.5	1.25	1.875	1.875
Maximum Performance Task Apraxia Score	0	0	0	2
Maximum Performance Task Dysarthria Score	0	0	0	0
Syllable Repetition Task PCC	88	90	84	86
Nonverbal				
WASI-2 Matrix Reasoning T-score*	55	48	41	47
Stimulability				
Miccio /r/ Stimulability (Max=33)	13	0	18	0

*indicates inclusionary assessments

Note: Standard scores are normed with a mean of 100 and standard deviation (SD) of 15. Scaled scores are normed with a mean of 10 and SD of 3. T-scores have a mean of 50 and SD of 10. See Participants section for descriptions of tasks.

Articulation. *The Goldman-Fristoe Test of Articulation-2* (*GFTA-2*; Goldman & Fristoe, 2000) assesses all consonants of English and provides a standard measure of speech production errors based on the frequency of articulation errors. The sounds-in words section was administered to elicit the productions of specific speech sounds. In order to characterize participants as having a Speech Delay for the present study, participants needed a standard score of below 80 with a percentile rank below 8% on the *GFTA-2*. A Sentence Imitation Task (Preston et al., 2014) which included 15 sentences containing many later developing sounds (e.g., /r, s/) was administered. Participants had to score below 20% accuracy on the rhotic portion of this task to be included in the present study.

The Training Probe Items and Generalization Probe Items were also administered (by imitation) to collect baseline data during pre-treatment sessions and to confirm participants had errors on rhotic diphthongs before beginning speech therapy. Participants had to exhibit errors on rhotic diphthongs in at least two of the following contexts: /ɔr/, /ɪr/, /ɛr/, and /ar/. In order to characterize rhotic diphthongs as being in error and to qualify for the study, participants had to score below 30% accuracy in at least two of these rhotic diphthong contexts when the Training Probe Items and Generalization Probe Items were summed at the completion of baseline.

If possible, word lists designed to assess accuracy of other specific phonemes were also administered for two other sounds the participant had in error. If participants had sound(s) in error, the administration of the Untreated Sound Probe List(s) allowed for the untreated sounds to serve as a control when examining the effects of treatment. The Untreated Sound Probe List(s) were administered using imitation (please see Appendix F for an example of an Untreated Sound Probe List). However, if a participant did not have two other sounds in error, they were not disqualified from the study.

Oral language. In order to avoid unclear interpretations of the relations between treatment outcomes and language abilities, participants' language skills were assessed. To determine participants' language abilities were within normal limits for their age range, the following scores had to be achieved on the following tests: a standard score of 80 or greater on the *Peabody Picture Vocabulary Test-4* (*PPVT-4*; Dunn & Dunn, 2007) and a standard score greater than 7 on the *Recalling Sentences* and *Formulated Sentences* subtests on the *Clinical Evaluation of Language Fundamentals-5* (*CELF-5*; Wig, Semel, & Secord, 2013).

Phonological processing. To characterize the participants' phonological processing skills prior to the onset of therapy, the *Elision* and *Blending Words* subtests of *The Comprehensive Test of Phonological Processing* (*CTOPP-2*; Wagner, Torgesen, Rashotte, 2013) were administered. The *Speech Assessment and Interactive Learning System* (SAILS) (Rvachew, 1994) was also administered to assess participants' ability to evaluate productions of phonemes as either correct or incorrect. Participants listened to the production of a single word and were then asked whether the production was a "good" or "not good" way to produce the target word where "not good" rhotics included derhoticized and glided distortions. A total of 100 trials from the highest level of difficulty available in the software program were presented where 20 words were presented for each of the following phonemes: /r, s, θ, f, ʃ/.

To evaluate phonological working memory through assessing sequencing and production of pseudowords, a *Nonword Repetition Task* (Dollaghan & Campbell, 1998) was administered. During this task, participants were asked to repeat 16 previously recorded nonwords containing one, two, three, and four syllables. A Percent Consonants Correct (PCC) out of 96 total phonemes was computed from the phonetically transcribed responses of each participant. Participants were not excluded from the study based on phonological processing abilities.

Motor speech assessment. A number of motor speech assessments were administered to determine participants had a true SD and to exclude participants who were suspected of having Childhood Apraxia of Speech (CAS). The *Maximum Performance Tasks* (cf. Thoonen et al. 1996, 1997, 1999; Rvachew, Hodge & Ohberg, 2005) was administered in order to distinguish children who have Childhood Apraxia of Speech (CAS) or who are Dysarthric. During this task the length of the following sustained vowel and phonemes were assessed: /a/, fricatives including /f, s, z/, rapid productions for single syllables /pʌ/, /tʌ/, and /kʌ/, and rapid productions of the syllable sequences /pʌtʌkʌ/ and /mama/. Praat (Boersma & Weeninck, 2013) was used to obtain the duration of each sound or syllable. As described by Rvachew, Ohberg and Savage (2006), a score of 0 on the dysarthria scale represents "not dysarthric" and a score of 0 on the apraxia scale represents "not apraxic."

Other assessments administered to evaluate participants' motor speech abilities included: *Challenging Word Task* (e.g., 8 repetitions of 8 words) to assess token-to-token consistency; *Multisyllabic Word Repetition Task* (Preston & Edwards, 2007) to assess lexical stress, sequencing, and transitioning of words; and *Emphatic Stress Task* (cf. Shriberg et al., 2010) to elicit phrase-level stress. Participants' performance on all the above motor speech assessments was used to rule out motor speech impairment.

Exclusion criteria. Individuals with the following diagnosed disabilities were excluded from this study: developmental disabilities such as Autism or Cerebral Palsy; known or likely brain injury such as head trauma or meningitis; underlying structural or functional abnormalities such as cleft palate; known genetic syndromes such as Down syndrome; and severe psychiatric disorders such as Obsessive-Compulsive disorder or Tourette's syndrome. Individuals who were diagnosed or suspected of having CAS or being dysarthric during qualifying assessments were also excluded from the present study. These exclusion criteria were provided in order to avoid unclear interpretations of the relations between treatment outcomes and potential cognitive or motor abilities. No participants that were assessed during baseline were excluded from the present study.

Treatment

All four participants were treated by the same graduate student enrolled in the Speech-Language Pathology Master's Program at Syracuse University who was familiar with ultrasound treatment procedures. Baseline or follow-up sessions that included only testing were conducted by this student or by an ASHA certified speech-language pathologist who was also familiar with the procedures.

Baseline collection. Participants were randomly assigned a baseline number of 3-5. Three baseline sessions, lasting approximately one hour, were conducted at the Gebbie Clinic during initial qualification testing for participants. The first baseline session included the following: consent/assent forms completed by parents and participants; Background Information form completed by parents; hearing and vision screening completed by participants. The *WASI-2 Matrix Reasoning Index*, *PPVT-4*, *GFTA-2* and *CELF-5 Recalling Sentences* and *Formulated Sentences* subtests were administered to participants. The Baseline Training Probe List (please see Appendix G for Baseline Training Probe List), Generalization Probe List, and Untreated Sound Probe List(s) were administered to participants for collection of the first baseline data point.

Provided that participants met all eligibility criteria to be included in the study during the first baseline session, participants were scheduled to attend a second baseline session. During the second baseline session the following assessments were completed: The *CTOPP-2* (Wagner et al., 2013); *Maximum Performance Tasks* (cf. Thoonen et al. 1996, 1997, 1999; Rvachew et al., 2005); and a 50 utterance conversational speech sample to observe accuracy of speech sounds in conversational speech. The Baseline Training Probe List and Generalization Probe List were administered to participants for collection of the second baseline data point.

During the third baseline session, the following tasks were administered: *Challenging Word Task*; *Multisyllabic Word Repetition Task* (Preston & Edwards, 2007); *Nonword Repetition Task* (Dollaghan & Campbell, 1998); and *Emphatic Stress Task* (cf. Shriberg et al., 2010). Participants were also introduced to the ultrasound image display, and an informal analysis of tongue shape was completed by the clinician. An /r/ stimulability probe (adapted from Miccio, 2002) was administered to determine participants' rhotic stimulability. According to Miccio (2002), a child is considered stimulable for a sound if they can correctly imitate the sound target in 30% of contexts. Participants were asked to imitate target syllables containing rhotics three times each with the ultrasound beneath the chin. This further allowed the clinician to determine what type of tongue placement cues were necessary during future therapy sessions. The Baseline Training Probe List and Generalization Probe List were administered to participants for collection of the third baseline data point.

Depending on the number of baseline sessions randomly assigned to participants, the fourth baseline data collection point was gathered during the administration of the Baseline Training Probe List and Generalization Probe List during the beginning of the first treatment session before treatment began. If participants were assigned five baseline sessions, the fifth baseline was collected by having participants return for a separate baseline session where the ultrasound was introduced instead of introducing the ultrasound during the third baseline session. After the ultrasound was introduced, the Baseline Training Probe List and Generalization Probe

List were administered to participants for collection of the fourth baseline data point whereas the fifth baseline data point was still collected before the first therapy session. It should be noted that while generally most tests were administered during the correct baseline session, if a participant had difficulty attending to a task toward the end of a baseline session an assessment requiring less effort was given which may have been from a different baseline testing session or the task may have been saved for the next baseline session if possible.

Target Selection

The Baseline Training Probe List was administered during baseline data collection which contained five words for each of the following /Vr/ (vowel plus /r/) contexts occurring in any word position: /ɪr/, /ɔr/, /ar/, and /ɛr/. The two rhotic contexts to be trained during therapy were selected based on the participant's performance on the Baseline Training Probe List and the Generalization Probe List administered during baseline data collection. Two rhotic contexts were selected from the Baseline Training Probe Items and Generalization Probe Items a participant produced with 30% accuracy or lower when accuracy on all baseline probes was summed. It should be noted the final baseline data reported for acquisition and retention is reported as a percent correct (the average taken from four listeners) of the two rhotic contexts (out of 10) selected for training from the Baseline Training Probe List for each participant in order to demonstrate the level of accuracy for the two rhotic diphthongs chosen for training.

The Training Probe List administered at the beginning and end of sessions for each participant was then created by taking the 10 Baseline Training Probe Items for the two rhotic contexts selected for training in addition to 10 other words containing the target sounds. These 20 words were then divided into 10 Training Probe Items for US sessions and 10 Training Probe

Items for NoUS sessions. Given the changing nature of the Training Probe Items, participants were not always administered the same Training Probe List.

Treatment Sessions 1-14

Treatment structure. Participants attended two sessions a week for a total of 14 sessions over a seven week period (actual time frame was 7-8 weeks). Four participants were included in the present study. All sessions were conducted at Dr. Preston's Speech Production Lab. A Seemore PI 7.5 MHz probe or Echo Blaster 128 was used during all baseline, treatment, and follow-up sessions.

A blocked phase randomization scheme was used for this study. Each week involved two different treatment sessions: one US and one NoUS. The order of one US treatment session and one NoUS treatment session was randomly assigned to each week (see Appendix H for each participant's randomization schedule). US treatment and NoUS treatment were identical in every aspect, except for the use of ultrasound. Table 2 outlines US treatment and NoUS treatment.

Table 2: US and NoUS Treatment Schedule	
US Treatment Session	**NoUS Treatment Session**
4-6 minutes Generalization and Training Probe Master List	4-6 minutes Generalization and Training Probe Master List
2-3 minutes Untreated Sound Probe List(s) (completed every 2 sessions)	2-3 minutes Untreated Sound Probe List(s) (completed every 2 sessions)
Block A: 13 minutes therapy with *ultrasound*	Block A: 13 minutes therapy with *no ultrasound*
2 minute break	2 minute break
Block B: 13 minutes therapy with no ultrasound	Block B: 13 minutes therapy with no ultrasound
2 minute break	2 minute break
Block C Part I: 7 minutes therapy with *ultrasound*	Block C Part I: 7 minutes therapy with *no ultrasound*
Block C Part II: 6 minutes therapy with *ultrasound*	Block C Part II: 6 minutes therapy with *no ultrasound*
2-3 minutes Training Probe List	2-3 minutes Training Probe List

Treatment sessions were similar to those outlined by Preston et al. (2014, 2015). US treatment and NoUS treatment always began with administration of the randomized master list containing 45 Generalization Probe Items and Training Probe Items. Each session contained three 13 minute blocks of therapy (A, B, C). A timer was used to ensure allotted time for each block was met. However, participants finished the block of six trials they were in when the timer went off. Two target sounds (e.g., /ɪr, ar/) were practiced throughout the 14 sessions. Ten words were trained per session where each target sound was in five words. A different set of 10 words were trained during US sessions and NoUS sessions.

Following principles of The Challenge Point Framework (Guadagnoli & Lee, 2004; Hitchcock & McAllister-Byun, 2015) and guided by instructional steps for production training (Van Riper & Erickson, 1996**)**, each target sound was trained using the following principles of chaining to increase complexity. Chaining requires building from simple to complex stimuli; the sounds must be produced correctly 5/6 times in isolation to move to syllables; 5/6 times in syllables to move to monosyllabic words; 5/6 times in monosyllabic words to move to multisyllabic words; 5/6 times to move to the short phrase level; and 5/6 times at the short phrase level to move to the varied longer phrase level. Within each chain, multisyllabic words were defined as any word containing two or more syllables; short phrases were defined as phrases of 4-5 syllables that contained the target multisyllabic word; and varied longer phrases were defined as 6-7 syllable phrases that contained the target multisyllabic word. The majority of the first 10 chains used for Training Probe Items did not contain additional rhotics: if an additional rhotic did occur in any phrase, only the target rhotic was scored. Once participants reached a high level of success in therapy, additional rhotics may have appeared more frequently in some chains. During each block of therapy, participants practiced target sounds increasing in complexity when correct production criteria were met.

Guided by principles of motor learning, both US treatment and NoUS treatment used blocked practice intervals of six when establishing the target sound and training the target sound in syllables, monosyllabic words, multisyllabic words, and short phrases. When participants displayed high performance of the target sound and reached the varied long phrase level, variability was added by varying the context of the target sound. For example, if the target sound was /ɔr/ and the trained word was *four*, once the varied long phrase level was reached the following phrases or short sentences were included in training: *on channel twenty-four; she is twenty-four today; I bought twenty-four of those; did she get twenty-four; is he only twenty-four; I am in seat twenty-four*. This allowed for practice variability without changing the target word within a chain.

Elicitation. Regardless of if the session was US or NoUS, each session began in the elicitation phase during block A. Before the structured practice could begin the participant had to produce each of the two target sounds six times that were judged to be accurate by the clinician. During elicitation, to achieve six correct productions of each rhotic diphthong (e.g., /ar/ and /ɔr/), any amount and type of cueing or feedback was allowed (whereas during structured practice, the amount, type of feedback, and practice was predetermined). Therefore, a client may attempt 30 /ar/ sounds in a row with feedback on every trial in order to elicit the target sound six times. If a participant did not produce six accurately judged target sounds of each context by the clinician during block A, then the elicitation phase continued during the next blocks until this criteria was met. A participant was not allowed to leave the elicitation phase unless six accurately judged target sounds of both contexts were produced: eliciting six accurately judged target sounds of only one context was not enough to leave the elicitation phase. Therefore, it sometimes occurred that a participant had more than six accurately judged target sounds of only one context as the clinician continued to switch between training both contexts so the participant did not get discouraged by continuous unsuccessful productions on one of the contexts. It was possible for a participant to remain in the elicitation phase for the entire duration of a therapy session. However, irrespective of if the participant remained in the elicitation phase, the specifications regarding each block of an US or NoUS session were upheld (e.g., time duration, use of US/NoUS, and game allotment).

Condition differences. During blocks A and C of US treatment, 13 minutes with the ultrasound was provided where a tongue image display allowed the clinician to provide more explicit verbal KP feedback for participants and allowed participants to view visual feedback of tongue movements. The clinician instructed participants to use the visual display screen to move the front or back of the tongue in the appropriate directions to achieve correct production of the target sound. For example, the participant held the ultrasound probe below the chin and the clinician used the ultrasound image to cue the participant to change tongue positions (e.g., "bring your tongue into your throat"). Participants were given the option of either holding the ultrasound probe below the chin or using a stand for the ultrasound probe where they could rest the chin on the ultrasound probe.

During block B of US treatment sessions where no ultrasound was provided, drill-like activities were used where verbal tongue placement cues (e.g., "touch the sides of your tongue to the inside of your top teeth") were provided. Blocks A, B, and C of NoUS treatment were all structured as block B of US treatment. During block B in US treatment and NoUS treatment, a game requiring a turn of less than a few seconds (e.g., turn playing Angry Birds) was provided when necessary after two chains had been completed. In both US treatment and NoUS treatment sessions, two minute breaks were always provided after block A and after block B to provide motivation and allow participants a short time to relax. During these breaks, participants had a choice of playing a game on the computer or playing a short active game (e.g., ring toss).

During block C, random practice was provided during the last six minutes (see Table 1). Thus, block C in all US sessions and NoUS sessions was split into two parts where part I included typical chaining procedures and part II included random practice. However, if the participant did not move past the elicitation phase, elicitation continued during block C part II. If the participant did not move beyond the syllable level, random practice was provided at the syllable level only. If participants who had progressed past the syllable level in chains, random practice was provided by taking the highest point reached in that chain during training (i.e., monosyllabic, multisyllabic, short phrase, or varied long phrase) and randomly presenting all targets. A timer was set for seven minutes that included typical chaining procedures, the treating clinician then reset the timer for six minutes to include random practice.

If participants became restless during any block containing NoUS they were allowed to practice while standing up or playing movement games (e.g., take a step after five trials to reach the other side of the room, hop as far as you can after ten trials, etc.); however, no games requiring a stimulus were allowed during this time (e.g., computer games, ring toss, etc.). If participants became restless during an US block they were allowed to stand up and practice with the ultrasound; however, they were not allowed to move around the room given the limitations of the ultrasound. Restlessness was managed by the clinician during US and NoUS blocks by providing time reinforcement (e.g., only four minutes until break time) or games to inspire motivation (e.g., let’s see if we can get 15 new sounds in the next minute, let’s try for 20 more trials, etc.). When participants had to use the restroom while in the middle of a block, the timer was stopped and then started from the remaining time to ensure no practice time was lost while using the restroom.

Feedback schedules. The frequency of verbal KP feedback and verbal KR feedback was guided by the principles of motor learning. Immediate high frequency verbal KP and KR feedback was used when establishing the novel target sound at the syllable level. Due to the nature of feedback in speech therapy, it is difficult to provide KP feedback without providing KR feedback first; thus, verbal KP feedback was always provided with verbal KR feedback. However, verbal KR feedback was provided alone when higher levels were reached within a chain. The frequency of verbal KP feedback decreased as participants progressed to higher levels within a chain. Once participants reached the varied longer phrase level, no verbal KP feedback was provided; only verbal KR feedback was given. The total number of trials receiving verbal KP feedback and verbal KR feedback decreased as participants progressed to higher levels within a chain on a target word to promote self-monitoring. Feedback during random practice in block C part II was consistent with feedback provided at the short phrase level (33% KR and 17% KR and KP).

The treating clinician encouraged self-monitoring at all levels within a chain by randomly asking participants what they thought of a production (e.g., “new way” vs. “old way”). Regardless of the participants’ answer, the treating clinician always provided the necessary KP and/or KR for a trial. The percent of verbal KP feedback and verbal KR feedback provided for each level of training is shown in Table 3 (see Appendix I for Data Collection Form).

Table 3: US and NoUS Treatment Feedback Schedule

Feedback	**Syllable**	**Mono-syllabic Word**	**Multi-syllabic Word**	**Short Phrase Level**	**Varied Longer Phrase Level**	**Random Practice**
KR only	0% (0/6)	17% (1/6)	33% (2/6)	33% (2/6)	33% (2/6)	33% (2/6)
KR/KP	83% (5/6)	67% (4/6)	33% (2/6)	17% (1/6)	0% (0/6)	17% (1/6)
Total Feedback	83% (5/6)	83% (5/6)	67% (4/6)	50% (3/6)	33% (2/6)	50% (3/6)

Verbal KP feedback and verbal KR feedback were systematically provided for blocks A, B, C part 1, and C part II of all treatment sessions. Verbal KP feedback and verbal KR feedback were randomly assigned to each trial (e.g., KR and KP provided for first trial vs. only KR provided for first trial) within each block of six trials. Therefore, variation of feedback occurred while still ensuring that the frequency of verbal KR feedback and verbal KP feedback was met during each training block of six trials.

US treatment and NoUS treatment were identical in target selection methods, treatment duration, practice schedule, and feedback frequency. The only difference was the use of the ultrasound in US treatment which increased the specificity of KP feedback provided by allowing the clinician to provide specific tongue placement cues when using verbal KP feedback and allowing participants access to KP feedback that included visual feedback.

Post-treatment Follow-up Sessions

Participants were asked to return for two follow-up sessions. One session took place approximately one week after treatment ended to collect post-treatment data. The final Training Probe Items a participant was trained on were administered to evaluate retention from the previous session. The Generalization Probe List was also administered to observe generalization from the last session. The Untreated Sound Probe List(s) were administered to observe if progress was made on the untreated phonemes since the previous session.

Participants were asked to return approximately one month after treatment ended to collect additional post-treatment data. The Generalization Probe List and the final Training Probe List were administered to evaluate if long-term retention and generalization occurred within the past month (acquisition could not be observed during either follow-up session because treatment did not occur during these sessions). The Untreated Sound Probe List(s) were administered to observe any progress made on the untreated phonemes since the beginning of the study.

Treatment Reliability and Fidelity

A master list containing Generalization and Training Probe Items was always administered in random order at the beginning of each session. At the end of every session (except during the two follow-up sessions), only the Training Probe Items were administered. It occurred two times that a word from the Generalization and/or Training Probe List was not elicited and to account for this the denominator in the percentage calculation was adjusted. For example, if 1 of the 25 Generalization Probe Items was not elicited then the total number of Generalization Probe Items produced correctly for that day were taken out of 24 instead of 25. The Untreated Sound Probe List(s) were always administered during every other session starting with session 1 (and during both follow-up sessions). It occurred one time where one Untreated Sound Probe List was not administered and this error was noted in the participant's graph.

All scored Training Probe Items and Generalization Probe Items were scored point-by-point by four listeners (the treating clinician and three additional listeners who were unaware of treatment conditions) via audio recordings and averaged. All additional listeners were either ASHA certified speech-language pathologists or graduate students who were familiar with scoring rhotics. All listeners were provided with written guidelines on whether to score a rhotic as correct or incorrect. Listeners were instructed to still score rhotic diphthongs as correct if the following occurred: there was a pause during the transition between the vowel and the rhotic, but both the vowel and rhotic were still judged to be perceptually accurate, the voice quality of the rhotic diphthong was hoarse, but the target diphthong was judged to be perceptually accurate, and any additional error (e.g., omissions, substitutions, etc.) occurred in a syllable that did not contain the rhotic diphthong. Rhotic diphthongs were scored as incorrect if the following occurred: regardless of if the rhotic was perceptually judged as correct, the vowel of the rhotic diphthong was perceptually judged as incorrect, any error occurred in a syllable that contained the rhotic diphthong (e.g., addition, substitution, etc.), and too much pharyngeal constriction distorted the quality of the rhotic diphthong. In general, listeners were instructed to score "marginally correct" productions as incorrect.

A percentage of all correctly scored items was computed for each listener. The percentages of correctly scored items were then compared between listeners to determine the degree of agreement. The average of all four listeners' data for each probe was then used as the final data presented for the acquisition, retention, and generalization results.

To ensure all listeners' scores were in general agreement, an absolute value between all four listeners' scores was computed for acquisition, retention, and generalization for each participant (including baseline and follow-up items). All values from each participant were then averaged together to quantify the average value of the difference in scoring between each listener. Baseline and follow-up data scores were included in this calculation. The average pairwise absolute difference between all four pairs of listeners was 7.11%.

A timer was used to ensure that the duration of each block was approximately 13 minutes. Participants were allowed to finish the current block of six training items when the timer went off; thus, adherence to this time commitment was always controlled. Regardless if a client arrived late to a scheduled session, treatment was always provided for one hour sessions.

Additionally, treatment fidelity checks were performed by one graduate students in the Speech Pathology Program who had experience with scoring rhotics. This student reviewed two randomly selected sessions per participant (one in each treatment condition) via audio or video recordings to document adherence to the protocol with respect to percentage of verbal KR feedback and verbal KP feedback administered as well as agreement in rhotic scoring (KR feedback) during training in the chosen session. Fidelity checks were only selected from treatment sessions where the participant made it out of the elicitation phase since adherence to verbal KR feedback and KP feedback could not be scored since the feedback protocol was not followed during the elicitation phase. If a participant did not make it past the elicitation phase for each treatment condition then the fidelity check was completed for only one treatment condition.

A percentage of adherence to verbal KR feedback and KP feedback was calculated per trial of all eight sessions randomly chosen for fidelity checks across participants to ensure the treating clinician was adhering to the specified feedback protocol. The quantified verbal feedback as outlined by the feedback protocol was accurately supplied 96.27% of the time. Treatment fidelity checks also included a percentage of agreement in rhotic scoring calculated for every trial where explicit KR feedback was provided by the treating clinician. This ensured the treating clinician was in fact providing accurate reinforcement and diminishing true erred productions within a treatment session. Agreement on the appropriateness of KR feedback between the treating clinician and graduate student was 91.12%.

Moreover, for each session selected for the fidelity check the number of trials elicited from a participant during elicitation and training within a session was quantified and averaged across the eight sessions in order to demonstrate the numerous rhotic trials elicited from a participant during each session. The average number of rhotic trials elicited from a participant during each session was 215.

Results

The final data presented for the acquisition, retention, and generalization results were an average of the treating clinician and three additional listeners' data for each probe item given the specified time point. Results focused on the following data analyses: (a) descriptive statistics (i.e., means and standard deviations associated with each treatment condition) to complement visual inspections of single-subject graphs, (b) effect size calculation (raw and standardized difference between the two treatment conditions by computing Cohen's d), and (c) statistical results from a randomization test (Kratochwill & Levin, 2014). For the present study, a randomization test using all possible permutations of data was used to calculate p-values (significance values) conducted in R (R Core Team, 2015) where two directional p-values per condition were computed and then compared with a critical value of .05 for significance testing. Results are discussed using individual and summary analyses.

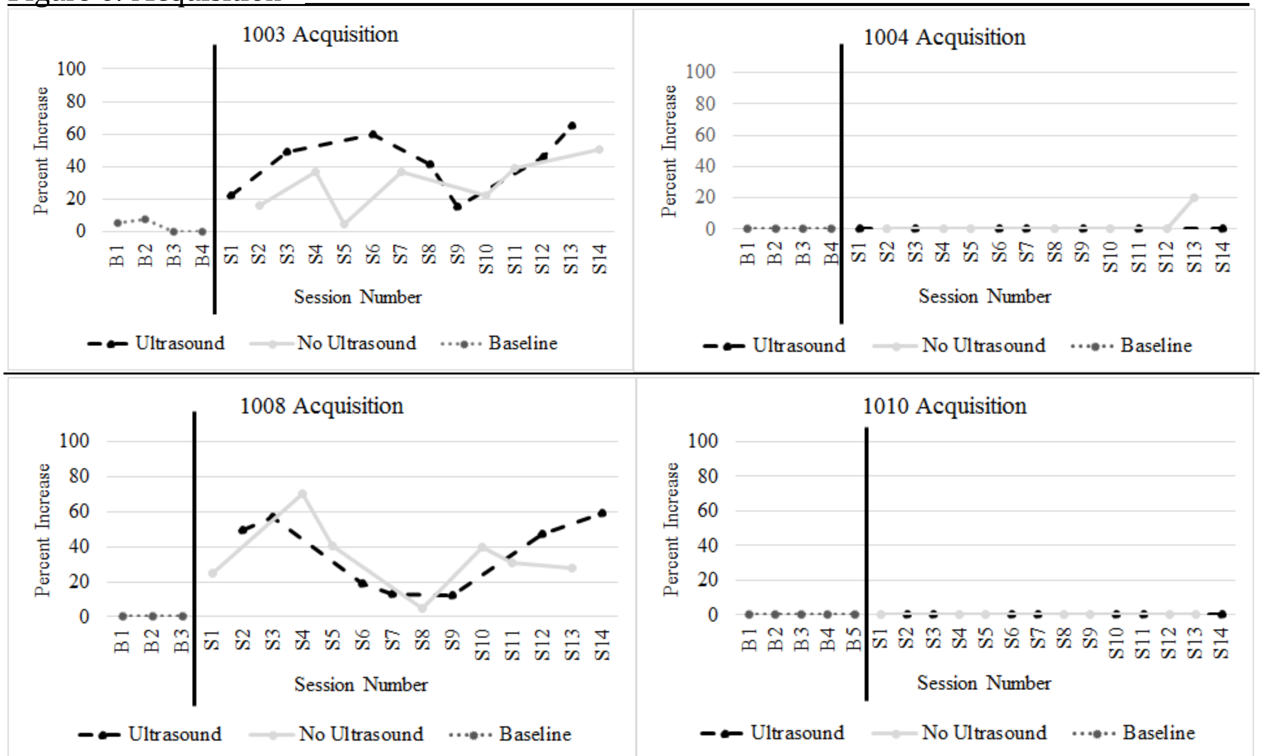
Acquisition

The administration of the Training Probe List during the beginning of the session and the end of the session allowed for acquisition to be quantified by computing the percent increase in accuracy from the beginning to the end of each session (i.e., acquisition rate). Figure 6 shows participants' percent increase of acquisition items (out of 100%) for all individual sessions.

Visual inspection of Figure 6 revealed most participants did not show a strong advantage for US sessions over NoUS sessions during acquisition. However, participant 1003's graph showed a generally consistent advantage for US sessions over NoUS sessions. Participant 1008's graph showed no consistent advantage for either US sessions or NoUS sessions; however, four of the seven blocks demonstrated a higher rate of acquisition with an US session.

Participant 1004 and 1010's graphs reveal both participants' acquisition rates primarily remained at 0%, demonstrating no advantage for either US sessions or NoUS sessions.

Figure 6: Acquisition



*B = Baseline; S = Session

**The baseline for Acquisition used in the above graphs is reported as a percent correct (the average taken from four listeners) of the two rhotic contexts chosen for training. Follow-up data was not able to be measured given the procedures required for measuring acquisition rate.

Descriptive statistics, effect sizes, and p-values were calculated to quantify and support the judgments made upon visual inspection. These results are shown below in Table 4. As randomization tests were computed for each individual, separate p-values were derived. Participant 1003 showed a significant advantage of US sessions over NoUS sessions in acquisition scores; however, the remaining three subjects did not show a significant advantage for either treatment.

Table 4: Acquisition

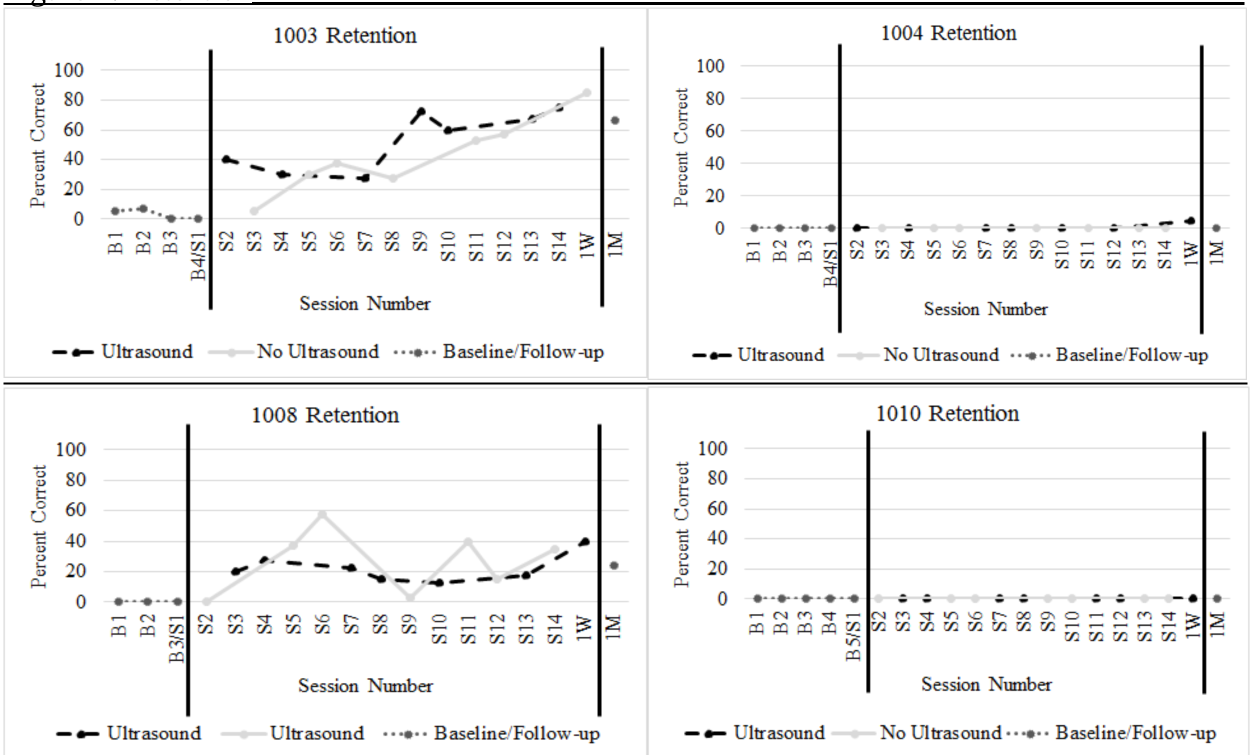
		Participants			
Condition	Calculation	1003	1004	1008	1010
US	$\bar{x}$ (SD)	42.6 (18.2)	0 (0)	36.8 (21.2)	0 (0)
NU	$\bar{x}$ (SD)	29.4 (15.6)	2.86 (7.56)	34.2 (19.8)	0 (0)
Difference	$\bar{x}_{US} - \bar{x}_{NoUS}$	13.2	-2.86	2.67	0
Effect Size (Cohen's d)	$d = \frac{\bar{x}_{US} - \bar{x}_{NoUS}}{SD_{pooled}}$	.783	-.756	.130	N/A
P-value (NoUS−US)	Randomization statistic	.969	.5	.492	1.0
P-value (US−NoUS)	Randomization statistic	.039*	1.0	.516	1.0

$\bar{x}$ = mean percent; SD = standard deviation; * denotes statistical significance

Retention

Retention was quantified by examining the percent of Training Probe Items produced correctly at the beginning of the next consecutive session. Individual participant data concerning retention are presented in Figure 7. Figure 7 shows participants' percent of retention (out of 100%) achieved for all individual sessions.

Figure 7: Retention



*B = Baseline; S = Session; 1W = 1 week follow-up session; 1M = 1 month follow-up session
**The baseline for retention used in the above graphs is reported as a percent correct (the average taken from four listeners) of the two rhotic contexts chosen for training.
***Follow-up data is reported as a percent correct of all trained items that comprised the last treatment session's Training Probe List (20) that were retained one month after treatment ended.
****The Training Probe List was administered at the beginning of the next consecutive session: thus, session 2 reflects the influence of an ultrasound/no ultrasound session that occurred in session 1.

Visual inspection of Figure 7 revealed there was a negligible difference between US sessions and NoUS sessions. Visual inspection of participant 1003's graph suggested an advantage for US sessions over NoUS sessions in four of the seven blocks when quantifying retention between two consecutive sessions. Further, visual inspection of participant 1008's graph revealed a trend for US sessions over NoUS sessions in four of the seven blocks. Graphs from participants 1004 and 1010 showed both participants' retention primarily remained at 0%: there was not an advantage for either US sessions or NoUS sessions.

Descriptive statistics, effect sizes, and p-values were calculated to quantify and support the judgments made upon visual inspection. These results are shown below in Table 5. Though visual inspection revealed a slight advantage for US sessions over NoUS sessions for participants 1003 and 1008, none of the participants showed a statistically significant advantage for one treatment over the other in retention scores.

Table 5: Retention

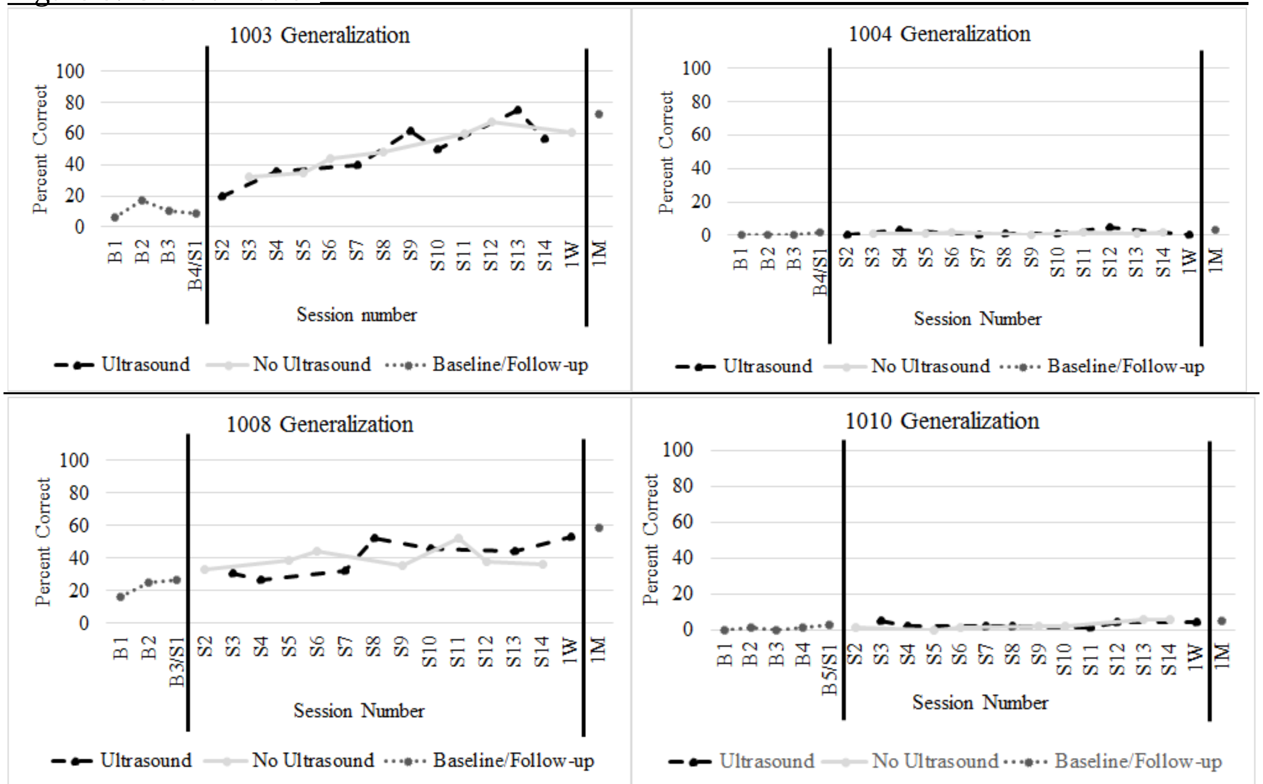
		Participants			
Condition	Calculation	1003	1004	1008	1010
US	$\bar{x}$ (SD)	53.2 (20.3)	0.71 (1.89)	22.1 (9.29)	0 (0)
NU	$\bar{x}$ (SD)	42.1 (25.6)	0 (0)	26.8 (21.4)	0 (0)
Difference	$\bar{x}_{US} - \bar{x}_{NoUS}$	11.1	0.71	-4.64	0
Effect Size	$d = \frac{\bar{x}_{US} - \bar{x}_{NoUS}}{SD_{pooled}}$	.482	.756	-.303	N/A
P-value (NoUS-US)	Randomization statistic	.875	1.0	.297	1.0
P-value (US-NoUS)	Randomization statistic	.172	.5	.711	1.0

$\bar{x}$ = mean percent; SD = standard deviation; * denotes statistical significance

Generalization

Generalization was quantified by examining the percent of Generalization Probe Items produced correctly at the beginning of each session. Individual participant data concerning generalization are presented in Figure 8. Figure 8 displays the percent correct of Generalization Probe Items (out of 100%) participants achieved for all individual sessions.

Figure 8: Generalization



*B = Baseline; S = Session; 1W = 1 week follow-up session; 1M = 1 month follow-up session
**The baseline and follow-up data for generalization used in the above graphs is reported as a percent correct (the average taken from four listeners) of the Generalization Probe List.
***The Generalization Probe List was administered at the beginning of the next consecutive session: thus, session 2 reflects the influence of an ultrasound/no ultrasound session that occurred in session 1.

Visual inspection of Figure 8 revealed no significant overall trend for an advantage for NoUS sessions or US sessions. Individually, participant 1003's graph revealed that generalization from both NoUS sessions and US sessions appeared to remain within close proximity of each other; however, a very minimal advantage for NoUS sessions over US sessions can be seen in four of the seven blocks. Participant 1008's graph further demonstrated a minimal advantage for NoUS sessions over US sessions in four of the seven blocks. Participants 1004 and 1010 made only negligible gains in generalization.

Descriptive statistics, effect sizes, and p-values were calculated to quantify and support the judgments made upon visual inspection. These results are shown below in Table 6. There was no statistically significant advantage of one treatment over the other in participants' generalization scores.

Table 6: Generalization

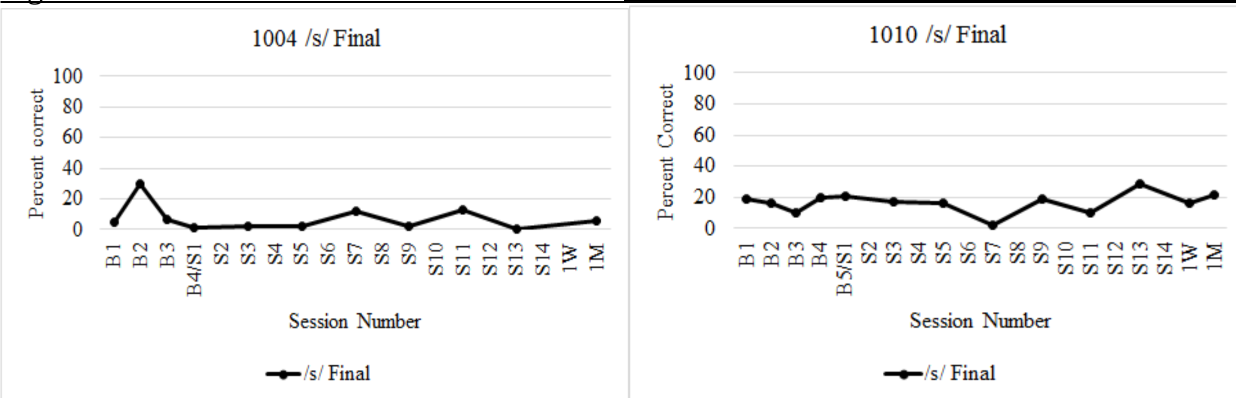
Condition	Calculation	Participants			
		1003	1004	1008	1010
US	$\bar{x}$ (SD)	48.6 (18.3)	1.43 (1.90)	40.7 (10.6)	2.86 (1.46)
NU	$\bar{x}$ (SD)	49.7 (13.7)	1.29 (0.76)	39.6 (6.46)	2.57 (2.44)
Mean Difference	$\bar{x}_{US} - \bar{x}_{NoUS}$	-1.14	0.14	1.08	0.29
Effect Size	$d = \frac{\bar{x}_{US} - \bar{x}_{NoUS}}{SD_{pooled}}$	-.071	.107	.127	0.15
P-value (NoUS-US)	Randomization statistic	.383	.570	.594	.688
P-value(US-NoUS)	Randomization statistic	.648	.5	.422	.344

$\bar{x}$ = mean percent; SD = standard deviation; * denotes statistical significance

Untreated Sound Probe List

Two participants had one non-rhotic sound in error: this allowed for an Untreated Sound Probe List to be administered. The Untreated Sound Probe Items served as a means to provide additional experimental control for the study. The Untreated Sound Probe List was administered during all baseline and follow-up sessions and during every other treatment session beginning with treatment session 1. Participants 1004 and 1010 both had final /s/ distortions; thus, the Untreated Sound Probe List administered to both participants was final /s/. Individual participant data concerning final /s/ are presented in Figure 9. Figure 9 displays the percent correct of Untreated Sound Probe Items (out of 100%) participants achieved for individual sessions.

Figure 9: Final /s/ as Untreated Sound Probe List



*B = Baseline; S = Session; 1W = 1 week follow-up session; 1M = 1 month follow-up session
**The baseline and follow-up data for the Untreated Sound Probe List used in the above graphs is reported as a percent correct (the average taken from four listeners) of the same probe list administered throughout treatment.
***Due to an administration error, there is no data point on 1004's graph for the 1 week follow-up session.

Visual inspection of Figure 9 revealed the accuracy achieved on the Untreated Sound Probe List for participants 1004 and 1010 remained consistent during treatment. Thus, variables that may have affected a participant's performance such as maturation effects were minimal when examining participants' accuracy achieved on the Untreated Sound Probe List.

Discussion

This study compared effects of US treatment and NoUS treatment for 7-9 year olds with rhotic distortions. Hypotheses derived from motor learning theory predicted the use of ultrasound as KP feedback would facilitate acquisition within a session; however, the use of ultrasound as KP feedback would hinder motor learning between consecutive sessions. A relatively novel single-subject design, randomized block design (Kratochwill and Levin, 2014) was used to evaluate participants' acquisition of rhotics within speech therapy while evaluating generalization and retention between consecutive sessions. The addition of randomization to the study allowed for a variety of data-analysis strategies to be implemented including effect size and statistical significance (Rvachew, 1988).

Acquisition

The acquisition rate served as the dependent variable to address Hypothesis 1: frequent exposure to ultrasound as visual feedback of the tongue will better facilitate acquisition of rhotic diphthongs compared to no ultrasound within treatment sessions. Participant 1003 showed a statistically significant advantage for US treatment over NoUS treatment relating to acquisition within a session. Although participant 1008 displayed a small advantage for US treatment over NoUS treatment relating to acquisition within a session, the advantage was not statistically significant. When comparing effect sizes across participants, participant 1003 demonstrated a larger effect size for US sessions over NoUS sessions with d=.78, while participant 1008's effect size for US sessions over NoUS sessions was relatively smaller at d=.13. It should be noted that participant 1004's effect size was inflated due to a small standard deviation as a minimal acquisition increase was observed in only 1 of 14 sessions; therefore, participant 1004's effect size cannot reliably be qualified. Neither participant 1004 nor 1010 made notable progress in acquisition throughout the 14 treatment sessions.

Consequently, Hypothesis 1 was true for only one of the four participants. Therefore, these findings provided some support for the observation derived from motor learning theory: KP feedback facilitates acquisition. In non-speech tasks that are not well established KP feedback may facilitate the rate of acquisition and overall performance levels (Newell et al. 1990); additionally, the present study suggested KP feedback may be useful for some individuals in facilitating acquisition of speech tasks.

Retention

Retention served as the dependent variable to address Hypothesis 2: frequent exposure to ultrasound as visual feedback of the tongue may impede retention of rhotic diphthongs compared to no ultrasound treatment. No participants showed a statistically significant result to support Hypothesis 2, and visual inspection of the data also supported the conclusion that there were no differences in retention between the two treatments. While the results were not statistically significant, participants 1003 and 1008 visually demonstrated a slight advantage for higher retention rates in more US sessions than NoUS sessions. Neither participant 1004 or 1010 made notable progress relating to retention throughout the 14 treatment sessions. When the effect sizes were examined, participant 1003 showed a medium effect size difference of d=.48 for US sessions over NoUS sessions. Participant 1008 showed a small effect size difference of d=-.30 for NoUS sessions over US sessions. Again, participant 1004's retention rates were negligible making a comparison between conditions difficult.

It was notable that participant 1003, who showed a strong advantage for US treatment during acquisition, also showed the strongest (but not statistically significant) advantage for US treatment in retention. Concerning non-speech tasks, an increase in KP feedback may impede motor learning (Hodges & Franks, 2001); however, participant 1003 demonstrated that an increase in KP feedback facilitated motor learning by moderately increasing retention between consecutive sessions. Thus, concerning participant 1003, increased KP feedback did not impede motor learning. Similar findings were observed in participant 1008 who did not demonstrate a decline in motor learning due to increased KP feedback. Further, this may suggest that an increased rate of acquisition due to US treatment may also facilitate (and/or may not hinder) retention. Additional studies with larger sample sizes are needed to further examine this claim.

Generalization

Accuracy on the Generalization Probe List served as the dependent variable addressing Hypothesis 3: frequent exposure to ultrasound as visual feedback of the tongue may impede generalization of rhotic diphthongs compared to no ultrasound between treatment sessions. There were no statistically significant results to support Hypothesis 3 and visual inspection of the graphs revealed similar generalization rates associated with the two treatments. However, pertaining to speech tasks, this observation contradicted the suggestion that an increase in KP feedback may impede motor learning as similarly found in non-speech tasks (Hodges and Franks, 2001). The means of Generalization Probe Items produced correctly during US sessions and NoUS sessions showed very little variability for all participants. Further, the standard deviations observed for US sessions compared to NoUS sessions were similar in value for all participants, demonstrating neither US sessions nor NoUS sessions contributed to more variability in generalization.

Visual analysis did not reveal that an increase in KP feedback using US treatment impeded motor learning when compared to NoUS treatment. The observed effect sizes for each participant further support this claim as all effect sizes were qualified as small: three participants showed a small effect for US sessions over NoUS sessions and one participant showed a small effect for NoUS sessions over US sessions.

Pre/Post Data

The effectiveness of treatment that includes an ultrasound as visual feedback component can be examined as a whole by using pretreatment and posttreatment data. Figure 10 displays the average of all pre-treatment data collected at baseline for the Generalization Probe Items and the data collected at the 1 month follow-up.

Figure 10: Generalization Data Comparison

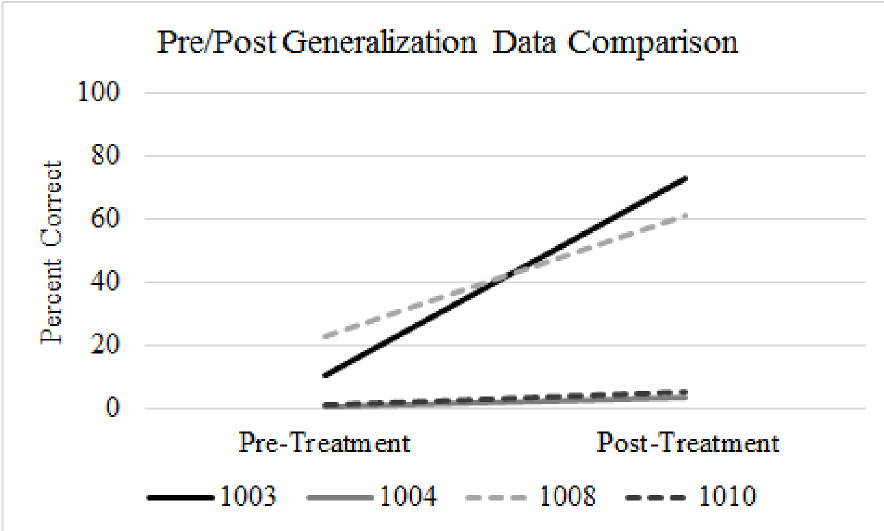


Figure 10 demonstrated the increase in motor learning measured by the overall increase in the percent of Generalization Probe Items from baseline to the one-month follow-up for all participants. Thus, the treatment as a whole, which contained sessions with and without ultrasound visual feedback, was successful for participants 1003 and 1008 as they both made noticeable gains in correct productions of untrained rhotic sounds over 14 sessions that remained one-month after therapy ceased. Participants 1004 and 1010 did not establish the target sounds in isolation until the second half of therapy; thus, acquisition performance at the word level was not achieved. Subsequently, motor learning could not take place as acquisition must precede learning in order to maximize learning (retention and generalization) during speech therapy (Maas et al., 2008; Schmidt & Lee, 1999).

Additional Considerations

The use of a single-subject randomized block design replicated across four subjects provided multiple opportunities for the intervention to demonstrate an effect. However, treatment dose (i.e. fourteen sessions) may not have been enough for all participants to demonstrate an effect for the treatment. Both participants 1004 and 1010 demonstrated establishment of the trained sound in isolation within the last five treatment sessions. Thus, had the treatment dose been longer than 14 sessions, these participants may have revealed additional information concerning the rate of acquisition within US sessions and NoUS sessions. Further research evaluating the effectiveness of ultrasound visual feedback therapy given a larger dose is needed.

Both participants 1003 and 1008, who demonstrated a response to the treatment as a whole, were older than participants 1004 and 1010, who did not demonstrate a response to treatment. Thus, the following may be hypothesized from the above small-scale results: treatment containing an ultrasound as visual feedback component for rhotics is more effective for those age 9;0 and above. One speculated reason for this is that the older participants were better able to interpret the visual feedback the ultrasound provided and incorporate the feedback into their own rhotic production. Another reason may be that since rhotics are one of the last phonemes to be acquired in children when compared to all other English phonemes that therapy targeting rhotics will be more effective once participants have reached age 9;0 where it is generally accepted that the production of American English speech sounds should be mastered (Arlt & Goodban, 1976; Shriberg 1994; Smit et al., 1990; Templin, 1957). Further, if the sound should be mastered at age 9;0, but has yet to be acquired, perhaps the stimulability for rhotics may have at least improved with a more mature articulatory system in place at age 9;0 compared to 7;0.

Interestingly, both participants 1003 and 1008 were stimulable (i.e., minimal establishment) for rhotics in some contexts (e.g., initial syllable position) scoring above 30% on the Miccio stimulability probe. Thus, both participants who were stimulable for rhotic targets demonstrated enhanced acquisition rates with the use of the ultrasound. In contrast, prior to therapy, participants 1004 and 1010 demonstrated they were not stimulable (i.e., no establishment) for rhotics in any context, scoring 0% on the Miccio stimulability probe. Consequently, the acquisition rates for participants 1004 and 1010's sessions were consistently below 10%. Thus, stimulability may be an important factor when examining the success of acquisition within US and NoUS rhotic speech therapy where those who are stimulable for a target sound will experience a higher success rate during acquisition.

It is important to note that all participants did establish the target sounds in isolation during therapy. Even participants 1004 and 1010, who were not stimulable for any rhotic before therapy, were stimulable for certain rhotic contexts after the conclusion of therapy. Comparisons of participants' scores on the pre-therapy Miccio stimulability probe and scores on the 1 week follow-up Miccio stimulability probe revealed that all participants' scores increased by at least 6%. Thus, therapy as a whole that contains an ultrasound as visual feedback component was successful in establishing stimulability for participants. Ultrasound has been found to be particularly useful at the beginning of therapy to help establish the sound in error by allowing participants to view their tongue configuration during a correct rhotic production (Lee et al., 2015). Further research evaluating the effectiveness of when an ultrasound visual feedback component is introduced in therapy is needed.

Additionally, both participants 1004 and 1010 had other consistent, noticeable sounds in error, while 1003 and 1008 did not. Thus, perhaps it is noteworthy that participants who only had one sound in error experienced a higher success rate during treatment containing an ultrasound as visual feedback component. This could again be related to age or severity, as an older participant had more time to stabilize other sounds in error. Further research evaluating the effectiveness of ultrasound visual feedback therapy on differing age groups is needed.

Caveats and Limitations

The focus of this study was on quantifying acquisition, generalization, and retention for sessions that involved ultrasound compared to performance from sessions using no ultrasound for children ages 7-9 years. Therefore, a randomized within block single-subject design was used to avoid cumulative effects on one treatment type occurring three or more times together, and to control for rising trends in the data. However, this design was not able to control further for cumulative effects that may occur from effective treatment as a whole pertaining to US and NoUS treatment sessions. Thus, the assumption was that generalization was attributed to the prior treatment session, though it is possible that the combined effects of all prior treatment could influence generalization. This design did not allow for unbiased control of a cumulative effect. Still, effect size calculations from the Generalization Probe List and Training Probe List revealed how US treatment or NoUS treatment affected the acquisition rate within a session, as well as generalization and retention in the subsequent session. Group designs such as a randomized control trail would be required to overcome this limitation of multiple treatments being administered to an individual.

It is noteworthy that this study was primarily concerned with quantifying performance, retention, and generalization: performance was examined within treatment sessions comparing ultrasound visual feedback sessions versus those with no ultrasound; retention and generalization were examined between two consecutive treatment sessions wherein the use of ultrasound visual feedback versus no ultrasound in the previous session was compared. Therefore, while visual inspection of the collected data allowed for general conclusions to be drawn about the effectiveness of overall treatment containing US sessions and NoUS sessions, no conclusions could be drawn concerning the effectiveness of only US treatment or only NoUS treatment over the length of the study. Pre-to-post-treatment comparisons only allowed for evaluation of the treatment program as a whole.

Conclusion

This study allowed for the first comparisons to be drawn between treatment using ultrasound as visual feedback and treatment with no ultrasound: informing the field of speech-language pathology of important considerations concerning how exposure to ultrasound as visual feedback of the tongue facilitated acquisition of rhotic diphthongs compared to no ultrasound during treatment sessions and how frequent exposure to ultrasound as visual feedback of the tongue may impede retention and generalization of rhotic diphthongs. By addressing the three hypotheses derived from motor learning theory, this study built upon existing research to establish new findings in the efficacy of ultrasound that will support and guide future research. In particular, for some children, acquisition may be facilitated by ultrasound visual feedback, and there was no evidence that ultrasound visual feedback inhibited retention or generalization.

References:

Adler-Bock, M., Bernhardt, B. M., Gick, B., & Bacsfalvi, P. (2007). The use of ultrasound in remediation of North American English /r/ in 2 adolescents. *American Journal of Speech-Language Pathology*, 16, 128–139.

American Speech-Language-Hearing Association (2014). *Steps in the Process of Evidence-Based Practice.* Retrieved from http://www.asha.org/members/ebp/assessing.htm.

Aram, D. M., & Hall, N. E. (1989). Longitudinal follow-up of children with preschool communication disorders: Treatment implications. *School Psychology Review*, 18(4), 487-501.

Arlt, P. B., & Goodban, M. T. (1976). A comparative study of articulation acquisition as based on a study of 240 normals, aged three to six. *Language, Speech, and Hearing Services in Schools*, *7*(3), 173-180.

Austermann Hula, S. N., Robin, D. A., Maas, E., Ballard, K. J., & Schmidt, R. A. (2008). Effects of feedback frequency and timing on acquisition, retention, and transfer of speech skills in acquired apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 51(5), 1088-1113.

Bacsfalvi, P. (2010). Attaining the lingual components of /r/ with ultrasound for three adolescents with cochlear implants. *Canadian Journal of Speech Language Pathology and Audiology*, 34, 206–217.

Bacsfalvi, P., & Bernhardt, B.M. (2011). Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography. *Clinical Linguistics & Phonetics, 25*(11-12), 1034-1043.

Bacsfalvi, P., Bernhardt, B.M., & Gick, B. (2007). Electropalatography and ultrasound in vowel remediation for adolescents with hearing impairment. *International Journal of Speech-Language Pathology*, *9*(1), 36-45.

Baker, E., & McLeod, S. (2011). Evidence-based practice for children with speech sound disorders: Part 1 narrative review. *Language, Speech, and Hearing Services in Schools*, *42*(2), 102-139.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology review*, *16*(4), 161-169.

Bernhardt, B., Bacsfalvi, P., Adler-Bock, M., Shimizu, R., Cheney, A., Giesbrecht, N., . . . Radanov, B. (2008). Ultrasound as visual feedback in speech habilitation: Exploring consultative use in rural British Columbia, Canada. *Clinical Linguistics & Phonetics*, 22(2), 149–162.

Bernhardt, B., Bacsfalvi, P., Gick, B., Radanov, B., & Williams, R. (2005a). Exploring the Use of electropalatography and ultrasound in speech habilitation. *Revue d'orthophonie et d'audiologie*, 29 (4), 169-182.

Bernhardt, B., Gick, B., Bacsfalvi, P., & Alder-Bock, M. (2005b). Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics & Phonetics*, 19 (6/7), 605-617.

Bernhardt, B., Gick, B., Bacsfalvi, P., & Ashdown, J. (2003). Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners. *Clinical Linguistics & Phonetics*, 17(3), 199–216.

Bernthal, J.E., Bankson, N. W., & Flipsen Jr., P (6^{th}) (2009). *Articulation and Phonological Disorders*. Boston: Pearson Education, Inc.

Boersma, P., & Weeninck, D. (2013). Praat v 5.3.49: www.praat.org.

Boyce, S. E., Combs, S., & Rivera-Campos, A. (2011). Acoustic and articulatory characteristics of clinically resistant/r/. *The Journal of the Acoustical Society of America*, *129*(4), 2625-2625.

Bulte, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods*, 40, 467-478.

Cleland, J., Scobbie, J. M., & Wrench, A. A. (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics*, (0), 1-23.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dent, H., Gibbon, F., & Hardcastle, B. (1995). The application of electropalatography (EPG) to the remediation of speech disorders in school-aged children and young adults. *International Journal of Language & Communication Disorders*, *30*(2), 264-277.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1136-1146.

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis, MO: Elsevier Health Sciences.

Dunn, L. M. & Dunn D. M. (2007). *Peabody Picture Vocabulary Test* (4 ed.). Minneapolis, MN: Pearson.

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology*, *34*(4), 437.

Fey, M. E., Finestack, L. H., & Schwartz, R. G. (2009). Research and development in child language intervention: A five-phase model. In R. G. Schwartz (Ed.), *Handbook of child language disorders* (513-529). New York, NY: Psychology Press.

Fletcher, S. G. (1982). Medical electronics II: Seeing speech in real time: The deaf can now view tongue, jaw, and other vocal-tract movements on a CRT display. *Spectrum, IEEE*, *19*(4), 42-45.

Fletcher, S. G. (1983). New prospects for speech by the hearing impaired. *Speech and language: Advances in basic research and practice*, 9, 1-41.

Fletcher, S. G., Dagenais, P. A., & Critz-Crosby, P. (1991). Teaching consonants to profoundly hearing-impaired speakers using palatometry. *Journal of Speech, Language, and Hearing Research*, *34*(4), 929-943.

Fletcher, S. G., McCutcheon, M. J., & Wolf, M. B. (1975). Dynamic palatometry. *Journal of Speech, Language, and Hearing Research*, *18*(4), 812-819.

Gibbon, F., & Hardcastle, W. (1987). Articulatory description and treatment of “lateral/s/” using electropalatography: A case study. *International Journal of Language & Communication Disorders*, *22*(3), 203-217.

Gibbon, F., & Hardcastle, W. (1989). Deviant articulation in a cleft palate child following late repair of the hard palate: a description and remediation procedure using electropalatography (EPG). *Clinical linguistics & phonetics*, *3*(1), 93-110.

Gibbon, F. E., Stewart, F., Hardcastle, W. J., & Crampin, L. (1999). Widening access to electropalatography for children with persistent sound system disorders. *American Journal of Speech-Language Pathology*, 8, 319–334.

Gick, B., Bacsfalvi, P., Bernhardt, B. M., Oh, S., Stolar, S., & Wilson, I. (2007. A motor differentiation model for liquid substitutions in children's speech. In *Proceedings of Meetings on Acoustics*: Vol. 1, No. 1, p. 060003. http://dx.doi.org/10.1121/1.2951481.

Gick, B., Bernhardt, B., Bacsfalvi, P., & Wilson, I. (2004). Ultrasound imaging applications in second language acquisition. Unpublished manuscript.

Gick, B., Campbell, F., Oh, S., & Tamburri-Watt, L. (2006). Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. *Journal of Phonetics*, *34*(1), 49-72.

Gierut, J. A., Morrisette, M. L., & Dickinson, S. L. (2015). Effect Size for Single-Subject Design in Phonological Treatment. *Journal of Speech, Language, and Hearing Research*.

Goldman, R. & Fristoe, M (2000). *Goldman Fristoe Test of Articulation* (2 ed.). Minneapolis, MN: Pearson.

Gruber, F.A. (1999). Probability estimates and paths to consonant normalization in children with speech delay. *Journal of Speech, Language, and Hearing Research*, *42*, 448-459.

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of motor behavior*, *36*(2), 212-224.

Haskins Laboratories (2014). *Ultrasound Visual Feedback*. Retrieved from http://www.haskins.yale.edu/uvf/example_images.html.

Hitchcock, E. R., & Mcallister Byun, T. (2015). Enhancing generalisation in biofeedback intervention using the challenge point framework: A case study. *Clinical linguistics & phonetics*, *29*(1), 59-75.

Hodges, N.J., & Franks, I.M. (2001). Learning a coordination skill: Interactive effects of instruction and feedback. *Research Quarterly for Exercise and Sport, 72*(2), 132-142.

Hodson, B., & Paden, E. (1991). *Targeting intelligible speech: A phonological approach to remediation.* Austin, TX: Pro-Ed.

Klein, H. B., McAllister-Byun, T., Davidson, L., & Grigos, M. I. (2013). A Multidimensional Investigation of Children's /r/ Productions: Perceptual, Ultrasound, and Acoustic Measures. *American Journal of Speech-Language Pathology*, 22, 540-553.

Kratochwill, T. R. & Levin, J.R (eds.). (2014). *Single-Case Intervention Research*. Washington, DC: American Psychological Association.

Lai, Q., & Shea, C. H. (1998). Generalized motor program (GMP) learning: Effects of reduced frequency of knowledge of results and practice variability. *Journal of motor behavior*, *30*(1), 51-59.

Lai, Q., Shea, C. H., Wulf, G., & Wright, D. L. (2000). Optimizing generalized motor program and parameter learning. *Research quarterly for exercise and sport*, *71*(1), 10-24.

Lee, S. A. S., Wrench, A., & Sancibrian, S. (2015). How to Get Started with Ultrasound Technology for Treatment of Speech Sound Disorders. *SIG 5 Perspectives on Speech Science and Orofacial Disorders*, *25*(2), 66-80.

Lewis, B. A., Freebairn, L. A., & Taylor, H. G. (2000). Academic outcomes in children with histories of speech sound disorders. *Journal of Communication Disorders*, *33*(1), 11-30.

Lewis, B. A., Shriberg, L. D., Freebairn, L. A., Hansen, A. J., Stein, C. M., Taylor, H. G., & Iyengar, S. K. (2006). The genetic bases of speech sound disorders: Evidence from spoken and written language. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1294-1312.

Maas, E., Robin, D. A., Hula, S. N. A., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology*, *17*(3), 277-298.

McAllister-Byun, T., Hitchcock, E. R., & Swartz, M. T. (2014). Retroflex Versus Bunched in Treatment for Rhotic Misarticulation. Evidence From Ultrasound Biofeedback Intervention. *Journal of Speech, Language, and Hearing Research*, 1-15.

McAllister-Byun, T. M., & Hitchcock, E. R. (2012). Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation. *American Journal of Speech-Language Pathology*, *21*(3), 207-221.

McCormack, J., McLeod, S., McAllister, L., & Harrison, L. J. (2009). A systematic review of the association between childhood speech impairment and participation across the lifespan. *International Journal of Speech-Language Pathology*, *11*(2), 155-170.

McReynolds, L. V., & Thompson, C. K. (1986). Flexibility of Single-Subject Experimental Designs. Part I Review of the Basics of Single-Subject Designs. *Journal of Speech and Hearing Disorders*, *51*(3), 194-203.

Ménard, L., Aubin, J., Thibeault, M., & Richard, G. (2012). Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model. *Folia Phoniatrica et Logopaedica*, 64(2), 64–72.

Miccio, A. W. (2002). Clinical Problem Solving Assessment of Phonological Disorders. *American Journal of Speech-Language Pathology*, *11*(3), 221-229.

Modha, G., Bernhardt, B. M., Church, R., & Bacsfalvi, P. (2008). Case study using ultrasound to treat /r/. *International Journal of Language & Communication Disorders*, 43(3), 323–329.

Nathan, L., Stackhouse, J., Goulandris, N., & Snowling, M. J. (2004). The Development of Early Literacy Skills Among Children With Speech Difficulties: A Test of the Critical Age Hypothesis. *Journal of Speech, Language, and Hearing Research*, *47*(2), 377-391.

Newell, K., Carlton, M., & Antoniou, A. (1990). The interaction of criterion and feedback information in learning a drawing task. *Journal of Motor Behavior*, *22*(4), 536-552.

Preston, J.L., Brick, N., & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology, 22*(4), 627-643.

Preston, J. L., & Edwards, M. L. (2007). Phonological processing skills of adolescents with residual speech sound errors. *Language, speech, and hearing services in schools*, *38*(4), 297-308.

Preston, J. L., & Edwards, M. L. (2009). Speed and accuracy of rapid speech output by adolescents with residual speech sound errors including rhotics. *Clinical Linguistics & Phonetics*, 23(4), 301–318.

Preston, J.L., & Leaman, M. (2014). Ultrasound visual feedback for acquired apraxia of speech: A case report. *Aphasiology*, *28*(3), 278-295.

Preston, J. L., Maas, E., Whittle, J., Leece, M. C., & Mccabe, P. (2015). Limited acquisition and generalisation of rhotics with ultrasound visual feedback in childhood apraxia. *Clinical linguistics & phonetics*, (ahead-of-print), 1-17.

Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., & Maas, E. (2014). Ultrasound visual feedback treatment and practice variability for residual speech sound errors. *Journal of Speech, Language, and Hearing Research*.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robey, R. R. (2004). A five-phase model for clinical-outcome research. *Journal of communication* *disorders, 37*(5), 401-411.

Ruscello, D. (1995a). Speech appliances in the treatment of phonological disorders. *Journal of Communication Disorders*, 28, 331–353.

Ruscello, D. (1995b). Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders*, 28, 279–302.

Rvachew, S. (1988). Application of single subject randomization designs to communicative disorders research. *Human Communication Canada, 12*(4), 7-13.

Rvachew, S. (1994). Speech perception training can facilitate sound production learning. *Journal of Speech Language and Hearing Research, 37*, 347-357.

Rvachew, S., Hodge, M., & Ohberg, A. (2005). Obtaining and interpreting maximum performance tasks from children: A tutorial. *Journal of Speech Language Pathology and Audiology, 29*(4), 146.

Sax, M. R. (1972). A longitudinal study of articulation change. *Language, Speech, and Hearing Services in Schools, 3*(1), 41-48.

Schmidt, A. M. (2007). Evaluating a new clinical palatometry system. *International Journal of Speech-Language Pathology, 9*(1), 73-81.

Schmidt, A., & Beamer, J. (1998). EPG treatment for training Thai speakers of English. *Clinical Linguistics & Phonetics, 21*, 389-403.

Schmidt, R. A. & Lee, T. D. (1999). *Motor Control and Learning* (3rd ed.). Champaign, IL: Human Kinetics.

Secord, W.A., Boyce, S.E., Donohue, J.S., Fox, R.A., & Shine, R.E. (2007). *Eliciting sounds: Techniques and strategies for clinicians* (2nd ed.). Clifton Park, NY: Thomson Delmar Learning.

Shriberg, L. D. (1975). A response evocation program for /ɝ/. *Journal of Speech and Hearing Disorders,* 40, 92–105.

Shriberg, L. D. (1980). An intervention procedure for children with persistent /r/ errors. *Language, Speech, and Hearing Services in Schools,* 11, 102–110.

Shriberg, L. D. (2010). Childhood speech sound disorders: From postbehaviorism to the postgenomic era. In Paul, R. & Flipsen, P. (Eds.), *Speech sound disorders in children*, 1-34. San Diego, CA: Plural Publishing.

Shriberg, L. D. (1997). Developmental phonological disorders: One or many. In B. Hodson & M. Edwards (Eds.), *Perspectives in applied phonology* (pp. 105–127). Gaithersburg, MD: Aspen.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The Speech Disorders Classification System (SDCS) Extensions and Lifespan Reference Data. *Journal of Speech, Language, and Hearing Research*, *40*(4), 723-740.

Shriberg, L. D., Fourakis, M., Hall, S. D., Karlsson, H. B., Lohmeier, H. L., McSweeny, J. L., ... & Wilson, D. L. (2010). Extensions to the speech disorders classification system (SDCS). *Clinical linguistics & phonetics*, *24*(10), 795-824.

Shriberg, L. D., & Kwiatkowski, J. (1988). A follow-up study of children with phonologic disorders of unknown origin. *Journal of Speech and Hearing Disorders*, *53*(2), 144-155.

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

Shuster, L. I., Ruscello, D. M., & Smith, K. D. (1992). Evoking /r/ using visual feedback. *American Journal of Speech-Language Pathology*, 1(3), 29–34.

Shuster, L. I., Ruscello, D. M., & Toth, A. R. (1995). The use of visual feedback to elicit correct /r/. *American Journal of Speech-Language Pathology*, 4(2), 37–44.

Skelton, S. L., & Hagopian, A. L. (2014). Using Randomized Variable Practice in the Treatment of Childhood Apraxia of Speech. *American Journal of Speech-Language Pathology*, *23*(4), 599-611.

Smit, A. B. (1986). Ages of Speech Sound Acquisition Comparisons and Critiques of Several Normative Studies. *Language, Speech, and Hearing Services in Schools*, *17*(3), 175-186.

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa Articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders*, 55(4), 779–798.

Templin, M. (1957). Certain language skills in children: Their development and interrelationships (Monograph Series No. 26). *Minneapolis: University of Minnesota, The Institute of Child Welfare*.

Thoonen, G., Maassen, B., Gabreels, F., Schreuder, R., & Swart, B. (1997). Towards a standardised assessment procedure for developmental apraxia of speech. *International Journal of Language & Communication Disorders*, *32*(1), 37-60.

Thoonen, G., Maassen, B., Gabreels, F., & Schreuder, R. (1999). Validity of maximum performance tasks to diagnose motor speech disorders in children. *Clinical Linguistics & Phonetics*, 13, 1-23.

Thoonen, G., Maassen, B., Wit, J., Gabreels, F., & Schreuder, R. (1996). The integrated use of maximum performance tasks in differential diagnostic evaluations among children with motor speech disorders. *Clinical Linguistics & Phonetics*, *10*(4), 311-336.

Van Riper, R & Erickson, R. L. (1996). *Speech Correction: An Introduction to Speech Pathology and Audiology* (9th ed.). Boston, MA: Allyn and Bacon.

Volin, R. A. (1998). A relationship between stimulability and the efficacy of visual biofeedback in the training of a respiratory control task. *American Journal of Speech-Language Pathology*, *7*(1), 81-90.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *Comprehensive Test of Phonological Processing* (2 ed.). Austin, TX: Pro-ed.

Wechsler, D. (2011). *Wechsler abbreviated scales of intelligence* (2 ed.). Minneapolis, MN: Pearson.

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical Evaluation of Language Fundamentals* (5 ed.). Bloomington, MN: Pearson.

Zakzanis, K. K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of clinical neuropsychology*, *16*(7), 653-667.

Zharkova, N. (2013). Using Ultrasound to Quantify Tongue Shape and Movement Characteristics. *The cleft palate-craniofacial journal*, 50 (1), 76-81.

Appendix A
Training Probe List Sample

US words
/ar/ arm
/ar/ car
/ar/ far
/ar/ star
/ar/ par
/ɪr/ deer
/ɪr/ near
/ɪr/ fear
/ɪr/ tear
/ɪr/ gear

NoUS words
/ar/ mar
/ar/ tar
/ar/ jar
/ar/ spar
/ar/ ark
/ɪr/ clear
/ɪr/ cheer
/ɪr/mere
/ɪr/ hear
/ɪr/ peer

Appendix B

Training Probe Chains

/ar/ Ultrasound

1.	/ar/	car	cartoons	old Sunday cartoons	She saw new Monday cartoons. Watching funny cartoons. Beth hates watching cartoons. I laughed at the cartoons. The cartoons put us asleep. Cartoons begin at six.
2.	/ar/	par	partake	she will partake	She will partake in that. He will partake in this. She will partake, if you do. When you partake, she will. If you partake, I will. I will partake to win.
3.	/ar/	far	farmer	old farmer Joe	The farmer has many pigs. Farmer Jim has donkeys. The farmer will plow tonight. Farmer Bob has two cats. I know a nice farmer. Who is that young farmer?
4.	/ar/	arm	armload	she has an armload	She held onto an armload. They had an armload of things. His armload was too heavy. He has an armload of food. His armload of books fell. The pen was in his armload.
5.	/ar/	star	starter	small starter plate	He ate the starter plate. She wanted one starter plate. The starter plate had salad. No starter plate was given. Ben asked for a starter plate. The starter plate looked nice.

/ar/ No Ultrasound

1.	/ar/	jar	ajar	closet was ajar	The window was ajar. Did you leave the case ajar? Close the book that is ajar.

				Who left the box ajar? The oven was left ajar. Do not leave the chest ajar.
2. /ar/	ark	arcsine	find the arcsine	What arcsine did you get? The arcsine is negative. I did not know the arcsine. That arcsine was complex. What is the arcsine function? Did you get the arcsine?
3. /ar/	tar	guitar	yellow guitar	He is playing the guitar. The guitar is white and blue. Can you play the guitar? I bought a new guitar. The shop has one guitar. I asked to play the guitar.
4. /ar/	spar	Sparta	king of Sparta	Sparta used to be a place. I have not seen Sparta. The queen lived in Sparta. The movie shows Sparta. They made pots in Sparta. Sparta had a temple.
5. /ar/	mar	market	hectic market	The market had candy. The market was on Monday. Can we go to the market? Buy milk at the market. When does the market open? The market sells donuts.

/ɪr/ Ultrasound

1. /ɪr/	deer	deerfly	big black deerfly	I saw a huge deerfly. The boy found a deerfly. The deerfly sat on a leaf. Did you see the deerfly? I do not like that deerfly. How big is that deerfly?
2. /ɪr/	gear	headgear	ugly headgear	I do not have headgear. His headgear was yellow. She got headgear last night. Ben has old blue headgear.

					He will not use his headgear. He lost his new headgear.
3.	/ɪr/	tear	tearful	tearful weeping	She was sad and tearful. Why is she so tearful? He was mad and tearful. I was tearful today. I hate being tearful. No one was tearful at lunch.
4.	/ɪr/	near	engineer	gifted engineer	My mom is an engineer. The engineer made a plan. He paid the engineer. The engineer is late. The engineer built the pond. I saw the engineer.
5.	/ɪr/	fear	hemisphere	south hemisphere	hot, humid, south hemisphere Which hemisphere is that? Cold, snowy hemisphere Let's go to that hemisphere. Stay in the south hemisphere. That hemisphere is windy.

/ɪr/ No Ultrasound

1.	/ɪr/	Mere	meercat	one black meercat	I own a white meercat. Have you seen a meercat? lovely, fun, playful meercat Does the meercat live outside? The zoo has a meercat. I named my meercat Tim.
2.	/ɪr/	Clear	clearly	clearly happy	He was clearly sad to lose. Can you see clearly outside? The window is clearly clean. You can clearly tell time. Did she clearly see him? She was clearly upset.
3.	/ɪr/	Hear	hereby	I hereby state	I hereby state the law. You will hereby follow me. Changes are hereby allowed. I hereby quit the team. Sue hereby denies the claim. I hereby join the club.

4.	/ɪr/	Cheer	cheerlead	loudly cheerlead	Ten boys will cheerlead today. I did not cheerlead tonight. The coach watched them cheerlead. They had to cheerlead last night. You have to shout to cheerlead. I have fun when I cheerlead.
5.	/ɪr/	Peer	appear	quickly appear	She watched the ghost appear. Swiftly and quickly appear. I did not see him appear. Did she appear to be home? What did he appear to do? It did not appear so.

/ɛr/ Ultrasound

1.	/ɛr/	bear	barefoot	walking barefoot	I like being barefoot. I always walk barefoot. Don’t walk barefoot in the sand. Walk barefoot in the house. Only kids play barefoot. Jake was barefoot at school.
2.	/ɛr/	pear	compare	compare cool toys	The siblings compare outfits. Let’s compare the lunch boxes. I want to compare notebooks. Do not compare assignments. His mom will compare the shoes. Compare the penny and dime.
3.	/ɛr/	care	careful	be careful outside	My mom said be careful. You must be careful in class. Be careful when jogging. Take notice and be careful. Be careful when swinging. Slow down and be careful.
4.	/ɛr/	wear	swimwear	new pink swimwear	He loved his old swimwear. She bought new swimwear today. He lost his swimwear last night. What size is that swimwear? Put on the black swimwear. You need swimwear in the pool.
5.	/ɛr/	chair	highchair	old white highchair	Mom found an old highchair. I do not use a highchair.

				Jen sat in the highchair. Do you have a highchair? The highchair is for babies. The new highchair was yellow.

/ɛr/ No Ultrasound

1. /ɛr/	share	timeshare	Disney timeshare	Do you have a timeshare? They had to use the timeshare. The timeshare was neat and cool. Use the timeshare this month. Make sure to book the timeshare. When did you use the timeshare?
2. /ɛr/	fair	unfair	the unfair test	The test was so unfair. She took the unfair quiz. The baseball game was unfair. Do not be unfair to them. His mom was being unfair. Tom acted unfair to Ben.
3. /ɛr/	stair	staircase	the long staircase	The staircase was too steep. Do not fall on the staircase. Only walk up the staircase. It was a wooden staircase. Do not spill on the staircase. No one used the staircase.
4. ɛr/	mare	nightmare	daily nightmare	The nightmare was too odd. He had a nightmare today. I did not have a nightmare. His nightmare was happy. The nightmare woke him up. The nightmare was funny.
5. /ɛr/	dare	daredevil	foolish daredevil	The daredevil was silly. I know a daredevil. The daredevil has a dog. The daredevil took the stage. The daredevil has a cape. Did you see the daredevil?

/ɔr/ Ultrasound

1. /ɔr/	core	encore	standing encore	The encore was too loud. They will give one encore. Did they play an encore?

					How was the encore tonight? John asked for an encore. The band gave an encore.
2.	/ɔr/	shore	shoreline	sandy shoreline	Swim close to the shoreline. Do not leave the shoreline. The shoreline is sandy. I will tan by the shoreline. I will walk to the shoreline. The shoreline is nice and cool.
3.	/ɔr/	door	doormat	welcome doormat	They got a new doormat. The doormat was navy blue. The doormat said hello. We bought a pink doormat. Shoes go on the doormat. The old doormat was muddy.
4.	/ɔr/	sore	soreness	painful soreness	She has a bad soreness. The soreness did not leave. She felt the soreness daily. Annoying, stabbing, soreness His soreness was aching. Walking caused him soreness.
5.	/ɔr/	four	twenty-four	twenty-four minutes	On channel twenty-four She is twenty-four today. I bought twenty-four of those. Did she get twenty-four? Is he only twenty-four? I am in seat twenty-four?

/ɔr/ No Ultrasound

1.	/ɔr/	More	mortified	highly mortified	Pat and Bob feel mortified. I was not mortified. The kids felt mortified. Does he look mortified? Why was she mortified? I hate feeling mortified.
2.	/ɔr/	war	warship	blue steel warship	I saw the huge warship. The movie had a warship. The warship had a cool name. The game used a warship. The warship was too big.

					Have you seen a warship?
3.	/ɔr/	bore	boredom	long-lasting boredom	Boredom got the best of him. Boredom makes me feel sad. She was sitting in boredom. No end in sight boredom He sat filled with boredom. Keep busy to stop boredom.
4.	/ɔr/	store	bookstore	giant bookstore	When does the bookstore open? I love the new bookstore. The bookstore sells coffee. I study in the bookstore. I went to the bookstore. Lovely, quiet bookstore
5.	/ɔr/	floor	seafloor	sandy seafloor	Plants sit on the seafloor. I have not seen the seafloor. The seafloor only looks black. You can’t swim by the seafloor. The deep seafloor is clean. Fish live in the seafloor.

Appendix C
Target Sound Word Bank

/ar/ Ultrasound Word Bank

1.	/ar/	part	partly	partly finished	He partly finished cleaning. The partly shaded pool. That is only partly fixed. She was partly finished. The leaf turned partly yellow. The steak was partly cooked.
2.	/ar/	bar	spacebar	key the spacebar	Tap the spacebar to type. Both thumbs go on the spacebar. Do you see the spacebar? The spacebar can help you. The spacebar is all black. The spacebar keeps getting stuck.
3.	/ar/	harm	harmless	harmless puppy	The fly outside was harmless. He was big, but harmless. She watched the harmless film. She will pet the harmless cat. The harmless fish looked mad. The tall boy was harmless.

/ar/ No Ultrasound Word Bank

1.	/ar/	heart	heartbeat	steady heartbeat	He has a stable heartbeat. He felt his quick heartbeat. A heartbeat does not stop. She had a slow heartbeat. The man took his heartbeat. His heartbeat was loud and fast.
2.	/ar/	lark	skylark	singing skylark	The sweet song of a skylark. The skylark was yellow. Do you see the skylark? The skylark is in the bush. The skylark built a nest? The skylark flew away.
3.	/ar/	sharp	sharpness	sharpness of glass	The sharpness of his song. His mental sharpness was odd. The sharpness of his mind showed. Test his mental sharpness. The photo’s sharpness was key.

Intellectual sharpness

/ɪr/ Ultrasound Word Bank

1.	/ɪr/	leer	chandelier	fancy chandelier	The house has a chandelier. The chandelier was heavy. I got a new chandelier. The chandelier has diamonds. Did you buy a chandelier? He put up the chandelier.
2.	/ɪr/	spear	spearfish	ugly spearfish	Spearfish are in the lake. Have you eaten spearfish? I do not like spearfish. They caught a huge spearfish. Spearfish is for lunch today. I saw fifty spearfish.
3.	/ɪr/	beard	beardless	beardless old man	The beardless man ate alone. He liked being beardless. He was always beardless. Being beardless is awesome. Shaving made him beardless. The beardless man said hello.

/ɪr/ No Ultrasound Word Bank

1.	/ɪr/	weird	weirdness	absolute weirdness	The weirdness was silencing. His weirdness made it awkward. The book was full of weirdness. Total lack of weirdness Complete absence of weirdness. His weirdness was engaging.
2.	/ɪr/	smear	smearcase	stinky smearcase	Would you like some smearcase? The smearcase is too soft. I bought some new smearcase. Have you eaten the smearcase? The smearcase was homemade. Put the smearcase on toast.
3.	/ɪr/	year	midyear	midyear budget	Midyear evaluation When will it be the midyear? Midyear analysis The midyear is coming soon. Finalize the midyear task.

Hand in the midyear study.

/ɛr/ Ultrasound Word Bank

1.	/ɛr/	there	thereby	thereby announce	I thereby state my claim. I thereby absolve the dues. It is thereby concluded. I thereby give consent. Thereby hangs my fall coat. He thereby deceived me.
2.	/ɛr/	clair	éclair	vanilla éclair	He ate an éclair today. The éclair was tasty. I ate the éclair second. The éclair looked delicious. Have you had an éclair? Eat the éclair with coffee.
3.	/ɛr/	glare	aglare	somewhat aglare	The sky was aglare tonight. The window looked aglare. The noise made him stand aglare. His eyes were aglare with love. His speech was aglare with hope. The jewel was aglare with light.

/ɛr/ No Ultrasound Word Bank

1.	/ɛr/	tear	teardown	teardown inspection	She wanted the teardown. They bought the old teardown. The teardown was full of mold. The teardown was easy. The teardown site was close by. The teardown had to happen.
2.	/ɛr/	flare	flare-up	sudden flare-up	His asthma would flare-up. The headaches may flare-up. I had a flare-up today. That may cause a flare-up. I had an acute flare-up. The flare-up has not stopped.
3.	/ɛr/	scare	scarehead	back page scarehead	The scarehead was too small. The scarehead caught my eye. Put the scarehead up top. Type the scarehead in big font.

Did you see the scarehead?
How big is that scarehead?

/ɔr/ Ultrasound Word Bank

1.	/ɔr/	pore	polypore	polypore fungus	Stinky polypore fungus Polypore is a fungus. The polypore was huge. A polypore collection. The polypore smelled bad. The polypore looked odd.
2.	/ɔr/	corn	popcorn	salty popcorn	I love the smell of popcorn. Popcorn is the best snack. Please pass the bowl of popcorn. I like movie popcorn. Do you want some popcorn? The popcorn tastes fabulous.
3.	/ɔr/	sort	resort	resort the deck	You must resort the tokens. Resort the deck and don't cheat. Did you resort the books? I will do the resort. The resort will take place soon. Resort the deck and begin.

/ɔr/ No Ultrasound Word Bank

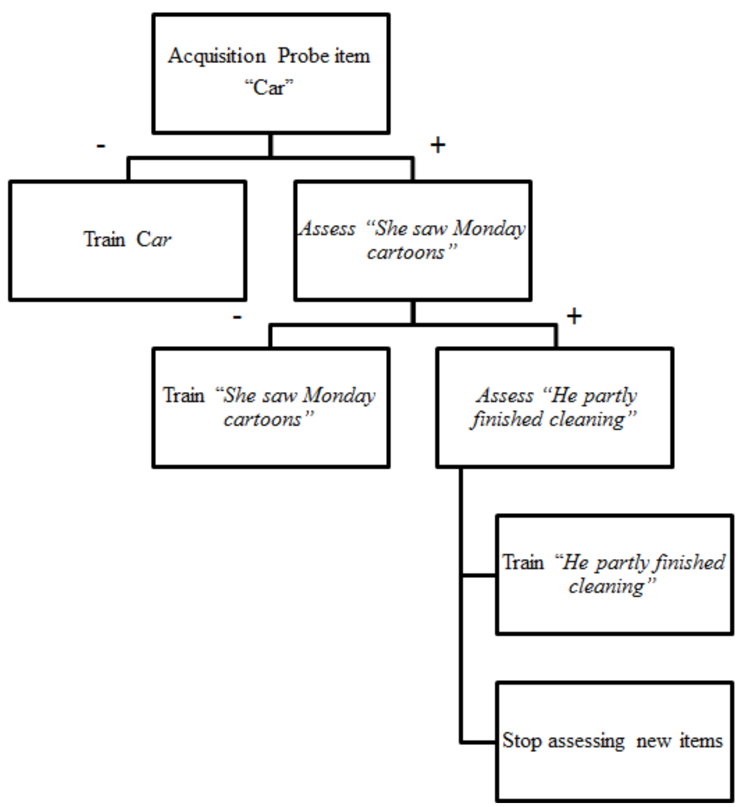
1.	/ɔr/	lore	galore	movies galore	You can buy gadgets galore. They have tablets galore. Did you see the shoes galore? They had sales galore today. We went swimming galore. We found tickets galore.
2.	/ɔr/	thorn	hawthorn	hawthorn bushes	I saw the big hawthorn bush. The hawthorn was in bloom. I planted one hawthorn. The hawthorn cut me today. The hawthorn looked lovely. She bought me a hawthorn.
3.	/ɔr/	Port	import	Spanish import	He liked Polish import food. The import candy went bad.

Please import the old files.
We will import the coffee.
You must pay the import tax.
I will import the photos.

Appendix D
Procedures for Adapting the Training Probe List

Each participant began the first therapy session with all monosyllabic Training Probe Items. When a monosyllabic Training Probe Item was produced correctly at the beginning of a session, a longer phrase containing 6-7 syllables (i.e., the highest level of a chain) was introduced as the Training Probe Item. When that phrase was produced correctly, a new 6-7 syllable phrase containing a different multisyllabic word that included the target sound was introduced. Up to three new phrases were introduced per Training Probe Item in order to find a phrase where the participant produced the target sound incorrectly. Figure 11 below displays the process used when selecting the appropriate Training Probe Item to be trained within a session given a participant's production.

Figure 11: Training Probe Item Selection Process



A Training Probe Item was only changed if it was produced correctly before a session where it was going to be trained; if the Training Probe Item was correctly produced before a session where it was not going to be trained, it was not changed. This structure helped to ensure a true measure of performance during acquisition occurred in each session as participants were not practicing sound targets in words or phrases that had already been correctly produced. However, if a participant produced *chair* (i.e., a trained word) correctly at the end of session 9, but incorrectly at the beginning of session 10, *chair* would still be trained during session 10 because although acquisition was reached in session 9, retention was not observed. Since acquisition and retention both precede generalization, retention must be achieved before moving onto a new target because learning is the ultimate goal of speech therapy. However, if a participant incorrectly produced *chair* (i.e., a trained word) at the end of session 9, but correctly produced chair at the beginning of session 10, *chair* would be replaced and not trained during session 10 since generalization of the trained word must have occurred between the sessions. When possible, Training Probe Items produced correctly at the beginning of a session would not be trained to help ensure each participant began every session at 0% accuracy on Training Probe Items.

However, a finite number of chains were available for each sound context. If a participant worked through all word bank Training Probe Items for a sound target, a phrase or short sentence training target they correctly produced at the beginning of a session was administered again as a Training Probe Item. Although a Training Probe Item had previously been correctly produced, it may have been used again as a training target if the participant produced it incorrectly at the beginning of a session. Up to three new phrases were introduced per Training Probe Item in order to find a phrase where the participant produced the target sound incorrectly and sometimes the three phrases included previously trained Training Probe Items.

Appendix E

Generalization Probe List

/ɚ/ doctor
/ɚ/ beautiful flower
/ar/ barn
/ar/ big yard
/ɛr/ arrow
/ɛr/ white hair
/ɝ/ birthday
/ɝ/ purple coat
/ɪr/ pioneer
/ɪr/ steering wheel
/ɔr/ score
/ɔr/ making s'mores
/r/ rice
/r/ red shoes
/gr/ sweet grapes
/dr/ drinking milk
/fr/ frying pan
/pr/ pretty sky
/str/ straw
/spr/ yellow sprinkles
/θr/ three
/r/ Tina loves rain.
/ɝ/ I got a fern.
/ɚ/ Superman is awesome.
/r/ The room is painted pink.

Appendix F

Untreated Sound Probe List for /s/ Final

1. Geese
2. Kiss
3. Face
4. Chess
5. Gas
6. Bus
7. Purse
8. Nurse
9. Boss
10. Moose
11. Voice
12. Dice
13. Rice
14. House
15. Police
16. Necklace
17. Walrus
18. Shoelace
19. Princess
20. Hourglass
21. Universe
22. Lacrosse
23. Octopus
24. Caboose
25. Tree house

Appendix G

Baseline Training Probe List

/ar/ car
/ar/ ark
/ar/ arm
/ar/ tar
/ar/ park
/ɪr/ deer
/ɪr/ gear
/ɪr/ tear
/ɪr/ near
/ɪr/ fear
/ɛr/ bear
/ɛr/ pear
/ɛr/ care
/ɛr/ wear
/ɛr/ chair
/ɔr/ corn
/ɔr/ shore
/ɔr/ door
/ɔr/ sore
/ɔr/ four

Appendix H

Randomization Schedule for Participants

Participant	Random Baseline #	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7
1003	4	US NoUS	US NoUS	NoUS US	NoUS US	US NoUS	NoUS US	US NoUS
1004	3	US NoUS	US NoUS	NoUS US	US NoUS	US NoUS	US NoUS	NoUS US
1005	5	NoUS US	US NoUS	NoUS US	US NoUS	NoUS US	US NoUS	NoUS US
1008	3	NoUS US	US NoUS	NoUS US	US NoUS	US NoUS	NoUS US	NoUS US

Appendix I

Data Collection Form (cf. Preston et al., 2014)

ELICITATION for S1 ______________________________ Block: __________

ELICITATION for S2 ______________________________

Syl.	**Feedback**	**Score**	**Monosyl. Wd**	**Feedback**	**Score**	**Multisyl. Wd**	**Feedback**	**Score**	**Short Phr**	**Feedback**	**Score**	**Var Long Phr**	**Feedback**	**Score**
S1	KR,KP		Mono1	KR,KP		Multi1			Phr 1			1		
S1	KR,KP		Mono1	KR,KP		Multi1	KR,KP		Phr 1	KR		2	KR	
S1			Mono1	KR,KP		Multi1	KR,KP		Phr 1	KR,KP		3	KR	
S1	KR,KP		Mono1	KR,KP		Multi1	KR		Phr 1			4		
S1	KR,KP		Mono1			Multi1			Phr 1	KR		5		
S1	KR,KP		Mono1	KR		Multi1	KR		Phr 1			6		

BIOGRAPHICAL DATA

NAME OF AUTHOR: Greta Marie Sjolie
PLACE OF BIRTH: Yankton, South Dakota
DATE OF BIRTH: March 16, 1989
UNDERGRADUATE SCHOOL ATTENDED:
Simpson College, Indianola, Iowa

DEGREE AWARDED:
Bachelor of Arts in English and Political Science, 2011, Simpson College

PROFESSIONAL EXPERIENCE:
Research Assistant, Speech Production Laboratory, Syracuse University