JSLHR

Research Article

# Treatment for Residual Rhotic Errors With High- and Low-Frequency Ultrasound Visual Feedback: A Single-Case Experimental Design

Jonathan L. Preston,[a,b] Tara McAllister,[c] Emily Phillips,[b] Suzanne Boyce,[d,b] Mark Tiede,[b] Jackie S. Kim,[e] and Douglas H. Whalen[f]

**Purpose:** The aim of this study was to explore how the frequency with which ultrasound visual feedback (UVF) is provided during speech therapy affects speech sound learning.
**Method:** Twelve children with residual speech errors affecting /ɹ/ participated in a multiple-baseline across-subjects design with 2 treatment conditions. One condition featured 8 hr of high-frequency UVF (HF; feedback on 89% of trials), whereas the other included 8 hr of lower-frequency UVF (LF; 44% of trials). The order of treatment conditions was counterbalanced across participants. All participants were treated on vocalic /ɹ/. Progress was tracked by measuring generalization on /ɹ/ in untreated words.
**Results:** After the 1st treatment phase, participants who received the HF condition outperformed those who received LF. At the end of the 2-phase treatment, within-participant comparisons showed variability across individual outcomes in both HF and LF conditions. However, a group level analysis of this small sample suggested that participants whose treatment order was HF–LF made larger gains than those whose treatment order was LF–HF.
**Conclusions:** The order HF–LF may represent a preferred order for UVF in speech therapy. This is consistent with empirical work and theoretical arguments suggesting that visual feedback may be particularly beneficial in the early stages of acquiring new speech targets.

Children with speech sound disorders exhibit errors on speech sounds that continue beyond the typical developmental window. Although most young children develop typical speech sound production by ages 8–9 years, a subset of children show persisting speech errors, typically substitutions or distortions of later developing sounds such as /ɹ, l, s, z, ʃ, and θ/. These unresolved speech sound disorders are often described as residual speech sound errors (RSEs). In American English, approximately 1% to 2% of high school- and college-aged individuals have unresolved RSEs (Culton, 1986; Flipsen, 2015), and distortion of /ɹ/ is among the most common errors (Shriberg, 2009). Even when impacts on intelligibility are minor, RSEs may compromise the naturalness or social acceptability of speech output, leading to negative social consequences (Crowne Hall, 1991; Silverman & Paulus, 1989) and impacts on socioemotional well-being (Hitchcock, Harel, & McAllister Byun, 2015). Importantly, some children with RSEs show limited progress with traditional articulatory treatment (McAllister Byun & Hitchcock, 2012; Shriberg, 1975), suggesting that it is critical to explore alternative treatment options. One approach that has been shown to be efficacious for some children with RSEs is ultrasound visual feedback (UVF), which allows for the display of tongue movements in real time (Adler-Bock, Bernhardt, Gick, & Bacsfalvi, 2007; Modha, Bernhardt, Church, & Bacsfalvi, 2008; Preston et al., 2014).

[a]Department of Communication Sciences and Disorders, Syracuse University, NY
[b]Haskins Laboratories, New Haven, CT
[c]Department of Communicative Sciences & Disorders, New York University, New York
[d]Department of Communication Sciences and Disorders, University of Cincinnati, OH
[e]Department of Communication Sciences and Disorders, Columbia University, New York, NY
[f]Program in Speech-Language-Hearing Sciences, City University of New York Graduate Center, New York

Correspondence to Jonathan L. Preston: jopresto@syr.edu

Editor-in-Chief: Julie Liss
Editor: Bharath Chandrasekaran

Received November 27, 2017
Revision received February 8, 2018
Accepted April 3, 2018
https://doi.org/10.1044/2018_JSLHR-S-17-0441

*Disclosure: The authors have declared that no competing interests existed at the time of publication.*

However, additional research is needed to optimize the intervention for maximally beneficial improvements in speech. In this study, therefore, we explore how the frequency and order with which visual articulatory feedback is delivered during speech therapy influence treatment outcomes for children with RSEs.

Children with RSE-/ɹ/, by definition, are beyond the typical age of acquisition for /ɹ/ (approximately 8–9 years; Smit, Hand, Freilinger, Bernthal, & Bird, 1990), but they have not achieved the proper articulatory configuration to produce an acoustically acceptable /ɹ/. The American English rhotic is considered especially challenging in speech sound acquisition because of its variability and complexity in articulation. Unlike other English speech sounds that require only one major lingual constriction or narrowing of the vocal tract, speakers must form two major lingual constrictions (anterior and posterior) to produce /ɹ/ (Alwan, Narayanan, & Haker, 1997; Delattre & Freeman, 1968; Klein, McAllister Byun, Davidson, & Grigos, 2013). Speakers use a range of tongue shapes for /ɹ/, and speech-language clinicians vary in the tongue shapes they cue (Ball, Müller, & Granese, 2013). These include shapes that have classically been termed *retroflex*, where the tongue tip raises near the alveolar ridge, and shapes that have classically been called *bunched*, where the tongue tip lowers while the anterior tongue body raises to approximate the hard palate. In addition, there are anatomic variations in tongue shape that are neither classically "bunched" nor "retroflex" but that combine some features of either (Boyce, 2015; Tiede, Boyce, Holland, & Choe, 2004). These shapes also are utilized differently across contexts, with some speakers using a particular tongue shape consistently across contexts and others using different tongue shapes in different phonetic or prosodic environments (Mielke, Baker, & Archangeli, 2016). This variability may contribute to the challenge of teaching an acoustically acceptable /ɹ/.

Traditional treatment approaches to elicit /ɹ/ commonly involve several techniques. An auditory model of correct /ɹ/ is often provided to the client to imitate. The clinician may also provide specific verbal cues encouraging the child to modify tongue shape and/or the location of vocal tract constrictions (e.g., Ruscello & Shelton, 1979; Secord, Boyce, Donohue, Fox, & Shine, 2007). In addition, attempts may be made to shape another phoneme such as /l/ into a perceptually acceptable /ɹ/ (e.g., Shriberg, 1975). Although these techniques are successful for some individuals, not all children respond to traditional treatment methods (e.g., McAllister Byun & Hitchcock, 2012; Shriberg, 1975). Traditional techniques may be limited because of the clinical challenges of verbally describing the relatively complex, visually concealed articulatory positions for /ɹ/ (Delattre & Freeman, 1968; Guenther et al., 1999; Tiede et al., 2004) and/or because of children's difficulties auditorily recognizing their errors (Shuster, 1998). However, when real-time articulatory information is available to both the clinician and the client, two advantages may apply: (a) A visual referent may enable more explicit instructions to be delivered by the clinician, and (b) the visual referent may provide an additional sensory modality to enable self-monitoring of articulatory movement by the client. UVF represents one noninvasive option to provide a visual display of articulation.

## UVF in Speech Therapy for /ɹ/

By holding an ultrasound transducer beneath the chin, real-time images can be generated to visualize the tongue's shape during speech (Preston, McAllister Byun, et al., 2017; Shawker & Sonies, 1985). Sagittal images can show the tongue from anterior (tip or blade) to posterior (root), depending on the field of view and the location of the sublingual cavity. Characteristics of an acoustically acceptable /ɹ/ in sagittal view may include elevation of the tongue tip or blade, lowering of the posterior tongue dorsum, and posterior movement of the tongue root toward the pharynx. For most acoustically acceptable productions, a groove in the tongue dorsum may also be apparent (Preston, McAllister Byun, et al., 2017). UVF may enhance the clinician's ability to recognize and describe to the child specific aspects of their /ɹ/ distortion, which may include a low tongue blade, high tongue dorsum, or a lack of tongue root retraction (Klein et al., 2013). As the child learns to recognize these articulatory components, they may be able to use the real-time visual display to self-monitor their articulation and modify their tongue movements for a more acceptable production.

To date, numerous case studies and single-subject experimental designs have suggested that UVF may facilitate improved speech sound accuracy for a variety of lingual phonemes (e.g., Cleland, Scobbie, & Wrench, 2015), although much research has focused on /ɹ/. In a case study, Adler-Bock et al. (2007) reported on two children with RSE-/ɹ/ who were treated with UVF over 14 sessions. From pretreatment to posttreatment, one child improved from approximately 2% to 64% accuracy, whereas another improved from 5% to 54%. Modha et al. (2008) reported a case study of a child with RSE-/ɹ/ who improved from 0% to 100% accuracy in only nine sessions. Single-subject experimental studies have likewise reported positive results in most, but not all, participants. For example, Preston and colleagues found that treating /ɹ/ with UVF was associated with an increased accuracy of untrained /ɹ/ words, with an improvement of approximately 35% over seven sessions (Preston, Leece, & Maas, 2017; Preston et al., 2014). Other research reported that UVF may be somewhat more effective when children with RSEs are permitted to explore variations in tongue shape to achieve an acceptable /ɹ/, rather than being required to aim for a prespecified shape (McAllister Byun, Hitchcock, & Swartz, 2014). UVF has also been used to remediate /ɹ/ distortions in individuals with speech sound errors associated with childhood apraxia of speech (Preston, Brick, & Landi, 2013; Preston, Leece, & Maas, 2016; Preston, Maas, Whittle, Leece, & McCabe, 2016) and hearing loss (Bacsfalvi, 2010;

Bacsfalvi & Bernhardt, 2011; Bacsfalvi, Bernhardt, & Gick, 2007).

### *Amount of Visual Feedback*

Variations in treatment outcomes may be explained, to some extent, by procedural differences in treatment delivery. For example, some studies have made UVF available for the entire duration of each session, typically around 30–60 min (McAllister Byun et al., 2014). Other studies have explicitly included blocks of practice with and without ultrasound. For example, Preston and colleagues (2014; Preston, Leece, & Maas, 2017) described procedures in which treatment sessions were subdivided into four 13-min blocks of practice; two of the four blocks featured UVF, resulting in UVF for 50% of the practice time.

In addition to obtaining empirical evidence of the efficacy of UVF, it is important to pursue a theoretical understanding of how some procedural differences in ultrasound biofeedback delivery may impact learning. Recent treatment research on UVF stems from the literature on the principles of motor learning (Bislick, Weir, Spencer, Kendall, & Yorkston, 2012; Maas et al., 2008). Within this framework, *skill acquisition*, the learner's preliminary success in achieving a motor skill through guided and structured practice, is distinguished from *skill learning*, the consequent retention and generalization of targeted skills to other contexts (Schmidt & Lee, 2011). The frequency and type of feedback provided by the treating clinician are parameters that may affect a client's success in acquiring and/or generalizing a new general motor plan (Maas et al., 2008; Schmidt & Lee, 2011). Thus, there is a theoretical reason to expect that the relative amount of visual feedback will influence speech motor learning.

With respect to feedback type, UVF may be characterized as a form of *knowledge of performance* feedback, which involves detailed information about the movements executed (Maas et al., 2008). This can be contrasted with *knowledge of results* feedback, which simply classifies a motor action as correct or incorrect. Knowledge of performance feedback provides additional information about how to achieve unfamiliar movement targets, which may assist in establishing a new generalized motor program, such as a different tongue configuration to produce /ɹ/ (Maas et al., 2008; Preston et al., 2014; Schmidt & Lee, 2011). There is empirical evidence suggesting that, when a motor pattern is unknown or is particularly complex, detailed feedback may aid in the rate of acquisition of that motor pattern in nonspeech tasks (Newell, Carlton, & Antoniou, 1990) and, at least for some individuals, in speech tasks (Sjolie, Leece, & Preston, 2016). However, increased dependence on external feedback has been reported to have neutral to negative long-term effects on the retention and/or generalization of learned skills (Hodges & Franks, 2001). Thus, excessive knowledge of performance feedback could hinder rather than facilitate later stages of speech motor learning.

In addition to the type of feedback that is available, the frequency with which feedback is provided may influence learning. Studies have reported that reduced feedback frequency can have a beneficial effect on speech motor learning (Austermann Hula, Robin, Maas, Ballard, & Schmidt, 2008; Steinhauer & Grayhack, 2000), perhaps because learners begin to rely more on intrinsic than extrinsic feedback. Maas, Butalla, and Farinella (2012) explored the influence of feedback frequency in a therapy program for four children with childhood apraxia of speech. Within-participant comparisons yielded mixed results, however, with some children showing greater gains with frequent feedback and other children showing greater gains under conditions with less feedback.

Although most previous studies have compared only single conditions of learning (e.g., high-frequency [HF] vs. low-frequency [LF] feedback), learning may also be viewed as a dynamic process in which the information presented to the learner should be adapted on the basis of their evolving level of ability (Maas et al., 2008). Both theoretical models of motor learning (Guadagnoli & Lee, 2004; Maas et al., 2008) and previous empirical evidence (McAllister Byun & Campbell, 2016) suggest that it may be optimal to provide frequent visual feedback early in the learning process and then withdraw the feedback. Thus, participants in the current study received both a period of UVF speech therapy in which visual feedback was made available in a preponderance of trials (HF of feedback) and a period featuring feedback in a smaller subset of trials (LF of feedback). The order in which HF and LF treatment conditions were provided was counterbalanced across participants to keep the overall proportion of feedback constant and create an opportunity to observe an effect of order on treatment response.

## Purpose and Hypotheses

The purpose of this study was to explore the differential effects of HF and LF UVF in the remediation of /ɹ/ distortions in children with RSEs. The following hypotheses were addressed: (a) Treatment that includes UVF will result in improvement in speech sound accuracy on untreated words; (b) On average, LF UVF will facilitate speech sound learning better than HF UVF because schema-based models of motor learning predict that generalization learning is best facilitated with less frequent feedback (Austermann Hula et al., 2008; Steinhauer & Grayhack, 2000); and (c) With the total amount of UVF held constant, the order HF–LF would better facilitate speech sound learning than LF–HF because the former provides the optimal alignment of detailed feedback with early phases of learning and reduced feedback with later stages.

## Method

### *Participant Characteristics*

The study included 12 native speakers of rhotic dialects of North American English with RSE-/ɹ/ between the ages of 8 and 16 years.[1] Children were referred by local speech-language pathologists (SLPs) or by parental response to flyers posted throughout the community. All participants had RSE-/ɹ/ in the absence of any identifiable etiology (such as Down syndrome, autism, or hearing loss). Children were required to pass a pure-tone hearing screening at 20 dB at 1000, 2000, and 4000 Hz to be eligible for the treatment study. Adequate receptive language scores as defined by standard scores above 80 on the Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007) were also prerequisites.

## Speech Assessments

To be eligible for the study, participants had to score below the seventh percentile on the Goldman-Fristoe Test of Articulation–Second Edition (Goldman & Fristoe, 2000). Participants were further assessed using researcher-developed probes to evaluate rhotic accuracy at the word level (50 words) and sentence level (five sentences), and they were required to score below 25% accuracy at the word level (on the basis of ratings from a certified SLP) to be eligible for the study. A brief conversational speech sample was also collected to confirm the presence of /ɹ/ errors. Additional sound errors are shown in Table 1.

A maximum performance task was administered to assess speech motor functioning. The procedures followed those outlined in previous studies (Rvachew, Hodge, & Ohberg, 2005; Thoonen, Maassen, Gabreels, & Schreuder, 1999; Thoonen, Maassen, Wit, Gabreels, & Schreuder, 1996). Duration measures were recorded for sustained phonemes /f/, /s/, /z/, and /ɑ/, and syllable rate was measured for repeated syllables /pɑ/, /tɑ/, and /kɑ/ and the syllable sequences /pɑtɑkɑ/; the accuracy of the /pɑtɑkɑ/ sequence was also scored. These measures were used to derive separate scores for apraxia (based on slow and inaccurate trisyllables) and dysarthria (based on a short duration of sustained phonemes or slow syllables), whereby 0 represents "not dysarthric/apraxic," 1 is "undefined," and 2 represents "probable dysarthria/apraxia." No participants received a score of 2 on either the dysarthria or apraxia scale.

Stimulability was assessed using a task adapted from Miccio (2002). Participants imitated 11 different syllables (e.g., /ɑɹ, ɹɑ, ɪɹ, ɹi/) three times each, for a total of 33 productions. Percent /ɹ/ correct was computed on the basis of ratings by a certified SLP.

## Additional Descriptive Assessments

Before treatment, children completed several language and cognitive assessments for descriptive purposes. These tasks included the Recalling Sentences and Formulated Sentences subtests of the Clinical Evaluation of Language Fundamentals–Fifth Edition (Wiig, Semel, & Secord, 2013), the Phonological Awareness subtests (Elision, Blending Words, and Phoneme Isolation) of the Comprehensive Test of Phonological Processing–Second Edition (Wagner, Torgesen, Rashotte, & Pearson, 2013), and the Matrix Reasoning subtest of the Wechsler Abbreviated Scales of Intelligence–Second Edition (Wechsler, 2011). Scores on these assessments are presented in Table 1.

## Study Design

This single-subject experimental study followed a multiple-baseline across-subjects design with a baseline phase, two treatment phases, and a maintenance phase after each treatment. Participants were randomly assigned to receive three, four, or five baseline probes, followed by Treatment Phase 1 (eight sessions), three midpoint probes, Treatment Phase 2 (eight sessions), and three maintenance probes. For participants who were available, follow-up probe data were collected 2 months after the final maintenance session to track continued progress. Each participant was exposed to two treatment conditions, with order counterbalanced across participants. Treatment Phases 1 and 2 were randomly assigned through concealed envelopes such that the order for six children was HF–LF and the order for the other six children was LF–HF. Sessions were scheduled to occur twice per week throughout the study. A manual was developed to guide the implementation of the study procedures and is freely available (Preston & McAllister, 2017).

### Condition Differences

Practice during each treatment session included 27 blocks, each consisting of six trials on vocalic /ɹ/ items. In the HF condition, visual feedback was made available in 24 of 27 blocks (89%), whereas in the LF condition, visual feedback was provided in only 12 of 27 blocks (44%).

### Probe Data

Each treatment session began and ended with administration of a 25-item probe that elicited each of the vocalic targets /ɝ/, /ɑɹ/, /ɔɹ/, /ɪɹ/, and /ɛɹ/[2] in five words apiece. These probes were used for monitoring progress and consisted of /ɹ/ words that were not treated. Words were presented in random order on a computer screen using conventional orthography. Probe data were recorded using a lapel microphone at a 44.1-kHz sampling rate. No verbal or visual feedback was provided during administration of probes. Change on this task was used as the primary outcome measure.

[1] One of the original 12 participants withdrew because of scheduling conflicts with extracurricular activities, with the result that no effect size could be calculated for this participant. Therefore, a new (13th) participant was recruited and treated, yielding a balanced set with complete data from 12 individuals.

[2] We follow the clinically common convention of classifying syllabic and postvocalic rhotics as "vocalic," in contrast with the "consonantal" rhotic in syllable-onset position.

**Table 1.** Participant characteristics.

| Participant | 111 | 112 | 113 | 114 | 115 | 119 | 122 | 124 | 125 | 126 | 127 | 128 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | M | M | M | M | M | M | M | M | M | M | M | M |
| Age (years;months) | 15;11 | 10;9 | 12;11 | 8;2 | 10;3 | 9;7 | 9;11 | 11;0 | 11;1 | 16;10 | 10;3 | 9;4 |
| GFTA-2 std score | < 40 | 83 | < 40 | 83 | 84 | 76 | 65 | 72 | 83 | 80 | 81 | 81 |
| GFTA-2 percentile | < 1 | 5 | < 1 | 4 | 5 | 4 | < 1 | 2 | 2 | 3 | 4 | 6 |
| PPVT-4 std score | 99 | 114 | 119 | 120 | 131 | 149 | 100 | 135 | 114 | 152 | 152 | 99 |
| CELF-5 RS scaled score | 10 | 9 | 9 | 13 | 12 | 17 | 13 | 8 | 9 | 17 | 10 | 15 |
| CELF-5 FS scaled score | 9 | 10 | 9 | 7 | 15 | 16 | 13 | 11 | 8 | 11 | 8 | 11 |
| CTOPP-2 PA composite | 62 | 90 | 96 | 92 | 86 | 98 | 90 | 128 | 77 | 100 | 98 | 96 |
| WASI-2 MR *T* score | 20 | 47 | 49 | 56 | 63 | 52 | 33 | 64 | 54 | 74 | 36 | 41 |
| Max performance task dysarthria score | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Max performance task apraxia score | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Stimulability | 0% | 0% | 0% | 54.5% | 0% | 0% | 0% | 0% | 66.5% | 21% | 0% | 0% |
| Ages at which child received speech therapy | 6 months to present | 2 to present | None | None | 5 to present | 2 to present | 5 to present | 5 to present | 3 to present | 11–12 | 4 and 6 | 5 to present |
| Additional sound errors | /s, z/ | None | /s, z/ | None | None | None | /l/ | None | None | /s/ | None | None |

*Note*. Standard scores have a mean of 100 and a standard deviation of 15. Scaled scores have a mean of 10 and a standard deviation of 3. Stimulability is the percentage of correct /r/ productions during 33 imitative attempts. M = male; GFTA-2 = Goldman-Fristoe Test of Articulation–Second Edition; std = standard; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition; CELF-5 = Clinical Evaluation of Language Fundamentals–Fifth Edition; RS = Recalling Sentences; FS = Formulated Sentences; CTOPP-2 = Comprehensive Test of Phonological Processing–Second Edition; PA = Phonological Awareness; WASI-2 = Wechsler Abbreviated Scales of Intelligence–Second Edition; MR = Matrix Reasoning; Max = maximum.

## *Treatment Procedure*

One of two certified SLPs conducted all sessions. No participants received additional speech therapy for /ɹ/ at the time of the study. Audio and video recordings of the ultrasound images were collected for each session.

### Target Selection

Treatment targets were selected on the basis of the clinician's rating of each participant's performance on the 50-word probes administered during baseline sessions. For consistency across participants, only the five vocalic targets /ɝ/, /ɑɹ/, /ɔɹ/, /ɪɹ/, and /ɛɹ/ were eligible for treatment. For each participant, the three vocalic targets that were the least accurate were selected for treatment. These targets remained the same for both treatment phases. In each session, three words were randomly chosen from a list for each of the three targets.

### Treatment Session Prepractice

After the administration of a probe (described below), each session began with a period of prepractice that was timed to have a 5-min duration. During the first session, tongue anatomy was discussed and the ultrasound images were explained. Features of correct /ɹ/ were described and demonstrated. The participants then had to demonstrate that they could trace the tongue contour on the image, identify the side that represents the "front" and "back," and discuss the major features of /ɹ/ articulation. In subsequent sessions, prepractice was intended to be relatively unstructured to allow explanations about /ɹ/ articulation that the clinician deemed to be most helpful for the participant. The ultrasound was used during prepractice to describe the participant's tongue shape, but additional visual information was also provided. This included a poster with 22 magnetic resonance (MR) images of various adult speakers producing /ɹ/ (Boyce, 2015), and comparisons were made between the child's tongue shape and some of the MR images. In addition, during prepractice, the clinician selected targets at syllable, word, and sentence levels, based on the child's skill level and stimulability. Traditional shaping strategies (e.g., shaping /ɹ/ from /l/ or /ɑ/) and phonetic placement cues were used at the clinician's discretion.

### Treatment Session Practice Schedule

Prepractice was followed by a period of structured practice. The Challenge Point Framework (CPF; McAllister Byun, Ortiz, & Hitchcock, 2016), a researcher-developed open-source software program, was used to guide stimulus presentation and clinician feedback. Each participant practiced three vocalic /ɹ/ contexts, and the program randomly selected three words per context, for a total of nine target words each session. Each word was practiced 18 times (in three blocks of six attempts). The practice component of the session therefore lasted for 162 practice attempts or 45 min, whichever occurred first. During practice, an MR image of correct /ɹ/ was placed adjacent to the ultrasound display so that a comparison could be made between the MR and ultrasound tongue shapes.

A verbal model was provided at the beginning of each block of six trials, and verbal knowledge of performance feedback (information related to articulator movement) was provided at the end of each block. The clinician recorded the participant's response on each trial as 0 or 1 on the basis of their clinical impression of an incorrect (substituted or distorted) or correct production of vocalic /ɹ/. The CPF software automatically tallied the scores entered by the clinician and used them to make adaptive changes in practice difficulty, as detailed below. The CPF program also indicated to the clinician the type of feedback that was required for each trial: no feedback, knowledge of results feedback (indicating whether the response was correct or incorrect), or knowledge of performance feedback (verbal description of aspects of the articulatory movements).

Difficulty was adjusted adaptively by the CPF software as follows: Five or more correct responses in a block triggered an increase in difficulty, three or fewer correct responses triggered a decrease in difficulty, and four correct responses resulted in no change. Changes in practice difficulty were made to two parameters on a rotating basis. Stimulus complexity was adjusted by manipulating the number of syllables per word, the presence or absence of the competing phonemes /l/ and /w/ (where a competing phoneme is defined as a consonant target that shares a major articulatory gesture with /ɹ/), and the presence or absence of a carrier phrase or sentence context. Performance-driven changes in verbal feedback frequency included a reduction of knowledge of results feedback from four to three to two trials per block of six trials. In addition, between-session modifications included changes from fully blocked practice (three consecutive blocks of the same stimulus item) to random-blocked practice (a new stimulus item randomly selected for each block of six trials) to fully random (each trial within a block featured a randomly selected stimulus item). The parameters of the hierarchy at the end of each session were used as the starting point for the next session for that participant.

### Trials Using Ultrasound

The CPF software provided prompts indicating whether UVF should be provided or withheld in a given trial. During trials in which UVF was available, a Siemens Acuson X300 ultrasound with C8-5 ($n$ = 8) or C6-8 ($n$ = 4) transducer was used. To limit variability between participants, ultrasound feedback was provided in the sagittal section only. A single MR image adjacent to the ultrasound display was referenced to highlight similarities and differences between the participant's tongue shape and the target shape as part of knowledge of performance feedback. The MR image used for a given participant was selected to highlight specific components of tongue shape (e.g., elevation of tongue tip, lowering of dorsum) that were judged to be facilitative or important for improving that individual's rhotic production. Determinations about whether ultrasound feedback was provided or withdrawn occurred after every three blocks (18 trials).

### Probe Measurement

Changes in accuracy were assessed using /ɹ/ word probes elicited during baseline sessions, midpoint sessions, and maintenance sessions as well as before and after each treatment session. Audio files from all sessions were segmented into individual words and aggregated with recordings from all subjects. These audio files were then uploaded to Amazon's Mechanical Turk crowdsourcing platform where naive listeners made binary judgments of accuracy (correct/incorrect) for each token. Files were randomized, and listeners were blind to the treatment phase of each recording. Previous research validating the use of crowdsourced listeners' ratings of child speech data (McAllister Byun, Halpin, & Szeredi, 2015) found that binary ratings aggregated across at least nine naive listeners recruited online converged with ratings aggregated across the "industry standard" of three expert listeners. Accordingly, the current study collected binary ratings of each speech token from at least nine online listeners, following the protocol introduced in McAllister Byun et al. (2015).[3] All participants had U.S.-based IP addresses and, per self-report, were native speakers of English with no history of speech or hearing impairment. When aggregating accuracy ratings across listeners, we use $\hat{p}_{\text{correct}}$, defined as the percentage of "correct" ratings out of all ratings, pooled across listeners (McAllister Byun, Harel, Halpin, & Szeredi, 2016).

### Treatment Fidelity

To track fidelity of treatment implementation, video recordings from two sessions per participant were reviewed, with one session selected from each treatment phase. A verbal model was expected at the beginning of each block but not on subsequent trials within a block; modeling was provided as prescribed on an average of 94% of the trials (*SD* = 8%; Clinician 1: *M* = 98%, Clinician 2: *M* = 88%). The amount and type of verbal feedback expected depended on the practice level. When verbal knowledge of results feedback was expected, the clinician provided the appropriate type of feedback in 91% of the trials (*SD* = 5%; Clinician 1: *M* = 93%; Clinician 2: *M* = 89%). Conversely, when verbal knowledge of performance feedback was expected, the appropriate feedback type was provided 98% of the time (*SD* = 2%; Clinician 1 and 2 means: 98%).

### Analyses

Triangulating across multiple analysis methods is a recommended strategy to improve the robustness of conclusions drawn from single-subject experimental research (Kratochwill & Levin, 2014). Accordingly, data from this study were analyzed both within and across participants using visual inspection, effect sizes, and a mixed-effects logistic regression model.

Standardized effect sizes were computed using Busk and Serlin's $d_2$ statistic (Beeson & Robey, 2006), in which standard deviations are pooled across the two phases being compared. Following Maas and Farinella (2012), we adopt the effect size of 1.0 as the minimum $d_2$ that can be considered clinically relevant; that is, the change in accuracy from pretreatment to posttreatment must exceed the pooled standard deviation. Standardized measures like $d_2$ can yield an inflated estimate of treatment effect when variance is low, so unstandardized effect sizes were also calculated and taken into consideration in the interpretation of participants' response to treatment. Because this study involved two phases of treatment, three effect sizes were calculated for each participant: for Phase 1 of treatment, for Phase 2 of treatment, and for both phases taken jointly. Effect sizes were calculated using $\hat{p}_{\text{correct}}$ pooled across all vocalic /ɹ/ variants.

Two logistic mixed-effects models were used for a quantitative comparison of outcomes across individuals (see Rindskopf & Ferron, 2014, for a discussion of mixed models in single-subject designs). For these analyses, we used an uncollapsed data set in which each data point was a single listener's rating of a single token. The binary rating assigned by each listener (correct/incorrect) served as the dependent variable. The first model examined only data from the midpoint phase. Recall that participants were randomly assigned to receive either HF or LF feedback, followed by a switch in condition at midpoint. Thus, by examining participants' performance at midpoint, after only Phase 1 of treatment, we can assess the effect of treatment condition in the absence of any confounding influence of treatment order. Fixed effects included treatment condition (HF vs. LF) and mean percentage of tokens rated correct during the baseline interval as well as the interaction between those predictors. Baseline accuracy was included on the hypothesis that accuracy at the outset of treatment can influence the rate and/or magnitude of response to treatment. Random intercepts were included to reflect the fact that data points were nested within raters and words, and random slopes were examined as permitted by model convergence. Model selection was performed using log-likelihood ratio tests, and only those predictors, interactions, and random slopes that yielded a significant difference in likelihood relative to a reduced model were retained.

The second model examined only data from the final maintenance phase to test for an association between order of treatment delivery (HF-first vs. LF-first) and magnitude of change after all 16 sessions. Fixed effects included

[3] Because of data loss, such as cases in which a sound file failed to play, fewer than nine responses were collected for a subset of items. Items rated by eight unique listeners were considered adequate for inclusion in the analysis; items with seven or fewer ratings were recycled in clean-up blocks to collect additional ratings. Items that did not achieve at least eight ratings after three clean-up rounds were retained as follows: with five ratings, *n* = 189 (< 1% of total); with six ratings, *n* = 756 (4% of total); and with seven ratings, *n* = 849 (4.5% of total).

treatment order (HF-first vs. LF-first) and baseline accuracy (mean percentage of tokens rated correct during the baseline interval) as well as their interaction. As in the previous model, random intercepts were included and random slopes were examined to capture the nesting of data within raters and words. Model selection and reduction were performed as described for the first model.

All computation was carried out in the R software environment (R Core Team, 2015). Data wrangling and plotting were carried out using the packages *tidyr* (Wickham, 2016), *dplyr* (Wickham & Francois, 2015), and *ggplot2* (Wickham, 2009), and mixed models were fit using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015).

## Results

### Individual Results

Effect sizes representing change in $\hat{p}_{\text{correct}}$ for vocalic rhotics are reported in Table 2. The first column shows the mean and standard deviation of $\hat{p}_{\text{correct}}$ in the baseline period, averaged across all rated items from all baseline sessions. The second column presents the equivalent mean and standard deviation across all three midpoint sessions (between the two phases of treatment), and the third indicates the three posttreatment maintenance sessions. The next three columns report three standardized effect sizes: $\text{ES}_{\text{Phase1}}$ compares baseline versus midpoint scores, $\text{ES}_{\text{Phase2}}$ compares midpoint versus maintenance scores, and $\text{ES}_{\text{all}}$ compares baseline versus maintenance scores, reflecting overall gains across both phases of treatment. Participants are blocked by the order in which they received treatment (HF-first or LF-first), and the effect size for each phase (HF vs. LF) is reported in the next two columns. The second last column reports the difference in effect sizes between the two conditions (HF − LF), independent of the order in which they were administered. The final column shows the difference in effect sizes between the first and second phases (Phase 2 − Phase 1), independent of what treatment was administered in each phase. The effect sizes in Table 2 show a wide range of variability in overall response to treatment across individuals. Averaging across all 12 participants yields a mean increase in $\hat{p}_{\text{correct}}$ of 33, with a mean effect size of 13.31, suggesting that, taken collectively, participants' response to the treatment package was positive and exceeded the minimum value of 1.0 considered to be clinically significant (Maas & Farinella, 2012).

Figures 1 and 2 represent participants' patterns of change in vocalic rhotic accuracy $\hat{p}_{\text{correct}}$ over time, which can be visually inspected to corroborate the effect sizes reported in Table 2. The single-subject plots represent each child's performance across the two treatment phases as well as baseline, midpoint, and maintenance probe stages. The plots are grouped by phase condition (HF-first in Figure 1 and LF-first in Figure 2). Within each group, participants are ordered by increasing length of the baseline phase (3–5).[4] The *y*-axis represents $\hat{p}_{\text{correct}}$ aggregated across all vocalic /ɹ/ items in a probe.[5] In each session, a black circle represents performance on the presession probe measure, and a red asterisk represents performance on the postsession probe. Thus, the distance between the two probes in a session provides an index of the participant's progress during that treatment session. Finally, a dashed horizontal line tracks the participant's mean $\hat{p}_{\text{correct}}$ from the baseline interval, so that subsequent scores can be compared with the baseline mean.

All participants were judged to demonstrate a sufficiently low level of baseline variability (i.e., < 10% mean session-to-session variability). The greatest mean session-to-session variability in the baseline phase (7%) was observed in Participant 124. Visual inspection of baseline data raised no questions of extreme outliers or a possible rising trend for any participant.

### Visual Inspection

Figure 1 displays data from the six participants who were randomly assigned to receive HF treatment followed by LF treatment. Two participants in this group, 122 and 113, did not show significant evidence of improvement in either treatment condition. The remaining children in this group showed a robust effect of treatment, with overall effect sizes ranging from 5.6 to 94.2. Participant 127 showed the largest overall gains, exceeding 75% accuracy by the midpoint phase and approximating ceiling level accuracy in the posttreatment maintenance phase. Participant 126, who started with the highest pretreatment accuracy, showed large within-session gains (i.e., scores on postsession probes significantly exceeded scores on presession probes) during Phase 1 of treatment (HF), but these gains did not consistently carry over to the presession probe of the following session until the second phase of treatment. Participants 125 and 124 showed no change in the first five sessions of treatment but exhibited variable gains thereafter. In the case of Participant 125, a sizable increase in accuracy during the no-treatment midpoint phase suggests a potential delayed generalization effect; ongoing gains in Phase 2

[4] Each phase had a minimum duration of 3 sessions. However, the final probe in the baseline and midpoint phases additionally served as the pre-treatment probe for the first session of that phase of treatment. In Figures 1 and 2, these probes have been plotted with the corresponding treatment session to allow visual inspection of within-session change. However, these probes elicited 50 items (rather than the 25 items associated with a regular pre-treatment probe) and were grouped with baseline and midpoint phases for the purpose of descriptive and inferential statistics.

[5] After the exclusions described previously, the mean number of probe words on which $\hat{p}_{\text{correct}}$ scores are based was 18.11 for presession and postsession probes and 48.11 for baseline, midpoint, and maintenance probes. Because each item was rated by multiple listeners, the number of ratings collected in connection with a given probe session (i.e., the denominator in $\hat{p}_{\text{correct}}$) was roughly nine times the number of items in that probe.

**Table 2.** Proportion of /ɹ/ tokens rated correct at baseline, midpoint (after Phase 1), and maintenance (after Phase 2).

| Condition order | Subject | Baseline, *M* (*SD*) | Midpoint, *M* (*SD*) | Maintenance, *M* (*SD*) | $ES_{Phase1}$ | $ES_{Phase2}$ | $ES_{all}$ | HF–LF difference | Phase 2–Phase 1 difference |
|---|---|---|---|---|---|---|---|---|---|
| HF First | 125 | 5.65 (1.22) | 44.97 (27.58) | 89.07 (4.68) | 2.01 | 2.23 | 24.41 | −0.22 | 0.22 |
| | 126 | 38.91 (6.72) | 78.2 (2.96) | 69.29 (3.82) | 7.57 | −2.61 | 5.56 | 10.18 | −10.18 |
| | 127 | 2.74 (1.11) | 77.62 (5.2) | 94.33 (0.81) | 23.88 | 5.44 | 94.20 | 18.44 | −18.44 |
| | 122 | 5.76 (1.41) | 6.72 (1.02) | 7.61 (1.16) | 0.76 | 0.81 | 1.40 | −0.05 | 0.05 |
| | 113 | 6.7 (1.92) | 5.51 (1.68) | 7.87 (2.7) | −0.64 | 1.05 | 0.53 | −1.69 | 1.69 |
| | 124 | 18.91 (5.36) | 39 (5.41) | 72.65 (5.07) | 3.74 | 6.42 | 10.20 | −2.68 | 2.68 |
| LF First | 119 | 15.71 (2.6) | 56.79 (6.16) | 75.46 (8.81) | 8.69 | 2.38 | 8.51 | −6.31 | −6.31 |
| | 111 | 12.75 (2.44) | 12.17 (6.6) | 9.83 (0.08) | −0.12 | −0.43 | −1.47 | −0.31 | −0.31 |
| | 112 | 29.74 (4.38) | 75.53 (3.89) | 60.25 (4.5) | 10.93 | −3.63 | 6.89 | −14.56 | −14.56 |
| | 114 | 34.92 (4.12) | 49.51 (3.3) | 65.69 (4.95) | 3.83 | 3.70 | 6.76 | −0.13 | −0.13 |
| | 115 | 2.78 (1.12) | 2.95 (1.47) | 6 (2.61) | 0.13 | 1.44 | 1.73 | 1.31 | 1.31 |
| | 128 | 14.71 (3.67) | 18.74 (1.51) | 26.31 (19.41) | 1.29 | 0.55 | 1.00 | −0.74 | −0.74 |

*Note.* ES = effect size; HF = high-frequency ultrasound visual feedback; LF = low-frequency ultrasound visual feedback.

(HF) brought this participant to near-ceiling accuracy by the maintenance phase. Participant 124 showed modest gains in Phase 1 of treatment (LF), followed by more consistent progress in Phase 2 (HF).

Figure 2 illustrates data from the six participants who received treatment in the opposite order, with LF followed by HF treatment. As in the previous group, there were two nonresponders who showed no sustained change in either phase of treatment, Participants 111[6] and 115. In the case of Participant 115, the overall effect size of 1.73 exceeds the minimum to be considered clinically significant, but visual inspection makes it clear that this is a case where the standardized effect size has been inflated by very low variance. A third participant, 128, exhibited an overall effect size of exactly 1.0. Unlike the two participants deemed to be nonresponders, however, Participant 128 exhibited clear within-session gains, achieving accuracy scores above 60% on postsession probes in the final three sessions of Treatment Phase 2 (HF). However, these gains had not yet been solidified and carried over minimally into the posttreatment maintenance phase. The remaining three participants exhibited overall effect sizes ranging from 6.76 to 8.51. Participant 119, who made the largest gains overall, demonstrated consistent progress in Phase 1 (LF); in Phase 2, he made smaller gains but maintained an overall high level of accuracy. Participant 112 made strong gains in Phase 1 (LF) but exhibited an unexpected decline in performance after two sessions in Phase 2 (HF). Overall, this participant's accuracy during the maintenance phase was higher than it had been at baseline but not as accurate as it was at midpoint between treatment phases. Finally, Participant 114 made variable gains that were similar in magnitude across both phases of treatment; his overall effect size of 6.76 reflects a meaningful increase in accuracy from baseline to maintenance.

### *Across-Subjects Comparisons: Effect Sizes*

The boxplots in Figures 3–5 represents the distribution of effect sizes ($d_2$, as described above and reported in Table 2) that can be observed when the data are partitioned in different ways. In Figure 3a, effect sizes associated with HF treatment phases ($ES_{HF}$) are compared against effect sizes from LF treatment phases ($ES_{LF}$), independent of the order in which the two types of treatment were delivered. The boxplots in Figure 3a overlap to a large extent and share similar median values. Figure 3b examines a possible order effect, comparing the distribution of $ES_{Phase1}$ versus $ES_{Phase2}$, independent of the type of treatment delivered in each phase. Figure 3b shows that effect sizes observed in the first phase of treatment tended to be slightly larger than those observed in the second phase, although again there is substantial overlap.

Figure 4 considers a possible interaction between treatment type and order of treatment delivery. The boxplots in Figure 4 support the impression that phase order (first vs. second phase) is more prominent than treatment condition (HF vs. LF) in influencing effect sizes.

Finally, Figure 5 shows the distribution of overall effect sizes (from baseline to posttreatment) for children who received HF treatment first versus children who received LF treatment first. After all 16 sessions of treatment, effect sizes tended to be greater for children who received HF treatment before LF treatment, compared with those who received the reverse order of treatment conditions. Because of the small number of data points, hypothesis tests were not conducted on these comparisons of effect sizes across conditions. Instead, the logistic mixed model reported in the next section examines these effects and their interactions in greater detail.

### *Across-Subjects Comparisons: Mixed Logistic Model*

The first logistic mixed model examined midpoint accuracy as predicted by treatment condition (HF vs. LF)

[6] For Participant 111, data from probes during the first two treatment sessions were lost because of an error in the use of the recording equipment.

**Figure 1.** Individual plots for six participants who received high-frequency ultrasound treatment followed by low-frequency ultrasound treatment. *y*-Axis represents the proportion of probe words rated as correct. *x*-Axis represents time (BL = baseline; Tx = treatment session; MP = midpoint; MN = maintenance). During days on which treatment occurred, probes were administered before the session (circles) and after the session (asterisks). Dashed line represents the participant's mean baseline accuracy.

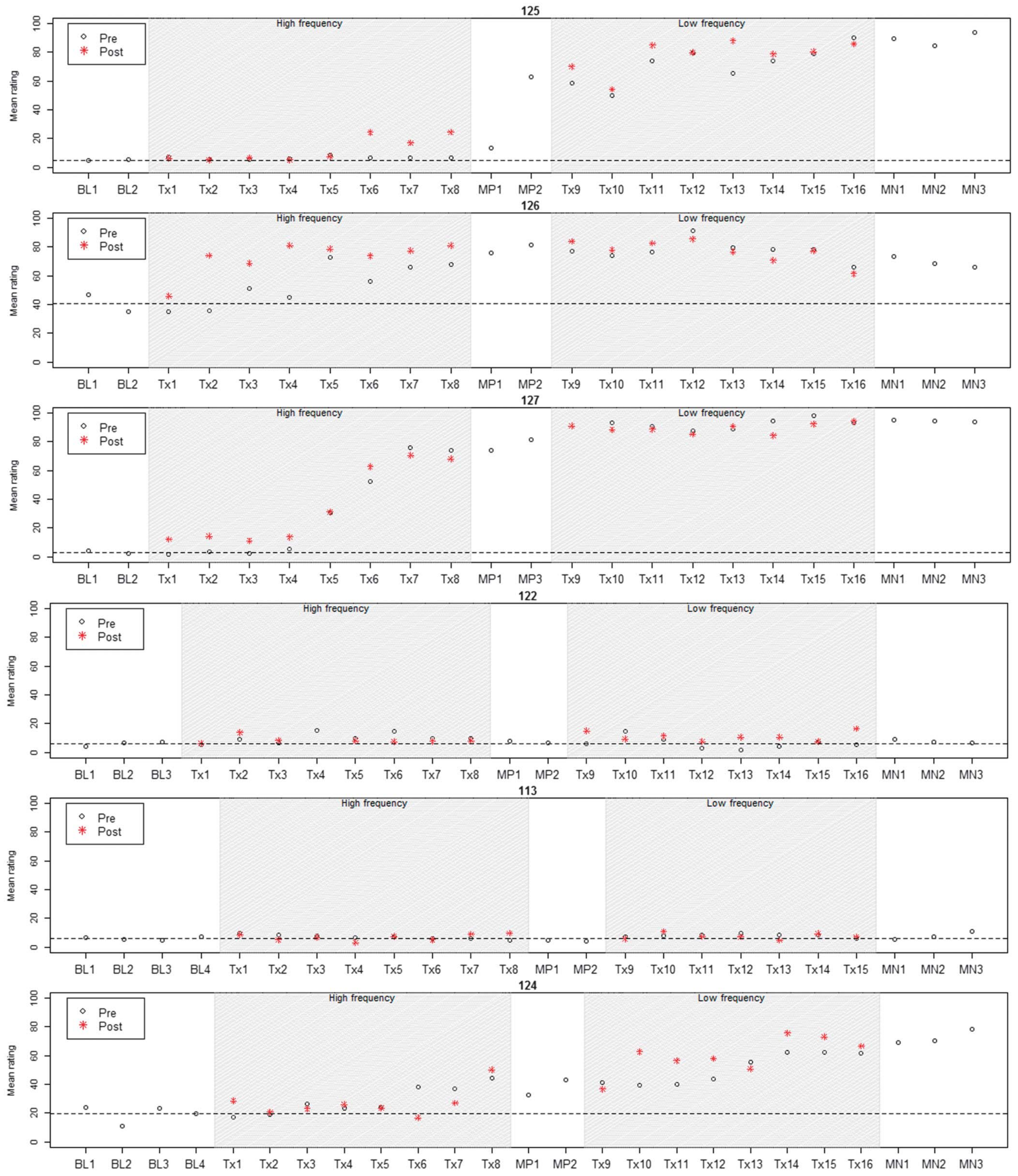


**Figure 2.** Individual plots for six participants who received low-frequency treatment followed by high-frequency treatment. *y*-Axis represents the proportion of probe words rated as correct. *x*-Axis represents time (BL = baseline; Tx = treatment session; MP = midpoint; MN = maintenance). During days on which treatment occurred, probes were administered before the session (circles) and after the session (asterisks). Dashed line represents the participant's mean baseline accuracy.

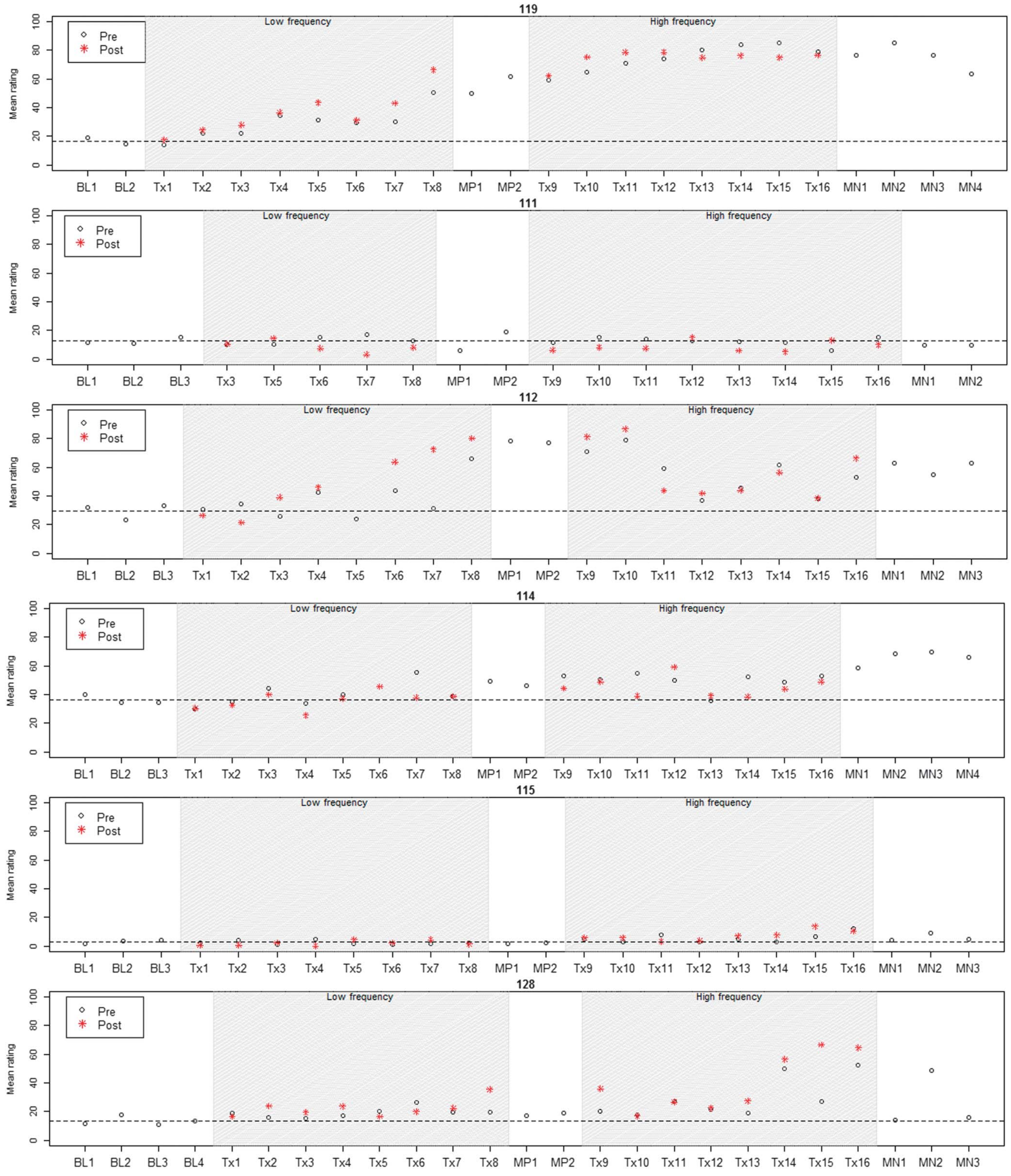


**Figure 3.** Boxplots depicting the distribution of effect sizes observed in connection with (a) high-frequency (dark gray) versus low-frequency (light gray) treatment, independent of phase order, and (b) Phase 1 versus Phase 2 of treatment, independent of treatment condition.

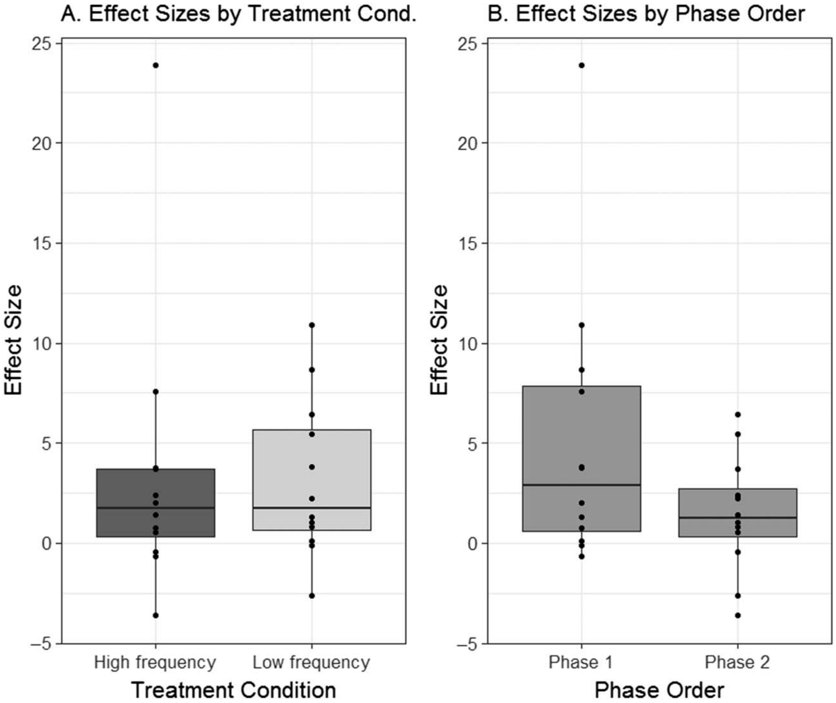


**Figure 4.** Boxplots depicting the distribution of overall effect sizes observed in connection with high-frequency (dark gray) versus low-frequency (light gray) treatment when HF treatment was provided first, versus the opposite order.

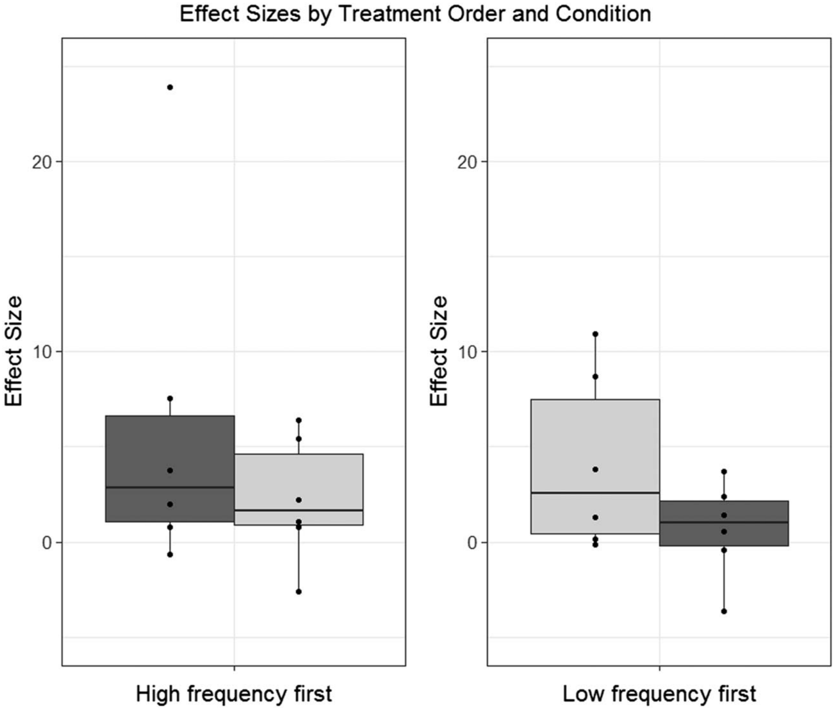
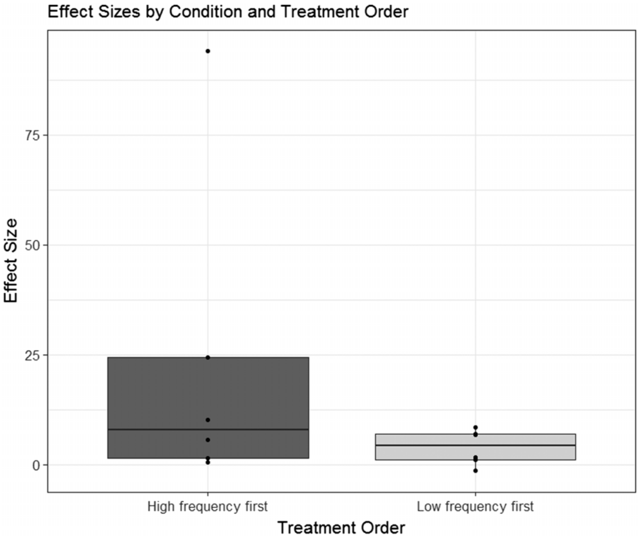


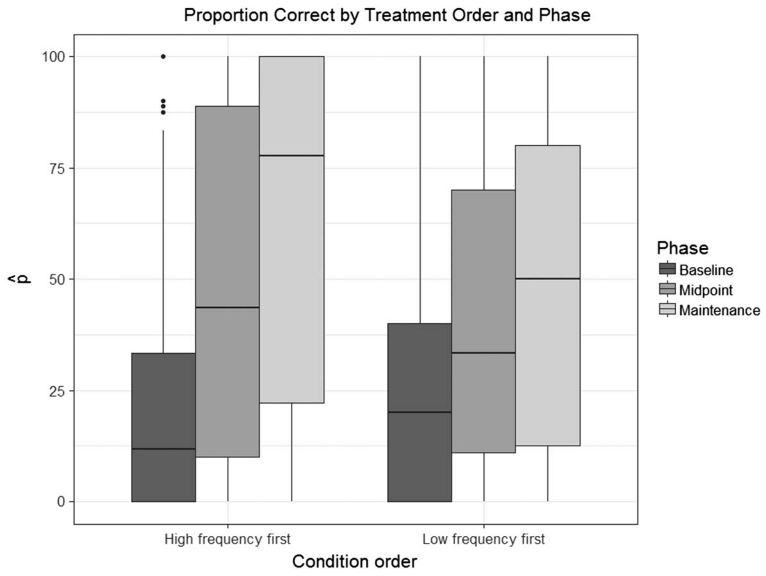
**Figure 5.** Boxplots depicting the distribution of overall effect sizes observed when high-frequency treatment was provided first, versus the opposite order.



and baseline accuracy. The final reduced model included both fixed effects and the interaction between them as well as random intercepts of rater (with a random slope by baseline accuracy) and word. The main effect of condition was significant (β = −1.17, *SE* = 0.09, $p < .001$). The direction of the coefficient indicates that the LF treatment condition was associated with significantly lower performance at midpoint than the HF condition. This difference can be seen in Figure 6, where the middle boxplot in each set of three represents the percentage of "correct" ratings for tokens elicited in the midpoint phase. Mean accuracy across the pretreatment baseline phase was also a significant predictor of accuracy in midpoint probes (β = 12.08, *SE* = 0.33, $p < .001$); unsurprisingly, higher accuracy at baseline was associated with higher accuracy at midpoint. Finally, the interaction between treatment condition and baseline accuracy was also significant (β = 3.21, *SE* = 0.32, $p \leq .001$), with a slightly stronger association between baseline accuracy and midpoint accuracy in the LF than the HF treatment condition. However, this interaction must be interpreted with caution in light of the small number of data points. Complete results of this mixed-effects model are reported in Appendix A.

The second logistic mixed model examined accuracy in the posttreatment maintenance phase as predicted by treatment order (HF-first vs. LF-first) and baseline accuracy. The final reduced model included both fixed effects and the interaction between them. The main effect of condition was significant (β = −0.96, *SE* = 0.08, $p < .001$), indicating that participants who received HF followed by LF treatment showed significantly higher accuracy in the maintenance phase than those who received treatment in the opposite order. This difference can be visualized in Figure 6, where the third boxplot in each set of three represents the proportion of "correct" ratings for tokens elicited in the maintenance phase. Baseline accuracy was a significant predictor of accuracy in posttreatment maintenance probes (β = 11.89, *SE* = 0.32, $p < .001$). Finally, the interaction between treatment condition and baseline accuracy was not significant (β = −0.23, *SE* = 0.29, $p = .43$). Complete results of this regression are reported in Appendix B.

**Figure 6.** Boxplots depicting proportion of "correct" ratings for tokens in each phase (baseline, midpoint, maintenance) when high-frequency treatment was provided first, versus low-frequency first.



## Discussion

This study extends previous single-subject experimental research documenting the efficacy of intervention incorporating UVF for residual rhotic errors. The specific goal was to compare the effects of intervention in which UVF was provided at a HF (89% of trials) versus at a LF (44% of trials). Participants completed two phases of intervention, receiving either eight sessions of HF followed by eight sessions of LF UVF, or the opposite order. Accuracy in producing rhotics in untreated words was probed at baseline, at the midpoint between the two phases of treatment, and after the end of treatment. This section will discuss (a) overall response to UVF intervention, independent of feedback frequency; (b) differences between the HF and LF conditions at midpoint, when participants had been exposed to only one frequency condition; and (c) differences at the end of the study between participants who received intervention in the order HF–LF versus LF–HF.

With respect to the first hypothesis, across the 12 participants in this study, the median effect size reflecting change in rhotic production accuracy after all 16 sessions of UVF intervention was 6.16 ($M$ = 13.31). Pooling across participants, mean accuracy calculated to be 16% in the baseline phase and 49% in the final maintenance phase, representing an overall increase of 33 percentage points. This supports previous studies in finding that UVF intervention can be an efficacious means to remediate RSEs for many children (e.g., Adler-Bock et al., 2007; Bacsfalvi, 2010; Preston et al., 2014). However, the magnitude of change varied across individuals, with four of 12 individuals (Participants 113 and 122 in the HF–LF order and Participants 111 and 115 in the LF–HF order) judged to be unresponsive to the 16 sessions of treatment. The presence of nonresponders is consistent with previous single-case studies of biofeedback intervention for children with RSEs. The following discussion considers several factors that might account for individual variability in response to UVF intervention.

At the midpoint of treatment, six participants had received eight sessions of HF UVF intervention, and six had received eight sessions of LF intervention. Previous theoretical and empirical work has suggested that generalization learning can be maximized when feedback is provided on a reduced schedule (Austermann Hula et al, 2008; Steinhauer & Grayhack, 2000). Therefore, the second hypothesis was that generalization gains at midpoint would be greater for the participants who received LF feedback in the first phase of treatment than participants who received HF feedback. The two treatment conditions showed roughly similar effect sizes after the first treatment phases (HF: $M$ = 6.2, median = 2.9; LF: $M$ = 4.1, median = 2.6), whereas the mixed-effects model suggested that, when controlling for other variables, /ɹ/ tokens produced by children who had only received eight sessions of HF treatment were significantly more likely to be rated correct than tokens produced by children who had received LF intervention. Thus, although schema-based motor learning theory led to the prediction that frequent knowledge of performance feedback could hinder learning (e.g., Maas et al., 2008), we observed the opposite effect—at least in the initial stages of treatment, generalization to untreated words was greater for children who had been provided with more UVF than children who had less UVF.

A plausible explanation for this finding may emerge when we consider it in conjunction with our third hypothesis pertaining to the order in which treatment conditions are delivered. Previous work in nonspeech motor learning (Hodges & Franks, 2001) suggests that knowledge of performance feedback may be most valuable in the early stages of treatment, when the learner is still establishing a new motor plan. In later stages, as the focus shifts from acquisition to generalization, knowledge of performance feedback loses its advantage and its impact may in fact be detrimental if the learner becomes overly dependent on detailed qualitative feedback. Thus, we hypothesized that overall effect sizes would be larger in individuals who received UVF in the sequence HF–LF as opposed to LF–HF. In keeping with this hypothesis, effect sizes from the beginning to end of the treatment program were generally larger for children in the HF–LF order (effect size: $M$ = 22.7, median = 7.9) than for those in the LF–HF order (effect size: $M$ = 3.9, median = 4.2). The raw percent change also depended on treatment order, with a mean increase of 43% above baseline levels for children in the HF–LF condition versus 22% in the LF–HF condition (see Table 2). Furthermore, the mixed-effects logistic regression analysis revealed a significant effect of condition on accuracy in posttreatment maintenance probes, indicating that /ɹ/ tokens produced by children in the HF–LF condition were significantly more likely to be rated correct than tokens from children who had undergone the LF–HF condition. These results are generally in line with the findings of McAllister Byun and Campbell (2016), who observed larger treatment gains when children received an initial phase of acoustic biofeedback treatment followed by a phase of traditional (no-biofeedback) treatment, versus the opposite order. Maas et al. (2012) offered a similar speculation, suggesting that LF may be beneficial for learning only once a child has already established a clear "reference of correctness," which, in this case, may be achieved during the HF condition.

Returning to our unexpected finding that larger gains were observed in the first phase in connection with the HF than the LF intervention condition, we note that the notion of an initial acquisition phase is not clearly operationalized in previous literature on speech motor learning, particularly among individuals with speech disorders. In the present case, the first eight sessions of UVF intervention may have represented the acquisition phase for many participants. Thus, rather than looking for a single condition to result in optimal outcomes, it may be particularly beneficial to consider the optimal sequencing of conditions as children progress through learning stages. Although empirical research will be needed to determine optimal criteria, we suggest that the transition from acquisition to generalization might be operationalized as the point at which a participant first begins to show gains that extend to untreated words in a probe measure elicited without feedback. Whatever specific benchmarks may ultimately be identified, our results make a general suggestion that HF visual feedback is likely to be most beneficial when applied in early stages of speech motor learning.

## Limitations and Future Directions

For the purpose of this study, HF feedback was operationalized as UVF in 88% of trials and LF feedback as UVF in 44% of trials. However, different results might have been obtained had the relative proportion of UVF been defined differently for the two conditions. In addition, it should be noted that the study included only children who had undergone either the order HF–LF or LF–HF to allow for total exposure to UVF to be similar after 16 sessions; no child received exclusively HF or exclusively LF treatment. Although HF–LF was the most effective option in this study, it is possible that the advantage observed for HF over LF over eight sessions could increase if children undergo the order of HF–HF for all 16 sessions. Thus, further research is needed to validate the claim that HF–LF is the optimal order for generalization learning. Moreover, it would be beneficial to conduct further research aimed at identifying indicators of the appropriate stage(s) during the learning process when a change in feedback frequency is most optimal (cf. Maas et al., 2012).

Another consideration arises from this study's use of adaptive difficulty in practice. All participants began practice at the same level and were held to the same criterion for advancement to higher levels of practice by the CPF software. However, because there were individual differences in the rate at which participants achieved successful productions, the amount of time spent practicing at a given level of complexity varied across participants. For example, because of the participants' differing accuracy levels, Participant 115 never advanced past blocked practice of monosyllabic words with a competing /l/ or /w/ (e.g., *leer*), whereas Participant 124 spent the last four sessions in randomized practice of sentences with multiple /r/ words (e.g., *He got in trouble for saying "steer."*). The adaptive structure of practice that we used is both theoretically grounded (e.g., Guadagnoli & Lee, 2004) and clinically defensible. However, a study design without this adaptive component might have provided a purer test of the influence of feedback frequency on treatment outcomes.

In addition, as noted above, there was considerable heterogeneity in individual outcomes across participants in the study, with four participants who were essentially nonresponders to both HF and LF treatment conditions. This variable response is in keeping with previous descriptions of the range of individual responses to biofeedback intervention (e.g., McAllister Byun et al., 2014; Preston et al., 2014). Although all children in this study had hearing, cognitive, and receptive language skills that were broadly within normal limits, additional child-specific characteristics presumably influence treatment outcomes. However, previous research raising this question has pointed out that it is difficult to draw strong conclusions from the small sample sizes typical of a single-subject experimental design (e.g., McAllister Byun & Campbell, 2016). In this study, there were no significant relationships between overall effect size and demographic variables including age ($\rho = -0.17$, $p = .59$) and score on the Phonological Awareness composite of the Comprehensive Test of Phonological Processing–Second Edition ($\rho = 0.13$, $p = .7$). One apparent similarity among the four nonresponders is that all had a stimulability score of 0% before treatment. However, there were five other participants who were initially not stimulable for /ɹ/ but were observed to progress in treatment, and stimulability rating did not emerge as a significant predictor of overall effect size ($\rho = 0.03$, $p = .92$). Furthermore, a child's natural history of speech impairment may relate to the quality of /ɹ/ distortions (Shriberg, Flipsen, Karlsson, & McSweeny, 2001) and therefore might influence treatment response, yet inspection of our data revealed relatively similar median effect sizes between the nine children with histories of speech disorder (6.89) and the three children identified after the age of 5 years (5.56). Thus, we echo previous studies in arguing that larger-scale research will be necessary to identify factors that adequately predict individual response to biofeedback (and other) interventions and that strategies should be developed to effectively manage these cases.

## Conclusions

This study supports previous research in finding that a speech therapy program that includes UVF can improve accuracy in the production of /ɹ/ in untreated words in many children with RSEs. Across 12 participants in this study, the median effect size after 16 sessions of UVF intervention was 6.16, representing a positive change of a clinically significant magnitude. However, four participants were judged to show no meaningful generalization in response to treatment, highlighting the need for larger-scale research investigating individual predictors of response to biofeedback intervention. With regard to the frequency of UVF, group level comparisons in this study suggested that children's /ɹ/ tokens were more likely to be rated correct after eight sessions of HF than eight sessions of LF ultrasound and that larger overall effect sizes were observed when treatment was provided in the order HF–LF as opposed to LF–HF. The small size of this study sample means that any across-subjects comparisons must be treated with caution. However, taking these findings in conjunction with previous theoretical and empirical research, we suggest that biofeedback intervention programs for individuals with RSEs should be structured to begin with a relatively higher level of biofeedback frequency and proceed to a lower level of biofeedback frequency.

## Acknowledgments

Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders under Award Number R01DC013668 (D. Whalen, PI). Thanks to Olivia Harold, Jacquie Fetzner, Jose Ortiz, and Jessica Whittle for assistance with this project. Siemens Corporation provided a long-term loan of the ultrasound equipment used in this study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

**Adler-Bock, M., Bernhardt, B., 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*(2), 128–139.

**Alwan, A., Narayanan, S., & Haker, K.** (1997). Toward articulatory–acoustic models for liquid approximants based on MRI and EPG data: Part II. The rhotics. *The Journal of the Acoustical Society of America, 101*(2), 1078.

**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. *Journal of Speech-Language Pathology and Audiology, 34*(3), 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.

**Ball, M. J., Müller, N., & Granese, A.** (2013). Towards an evidence base for /r/-therapy in English. *Journal of Clinical Speech and Language Studies, 20,* 1–23.

**Bates, D., Maechler, M., Bolker, B., & Walker, S.** (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

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

**Bislick, L. P., Weir, P. C., Spencer, K., Kendall, D., & Yorkston, K. M.** (2012). Do principles of motor learning enhance retention and transfer of speech skills? A systematic review. *Aphasiology, 26*(5), 709–728.

**Boyce, S. E.** (2015). The articulatory phonetics of /r/ for residual speech errors. *Seminars in Speech and Language, 36*(4), 257–270.

**Cleland, J., Scobbie, J. M., & Wrench, A. A.** (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics, 29*(8–10), 575–597.

**Crowne Hall, B. J.** (1991). Attitudes of fourth and sixth graders toward peers with mild articulation disorders. *Language, Speech, and Hearing Services in Schools, 22,* 334–340.

**Culton, G. L.** (1986). Speech disorders among college freshmen: A 13-year survey. *Journal of Speech and Hearing Disorders, 51,* 3–7.

**Delattre, P., & Freeman, D. C.** (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, 6*(44), 29–68.

**Dunn, L. M., & Dunn, D. M.** (2007). *Peabody Picture Vocabulary Test–Fourth Edition*. Minneapolis, MN: Pearson.

**Flipsen, P.** (2015). Emergence and prevalence of persistent and residual speech errors. *Seminars in Speech and Language, 36*(4), 217–223.

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

**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.

**Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., & Perkell, J. S.** (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *The Journal of the Acoustical Society of America, 105*(5), 2854–2865.

**Hitchcock, E. R., Harel, D., & McAllister Byun, T.** (2015). Social, emotional, and academic impact of residual speech errors in school-aged children: A survey study. *Seminars in Speech & Language, 36*(4), 283–294.

**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.

**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*(3), 540–553.

**Kratochwill, T. R., & Levin, J. R. (Eds.).** (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.

**Maas, E., Butalla, C. E., & Farinella, K. A.** (2012). Feedback frequency in treatment for childhood apraxia of speech. *American Journal of Speech-Language Pathology, 21*(3), 239–257.

**Maas, E., & Farinella, K. A.** (2012). Random versus blocked practice in treatment for childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research, 55*(2), 561–578.

**Maas, E., Robin, D. A., Austermann Hula, S. N., 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., & Campbell, H.** (2016). Differential effects of visual–acoustic biofeedback intervention for residual speech errors. *Frontiers in Human Neuroscience, 10,* 567.

**McAllister Byun, T., Halpin, P. F., & Szeredi, D.** (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders, 53,* 70–83.

**McAllister Byun, T., Harel, D., Halpin, P. F., & Szeredi, D.** (2016). Deriving gradient measures of child speech from crowdsourced ratings. *Journal of Communication Disorders, 64,* 91–102.

**McAllister Byun, T., & 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.

**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, 57*(6), 2116–2130.

**McAllister Byun, T., Ortiz, J., & Hitchcock, E. R.** (2016). *Challenge point (Version 1.5.8)*. Retrieved from http://cpp.umd.edu/

**Miccio, A. W.** (2002). Clinical problem solving: Assessment of phonological disorders. *American Journal of Speech-Language Pathology, 11*(3), 221–229.

**Mielke, J., Baker, A., & Archangeli, D.** (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language, 92*(1), 101–140.

**Modha, G., Bernhardt, B., Church, R., & Bacsfalvi, P.** (2008). Ultrasound in treatment of /r/: A case study. *International Journal of Language and Communication Disorders, 43*(3), 323–329.

**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., Leece, M., & Maas, E.** (2016). Intensive treatment with ultrasound visual feedback for speech sound errors in childhood apraxia. *Frontiers in Human Neuroscience, 10,* 440.

**Preston, J. L., Leece, M. C., & Maas, E.** (2017). Motor-based treatment with and without ultrasound feedback for residual speech-sound errors. *International Journal of Language and Communication Disorders, 52*(1), 80–94.

**Preston, J. L., Maas, E., Whittle, J., Leece, M. C., & McCabe, P.** (2016). Limited acquisition and generalisation of rhotics with ultrasound visual feedback in childhood apraxia. *Clinical Linguistics & Phonetics, 30*(3–5), 363–381.

**Preston, J., & McAllister, T.** (2017). *Ultrasound biofeedback for /r/*. Retrieved from https://osf.io/45bdz/

**Preston, J. L., McAllister Byun, T., Boyce, S. E., Hamilton, S., Tiede, M., Phillips, E., . . . Whalen, D. H.** (2017). Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors. *Journal of Visualized Experiments, 119,* e55123.

**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, 57*(6), 2102–2115.

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

**Rindskopf, D., & Ferron, J.** (2014). Using multilevel models to analyze single-case design data. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 221–246). Washington, DC: American Psychological Association.

**Ruscello, D. M., & Shelton, R. L.** (1979). Planning and self-assessment in articulatory training. *Journal of Speech and Hearing Disorders, 44*(4), 504–512.

**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), 146–157.

**Schmidt, R. A., & Lee, T. D.** (2011). *Motor control and learning: A behavioral emphasis* (5th 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.

**Shawker, T. H., & Sonies, B. C.** (1985). Ultrasound biofeedback for speech training: Instrumentation and preliminary results. *Investigative Radiology, 20*(1), 90–93.

**Shriberg, L. D.** (1975). A response evocation program for /er/. *Journal of Speech and Hearing Disorders, 40*(1), 92–105.

**Shriberg, L. D.** (2009). Childhood speech sound disorders: From postbehaviorism to the postgenomic era. In R. Paul & P. Flipsen (Eds.), *Speech sound disorders in children*. San Diego, CA: Plural.

**Shriberg, L. D., Flipsen, P., Karlsson, H. B., & McSweeny, J. L.** (2001). Acoustic phenotypes for speech-genetics studies: An acoustic marker for residual /ɝ/ distortions. *Clinical Linguistics & Phonetics, 15*(8), 631–650.

**Shuster, L. I.** (1998). The perception of correctly and incorrectly produced /r/. *Journal of Speech, Language, and Hearing Research, 41,* 941–950.

**Silverman, F. H., & Paulus, P. G.** (1989). Peer reactions to teenagers who substitute /w/ for /r/. *Language, Speech, and Hearing Services in Schools, 20,* 219–221.

**Sjolie, G. M., Leece, M. C., & Preston, J. L.** (2016). Acquisition, retention, and generalization of rhotics with and without ultrasound visual feedback. *Journal of Communication Disorders, 64,* 62–77.

**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.

**Steinhauer, K., & Grayhack, J. P.** (2000). The role of knowledge of results in performance and learning of a voice motor task. *Journal of Voice, 14*(2), 137–145.

**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), 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.

**Tiede, M. K., Boyce, S. E., Holland, C. K., & Choe, K. A.** (2004). A new taxonomy of American English /r/ using MRI and ultrasound. *The Journal of the Acoustical Society of America, 115*(5), 2633–2634.

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

**Wechsler, D.** (2011). *Wechsler Abbreviated Scales of Intelligence–Second Edition*. San Antonio, TX: Pearson.

**Wickham, H.** (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag. Retrieved from http://ggplot2.org

**Wickham, H.** (2016). *tidyr: Easily tidy data with 'spread()' and 'gather()' functions (R package version 0.5.1)*. Retrieved from http://CRAN.R-project.org/package=tidyr

**Wickham, H., & Francois, R.** (2015). *dplyr: A grammar of data manipulation (R package version 0.4.3)*. Retrieved from http://CRAN.R-project.org/package=dplyr

**Wiig, E., Semel, E., & Secord, W.** (2013). *Clinical Evaluation of Language Fundamentals–Fifth Edition*. Bloomington, MN: Pearson.

## Appendix A

Complete Results of Regression Model Examining Influences on Perceptually Rated Accuracy at Midpoint

| Term | Estimate | *SE* | Test statistic | *p* value |
|---|---|---|---|---|
| (Intercept) | −3.45 | 0.15 | −23.77 | < .001 |
| Condition (reference level: LF) | −1.17 | 0.09 | −12.47 | < .001 |
| Baseline accuracy | 12.08 | 0.33 | 36.83 | < .001 |
| Condition × Baseline Accuracy interaction | 3.21 | 0.32 | 9.91 | < .001 |

*Note.* LF = low-frequency ultrasound visual feedback.

## Appendix B

Complete Results of Regression Model Examining Influences on Perceptually Rated Accuracy at Posttreatment Maintenance

| Term | Estimate | *SE* | Test statistic | *p* value |
|---|---|---|---|---|
| (Intercept) | −2.11 | 0.11 | −20.06 | < .001 |
| Condition order (reference level: LF–HF) | −0.96 | 0.08 | −12.28 | < .001 |
| Baseline accuracy | 11.89 | 0.32 | 36.75 | < .001 |
| Condition Order × Baseline Accuracy interaction | −0.23 | 0.29 | −0.8 | .43 |

*Note.* LF–HF = low- followed by high-frequency ultrasound visual feedback.