AJSLP

Research Article

Remediating Residual Rhotic Errors With Traditional and Ultrasound-Enhanced Treatment: A Single-Case Experimental Study

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

Purpose: The aim of the study was to examine how ultrasound visual feedback (UVF) treatment impacts speech sound learning in children with residual speech errors affecting /ɹ/.
Method: Twelve children, ages 9–14 years, received treatment for vocalic /ɹ/ errors in a multiple-baseline across-subjects design comparing 8 sessions of UVF treatment and 8 sessions of traditional (no-biofeedback) treatment. All participants were exposed to both treatment conditions, with order counterbalanced across participants. To monitor progress, naïve listeners rated the accuracy of vocalic /ɹ/ in untreated words.
Results: After the first 8 sessions, children who received UVF were judged to produce more accurate vocalic /ɹ/ than those who received traditional treatment. After the second 8 sessions, within-participant comparisons revealed individual variation in treatment response. However, group-level comparisons revealed greater accuracy in children whose treatment order was UVF followed by traditional treatment versus children who received the reverse treatment order.
Conclusion: On average, 8 sessions of UVF were more effective than 8 sessions of traditional treatment for remediating vocalic /ɹ/ errors. Better outcomes were also observed when UVF was provided in the early rather than later stages of learning. However, there remains a significant individual variation in response to UVF and traditional treatment, and larger group-level studies are needed.
Supplemental Material: https://doi.org/10.23641/asha.8206640

Distortions of /ɹ/ are among of the most common speech sound errors in American English (Shriberg, 2009), with residual speech errors (RSEs) affecting /ɹ/ persisting in approximately 1%–2% of adolescents and young adults (Flipsen, 2015). RSEs may impact the intelligibility or social acceptability of speech production, potentially leading to a variety of negative social and socioemotional consequences (Hitchcock & McAllister Byun, 2015; Silverman & Paulus, 1989). Therefore, effective and efficient intervention options are essential. Although traditional articulation therapy involving auditory models and verbal cues for articulator placement can be effective in some cases, some individuals with RSEs affecting /ɹ/ make minimal progress with traditional therapy (McAllister Byun & Hitchcock, 2012; Shriberg, 1975). Thus, alternative interventions are needed for treating these errors. Ultrasound visual feedback (UVF), which provides information about lingual movements in real time, has proven to be efficacious for some children with RSE affecting /ɹ/ (Adler-Bock, Bernhardt, Gick, & Bacsfalvi, 2007; McAllister Byun, Hitchcock, & Swartz, 2014; Modha, Bernhardt, Church, & Bacsfalvi, 2008; Preston et al., 2014). The current study analyzes speech sound learning during periods of traditional therapy and UVF-enhanced treatment targeting vocalic /ɹ/ in children with RSE.

[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, NY
[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, NY
Correspondence to Jonathan L. Preston: jopresto@syr.edu

Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Mary Fagan
Received November 2, 2018
Revision received January 27, 2019
Accepted February 27, 2019
https://doi.org/10.1044/2019_AJSLP-18-0261

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

Rhotic sounds[1] emerge late in American English speech sound acquisition (Smit, Hand, Freilinger, Bernthal, & Bird, 1991). Rhotics can be especially difficult due to their complexity in articulation and variability in production. Whereas most speech sounds in English are articulated with only one major lingual constriction or narrowing of the vocal tract, /ɹ/ requires two major lingual constrictions: one with the anterior tongue approximating the palate and the other with the tongue root retracting toward the posterior pharyngeal wall (Alwan, Narayanan, & Haker, 1997; Delattre & Freeman, 1968). The characteristic acoustic and perceptual features of /ɹ/ can be achieved with a range of tongue shapes as long as these constrictions are achieved (Espy-Wilson, Boyce, Jackson, Narayanan, & Alwan, 2000). The two most common shapes are described as "bunched," where the tongue tip lowers while the anterior tongue body raises toward the hard palate, and "retroflex," where the tongue tip raises close to the alveolar ridge. Moreover, some /ɹ/ productions are neither classically "bunched" nor "retroflex," but a combination of features of both (Boyce, 2015; Tiede, Boyce, Holland, & Choe, 2004). The variability of tongue shapes for /ɹ/, both between speakers and across phonetic contexts within speakers, further adds to the difficulty of remediating this target in children with misarticulation.

Treatments for Residual /ɹ/ Errors

Several techniques are utilized in traditional articulatory treatment for /ɹ/. The clinician typically provides an auditory model of a correct /ɹ/ for the client to imitate (Van Riper & Erickson, 1996). Verbal and visual cues may also be provided to encourage the child to adjust the shape and/or location of the tongue (e.g., Ruscello & Shelton, 1979; Secord, Boyce, Donohue, Fox, & Shine, 2007). Another technique may involve cueing the child to produce a phoneme with a similar articulatory configuration such as /l/ and then shaping it into an acoustically acceptable /ɹ/ (Shriberg, 1975). Traditional treatment approaches are undoubtedly successful for some children with RSE affecting /ɹ/. However, some individuals do not respond to these methods and/or are not able to generalize proper articulatory configurations outside therapy (McAllister Byun & Hitchcock, 2012; Shriberg, 1975). A major clinical challenge is verbally describing complex articulatory configurations that vary across speakers and are visually concealed within the oral cavity (Cleland, Crampin, Wrench, Zharkova, & Lloyd, 2017; Delattre & Freeman, 1968). In addition, clients may have difficulty with phonetic perception and struggle to distinguish their output from the clinician's auditory model (Shuster, 1998).

Real-time visual feedback about articulation aims to allow the clinician to provide more explicit instructions while offering an additional sensory modality for the client to self-monitor and explore different articulatory positions. UVF represents a noninvasive means to provide an intraoral visual of the articulators. With UVF, real-time images of the tongue are generated by placing an ultrasound transducer against the skin beneath the chin (Preston, McAllister Byun, et al., 2017; Shawker & Sonies, 1985). Depending on the field of view and location of the sublingual cavity of the client, images in the sagittal (longitudinal) section may depict the tongue from anterior (tip or blade) to posterior (root). Ultrasound images may reveal elevation of the tongue tip or blade for an acoustically acceptable /ɹ/, as well as lowering of the posterior tongue dorsum and posterior movement of the tongue root toward the pharynx. UVF allows the clinician to target specific deviations that characterize a given child's distortion, such as an abnormally low tongue blade, an abnormally high tongue dorsum, or lack of tongue root retraction (Bacsfalvi, 2010; Boyce, 2015). Using these intraoral images, the child can observe the shape and movements of their own tongue and compare them to models representing correct production. Children may utilize the real-time visual display to explore different tongue shapes to achieve a perceptually acceptable production (McAllister Byun, Hitchcock, et al., 2014) and then self-monitor their articulatory configuration over the course of practice.

Improvements in articulation with UVF are presumed to be due to the detailed qualitative feedback it provides on movements of the tongue that are otherwise hidden. Such knowledge of performance (KP) feedback has been found to facilitate the acquisition of a new motor pattern, although it is argued that KP should be faded to encourage the learner to develop an internalized motor plan that can be readily generalized to new contexts (Hodges & Franks, 2001; Maas et al., 2008; McAllister Byun, 2017; Newell, Carlton, & Antoniou, 1990). In the case of articulatory feedback involving UVF, recent evidence suggests that the efficacy of treatment for residual /ɹ/ errors is influenced by the frequency and order of visual articulatory feedback, with greater gains observed when visual feedback is provided with a high frequency early in the treatment and reduced later (Preston et al., 2018). An analogous result was reported in the context of visual–acoustic feedback for RSE affecting /ɹ/, where children who received a period of treatment enhanced with a visual representation of the acoustic signal followed by a period of traditional treatment showed significantly greater gains than children who received the same treatments in the reverse order (McAllister Byun & Campbell, 2016). Therefore, visual feedback is hypothesized to confer greater benefit in early rather than later stages of learning.

Presently, over a dozen single-subject experimental studies and case studies have reported on the effects of UVF for dozens of individuals with RSEs, with the majority of children showing clear improvements in the accuracy of treated speech sounds. With respect to production of American English /ɹ/, improved accuracy has been reported in several dozen children attempting consonantal /ɹ/ (Shawker & Sonies,

[1]Among both clinicians and researchers, there is considerable diversity in the terminology and notation used for English rhotics in different syllable positions (see Lockenvitz, Kuecker, & Ball, 2015). Following a convention widely adopted among clinicians in North America, we use the term "consonantal /ɹ/" to refer to rhotics in onset position and "vocalic /ɹ/" to refer to rhotic offglides or nuclei (also transcribed /ɝ/ when stressed and /ɚ/ when unstressed).

1985), vocalic /ɹ/ (McAllister Byun, Hitchcock, et al., 2014; Sjolie, Leece, & Preston, 2016), and both (Adler-Bock et al., 2007; Bacsfalvi, 2010; Hitchcock & McAllister Byun, 2015; Preston et al., 2014, 2018; Preston & Leece, 2017; Preston, Leece, & Maas, 2017), although individual differences in treatment response are also commonly reported. Although UVF may provide visual information that is useful to both the clinician (for cueing) and client (for self-monitoring), traditional treatment is also effective for some children. With minimal research specifically comparing outcomes with and without UVF, it remains unclear whether the magnitude of progress is superior when UVF is included. Preston, Leece, et al. (2017) conducted a within-participant comparison in which each participant received seven sessions of traditional treatment without any visual feedback and seven sessions where treatment was enhanced with UVF in 50% of trials. In this cross-over design, six participants began in the UVF-enhanced treatment condition, and six began in the traditional treatment condition; all participants completed the same number of trials. Although within-session accuracy was greater in UVF-enhanced sessions, there was a negligible difference in generalization scores between conditions. However, children whose treatment order was UVF followed by traditional treatment showed slightly larger overall gains than children whose treatment order was the reverse. One limitation of that study was that, for each participant, different speech targets (i.e., consonantal and vocalic /ɹ/) were treated in the two conditions, which makes direct comparison across conditions difficult. Moreover, significant individual variation in response to both UVF-enhanced and traditional treatment conditions was observed (Preston, Leece, et al., 2017). In a different small-scale study, Bressmann, Harper, Zhylich, and Kulkarni (2016) compared changes in consonantal and vocalic /ɹ/ accuracy in four children ages 7–10 years receiving UVF and two children receiving traditional articulation therapy, reporting no statistically significant differences between groups. However, for children in the UVF group, only 10 min of each hour-long session included visual feedback, which is a smaller dosage of UVF than reported in most other studies. There is thus a need for further studies in which traditional treatment and UVF (in sufficient dose) are compared on the same speech target.

Purpose and Hypotheses

The purpose of this study was to compare gains in the production of vocalic /ɹ/ associated with eight sessions of traditional treatment and eight sessions of UVF-enhanced treatment, as well as the combination of these two treatments after 16 sessions. The order of the two phases of treatment was counterbalanced across participants. It was hypothesized that greater progress would be observed when visual feedback was included than when no visual feedback was included. Moreover, it was hypothesized that, with the total amount of UVF treatment held constant, better outcomes would be observed when treatment began with a phase of UVF followed by a phase of traditional treatment versus the reverse order.

Method

Participant Characteristics and Eligibility Criteria

Participants included 12 children (nine boys and three girls) who are native speakers of rhotic dialects of North American English and are between the ages of 9 and 14 years (M = 11 years 1 month). Children were referred by local speech-language pathologists (SLPs) or by parents from flyers posted throughout the community. To qualify for the study, participants were required to present with a primary RSE affecting /ɹ/ (i.e., RSE in the absence of any identifiable etiology or development disability such as cleft palate, autism, or hearing loss) characterized by a score below 85 on the Goldman–Fristoe Test of Articulation–Second Edition (Goldman & Fristoe, 2000), a score below 20% accuracy on a 50-item probe assessing /ɹ/ at the word level, and identifiable /ɹ/ distortions in a conversational speech, as assessed by a certified SLP. Participants passed a bilateral pure-tone hearing screening at 20 dB at 500, 1000, 2000, and 4000 Hz and demonstrated standard scores above 80 on the Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007) and scaled scores above 6 on the Recalling Sentences subtest of the Clinical Evaluation of Language Fundamentals–Fifth Edition (Wiig, Semel, & Secord, 2013).

In addition, speech motor functioning was assessed through administration of a maximum performance task, as follows (see Rvachew, Hodge, & Ohberg, 2005; Thoonen, Maassen, Gabreels, & Schreuder, 1999). Duration measures were recorded for sustained phonemes /f/, /s/, /z/, and /ɑ/, and syllable rate was measured for rapid production of the repeated syllables /pɑ/, /tɑ/, and /kɑ/ and the trisyllable sequence /pɑtɑkɑ/; the accuracy of the /pɑtɑkɑ/ sequence was also scored. These measures were used to derive separate scores for dysarthria (based on the short duration of sustained phonemes or slow syllables) and apraxia (based on slow and inaccurate trisyllables), whereby 0 represents *not dysarthric/apraxic*, 1 is *undefined*, and 2 represents *probable dysarthria or apraxia*. Inclusion in the study required that participants receive a score of less than 2 on both the dysarthria and apraxia scales.

Additional Descriptive Assessments

Before treatment, children completed additional assessments for descriptive purposes. These tasks included the Formulated Sentences subtest of the Clinical Evaluation of Language Fundamentals–Fifth Edition (Wiig et al., 2013) and the Phonological Awareness subtests of the Comprehensive Test of Phonological Processing–Second Edition (Wagner, Torgesen, Rashotte, & Pearson, 2013). Participants were also assessed for stimulability in /ɹ/ production using a probe adapted from Miccio's (2002) study, which required direct imitation of 11 different syllables (e.g., /ɹɑ, ɹu, ɹi/) three times each for a total of 33 productions. Finally, parent reports indicated that previous therapy had been provided for a median of 4 years prior to the study (range: 0–11 years). Pretreatment assessment data are presented in Table 1.

Table 1. Characteristics of 12 children participating in ultrasound visual feedback and traditional treatments.

Participant	129	130	131	132	134	135	136	139	140	141	142	143
Gender	M	F	M	M	M	M	F	M	M	M	F	M
Age (years;months)	9;10	11;8	12;5	11;6	11;0	14;3	8;11	11;11	9;1	12;6	10;2	11;5
GFTA-2 Std score	46	83	74	78	67	54	84	75	58	59	78	78
GFTA-2 Percentile	< 1	2	2	3	1	1	3	3	2	1	1	1
PPVT-4 Std score	105	128	114	121	103	132	118	124	104	139	127	121
CELF-5 RS Scaled score	13	16	13	8	13	8	16	10	16	12	15	11
CELF-5 FS Scaled score	8	9	9	9	12	10	14	9	11	15	12	16
CTOPP-2 PA Composite	84	120	96	96	103	84	125	88	107	103	90	107
MaxPT Apraxia score	1	1	0	0	1	0	0	1	0	1	1	1
MaxPT Dysarthria score	0	0	0	0	0	0	0	0	0	0	0	0
Stimulability (out of 33)	0	5	0	0	0	0	0	0	0	0	0	0
Ages at which child received speech therapy	3 years to present	8 years to present (inconsistently)	5 years to present	8 years present	7 years to present	3 years to present	None	In elementary school	None	3–4 years, 9–10 years	None	9–11 years
Approximate years of therapy	7	4	7	4	4	11	0	1	0	4	0	2
Additional sound errors	/s/	None	None	None	/s/	None	None	None	None	None	None	None

Note. M = male; F = female; GFTA-2 = Goldman–Fristoe Test of Articulation–Second Edition; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition; CELF-5 RS= Clinical Evaluation of Language Fundamentals–Fifth Edition, Recalling Sentences; CELF-5 FS: Clinical Evaluation of Language Fundamentals–Fifth Edition, Formulated Sentences; CTOPP-2 PA= Comprehensive Test of Phonological Processing, Phonological Awareness; MaxPT=maximum performance tasks.

Study Design

A multiple baseline across-subjects single-case experimental design was used. Each child participated in a pretreatment baseline phase (with three to five staggered baseline probes; see Byiers, Reichle, & Symons, 2012), Treatment Phase I (eight sessions), a midpoint no-treatment phase (three probes), Treatment Phase II (8 sessions), and a maintenance phase (three probes). All data collection sessions (baseline, Treatment Phase I, midpoint, Treatment Phase II, and maintenance) were scheduled to occur twice per week. Each participant was exposed to both treatment conditions, with the order counterbalanced across participants. The ordering of treatment conditions, as well as the number of baseline sessions, was randomly assigned through concealed envelope, such that six children received UVF followed by traditional treatment and the other six received traditional treatment followed by UVF.

Probe Data

A 50-word probe eliciting /ɹ/ in untrained words in a range of phonetic contexts was administered in all baseline, midpoint, and maintenance sessions (see Supplemental Material S3). The 50-item probe contained 25 instances of vocalic /ɹ/ in stressed syllables (/ɝ/, /ɑɹ/, /ɔɹ/, /ɪɹ/, and /ɛɹ/), five unstressed /ɚ/, and 20 consonantal /ɹ/. In addition, treatment sessions began and ended with a 25-item subset probe eliciting 18 vocalic and seven consonantal targets. Probe words were presented in random order on a computer screen, and responses were recorded with a Sennheiser lapel microphone with a sampling rate of 44.1 kHz and 16-bit encoding. No feedback was provided during the probes. The primary outcome measure was accuracy in vocalic /ɹ/ as rated by naïve listeners (see Probe Measurement below).

Treatment Procedures

The same American Speech-Language-Hearing Association–certified SLP conducted all therapy sessions. She was aware that the study was comparing the two treatment conditions, that both conditions have been shown to be effective, and that naïve listeners would be used to help determine whether or not the conditions were equally effective. However, she was not involved in the study design and was not made explicitly aware of the study's hypotheses, nor were any clinician ratings used to evaluate the efficacy of treatment in this study. Participants did not receive outside treatment for /ɹ/ during the study. For fidelity checking, audio recordings were collected in all sessions, and video recordings of the ultrasound images were collected for UVF sessions.

Target Selection

Three targets were selected for each participant from among the vocalic variants /ɝ/, /ɑɹ/, /ɔɹ/, /ɪɹ/, and /ɛɹ/. These were selected to reflect the three variants that were the least accurate on the pretreatment baseline probes for each child based on the study clinician's rating. These targets remained the same for both treatment phases. The rationale for targeting vocalic variants was that they have been argued to emerge earlier in development (McGowan, Nittrouer, & Manning, 2004) and are reported to be more stringently rated by nonexpert listeners, like those used in this study (Klein, Grigos, McAllister Byun, & Davidson, 2012). Choosing all targets from the same broad category (i.e., vocalic /ɹ/) was also intended to enhance internal validity.

Treatment Session Prepractice

Following the administration of the 25-item subset probe, each treatment session began with a timed 5-min prepractice period. During each participant's first session, tongue anatomy and articulatory requirements for /ɹ/ were discussed. In both UVF and traditional phases of treatment, a poster with 22 magnetic resonance (MR) images of various adult speakers producing /ɹ/ was used to describe various tongue shapes and positions for correct production (Boyce, 2015). Traditional shaping strategies (e.g., shaping /ɹ/ from /l/ or /ɑ/) and phonetic placement cues were used at the clinician's discretion. Prepractice was relatively unstructured to allow /ɹ/ to be practiced in whatever contexts were judged to be most helpful for the participant; targets in different phonetic contexts at the syllable, word, and/or sentence level could be selected based on the treating clinician's judgment.

In the UVF condition, the first treatment session featured an initial overview of how to interpret ultrasound images. During this training, each participant viewed an ultrasound image of their tongue and was taught to identify the side that represents the "front" and "back," identify the different parts of their tongue (tip, blade, dorsum, root), and discuss the major features of articulation of /ɹ/. In subsequent sessions, only a brief review during prepractice was needed to aid in interpreting ultrasound images. The child's tongue shape on the ultrasound display was compared against the MR images. Children were also encouraged to attempt to copy different tongue shapes from the MR images to achieve a perceptually accurate /ɹ/.

Treatment Session Structured Practice

Each prepractice period was followed by structured practice, which was designed to elicit 162 practice attempts. However, structured practice was terminated after 45 min, even if all 162 trials had not yet been completed. Stimuli and feedback prompts were presented using a researcher-developed open-source software program called Challenge-R (McAllister Byun, Ortiz, & Hitchcock, 2014). Drawing on motor learning research (Guadagnoli & Lee, 2004; Rvachew & Brosseau-Lapré, 2012), the Challenge-R aims to keep learners at an optimal level of difficulty ("challenge point") during speech practice. The software presents words from standard lists and records the treating clinician's ratings of accuracy and then uses the ratings as the basis for adaptive changes in practice complexity along multiple parameters, described in detail below. Each participant practiced three vocalic /ɹ/ contexts, and the program randomly selected three syllables or words per context for a total of nine targets

each session. Each target word or syllable was practiced up to 18 times per session (in three blocks of six attempts).

The clinician provided a verbal model prior to each block of six trials, and the Challenge-R software prompted her to provide verbal KP feedback at the end of each block (qualitative feedback such as "I like the way you kept the front of the tongue up" or "I see the dorsum lifting too high, remember to keep it down"). Each production of vocalic /ɹ/ was rated by the clinician as 0 (*substituted or distorted*) or 1 (*correct*) based on her clinical judgment, and accuracy was tallied by Challenge-R program. On a randomly selected subset of trials, the software also prompted the clinician to provide verbal knowledge of results feedback (feedback characterizing the accuracy of each production as "correct" or "not quite").

The Challenge-R software altered the intended difficulty of practice based on accuracy of the preceding six trials: five or more correct responses in a block prompted an increase in difficulty, three or fewer correct responses prompted a decrease in difficulty, and four correct responses resulted in no change in difficulty. Two parameters influencing task difficulty were adjusted: stimulus complexity (number of syllables per word, the presence or absence of the competing phonemes /l/ and /w/, and the presence or absence of a carrier phrase or sentence context) and frequency of verbal knowledge of results feedback (a reduction from four to three to two trials per block of six trials). In addition to these within-session changes, the schedule of stimulus presentation could be adjusted on a between-sessions basis. If cumulative accuracy in a session exceeded 80%, the next session changed 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). If the child again achieved > 80% accuracy at the session level, the schedule changed again to fully random practice (each trial within a block featured a randomly selected stimulus item). Each session began with the parameters that were used at the end of the participant's previous session.

During structured practice in either phase of treatment, an MR image of correct articulation of /ɹ/ was made available to instruct the client on correct positioning of the articulators. The image used for a given participant was selected to highlight specific aspects of tongue shape, such as elevation of tongue blade or lowering of dorsum, that were judged to be appropriate for improving that individual's rhotic production. Switching to a new MR image was possible if the clinician judged that the current target was not facilitating correct production.

Condition Differences

In the traditional treatment condition, visual feedback was not provided. In the UVF-enhanced condition, visual feedback was made available in 44% of trials (12 of 27 blocks). The Challenge-R software prompted whether ultrasound feedback should be provided or withheld in a given block. During trials in which UVF was available, a Siemens Acuson X300 ultrasound with C6-2 transducer was used. The participant was instructed on how to position the transducer beneath the chin to collect midsagittal images, and no head stabilization equipment was used. Participants had the option of holding the ultrasound transducer unassisted or placing it in a microphone stand. An MR image was positioned adjacent to the ultrasound display so that the clinician could discuss differences between the MR image and ultrasound tongue shapes produced by the child.

Treatment Fidelity

To monitor fidelity of treatment, research assistants reviewed video recordings from two sessions per participant, with one session selected from each treatment phase. Fidelity checks assessed whether the clinician provided verbal models at the beginning of each block (but not on subsequent trials within a block). Modeling was provided as prescribed on an average of 97% of trials (range per session: 89%–100%, SD = 2.9%). In addition, 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 97% of trials (range: 90%–100%, SD = 2.9%). When verbal KP feedback was expected, the appropriate feedback type was provided 95% of the time (range: 89%–100%, SD = 3.4%).

Although all sessions aimed to elicit the same number of practice trials, the actual number of trials elicited could be lower if the child did not get through all 162 trials within the 45-min period allotted for structured practice. To address this possible confound, the total number of practice trials completed was compared across conditions. The mean number of practice trials in the UVF-enhanced sessions was 139 (SD = 27), whereas the mean number of trials in traditional treatment sessions was 133 (SD = 29). A paired t test indicated no significant difference in the number of practice trials across conditions, $t(11) = 0.81$, $p = .438$.

Probe Measurement

The outcome variable in the study was perceptually rated accuracy of the treated variant, vocalic /ɹ/, in untreated words elicited during baseline, midpoint, and maintenance sessions, as well as probes administered at the start and end of each treatment session (Phase I and Phase II). Ratings from the treating SLP were not used as the outcome variable; instead, ratings were collected from naïve listeners. To prepare for these rating, audio files from all probes were segmented into individual words and normalized to a standard intensity of 70 dB. Audio files were then uploaded to Amazon Mechanical Turk crowdsourcing platform, where naïve listeners made binary judgments of accuracy (correct/incorrect) for the /ɹ/ sound in each token. Files were randomized within and across speakers, and listeners were blind to the treatment phase in which each recording was elicited. Previous research validating the use of crowdsourced listeners' ratings of children's /ɹ/ productions (McAllister Byun, Halpin, & Szeredi, 2015) found that binary ratings aggregated across at least nine naïve listeners recruited online converged with ratings aggregated across three expert listeners.

Accordingly, in this study, binary ratings of each speech token was collected from at least nine listeners. Listeners had United States–based IP addresses and, per self-report, were native speakers of English with no history of speech or hearing impairment. The total number of listeners was 51, with each listener rating an average of 639 files. Raters had a mean self-reported age of 40.8 years (SD = 9.8 years). When aggregating 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). As a measure of interrater reliability, the proportion of raters who agreed with the modal rating for each token was calculated. On average, agreement across raters was 81.8%.

Analyses

For maximally robust results, both within-participant and across-participants data were analyzed with multiple methods, including visual inspection, effect sizes, and a mixed-effects logistic regression model. Visual inspection is the conventional approach to the analysis of single-subject experimental data (Kratochwill & Levin, 2014), but limitations can include low interrater reliability (e.g., Brossart, Parker, Olson, & Mahadevan, 2006). Therefore, it is considered good practice to augment this approach with quantitative evidence, including effect sizes and hypothesis tests (Kratochwill & Levin, 2014).

Standardized effect sizes were computed using a modified version of Busk and Serlin's d_2 statistic (Beeson & Robey, 2006). In the typical calculation of effect size, the difference between phases is divided by the standard deviation (SD) pooled across the two phases being compared. However, when variance is very low in both phases included (which was often the case for baseline and maintenance phases), this method can result in inordinately high standardized effect sizes. To avoid inflated estimates, effect sizes were calculated using SD pooled across all three probe phases (baseline, midpoint, and maintenance) for a given participant. To be considered clinically relevant, effect sizes were required to exceed 1.0; that is, the change in accuracy from pre- to posttreatment must exceed the pooled SD (Maas & Farinella, 2012). Three standardized effect sizes were calculated for each participant: Phase I (from baseline to midpoint), Phase II (from midpoint to maintenance), and for both phases taken jointly (from baseline to maintenance). Effect sizes are computed from periods of no treatment because we are interested in measuring speech motor learning, which includes both generalization and retention. Effect sizes were calculated using $\hat{p}_{\text{correct}}$ pooled across all vocalic /ɹ/ variants. Unstandardized effect sizes (i.e., the raw change in accuracy across a treatment phase) were also considered to support interpretation of participants' response to treatment.

To compare outcomes across individuals, two logistic mixed-effects models were implemented (cf. Rindskopf & Ferron, 2014). An uncollapsed data set was used in which the binary rating (correct/incorrect) assigned to each token by each listener served as the dependent variable. The first model examined outcomes at midpoint. Recall that participants were randomly assigned to receive either traditional treatment or UVF in Phase I; thus, the effect of treatment condition could be assessed in the absence of any confounding influence of treatment order. Fixed effects included treatment condition and mean percentage of tokens rated correct during the baseline phase, as well as the interaction between those factors. Baseline accuracy was included as an individual-level characteristic based on prior evidence that accuracy at the outset of treatment can influence the magnitude of treatment response (e.g., Preston et al., 2018). Random intercepts were included to reflect the fact that data points were nested within raters and words.

The second model examined data at the end of Phase II to test whether the magnitude of change after all 16 sessions was impacted by the order of treatment delivery (traditional treatment followed by UVF or the reverse). Fixed effects included treatment order and baseline accuracy, as well as their interaction. As in the previous model, random intercepts were included to capture data nested within raters and words.

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

Results

Descriptive Results: Visual Inspection of Treatment Effects

Figures 1 and 2 show each participant's pattern of change in accuracy ($\hat{p}_{\text{correct}}$ aggregated across all items in a probe) before, during, and after the two treatment phases. Participants are grouped by treatment condition (UVF first in Figure 1 and traditional treatment first in Figure 2), and they are ordered by increasing length of the baseline phase. In each session, a black circle represents performance on the 25-item subset probe administered pretreatment, and a red star represents performance on the same probe administered at the end of the session. Thus, the distance between the two probes in a session indexes change over the course of that treatment session. The dashed horizontal line represents the participant's mean $\hat{p}_{\text{correct}}$ during the baseline phase, presented for comparison to subsequent phases.

Visual inspection of baseline data raised no questions of extreme outliers or rising baselines for any participant. Furthermore, all participants demonstrated < 10% mean session-to-session variability across the baseline phase. In summary, all participants were judged to demonstrate sufficiently stable baselines to serve as the basis for an evaluation of treatment effects.

Figure 1 shows data from the six participants who received UVF-enhanced treatment in Phase I followed by traditional Treatment Phase II. Participants 129 and 140

Figure 1. Individual plots for six participants who received ultrasound visual feedback treatment followed by traditional treatment. *Y*-axis represents proportion of probe words rated as correct. *X*-axis represents time (BL = baseline, Tx = treatment session, MP = midpoint, MN = maintenances). UVF = ultrasound visual feedback; Trad = traditional treatment. 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.

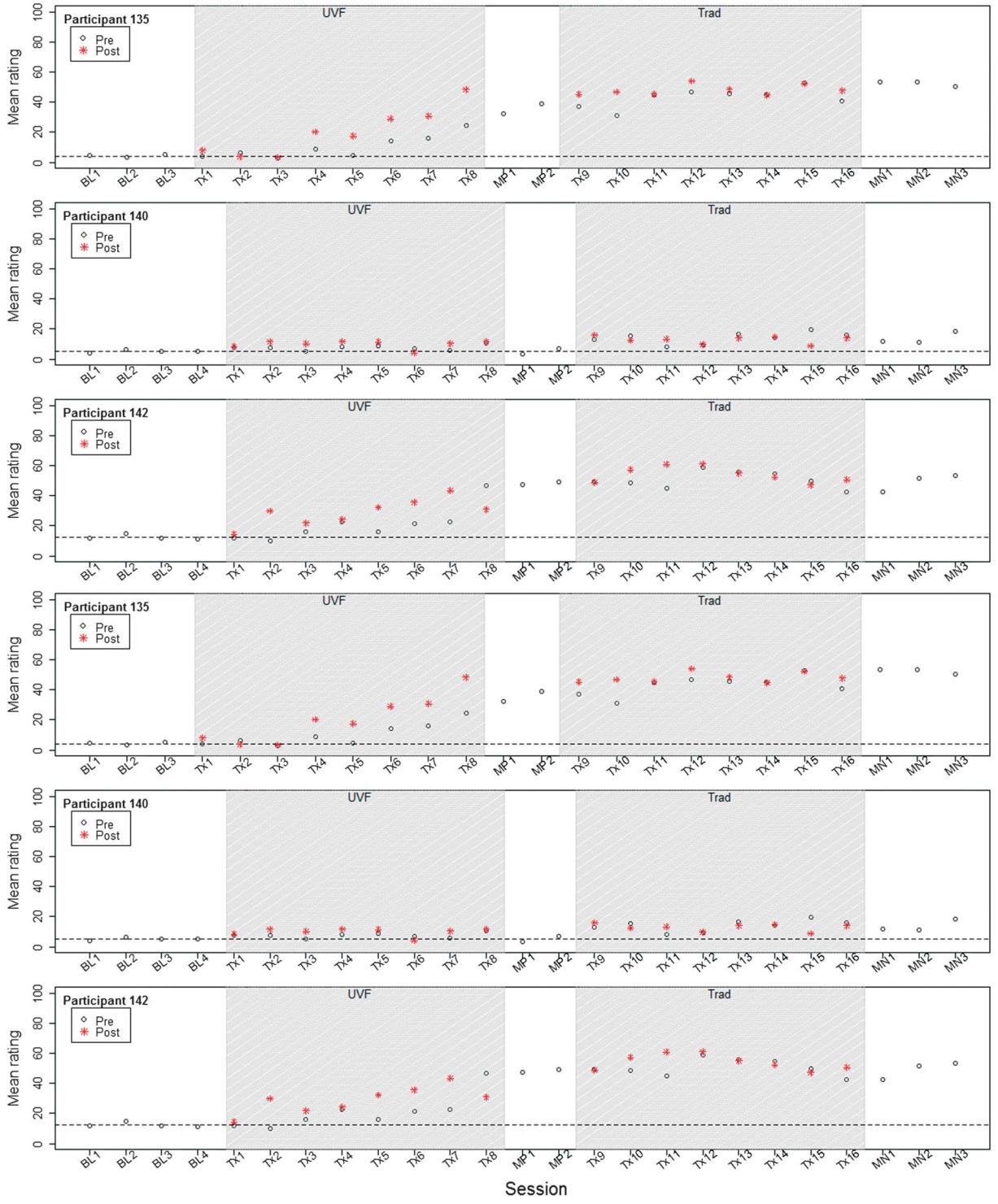
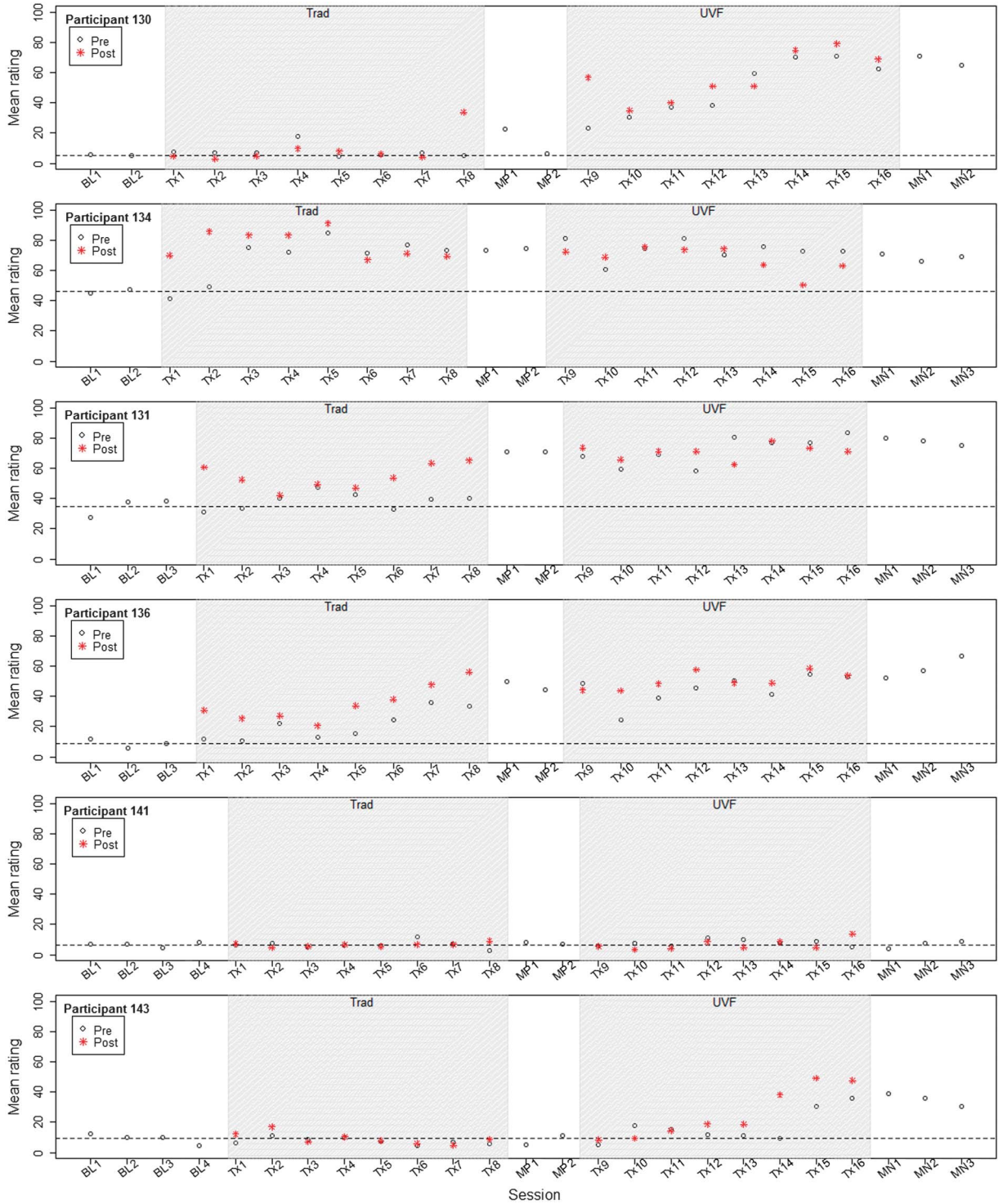


Figure 2. Individual plots for six participants who received traditional treatment followed by ultrasound visual feedback treatment. *Y*-axis represents proportion of probe words rated as correct. *X*-axis represents time (BL = baseline, Tx = treatment session, MP = midpoint, MN = maintenances). UVF = ultrasound visual feedback; Trad = traditional treatment. During days on which treatment occurred, probes were administered before the session (black circles) and after the session (asterisks). Dashed line represents the participant's mean baseline accuracy.



did not make significant progress in either treatment condition, although Participant 140 showed a very small degree of improvement at the end of Phase II and in the maintenance phase. The remaining four participants exhibited substantial gains during treatment and sustained improvements after treatment. Participant 132 exhibited a sizable response to treatment, starting near 0% accuracy at baseline, showing increased accuracy after the fourth session of UVF, and exceeding 80% accuracy in the midpoint phase; these gains were sustained through Phase II and in the posttreatment maintenance phase. Participants 135 and 142 steadily improved throughout treatment, with modest gains in Phase I (UVF) and increasingly accurate performance in Phase II (traditional treatment). Both participants showed the highest accuracy in the maintenance phase, which suggests continued generalization. Participant 139 started with the highest baseline accuracy in this group and showed variable and inconsistent /ɹ/ accuracy throughout both treatment phases.

Figure 2 depicts data from the six participants who received traditional treatment in Phase I followed by UVF-enhanced treatment in Phase II. Participant 141 did not demonstrate change in either treatment condition. The rest of the participants showed improvement above baseline performance in at least one condition. Participants 130, 131, and 136 exhibited the most significant gains, with more consistent progress in Phase II (UVF) and strong performance across the posttreatment maintenance phase. Participant 130 showed minimal response to treatment in Phase I (traditional treatment) but substantially increased accuracy Phase II (UVF). Participant 134 started with the highest overall baseline accuracy (over 40%) and made immediate progress but then showed inconsistent performance over the remainder of the treatment period, ultimately demonstrating a moderate degree of improvement. Participant 143 did not show changes in Phase I of treatment but demonstrated moderate gains at the end of Phase II, with gains mostly maintained in the posttreatment maintenance phase.

Descriptive Results: Individual Effect Sizes

Table 2 shows participants' accuracy and effect sizes based on change in $\hat{p}_{\text{correct}}$. The mean number of probe words on which $\hat{p}_{\text{correct}}$ scores are based was 15.67 (SD = 3.48) for pre- and posttreatment probes and 29.59 (SD = 0.66) for baseline and maintenance probes.[2] (Recall that only vocalic variants, which were targeted in treatment, were rated for calculation of effect sizes.) 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.

[2]Because only vocalic targets were examined for this study, the maximum number of possible words rated was 18 for the within-treatment probes and 30 for the baseline and maintenance probes. The number of items actually rated was somewhat lower due to data loss when ratings were not successfully collected from nine blinded listeners.

Participants in Table 2 are presented by the order in which they received treatment (UVF first or traditional treatment first), and for parallelism with figures above, they are ordered by increasing length of the baseline interval. The first column shows the mean and SD of $\hat{p}_{\text{correct}}$ in the baseline period, averaged across all vocalic /ɹ/ items from all baseline sessions (before Phase I). The second column shows the mean and SD across all midpoint sessions (between Phases I and II), and the third shows the mean and SD across the three maintenance sessions (following Phase II). The next three columns report three standardized effect sizes: baseline versus midpoint scores (ES_{PhaseI}), midpoint versus maintenance ($ES_{PhaseII}$), and baseline versus maintenance (reflecting overall gains across both phases of treatment, ES_{All}). Effect sizes for each condition (UVF vs. traditional treatment) are reported in the next two columns. The second-to-last column reports the difference in effect sizes between the two conditions independent of the order in which they were delivered. The final column shows the difference in effect sizes between the first and second phases (Phase II − Phase I), independent of which treatment was administered in each phase. Variability in overall response to treatment is evident across individuals.

After both phases of treatment, 10 of the 12 participants had achieved standardized effect sizes (ES_{All}) that exceed the minimum value (1.0) to be considered clinically significant. However, two participants (Participants 129 and 141) did not meet this criterion. A third participant (Participant 140) had an ES_{All} of 2.65, but the raw change in $\hat{p}_{\text{correct}}$ from baseline to maintenance was less than 10%, suggesting that the standardized effect size was somewhat inflated by low variance in this case.

Descriptive Results: Aggregated Effect Sizes

In this section, graphical representations and descriptive statistics are used to characterize the distributions that arise when standardized effect size data are partitioned in different ways. Inferential statistics are deferred until the logistic mixed models reported in the next section. In the boxplots that follow, middle bars represent medians, boxes represent the interquartile range (IQR; 25th–75th percentile), whiskers extend to the most extreme nonoutlier data point, and outliers (more than 1.5 × IQR outside the IQR) are represented as single points.

Figure 3a compares standardized effect sizes associated with UVF treatment phases versus traditional treatment phases, independent of the order in which the two types of treatment were administered. This figure reveals that the effect size distributions of the two treatments overlap, but the median value was higher for UVF (Mdn = 5.56) than traditional treatment (Mdn = 1.22). These relative magnitudes also hold when comparing unstandardized effect sizes: Calculations from the raw data in Table 2 indicate that participants who began in UVF treatment showed a median raw increase from baseline to midpoint in $\hat{p}_{\text{correct}}$ of 34.0 percentage points (range: −0.5 to 76.2), whereas those who

Table 2. Proportion of /ɹ/ tokens rated correct or $\hat{p}_{correct}$ at baseline, midpoint (after Phase I), and maintenance (after Phase II), along with associated standardized effect sizes.

Participant	Baseline *M* (*SD*)		Midpoint *M* (*SD*)		Maintenance *M* (*SD*)		ES_{PhaseI}	$ES_{PhaseII}$	ES_{all}	Treatment order	UVF advantage	Order effect
132	1.37	(0.90)	77.54	(5.84)	78.99	(2.83)	21.75	0.41	22.16	UVF_Trad	21.34	–21.34
139	21.12	(4.35)	59.22	(5.77)	62.67	(2.27)	9.41	0.85	10.26	UVF_Trad	8.56	–8.56
129	8.58	(3.14)	8.07	(0.60)	8.34	(2.28)	–0.22	0.12	–0.10	UVF_Trad	–0.34	0.34
135	4.09	(0.83)	36.02	(3.25)	52.48	(1.56)	17.03	8.78	25.81	UVF_Trad	8.25	–8.25
140	5.61	(1.31)	7.94	(4.74)	13.84	(4.21)	0.75	1.90	2.65	UVF_Trad	–1.15	1.15
142	12.39	(1.43)	48.53	(1.04)	49.12	(5.70)	12.49	0.21	12.70	UVF_Trad	12.28	–12.28
130	5.94	(1.34)	14.48	(11.67)	66.01	(4.22)	1.58	9.54	11.12	Trad_UVF	7.96	7.96
134	44.57	(3.19)	76.49	(4.28)	68.91	(2.51)	10.12	–2.40	7.72	Trad_UVF	–12.52	–12.52
131	33.59	(5.48)	69.96	(1.53)	77.86	(2.36)	10.00	2.17	12.17	Trad_UVF	–7.83	–7.83
136	9.64	(2.73)	47.60	(2.70)	58.73	(7.43)	8.85	2.60	11.44	Trad_UVF	–6.25	–6.25
141	6.55	(1.43)	7.09	(1.17)	6.74	(2.45)	0.34	–0.22	0.12	Trad_UVF	–0.56	–0.56
143	8.59	(3.00)	7.31	(3.46)	35.14	(4.25)	–0.39	8.52	8.13	Trad_UVF	8.91	8.91

Note. Participants are ordered based on the number of pretreatment baseline sessions (see Figures 1 and 2). *SD* = standard deviation; ES = effect size; UVF = ultrasound visual feedback; Trad = traditional treatment.

were assigned to traditional treatment showed a median increase of 20.2 (range: −1.3 to 38.0). 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 ES_{Phase1} tended to be larger than ES_{Phase2}, although there is considerable overlap between the two distributions.

Figure 4 examines the interaction between treatment type and order of treatment delivery (UVF treatment first versus traditional treatment first). These plots support the impression that both phase order (first vs. second phase) and treatment condition (UVF vs. traditional) played a role in influencing effect sizes.

Finally, Figure 5 contrasts the overall effect sizes (ES_{All}) for children who received UVF treatment first versus children who received traditional treatment first. Figure 5 shows that ES_{All} tended to be greater for children who received UVF before traditional treatment, compared to children who received the reverse order. With respect to unstandardized effect sizes derived from Table 2, participants whose treatment order was UVF then traditional showed a median raw increase in $\hat{p}_{correct}$ from baseline to posttreatment of

Figure 3. Boxplots depicting the distribution of effect sizes observed in connection with (A) ultrasound visual feedback versus traditional treatment, independent of phase order, and (B) Phase 1 versus Phase 2 of treatment, independent of treatment condition

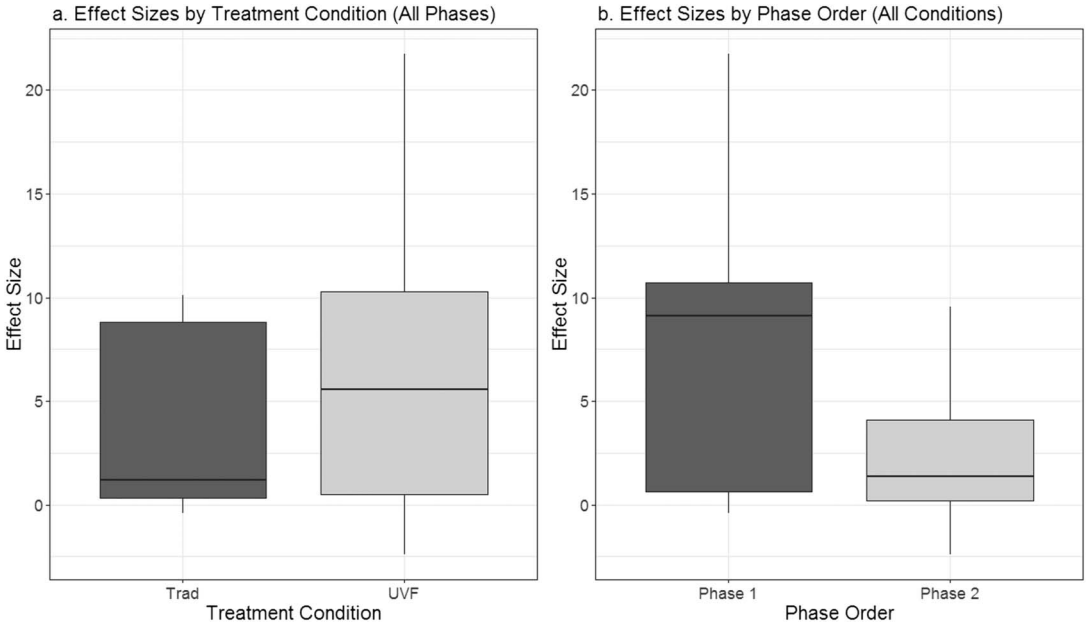
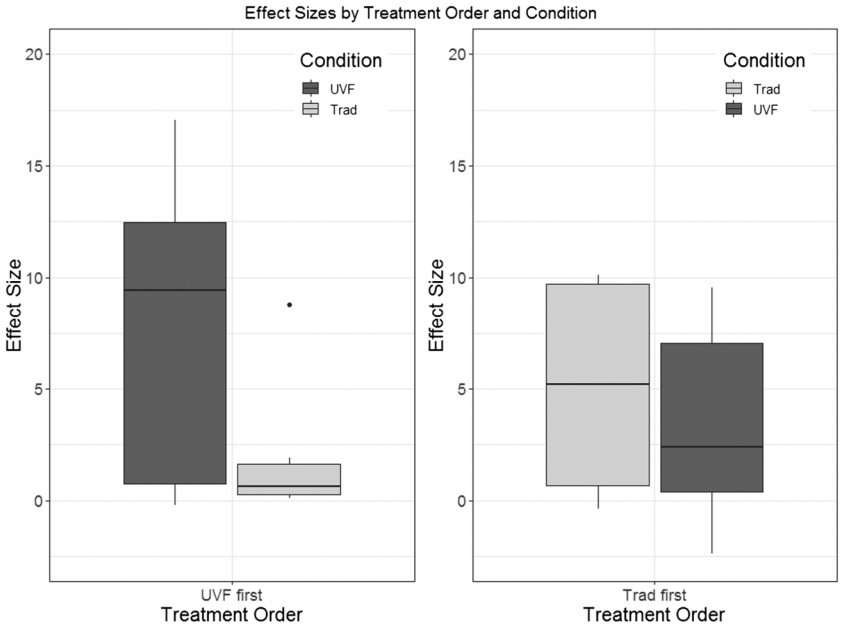


Figure 4. Boxplots depicting the distribution of effect sizes observed in connection with ultrasound visual feedback (UVF) versus traditional treatment when UVF was provided first, versus the opposite order.

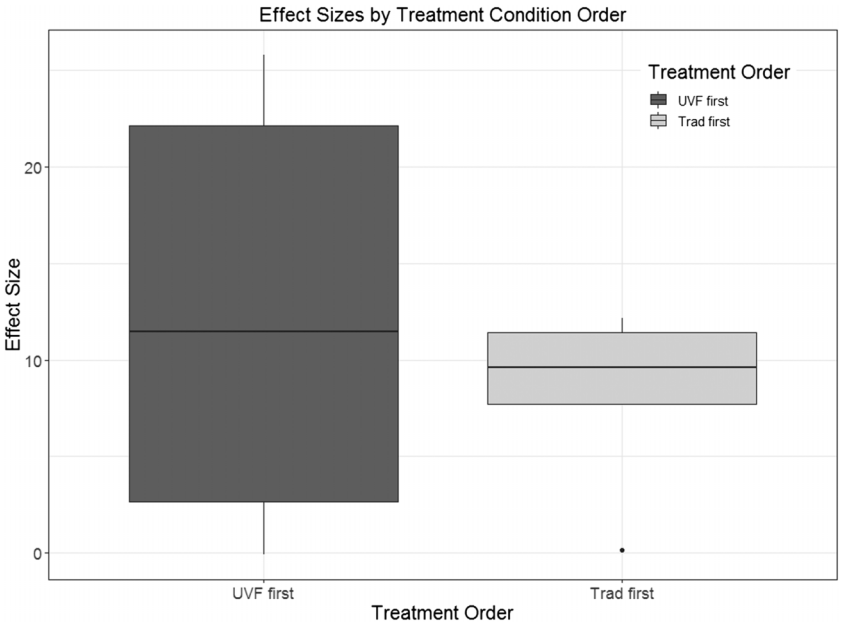


39.1% (range: −0.2 to 77.6), whereas those who were assigned to the order traditional treatment followed by UVF showed a median increase of 35.4% (range: 0.2–60.1). However, there was also greater variability in outcomes for children who received UVF before traditional treatment than the reverse order.

Across-Subjects Comparisons: Mixed-Effects Logistic Models

Quantitative comparison of outcomes across conditions was carried out using the two logistic mixed-effects models described above. The first model examined predictors of

Figure 5. Boxplots depicting the distribution of overall effect sizes observed when ultrasound visual feedback (UVF) treatment was provided first versus the opposite order.



accuracy at midpoint, after each participant had received a single phase of treatment (either UVF or traditional). Unsurprisingly, higher mean accuracy across the pretreatment baseline phase was associated with higher accuracy at midpoint ($\beta = 1.99$, $SE = 0.50$, $p < .001$). Importantly, treatment condition was also a significant predictor of /ɹ/ accuracy ($\beta = -1.70$, $SE = 0.07$, $p < .001$), with results indicating that, in Phase I, the traditional treatment condition was associated with significantly lower accuracy compared to the UVF condition. There was also a significant interaction between treatment condition and baseline accuracy ($\beta = 7.04$, $SE = 0.55$, $p < .001$). This interaction can be visualized in Supplemental Material S1, which shows that the association between baseline accuracy and midpoint accuracy was slightly steeper in the traditional than the UVF treatment condition. However, this interaction must be interpreted with caution in light of the small number of data points. Complete results of the regression are reported in Appendix A.

The second logistic mixed model had the same structure but evaluated accuracy following completion of both Phases I and II. There was no main effect of baseline accuracy ($\beta = 0.03$, $SE = 0.40$, $p = .94$). Importantly, there was a significant main effect of treatment order ($\beta = -0.51$, $SE = 0.06$, $p < .001$) indicating that beginning in the traditional treatment condition was associated with significantly lower posttreatment accuracy than beginning in the UVF condition. There was also a significant interaction between treatment condition and baseline accuracy ($\beta = 5.48$, $SE = 0.43$, $p < .001$). This interaction is depicted in Supplemental Material S2, which again shows the association between pre- and posttraining accuracy was slightly steeper in the children who began with traditional treatment compared to those who began with UVF treatment. Again, we avoid attaching a strong interpretation to this interaction in light of the small number of data points. Complete results of this regression are reported in Appendix B.

Discussion

This study compared outcomes following eight sessions of UVF-enhanced treatment and eight sessions of traditional treatment, counterbalanced in order, in 12 children with RSE. The present results confirm previous studies indicating that UVF can facilitate improved speech sound accuracy in many children with RSE. Specifically, the study revealed that, overall, larger gains in speech sound accuracy were observed following eight sessions of UVF (median increase of 34.0. percentage points in $\hat{p}_{\text{correct}}$) than after eight sessions of traditional treatment (median increase of 20.2 percentage points in $\hat{p}_{\text{correct}}$). In addition, when presented in sequence, larger gains were observed for children who began treatment with UVF followed by traditional treatment than for children who started in traditional treatment and were later exposed to UVF. This finding suggests that speech sound learning may be enhanced when treatment conditions are appropriately sequenced so that detailed feedback is initially provided to build a visuomotor representation of a sound and then faded to allow continued practice without visual support.

These results are in line with predictions from schema-based motor learning theory indicating that detailed feedback should be provided in the early stages of acquisition and then withdrawn to facilitate motor learning (Maas et al., 2008; Newell et al., 1990). Although the literature on generalization in motor learning in nonclinical populations indicates that learning may be inhibited when too much detailed feedback is provided (Ballard et al., 2012; Hodges & Franks, 2001; Maas et al., 2008), this appears not to have been the case for children with RSE in this study. Indeed, overall generalization to untrained words appears to have been facilitated, not inhibited, by treatment incorporating detailed articulatory feedback. Thus, for many children with RSE, it may be the case that the amount of information provided through visual feedback is favorable for successful speech sound learning. However, it is unclear whether earlier fading of UVF could yield greater gains in speech motor learning.

These results are in agreement with numerous previous studies on the efficacy of UVF for remediating /ɹ/ distortions (e.g., Adler-Bock et al., 2007; McAllister Byun, Hitchcock, et al., 2014; Preston et al., 2014, 2018), as well as other speech sound errors (Cleland, Scobbie, & Wrench, 2015), although not all studies have reported a clear advantage of biofeedback over traditional treatment (Bressmann et al., 2016; Preston, Leece, et al., 2017). Clinically, UVF is one viable treatment option for children with RSE, and the results of this small-scale study suggest that, on average, treatment effects may exceed those observed with traditional treatment. The results are also in line with findings from studies of other biofeedback types suggesting advantages derived from visual feedback-enhanced therapy (Gibbon et al., 2001; McAllister Byun & Campbell, 2016). Even though the current evidence base for UVF rests in small-scale experimental studies and case studies, the existence of numerous replications in the literature to date suggests that UVF should be considered an evidence-based approach for RSE. However, larger-scale studies including randomized controlled trials would help to adjudicate whether UVF clearly outperforms traditional treatment. Larger scale research is also needed to identify pretreatment factors associated with individual differences in response to treatment.

Three of the participants whose treatment order was traditional followed by UVF showed significant improvement during the initial phase of traditional treatment. This strong response in Phase I resulted in less room for further improvement in the UVF condition, and indeed, those three participants showed a larger standardized effect size in traditional than UVF treatment. The other three participants who received traditional followed by UVF treatment showed an advantage for UVF over traditional treatment. The three who improved the most with traditional treatment were also the individuals with the highest baseline accuracy levels, suggesting that traditional treatment may be best suited for individuals with initially higher accuracy levels. The three who responded well to traditional treatment had a range of

experience with previous treatments (0, 4, and 7 years); thus, the reason for their significant improvement in this study could differ across participants and could include factors such as readiness for change or exposure to a new SLP. The responders in the traditional condition confirm that well-structured traditional therapy can be effective even in some cases where previous treatment had been unsuccessful. Our results also suggest that UVF can be a viable option for some children who have not benefitted from previous traditional therapy.

Currently, UVF-enhanced treatment is most commonly adopted after other interventions have failed. This may be due to several factors, including the cost of the technology, access to training, and the cognitive requirements to benefit from visual feedback (i.e., older children may be more likely to benefit from the visual display than young children). Future studies may explore whether UVF can be a viable and cost-effective first-line intervention for some children with RSE, rather than waiting for other approaches to fail. In addition, the majority of published research to date has focused on RSEs affecting /ɹ/, and additional studies evaluating UVF for other sound errors would help to define the clinical populations most likely to benefit (Cleland et al., 2015).

Caveats and Limitations

This study compared outcomes from children who participated in eight sessions of UVF and eight sessions of traditional treatment, counterbalanced in order across participants. Although the sequence of UVF followed by traditional treatment was found to outperform the reverse order, numerous additional questions remain unanswered. For example, it remains unknown whether 16 sessions of UVF would yield larger treatment effects than the sequence of UVF followed by traditional treatment. In this study, among individuals who responded to UVF treatment, improvement began anywhere from Session 1 to Session 6 of the eight making up the UVF phase. For those who did not respond, it is unknown whether further sessions would have resulted in improvement. In addition, visual feedback was available for 44% of trials during UVF sessions. However, recent research suggests that providing visual feedback on more trials (e.g., 89% of trials) may be even more effective (Preston et al., 2018), suggesting that more intensive exposure to visual feedback may yield even larger differences as compared to traditional treatment, at least in the early phases of treatment. Many other parameters could also be manipulated for comparison to this study, such as the number of practice trials, scheduling of practice, and criteria for advancement across levels of adaptive difficulty. Thus, future research remains necessary to determine what exactly the optimal parameters of feedback and practice might be and how they may differ across individuals.

Another limitation is that the current study focused on generalization to untrained words containing the treated target, vocalic /ɹ/. Other levels of generalization were not assessed, such as generalization to sentences or conversation. Even generalization to untreated consonantal variants of /ɹ/ was not investigated as part of this study, primarily due to manuscript-length limitations. In addition, assessment of longer-term retention of skills (e.g., 6 months follow-up) could help to identify the distal effects of the treatments provided. Thus, there is a clear need for further research to fully define treatment outcomes.

Of the 12 participants, three were poor responders to both treatment conditions. Participants 129, 141, and 143 showed raw generalization gains of less than 10 percentage points ($\hat{p}_{\text{correct}}$) after both treatment conditions. This rate of nonresponse (three of 12) is similar to previous studies on RSE (McAllister Byun & Hitchcock, 2012; McAllister Byun, Hitchcock, et al., 2014; Preston, Leece, et al., 2017). Despite the persistent appearance of nonresponders in previous research, to date no clear profile has been identified to predict which children fail to respond to treatment. There were no obvious distinctions in these three nonresponders compared to the other participants with respect to speech, language, or demographic variables, which might predict their failure to respond (see Table 1). The presence of nonresponders is a cause for concern among clinicians because, for a subset of children, neither UVF nor traditional treatment would be a viable option as delivered in the manner and dose described here. It is possible, however, that other treatment options, such as a longer duration of treatment, more intensive scheduling (Preston & Leece, 2017), perceptual training methods, or different types of biofeedback (such as electropalatography or visual acoustic feedback) might enable improvements in speech production in these children. Future research should aim to identify attributes that may predict which children are unlikely to respond to either UVF or traditional treatment and seek to identify effective alternative approaches.

An additional limitation, common in many clinician-delivered intervention studies, is that the treating clinician could not feasibly be blinded to treatment condition (i.e., UVF vs. traditional). Although treatment fidelity was high in both conditions and equipoise was emphasized with the clinician, the potential for a clinician's conscious or unconscious bias could arise in treatment delivery. In this study, only the raters were blind to condition and time point.

Summary and Conclusions

Both UVF and traditional treatment can yield improved accuracy in vocalic / / in many children with RSE, but in this study, eight sessions of UVF treatment was found to facilitate a greater degree of change than an equivalent duration of traditional treatment. Moreover, following 16 sessions, UVF appears to be more effective when provided earlier rather than later in therapy, suggesting it may be clinically beneficial to consider UVF early in the therapy process. Further research is needed to understand the optimal parameters of UVF treatment and illuminate the causes of heterogeneity of response across individual participants.

Acknowledgments

Research reported in this publication was supported by National Institute on Deafness and Other Communication Disorders Award R01DC013668 (D. Whalen, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Jose Ortiz, Daphna Harel, Alyssa Skibo, Olivia Harold, and Heena Damania for assistance and the Siemens Corporation for loaning an ultrasound for use in this study.

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. https://doi.org/10.1044/1058-0360(2007/017)

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), 1078–1089.

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.

Ballard, K. J., Smith, H. D., Paramatmuni, D., McCabe, P., Theodoros, D. G., & Murdoch, B. E. (2012). Amount of kinematic feedback affects learning of speech motor skills. *Motor Control, 16*(1), 106–119. https://doi.org/10.1123/mcj.16.1.106

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Boyce, S. E. (2015). The articulatory phonetics of /r/ for residual speech errors. *Seminars in Speech and Language, 36*(4), 257–270. https://doi.org/10.1055/s-0035-1562909

Bressmann, T., Harper, S., Zhylich, I., & Kulkarni, G. V. (2016). Perceptual, durational and tongue displacement measures following articulation therapy for rhotic sound errors. *Clinical Linguistics & Phonetics, 30*(3–5), 345–362. https://doi.org/10.3109/02699206.2016.1140227

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30,* 531–563. https://doi.org/10.1177/0145445503261167

Byiers, B. J., Reichle, J., & Symons, F. J. (2012). Single-subject experimental design for evidence-based practice. *American Journal of Speech-Language Pathology, 21*(4), 397–414.

Cleland, J., Crampin, L., Wrench, A. A., Zharkova, N., & Lloyd, S. (2017). Visualising speech: Using ultrasound visual biofeedback to diagnose and treat speech disorders in children with cleft lip and palate. *Royal College of Speech and Language Therapists Conference, 2017*.

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. https://doi.org/10.3109/02699206.2015.1016188

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by X-ray motion picture. *Linguistics, 6*(44), 29–68. https://doi.org/10.1515/ling.1968.6.44.29

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

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., & Alwan, A. (2000). Acoustic modeling of American English /r/. *The Journal of the Acoustical Society of America, 108*(1), 343–356.

Flipsen, P. (2015). Emergence and prevalence of persistent and residual speech errors. *Seminars in Speech and Language, 36*(4), 217–223. https://doi.org/10.1055/s-0035-1562905

Gibbon, F., Hardcastle, W. J., Crampin, L., Reynolds, B., Razzell, R., & Wilson, J. (2001). Visual feedback therapy using electropalatography (EPG) for articulation disorders associated with cleft palate. *Asia Pacific Journal of Speech, Language and Hearing, 6*(1), 53–58. https://doi.org/10.1179/136132801805576798

Goldman, R., & Fristoe, M. (2000). *Goldman–Fristoe Test of Articulation–Second Edition (GFTA-2)*. 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. https://doi.org/10.3200/JMBR.36.2.212-224

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. https://doi.org/10.3109/02699206.2014.956232

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., Grigos, M. I., McAllister Byun, T., & Davidson, L. (2012). The relationship between inexperienced listeners' perceptions and acoustic correlates of children's /r/ productions. *Clinical Linguistics & Phonetics, 26*(7), 628–645. https://doi.org/10.3109/02699206.2012.682695

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

Lockenvitz, S., Kuecker, K., & Ball, M. J. (2015). Evidence for the distinction between "consonantal-/r/" and "vocalic-/r/" in American English. *Clinical Linguistics & Phonetics, 29*(8–10), 613–622. https://doi.org/10.3109/02699206.2015.1047962

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. https://doi.org/10.1044/1092-4388(2011/11-0120)

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. https://doi.org/10.1044/1058-0360(2008/025)

McAllister Byun, T. (2017). Efficacy of visual–acoustic biofeedback intervention for residual rhotic errors: A single-subject randomization study. *Journal of Speech, Language, and Hearing Research, 60*(5), 1175–1193. https://doi.org/10.1044/2016_JSLHR-S-16-0038

McAllister Byun, T., & Campbell, H. (2016). Differential effects of visual–acoustic biofeedback intervention for residual speech errors. *Frontiers in Human Neuroscience, 10,* 567. https://doi.org/10.3389/fnhum.2016.00567

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. https://doi.org/10.1016/j.jcomdis.2014.11.003

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. https://doi.org/10.1016/j.jcomdis.2016.07.001

McAllister Byun, T., & Hitchcock, E. R. (2012). Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation. *American Journal of Psychology, 21*(3), 207–221. https://doi.org/10.1044/1058-0360(2012/11-0083)

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. https://doi.org/10.1044/2014_JSLHR-S-14-0034

McAllister Byun, T., Ortiz, J., & Hitchcock, E. R. (2014). Challenge Point software (Version 1.5.8). Retrieved from http://cpp.umd.edu

McGowan, R. S., Nittrouer, S., & Manning, C. J. (2004). Development of /r/ in young, Midwestern, American children. *The Journal of the Acoustical Society of America, 115*(2), 871–884.

Miccio, A. W. (2002). Clinical problem solving: Assessment of phonological disorders. *American Journal of Speech-Language Pathology, 11*(3), 221–229. https://doi.org/10.1044/1058-0360(2002/023)

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

Newell, K. M., Carlton, M. J., & Antoniou, A. (1990). The interaction of criterion and feedback information in learning a drawing task. *Journal of Motor Behavior, 22*(4), 536–552. https://doi.org/10.1080/00222895.1990.10735527

Preston, J. L., & Leece, M. C. (2017). Intensive treatment for persisting rhotic distortions: A case series. *American Journal of Speech-Language Pathology, 26*(4), 1066–1079. https://doi.org/10.1044/2017_AJSLP-16-0232

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. https://doi.org/10.1111/1460-6984.12259

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, 2017*(119), e55123. https://doi.org/10.3791/55123

Preston, J. L., McAllister, T., Phillips, E., Boyce, S., Tiede, M., Kim, J. S., & Whalen, D. H. (2018). Treatment for residual rhotic errors with high- and low-frequency ultrasound visual feedback: A single-case experimental design. *Journal of Speech, Language, and Hearing Research, 61*(8), 1875–1892.

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. https://doi.org/10.1044/2014_JSLHR-S-14-0031

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R foundation for Statistical Computing.

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. https://doi.org/10.1044/jshd.4404.504

Rvachew, S., & Brosseau-Lapré, F. (2012). *Developmental phonological disorders: Foundations of clinical practice*. San Diego, CA: Plural.

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.

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. https://doi.org/10.1097/00004424-198501000-00022

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.

Shuster, L. I. (1998). The perception of correctly and incorrectly produced /r/. *Journal of Speech, Language, and Hearing Research, 41*(4), 941–950. https://doi.org/10.1044/jslhr.4104.941

Silverman, F. H., & Paulus, P. G. (1989). Peer reactions to teenagers who substitute /w/ for /r/. *Language, Speech, and Hearing Services in Schools, 20*(2), 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. https://doi.org/10.1016/j.jcomdis.2016.10.003

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1991). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Research, 34*(2), 446. https://doi.org/10.1044/jshr.3402.446

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.

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. https://doi.org/10.1121/1.4784878

Van Riper, C., & Erickson, R. L. (1996). *Speech correction: An introduction to speech pathology and audiology*. Allyn and Bacon.

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

Wickham, H. (2009). ggplot2: Elegant graphics for data analysis. https://doi.org/http://ggplot2.org

Wickham, H. (2016). tidyr: Easily tidy data with 'spread()' and 'gather()' functions. https://doi.org/http://cran.r-project.org/package=tidy

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

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

Appendix A

Complete Results of Mixed-Effects Logistic Regression Model Examining Influences on Perceptually Rated Accuracy at Midpoint

Term	Estimate	*SE*	Test statistic	*p*
(Intercept)	–0.81	0.09	–8.94	< .001
Condition (reference level: UVF)	–1.70	0.07	–23.10	< .001
Baseline accuracy	1.99	0.50	4.00	< .001
Condition × Baseline Accuracy interaction	7.04	0.55	12.91	< .001

Appendix B

Complete Results of Mixed-Effects Logistic Regression Model Examining Influences on Perceptually Rated Accuracy at Posttreatment

Term	Estimate	*SE*	Test statistic	*p*
(Intercept)	–0.40	0.07	–5.55	< .001
Condition (reference level: UVF first)	–0.51	0.06	–9.17	< .001
Baseline accuracy	0.03	0.40	0.08	.94
Condition × Baseline accuracy interaction	5.48	0.43	12.67	< .001