JSLHR

Research Article

Efficacy of Visual–Acoustic Biofeedback Intervention for Residual Rhotic Errors: A Single-Subject Randomization Study

Tara McAllister Byun[a]

Purpose: This study documented the efficacy of visual–acoustic biofeedback intervention for residual rhotic errors, relative to a comparison condition involving traditional articulatory treatment. All participants received both treatments in a single-subject experimental design featuring alternating treatments with blocked randomization of sessions to treatment conditions.
Method: Seven child and adolescent participants received 20 half-hour sessions of individual treatment over 10 weeks. Within each week, sessions were randomly assigned to feature traditional or biofeedback intervention. Perceptual accuracy of rhotic production was assessed in a blinded, randomized fashion. Each participant's response to the combined treatment package was evaluated by using effect sizes and visual inspection. Differences in the magnitude of response to traditional versus biofeedback intervention were measured with individual randomization tests.
Results: Four of 7 participants demonstrated a clinically meaningful response to the combined treatment package. Three of 7 participants showed a statistically significant difference between treatment conditions. In all 3 cases, the magnitude of within-session gains associated with biofeedback exceeded the gains associated with traditional treatment.
Conclusions: These results suggest that the inclusion of visual–acoustic biofeedback can enhance the efficacy of intervention for some individuals with residual rhotic errors. Further research is needed to understand which participants represent better or poorer candidates for biofeedback treatment.

The North American English rhotic is a late-emerging sound, with the normal developmental period estimated to extend as late as 8 years of age (Smit, Hand, Freilinger, Bernthal, & Bird, 1990). A small subset of speakers continue to show *residual errors* affecting rhotics through late childhood, adolescence, or even adulthood (Culton, 1986). In many cases, these individuals have received but have not responded to intervention aiming to normalize their production of rhotic sounds. Although it is certainly possible to function at a high level while exhibiting residual speech errors, some speakers experience rhotic misarticulation as a barrier to full participation in social and/or academic settings (Hitchcock, Harel, & McAllister Byun, 2015). Thus, more effective forms of intervention for residual rhotic errors could be beneficial to clients and clinicians alike. Recent evidence suggests that speech errors that do not respond to traditional methods may, in some cases, be eliminated through treatment incorporating *visual biofeedback*, the subject of this study.

It is widely agreed that children's difficulty acquiring English rhotics can be attributed, at least partly, to the complexity of its articulatory configuration. Whereas most speech sounds are articulated with a single major constriction of the vocal tract, the English rhotic is known to involve two major lingual constrictions; lip rounding is also typically present, at least in syllable-initial position (Bernhardt & Stemberger, 1998). The lingual components of English rhotics are typically described as an oral constriction in which the anterior tongue raises to a point near the palate and a pharyngeal constriction formed by retraction of the tongue root (e.g., Adler-Bock, Bernhardt, Gick, & Bacsfalvi, 2007; Boyce, 2015; Klein, McAllister Byun, Davidson, & Grigos, 2013). The shape of the anterior lingual constriction is highly variable both across and within speakers. Two major categories, *retroflex* (tip up) and *bunched* (tip down), are commonly recognized, although this binary classification represents an oversimplification of a continuum of tongue shapes (Tiede, Boyce, Holland, & Choe, 2004).

[a]Department of Communicative Sciences and Disorders, New York University

Correspondence to Tara McAllister Byun: tara.byun@nyu.edu

Editor: Julie Liss
Associate Editor: Tanya Eadie
Received January 31, 2016
Revision received May 20, 2016
Accepted August 28, 2016
https://doi.org/10.1044/2016_JSLHR-S-16-0038

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

Clinical practice has traditionally recognized a distinction between categories that can be termed *consonantal /ɹ/* and *vocalic /ɚ/* (Lockenvitz, Kuecker, & Ball, 2015). The vocalic category includes the syllabic rhotic found in words such as *bird* (/bɝd/) and *butter* (/bʌɾɚ/). On the basis of acoustic and articulatory evidence (e.g., McGowan, Nittrouer, & Manning, 2004), we will represent the postvocalic rhotic in words such as *hair, near,* and *door* as the offglide of a rhotic diphthong (/hɛɚ, nɪɚ, dɔɚ/), and we will include it in the vocalic category. However, this classification remains controversial; see discussion in Lockenvitz et al. (2015). Although many dialects of English (e.g., British Received Pronunciation) pronounce vocalic variants in a nonrhotic fashion (e.g., *bird* /bɜd/, *butter* /bʌtə/), this study specifically investigates children who are acquiring a dialect of American English in which rhotic pronunciation is standard in all positions in the syllable.

Visual Biofeedback Treatment for Residual Speech Errors

In a traditional articulatory approach to intervention (e.g., Van Riper & Erickson, 1996), the clinician produces an auditory model of the target speech behavior and uses verbal and visual cues to prompt the client to adopt a more appropriate articulatory posture. Although the efficacy of this approach has been documented (e.g., Hesketh, Nightingale, & Hall, 2000; Klein, 1996), a subset of children show difficulty following clinician cues to improve the phonetic accuracy of their speech (e.g., Ruscello, 1995). It remains unknown why some children respond to traditional methods and others do not, but one possible explanation stems from the observation that some children with residual rhotic errors exhibit a decreased ability to discriminate correct versus distorted /r/ sounds in their own output (Shuster, 1998). Children with perceptual deficits may struggle to match the clinician's auditory model for a target sound in a traditional treatment approach.

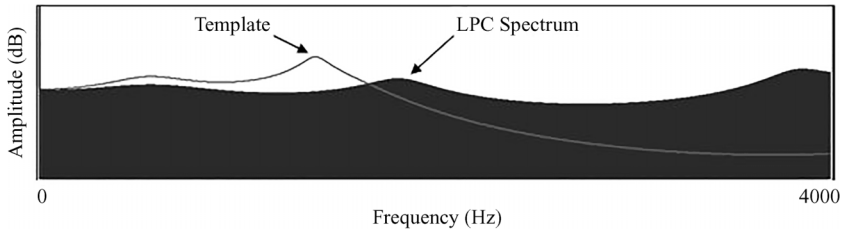
Visual biofeedback involves the use of instrumentation to capture measurements of some aspect of physiology or behavior, which can then be displayed in real time to the learner. This visual display provides a novel source of insight intended to help the learner exercise a higher level of conscious control over the process in question (Davis & Drichta, 1980; Volin, 1998). In the context of articulation, learners are provided with a model representing the target speech behavior, usually superimposed over or side-by-side with the real-time feedback display. Learners are encouraged to explore different production strategies in an effort to make their own output a closer match for the model. A range of biofeedback technologies have been used to provide real-time information about the location and movements of the articulators during speech, including electromagnetic articulography (e.g., Katz, McNeil, & Garst, 2010), ultrasound (e.g., Adler-Bock et al., 2007; McAllister Byun, Hitchcock, & Swartz, 2014; Preston, Brick, & Landi, 2013; Preston et al., 2014), and electropalatography (e.g., Gibbon, Stewart, Hardcastle, & Crampin, 1999). At the present time, all these methods are supported by evidence from case studies and/or single-subject experimental designs, which represent Phases I and II in Robey's (2004) five-phase model of clinical research. No research to date, unfortunately, has systematically compared different biofeedback methods against one another to establish their relative efficacy.

The present study focuses on a different type of enhanced feedback that will here be termed *visual–acoustic biofeedback*. In this approach, speakers view a dynamic image of the acoustic signal of their own speech, typically in the form of a linear predictive coding (LPC) spectrum (McAllister Byun & Hitchcock, 2012; McAllister Byun, Swartz, Halpin, Szeredi, & Maas, 2016) or a spectrogram (e.g., Shuster, Ruscello, & Smith, 1992; Shuster, Ruscello, & Toth, 1995). Certain sound contrasts are signaled by different locations of the resonant frequencies of the vocal tract, or *formants*, which appear as peaks in an LPC spectrum and as horizontal bands in a spectrogram. In visual–acoustic biofeedback, learners are cued to alter their speech output so that their formants line up with formant locations in a model representing correct production of the target sound. Visual–acoustic biofeedback may be particularly valuable for children with residual errors who exhibit deficits in auditory perception (e.g., Shuster, 1998) because it provides a clear, accurate visual target to replace an auditory target that may be imprecise or incorrect. Visual–acoustic biofeedback also represents a relatively low-cost alternative; the software WaveSurfer (Sjölander & Beskow, 2006), which can be used to provide real-time spectrographic feedback, is available as a free download.

The present study investigated the efficacy of visual–acoustic biofeedback by using the real-time LPC function of the Sona-Match module within the Computerized Speech Lab KayPENTAX, Lincoln Park, NJ). The visual display used to provide biofeedback is illustrated in Figure 1. The solid-colored wavelike shape represents a snapshot of the real-time LPC spectrum of the participant's speech; the peaks of the spectrum correspond to the formants of the participant's speech. The line labeled *Template* represents a trace of the LPC spectrum of a sustained rhotic produced by a speaker with typical articulation. The acoustic hallmark of rhoticity is a significant lowering of the third formant (F3); normative data from adults indicate that the mean height of F3 in /ɝ/ falls between 60% and 80% of its mean height in nonrhotic vowels (Hagiwara, 1995). Meanwhile, the second formant (F2) is relatively high in rhotics, resulting in a small distance between the two formants (Boyce & Espy-Wilson, 1997; Flipsen, Shriberg, Weismer, Karlsson, & McSweeny, 2001; Shriberg, Flipsen, Karlsson, & McSweeny, 2001). This can be seen in the template in Figure 1, where F2 and F3 are so close together that the LPC spectrum represents them with a single peak.

Formant frequencies are influenced not only by the sound being produced but also by speaker-level properties, such as height and sex; in the specific context of rhotics, this has been documented in normative studies by Lee, Potamianos, and Narayanan (1999) and Flipsen et al (2001). Clinicians who adopt visual–acoustic biofeedback are

Figure 1. Linear predictive coding (LPC) spectrum with a template representing vocalic /ɝ/ produced by a typical adult female. From McAllister Byun & Hitchcock (2012). Used with permission.



encouraged to assemble a library of templates from typical speakers representing a range of ages and physical sizes. Participants can then be paired with a template judged to represent the best match for their vocal tract size.[1] In the present study, participants were prompted to sustain /i/, /ɑ/, and /u/ vowels, while the clinician compared their formant patterns to templates of those vowels from our library of different speakers. The target used in treatment for a given participant was the rhotic template from the reference speaker whose formants were judged to most closely match those of the participant in nonrhotic vowel contexts.

Previous research has suggested that visual–acoustic biofeedback can establish perceptually accurate production of rhotics in children who have not responded to traditional forms of intervention. Two case studies (Shuster et al., 1992, 1995) described the successful application of spectrographic biofeedback treatment to remediate residual rhotic errors in three adolescents. McAllister Byun and Hitchcock (2012) conducted a quasi-experimental study in which 11 children received intervention for rhotic misarticulation in two phases: traditional intervention using verbal cues for articulator placement, followed by biofeedback intervention using a real-time LPC spectrum. Word-level generalization probes elicited after traditional treatment showed no change in accuracy relative to baseline, but probes elicited after biofeedback treatment showed a significant increase in acoustic and perceptual accuracy. McAllister Byun, Swartz, et al. (2016) evaluated the effects of 8 weeks of spectral biofeedback intervention in a single-subject experimental study of nine children with rhotic misarticulation who had previously received at least 5 months of traditional intervention without success. Six out of nine participants showed sustained improvement on at least one treated target, and a logistic mixed model indicated that rhotic-containing words produced at the end of the study were significantly more likely to be rated as perceptually correct by blinded listeners than the same words produced at baseline.

[1]Although the acoustics of rhotics also vary slightly across different phonetic contexts or positions in the syllable (e.g., Flipsen et al., 2001), previous studies of visual–acoustic biofeedback treatment have found that a single template can be used to cue rhotics across multiple contexts, suggesting that these differences are too minor to require separate targets (McAllister Byun, Swartz, et al., 2016; McAllister Byun & Hitchcock, 2012).

A limitation of the previous literature is the lack of a true experimental study comparing visual–acoustic biofeedback versus traditional articulatory intervention. The quasi-experimental design adopted by McAllister Byun and Hitchcock (2012) does not rule out the possibility that gains observed during the biofeedback treatment period could be partially driven by a late-emerging response to the initial phase of traditional intervention. In McAllister Byun, Swartz, et al. (2016), biofeedback was the only experimental intervention, with no comparison condition of traditional treatment. This study was designed to fill the need for a well-controlled comparison.

Single-Subject Randomization Designs

Because speech intervention can be labor intensive, and affected populations are often small, research in speech-language pathology frequently makes use of single-subject experimental designs (Byiers, Reichle, & Symons, 2012). However, certain properties of traditional single-subject research are not optimal for measuring the efficacy of speech interventions. Because these treatments specifically aim to produce lasting learned effects, they are not well suited to a design that requires gains to be reversed when treatment is discontinued, such as the ABAB or withdrawal design (Kratochwill & Levin, 2014). An alternative is the multiple-baseline across-behaviors design, in which progress is tracked across multiple targets as they are transitioned from a baseline to a treatment phase in a staggered fashion. However, this design is not suitable when there is only one target sound or behavior, or if there are multiple targets that are very similar to one another, because gains on a treated target may carry over to an untreated target and thus compromise experimental control (Rvachew & Brosseau-Lapré, 2012). In the context of speech intervention research, it would be ideal to use a single-subject experimental design that accommodates long-term trajectories of change over time and also allows for the possibility of generalization.

Rvachew (1988) suggested that the single-subject randomization design could fill this need. This design (Edgington, 1987) acknowledges that participant performance over the course of a study can be influenced by variables other than the immediate treatment context, including carryover from previous treatment. Instead of attempting to eliminate such factors, the randomization design aims to

make the treatment manipulation independent of these influences; this is accomplished by randomly assigning different sessions to different experimental conditions within each subject. Thus, in an eight-session study comparing treatments A versus B, one subject might be randomly assigned to the sequence ABBABAAB, while another could receive the random sequence BABBAABA. If there is a long-term trend reflecting learning over time, scores in both types of session will tend to go up over time. However, if one condition is playing a particularly important role in driving this long-term effect, the increments of progress observed in connection with sessions in that condition should be greater than the increments associated with the less effective condition. Probing performance both at the start and the end of a treatment session is a useful practice for this study design because it allows the experimenter to be precise about the timing of gains relative to the application of treatment (i.e., short-term gains can be measured from the start to the end of each treatment session and longer-term gains from the end of one treatment session to the start of the next).[2]

The single-subject randomization design offers an additional analytical benefit. Conventional single-subject designs do not meet standard assumptions for inferential statistical testing, and the outcomes of such studies are commonly evaluated by using visual inspection or other qualitative comparisons (Kratochwill, Levin, Horner, & Swoboda, 2014). However, the incorporation of randomization legitimizes the use of inferential statistics to test the significance of the difference between control and experimental conditions (Ferron & Levin, 2014), which could, in turn, enhance the interpretability and credibility of single-subject experimental research in the eyes of the broader research community (Kratochwill & Levin, 2014).

Research Questions

The present study used a single-subject randomization design to compare traditional and biofeedback approaches to intervention for residual errors affecting North American English rhotics. By using a combination of quantitative and qualitative analyses, this article addresses the following research questions:

1. Does a combined course of traditional and biofeedback intervention produce significant gains in production accuracy over 20 sessions? In particular, we will test whether gains for the treated rhotic variant generalize to untreated words elicited without feedback.
2. Does the magnitude of short-term progress on treated rhotic variants (change from pretest to posttest within a session) differ significantly between traditional and biofeedback treatment conditions?

[2]Even with these controls in place, we cannot rule out the possibility that learning in one condition could carry over to influence performance in another condition. This could pose a particular interpretive challenge if one approach is more facilitative of short-term learning, while the other favors long-term generalization gains.

Method

Participants

The study enrolled seven native speakers of American English ranging in age from 9;0 (years;months) to 15;0 years, with a mean age of 12;3 years (SD = 28.5 months). Five participants were male, and two were female, a ratio consistent with previous descriptions of the gender distribution of residual speech errors (Shriberg, 2010). Participants were recruited through flyers distributed to schools and community centers in New York City. Prior to the start of the study, participants completed an initial evaluation of speech and language function, and their parents completed a developmental history questionnaire; eligibility for inclusion in the study was determined on the basis of this combined information. Due to the within-subject nature of the comparisons reported here, a more diverse pool of candidates was considered for inclusion than would be typical for a randomized controlled trial or other study emphasizing between-subjects comparisons. Participant history information is summarized in Table 1; all names are pseudonyms.

The developmental history questionnaire collected information about language(s) and dialects spoken in the home, including a specific question asking whether r-deletion was part of the child's home dialect. In light of the linguistically diverse character of the urban population from which this sample was drawn, participants were not required to be monolingual speakers of English, but they were required to demonstrate native-level English proficiency (per parent and/or teacher report). One participant, Lucas, was an early sequential bilingual child who heard Spanish in the home and English at school. There were no cases in which the child's production of rhotics could reasonably be inferred to reflect a feature of the parent's language or dialect.

Parents were also asked to report the duration of any previous treatment targeting rhotic sounds. The estimated duration of previous treatment was heterogeneous, ranging from 0 months to 11 years (median duration 2.5 years). Participants were required to discontinue outside intervention targeting rhotics for the duration of their participation in this study. Parents were also asked to report what other speech sounds, if any, had been targeted in the child's previous treatment. It has been argued that research and clinical practice should distinguish between residual speech errors, which are found in children who presented with broader phonological delay in early childhood and gradually improved until only /r/ or /s/ distortions remained, versus persistent errors, in which distortions of late-developing phonemes represent the only errors in both early and late developmental stages (e.g., Flipsen, 2015). The data in Table 1 suggest that three participants (Evan, Ian, and Lucas) would likely be classified as presenting with residual speech errors, while three others might be better characterized as exhibiting persistent errors; no response was provided for one participant. Due to the incomplete and uncertain nature of this parent-reported data, we will not undertake more detailed investigation of the persistent–residual distinction in the context of the present study. We also

Table 1. Participant background data, per parent report.

Pseudonym	Gender	Age at enrollment	Duration of previous treatment targeting rhotics	Previous or current speech targets other than rhotics
Clara	F	10;6	2.5 years (school and private)	/s, l/
Evan[a,b]	M	9;0	2 years (school and private)	Multiple, especially /l/
Felix[a,b]	M	15;0	No response	No response
Garrett[a,b]	M	13;7	2.5 years (school)	/θ/
Ian[a,b]	M	14;11	11 years (private)	Multiple (specifics unknown)
Lucas	M	12;2	4 years (school)	Multiple (specifics unknown)
Piper	F	10;4	None	None

[a]Indicates parent report of comorbid attention-deficit/hyperactivity disorder. [b]Indicates parent report of comorbid language–based learning disability.

default to residual as the more conservative classification of children's errors (i.e., a previous history of errors cannot be ruled out on the basis of the evidence available).

Parents were asked to report any history of developmental or neurobehavioral disorder. Comorbid diagnoses of attention-deficit/hyperactivity disorder (ADHD) and/or language-based learning disability were permitted if they were not judged to actively interfere with the participant's ability to comprehend and follow instructions in the intervention setting. Four participants (Ian, Garrett, Evan, Felix) had a diagnosis of ADHD; these same four participants also had diagnoses of dyslexia or other language-based learning disability.

All participants passed a pure-tone hearing screening at 20 dB HL at 500, 1000, 2000, and 4000 Hz. They also showed no gross abnormalities on a screening examination of oral structure and function, adapted from Shipley and McAfee's (2008) checklist. Participants' receptive understanding of language was assessed with the Auditory Comprehension subtest of the Test of Auditory Processing Skills–Third Edition (Martin & Brownell, 2005). Participants were required to produce no more than three sounds other than rhotics in error on the Goldman-Fristoe Test of Articulation–Second Edition (Goldman & Fristoe, 2000).[3] Some participants did exhibit errors on late-developing nonrhotic sounds, such as /s/ and /l/. Test of Auditory Processing Skills–Third Edition scores and qualitative results of the Goldman-Fristoe Test of Articulation–Second Edition are reported in Table 2.

To select a sample of individuals who were relatively homogeneous with respect to baseline severity, participants were required to score below a fixed threshold of pretreatment accuracy in rhotic production. In particular, they were required to produce fully correct rhotics in no more than 30% of 70 word-level probe items. For this purpose, productions were rated in a binary fashion by consensus between two trained listeners. The trained listeners followed a strict criterion, such that distorted or intermediate productions were scored as incorrect. Individual scores are reported in Table 2. In addition to the overall number and percentage of items judged to be produced correctly, Table 2 reports each participant's accuracy when rhotics are subdivided into the categories consonantal and vocalic.

Readers may note some discrepancy between the baseline accuracy levels reported in Table 2 and blinded naïve listeners' ratings of baseline accuracy plotted in Figures 2–4, with the naïve listeners typically assigning the correct rating to more tokens than the clinician listeners. This is in keeping with previous research suggesting that naïve listeners tend to be more lenient than experts in their ratings of children's rhotic sounds (McAllister Byun, Halpin, & Szeredi, 2015). The fact that these listeners are slightly more lenient overall than experts does not affect the validity of our within-subject measures of treatment efficacy, which were uniformly rated by blinded naïve listeners (see details under Measurement section).

Design

In this single-subject randomization design, each participant received ten 30-min sessions of traditional intervention and ten 30-min sessions of biofeedback intervention. A randomized block schedule was adopted in which all participants received one session of each type in each week, but the order of treatment type within a week was determined randomly (Rvachew, 1988). The treatment period was immediately preceded by three baseline data collection sessions and immediately followed by three maintenance sessions; in both cases, the three sessions were collected over 1.5 to 2 weeks. A 50-word probe featuring rhotics in various phonetic contexts was administered in all baseline and maintenance sessions, and a 25-item subset probe was administered at the beginning and end of each treatment session. The same list of words was used for all probes of a given type; see complete lists in Appendix A. Probe measures featured a balance of words containing a rhotic in onset position and words with a rhotic in the rhyme of the syllable. The words elicited in probe measures were not targeted in treatment.

This study was structured to meet What Works Clearinghouse (WWC) standards for single-subject experimental design (Kratochwill et al., 2013). The alternating treatments design in which two treatments are compared

[3]If both members of a homorganic pair differing only in voicing (e.g., /s, z/) were affected in the same way, this was counted as a single error.

Table 2. Participant evaluation results.

Pseudonym	TAPS-3 standard score (percentile)	Nonrhotic error sounds (GFTA-2)	Rhotic word probe: Number of consonantal /r/ fully correct, out of 29 (%)	Rhotic word probe: Number of vocalic /r/ fully correct, out of 41 (%)	Rhotic word probe: Total number fully correct, out of 70 (%)
Clara	13 (84)	None	0 (0)	19 (46.3)	19 (27.1)
Evan	11 (63)	/s/ (minor distortion)	16 (55.2)	1 (2.4)	17 (24.3)
Felix	13 (84)	/ʃ/, /θ/	16 (55.2)	1 (2.4)	17 (24.3)
Garrett	11 (63)	/s/, clusters with /w/	7 (24.1)	0 (0)	7 (10)
Ian	11 (63)	None	0 (0)	0 (0)	0 (0)
Lucas	10 (50)	/ð,θ/, /l/, /s,z/	7 (24.1)	8 (19.5)	15 (21.4)
Piper	14 (91)	None	12 (41.4)	9 (22)	21 (30)

Note. TAPS-3 = Test of Auditory Processing Skills–Third Edition; GFTA-2 = Goldman-Fristoe Test of Articulation–Second Edition.

Figure 2. Longitudinal plots of $\hat{p}_{correct}$ for participants with large positive effect sizes. BL = baseline; Tx = treatment; MN = maintenance; and dashed line = mean across BL sessions.

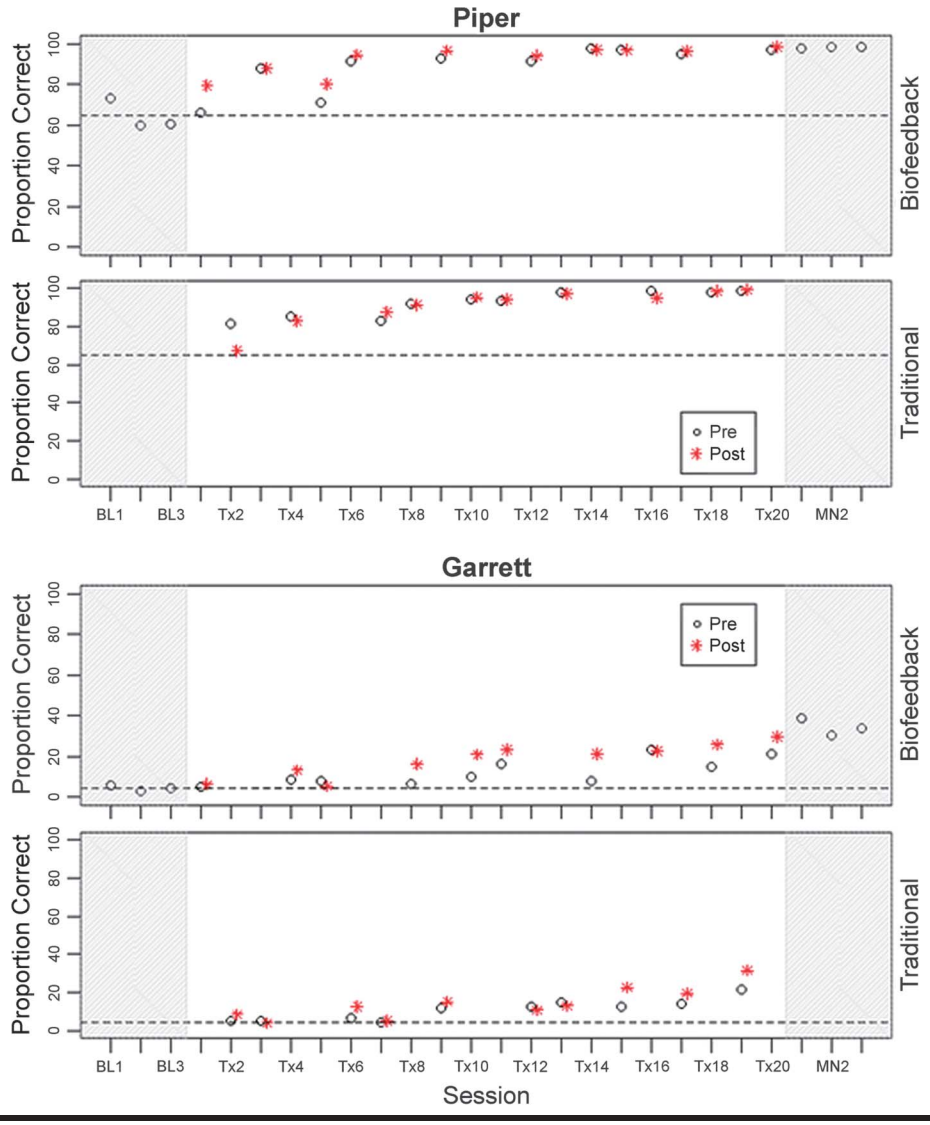
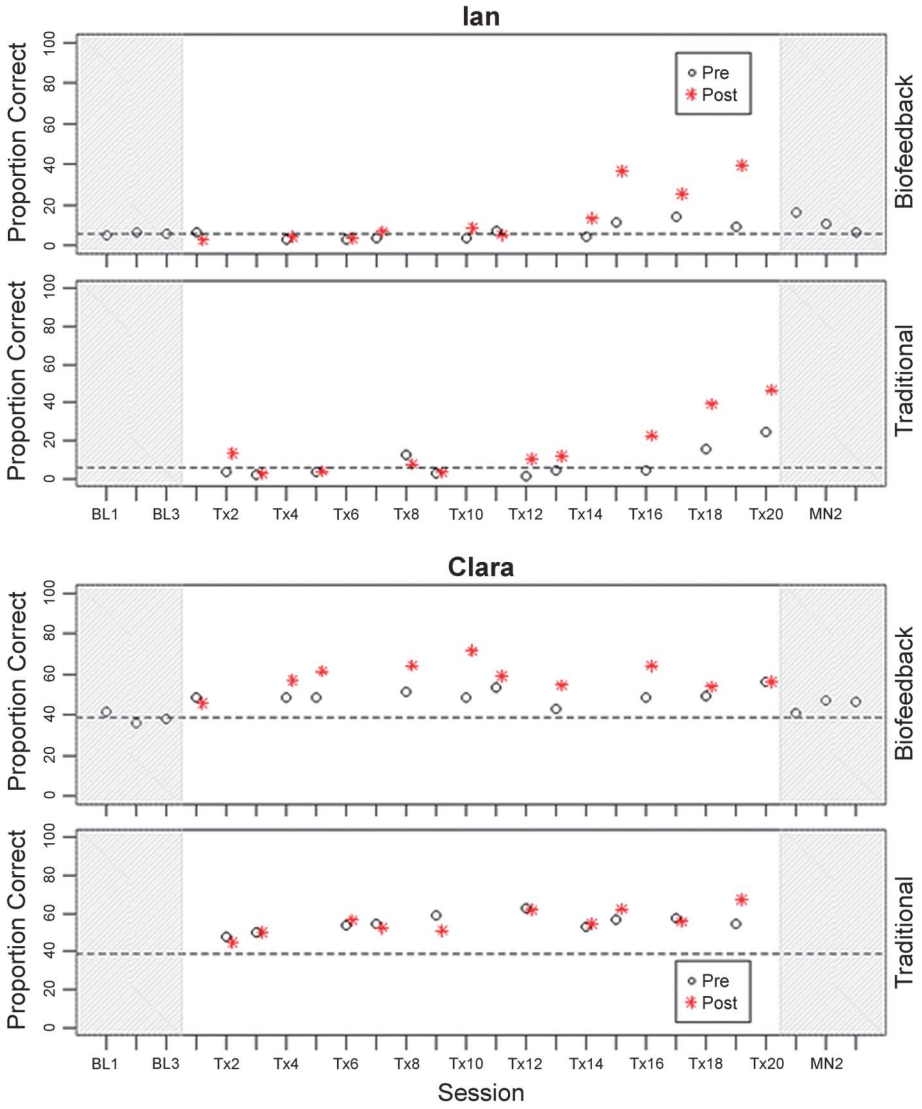


Figure 3. Longitudinal plots of $\hat{p}_{correct}$ for participants with small positive effect sizes. BL = baseline; Tx = treatment; MN = maintenance; and dashed line = mean across BL sessions.

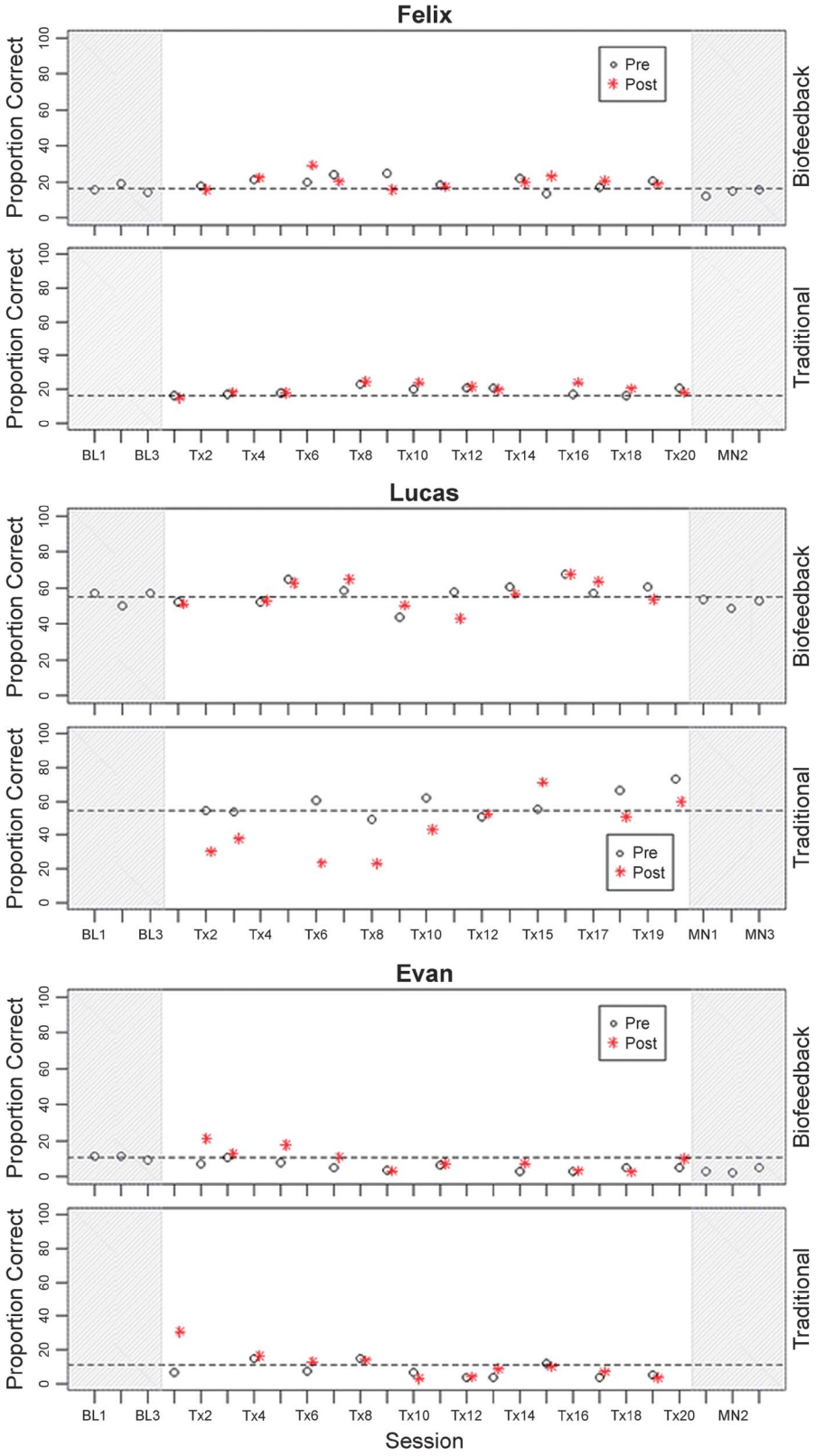


against one another meets the WWC requirement for a minimum of three opportunities to demonstrate an experimental effect at different points in time (Kratochwill et al., 2013). In addition, the inclusion of a minimum of three points of data collection in each phase of the study meets WWC standards *with reservations*; a minimum of five observations per phase would be needed to satisfy the standards in full.

Treatment

All treatments were administered by a single certified speech-language pathologist. The first two sessions of each type of treatment incorporated extended instructions. In the extended instructions for traditional treatment, line drawings and verbal descriptions were used to familiarize participants with three major components associated with accurate tongue placement for rhotics: retraction of the tongue root (Adler-Bock et al., 2007; Klein et al., 2013), elevation of the posterior lateral margins of the tongue (Bacsfalvi, 2010), and elevation of the anterior tongue. Explicit instructions to adopt a particular shape of the anterior tongue (e.g., retroflex versus bunched) were not provided (Klein et al., 2013; McAllister Byun, Hitchcock, & Swartz, 2014). In the extended instructions for biofeedback treatment, participants were familiarized with the real-time acoustic display and with the appearance of LPC spectra representing correct and incorrect rhotic productions. No articulatory instructions were provided during biofeedback sessions.

Figure 4. Longitudinal plots of $\hat{p}_{correct}$ for participants with null or negative effect size. BL = baseline; Tx = treatment; MN = maintenance; and dashed line = mean across BL sessions.



Subsequent sessions began with a 5-min period of free play, a relatively unstructured client–clinician interaction. In traditional sessions, this involved producing rhotics in various phonetic contexts while the clinician provided articulator placement cues. In biofeedback sessions, this involved attempting rhotics in different phonetic contexts while viewing the visual biofeedback display. Following the free play period, participants were cued to produce 60 trials of words containing rhotics in blocks of five trials. One verbal cue was provided prior to each block. The clinician was instructed to base her verbal input on separate lists of cues appropriate for traditional and biofeedback treatment conditions. A sample cue in the traditional condition is "Try to make the back part of your tongue go back, like for /ɑ/"; an example of a biofeedback cue is "Focus on making the third bump move over." A more detailed description of the nature of cueing for both traditional and biofeedback conditions can be found in McAllister Byun, Swartz, et al. (2016). In both conditions, the clinician provided verbal summary feedback after each block of trials; this took the form of a statement indicating which trial she had judged to be most accurate in the most recent block.

In light of evidence that vocalic /ɚ/ tends to be mastered earlier than consonantal /ɹ/ (Klein et al., 2013; MagLoughlin, 2016), only vocalic variants were used as targets at the outset of treatment. Each session featured a random selection of stimulus words from a master list, with stratification to ensure balanced representation across the categories /ɑɚ/, /ɔɚ/, /ɛɚ/, /ɪɚ/, and /ɝ, ɚ/. Provisions were made for participants to advance to consonantal /ɹ/ targets (specifically, it was stipulated that participants would need to sustain over 80% accuracy within treatment and also exhibit greater than 80% correct across all vocalic targets in two consecutive pretest probe measures), but no participant in the present study was judged to meet the criterion for advancement.

Adaptive Difficulty in Treatment

Previous studies of biofeedback intervention for speech have reported that although it is relatively straightforward to induce behavioral changes within the treatment setting, it is more difficult to ensure generalization of these gains to a setting in which biofeedback is not available (Gibbon & Paterson, 2006; McAllister Byun & Hitchcock, 2012). Hitchcock and McAllister Byun (2014) suggested that generalization of gains in biofeedback treatment for residual errors might be enhanced through the incorporation of principles in the *challenge point framework* (Rvachew & Brosseau-Lapré, 2012). Derived from motor learning research by Guadagnoli and Lee (2004), the challenge point is described as "the intersection of functional task difficulty and practice performance when the learner is receiving the optimum amount of information to promote learning" (Rvachew & Brosseau-Lapré, 2012, p. 773).

Practice difficulty was adjusted adaptively by using a structured implementation of challenge point principles through the custom software Challenge-R (McAllister Byun, Hitchcock, & Ortiz, 2014).[4] Challenge-R is a personal computer–based program that presents targets for production and records accuracy scores assigned by the treating clinician. After every 10 trials, the program tallies the participant's accuracy and adjusts one of three parameters on a rotating schedule. If a participant scores 80% or better in a block of 10, the program increases the difficulty of one parameter; a score below 50% triggers a decrease in difficulty. At the end of each session, the participant's current level in the challenge point hierarchy is saved so that the next session of the same type (biofeedback or traditional) can resume with the saved settings.

In the present study, the parameters adjusted within the session were the frequency of feedback (100%–50%–20%),[5] mode of elicitation (imitation, reading, or imitation with prosodic manipulations), and word shape complexity (one or two syllables, with or without competing /l/ or /w/ phonemes). In addition, the order of stimulus presentation was adjusted as an across-session parameter. If a participant's cumulative accuracy across a session exceeded 80%, the next session would feature an increase to the next level of complexity for order of presentation; cumulative accuracy below 50% could then trigger a downward adjustment in complexity. The lowest level of the hierarchy featured a fully blocked order of presentation in which one word was selected to represent each of the phonetic contexts /ɑɚ/, /ɔɚ/, /ɛɚ/, /ɪɚ/, and /ɝ, ɚ/; each word was elicited in two consecutive blocks of five trials. The next level of complexity was random-blocked: Each block elicited five trials of a single word, but the order of words was randomized across blocks. The highest level of complexity featured a fully random order, where a single block of five trials could contain different words representing different phonetic contexts.

Treatment Fidelity

To help the treating clinician maintain fidelity to the treatment protocol during sessions, both the clinician and an assistant made reference to a fidelity checklist during all sessions. In addition, 10% of all sessions were reviewed as a post hoc check of fidelity to the stated

[4]For the first three participants enrolled, the same adaptive schedule described for the Challenge-R program was implemented by using a customized script in E-Prime experiment presentation software (Psychology Software Tools, Inc., 2012). Participants' treatment experience was the same independent of the software used.

[5]In biofeedback sessions, adjustments in feedback frequency involved making biofeedback available in all trials, in 50% of trials, or in only the first trial in a block. In traditional treatment sessions, adjustments instead affected the frequency with which the clinician provided verbal summary feedback: after all blocks, after 50% of blocks, or after the first two blocks only. In blocks in which the clinician did not provide summary feedback, she instead prompted the participant to judge which trial in a block of five had been the most accurate.

protocol (Kaderavek & Justice, 2010). Sessions to be checked were pseudorandomly selected to feature a balance of participants, treatment types, and points in the course of treatment. To assess fidelity, the audio record of a treatment session was reviewed by research assistants not involved in treatment delivery. These raters completed a checklist to verify the following aspects of the study design: (a) each block of five trials was preceded by a reminder cue; (b) this cue was consistent with the stated treatment condition (biofeedback or traditional); (c) each block consisted of precisely five trials; (d) feedback or other interruptions did not occur within a block; and (e) summary feedback was provided after each block in which feedback was indicated.

Measurement

Ratings for speech samples collected in this study were obtained with the Amazon Mechanical Turk (AMT) crowdsourcing platform. In fields such as behavioral psychology (e.g., Paolacci, Chandler, & Ipeirotis, 2010) and linguistics (e.g., Sprouse, 2011), online crowdsourcing platforms have come to represent a valued resource for researchers to recruit large numbers of individuals to complete experimental tasks. Although online data collection is inherently noisier than lab-based data collection, computational models suggest that this noise can be overcome by aggregating responses across a large number of individuals (Ipeirotis, Provost, Sheng, & Wang, 2014). Numerous published studies have empirically validated results obtained through AMT against results collected in a typical laboratory setting (e.g., Crump, McDonnell, & Gureckis, 2013; Paolacci et al., 2010). The validity of crowdsourced data collection in the specific context of rating children's rhotic sounds was recently assessed by McAllister Byun, Halpin, & Szeredi (2015). Ratings aggregated across 250 naïve listeners on AMT were highly correlated with both expert listeners' ratings ($r = .92$) and with F3–F2 distance, an acoustic measure of rhoticity ($r = -.79$).[6] In addition, bootstrap analyses indicated that when responses were aggregated across subsamples of at least nine AMT listeners, performance relative to a gold standard measure was equivalent to that of samples of three trained listeners, which was treated as the industry standard. McAllister Byun, Swartz, et al. (2016), drawing on these findings, used responses aggregated across samples of nine AMT listeners to measure treatment outcomes in a single-subject study of rhotic misarticulation. Across the full set of data in that study, interrater agreement was calculated to be 80.7%.

The present study followed the protocol established in McAllister Byun, Halpin, and Szeredi (2015) and repeated in McAllister Byun, Swartz, et al. (2016), collecting binary ratings of each speech token from at least nine unique listeners recruited through AMT. Additional details on this protocol are provided in Appendix B. This use of AMT to obtain speech ratings was approved by the institutional review board at New York University. Participants and their parents gave consent for sound files to be shared with external listeners in an anonymized fashion for rating purposes.

Analyses

Recall that all participants began with treatment targeting vocalic rhotics, and none met the within-treatment accuracy criterion to advance to consonantal targets. Because limited generalization to the untreated variant was observed, the analyses reported below will focus on words representing the treated category, vocalic rhotics. Responses were aggregated across raters by using $\hat{p}_{\text{correct}}$, which represents the proportion of listeners who rated a token as a *correct r sound* in a binary forced-choice task. McAllister Byun, Harel, Halpin, and Szeredi (2016) examined the properties of $\hat{p}_{\text{correct}}$ calculated from crowdsourced listeners' rating of children's rhotic sounds. That study demonstrated that $\hat{p}_{\text{correct}}$, calculated over samples with $n = 9$ naïve listeners recruited online, correlated strongly with F3–F2 distance ($r = -0.76$, $p < .001$). In the plots that follow, $\hat{p}_{\text{correct}}$ has been averaged across all items in a session and multiplied by 100 to yield a percentage, which can be thought of as the percentage of times the *correct r* label was assigned out of the total number of ratings issued across all items in a given session. We will continue to use the label $\hat{p}_{\text{correct}}$ to remind the reader that these values are distinct from (although correlated with) conventional measures of percentage correct.

Effect sizes were calculated for each individual by comparing $\hat{p}_{\text{correct}}$ for single-word probes administered during the baseline phase versus the posttreatment maintenance phase. Effect sizes were standardized with Busk and Serlin's d_2 statistic (Beeson & Robey, 2006), which pools standard deviations across baseline and maintenance periods to minimize the occurrence of cases in which effect size cannot be calculated due to zero variance at baseline. This study follows Maas and Farinella (2012) in treating 1.0 as the minimum standardized effect size considered clinically relevant (i.e., the difference between pre- and posttreatment means exceeds the pooled standard deviation). However, d_2 has a limitation in that low variance at baseline can inflate the magnitude of effect sizes (Howard, Best, & Nickels, 2015). Therefore, any effect sizes with a magnitude between 1.0 and 2.5 will be visually inspected to confirm that the observed pattern is consistent with a meaningful effect of treatment.

Individual randomization tests evaluating the relative impact of traditional versus biofeedback treatment sessions were carried out with the R Package Single-Case Randomization Tests (Bulté & Onghena, 2008). Response to treatment within each session was quantified as the difference in $\hat{p}_{\text{correct}}$ between the posttreatment probe measure

[6]The correlation between perceptual ratings and F3–F2 distance is negative because accurate rhotic productions, which receive high perceptual ratings, are characterized by a small F3–F2 distance.

and the pretreatment probe measure.[7] The test statistic of interest was the difference between the mean within-session change across all biofeedback treatment sessions and the corresponding mean across traditional treatment sessions. The randomization algorithm compares this test statistic against the full set of possible differences between session types when the labels *traditional* and *biofeedback* are randomly assigned in all permutations permitted by the relevant blocking specifications. If there is a true difference between treatment conditions, the difference in means should be greater when the labels traditional and biofeedback are correctly aligned with sessions instead of being randomly assigned (Ferron & Levin, 2014). The p value is computed as the proportion of empirically derived test statistics that are equal to or greater than the observed test statistic (Rvachew, 1988).

Results

Treatment Fidelity

As indicated previously, recordings of 10% of all treatment sessions were reviewed to verify fidelity to the stated treatment protocol. The most frequent deviation from protocol was the absence of feedback following a block of trials when feedback should have occurred; this was reported in 11% of blocks reviewed. Verbal cues were generally consistent with the treatment condition designated for a session: Cues that were judged to be articulatory in nature were reported in only four of 72 biofeedback blocks observed. Within-block interruptions, often to redirect the child to the task, were an additional source of deviation from protocol. Such interruptions were observed in 16.5% of blocks in total, but they were not evenly distributed across participants; most were observed in connection with one of the younger participants with comorbid ADHD.

Individual Results: Effect Sizes

Effect sizes representing pre- to posttreatment change in $\hat{p}_{\text{correct}}$ for the treated category of vocalic rhotics are reported for all participants in Table 3. The first column shows participants' mean $\hat{p}_{\text{correct}}$ in the baseline period, averaged across all vocalic targets from all three sessions. The second column shows the equivalent mean across the three maintenance sessions. The third column shows the difference between these two values (i.e., the unstandardized effect size). This value was divided by the standard deviation pooled across baseline and maintenance sessions to arrive at the standardized effect size, d_2. Recall that because effect sizes are calculated before and after the full treatment period, they provide information about each participant's response to the complete treatment package, not their differential response to traditional versus biofeedback treatment. The effect sizes in Table 3 show a wide range of variability in overall response to treatment across individuals. Combining effect sizes across all participants yields a mean of 1.78 (median 1.5), suggesting that on average, participants' response to the combined biofeedback and traditional treatment package was positive and exceeded the minimum value considered clinically significant (Maas & Farinella, 2012). Individual patterns of response are examined in detail in the next section.

Individual Results: Visual Inspection

Figures 2–4, which represent each participant's pattern of change in accuracy ($\hat{p}_{\text{correct}}$) on treated items over time, can be visually inspected to corroborate the effect sizes reported in Table 3. Discussion of differences in relative response to biofeedback versus traditional treatment will be deferred until the next section. In the single-subject plots in Figures 2–4, each child is represented by two boxes. The top box reflects performance on probe measures administered before and after each biofeedback treatment session and additionally reports probe scores from the pretreatment baseline and posttreatment maintenance periods. Baseline and maintenance intervals are shaded gray. The lower box reflects performance during traditional treatment sessions. For each treatment session, a black circle represents performance on the pretreatment probe measure, and a red star represents performance on the posttreatment probe. The y-axis represents $\hat{p}_{\text{correct}}$ aggregated across all vocalic rhotic targets within a session.[8] Thus, the distance between the circle and the star within a session provides an index of the participant's progress during that treatment session. A dashed horizontal line tracks the participant's mean $\hat{p}_{\text{correct}}$ from the baseline interval, so subsequent scores can be compared with the baseline mean.

For convenience, the single-subject graphs have been grouped into three sets of two to three participants with similar effect sizes. Figure 2 depicts two participants, Piper and Garrett, who showed a robust response to treatment (effect sizes of 6.18 and 9.87, respectively). Piper began with relatively high accuracy (on the basis of naïve listeners' ratings) and was rapidly judged to attain ceiling-level performance, which she sustained throughout all treatment and maintenance sessions. Because Piper had not previously received treatment, there is a possibility that she could have been in a process of spontaneous resolution and would

[7]It is also possible to examine longer-term learning by using randomization tests to examine differences in the magnitude of change from the end of one session to the start of the next (e.g., the difference between Session 1 posttreatment probe score and Session 2 pretreatment probe score). In the present case, this measure was not optimal because the duration of time elapsed between the end of one session and the start of the next was not uniform either within or across children.

[8]The mean number of probe words on which $\hat{p}_{\text{correct}}$ scores are based was 14.59 (SD = 0.61) for pre- and posttreatment probes and 27.5 (SD = 4.0) for baseline and maintenance probes. The number of ratings collected in connection with a given probe session (i.e., the denominator in $\hat{p}_{\text{correct}}$) was roughly 9 times the number of items in that probe and often was larger.

Table 3. Individual effect sizes.

Pseudonym	Baseline mean $\hat{p}_{correct}$	Maintenance mean $\hat{p}_{correct}$	Unstandardized effect size	Pooled *SD*	d_2
Clara	38.57	45.09	6.52	3.06	2.13
Evan	10.8	3.35	–7.45	1.41	–5.3
Felix	16.26	13.96	–2.3	2.31	–0.99
Garrett	4.22	34.5	30.28	3.07	9.87
Ian	5.74	11.18	5.44	3.64	1.5
Lucas	54.92	51.55	–3.37	3.49	–0.96
Piper	64.74	98.1	33.36	5.4	6.18

have eliminated her rhotic misarticulation without treatment. However, she was 10 years old, past the age at which spontaneous resolution is normally expected (Gibbon & Paterson, 2006). Furthermore, the flat trajectory of scores during the pretreatment baseline phase suggests that her progress represents a response to treatment and not spontaneous improvement.

Garrett began with much lower accuracy and showed no meaningful progress on word probes for an extended period of time, but in the later part of the study, he made slow but continuous gains. His accuracy in the maintenance period was substantially higher than its level at baseline, although his scores remained well below ceiling-level accuracy.

Figure 3 shows two participants, Ian and Clara, whose effect sizes were smaller but still exceeded the minimum to be considered clinically relevant (1.5 and 2.13, respectively). Visual inspection reveals a pattern of change over time that is also consistent with a meaningful response to treatment. Ian, who had the longest duration of previous unsuccessful treatment out of all participants, exhibited virtually no change in accuracy in the first 15 sessions. In the final five sessions, though, his accuracy in pretreatment probes reliably exceeded baseline levels, with substantial additional improvement in each posttreatment probe. These gains, unfortunately, were tenuous, and scores drifted back down to baseline levels over the duration of the maintenance period. Clara reliably exceeded baseline-level performance in both pre- and posttreatment probes throughout the course of treatment, but like Ian, she showed weaker gains in the posttreatment maintenance interval, which may reflect that the full probe administered in baseline and maintenance sessions contained additional items that were less familiar than the fixed 25-word subset administered at the start and end of each treatment session.

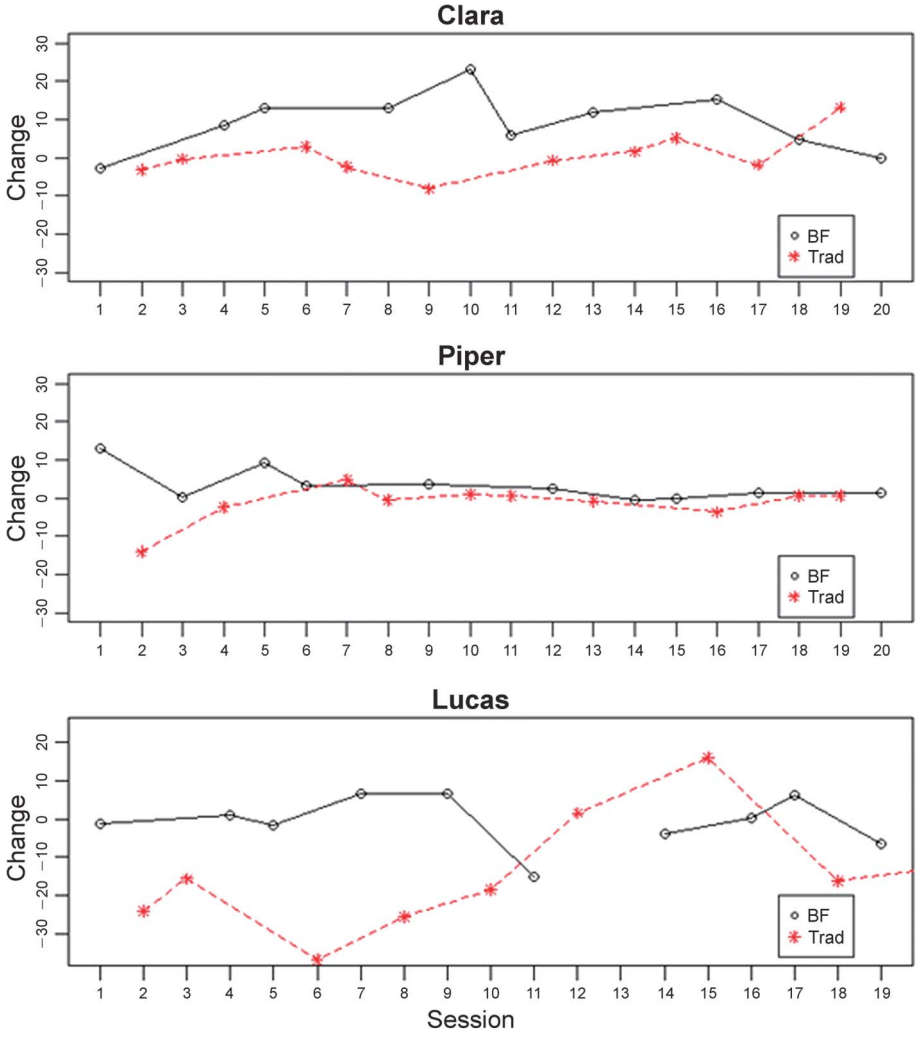
Figure 4 shows three participants who demonstrated a null or negative effect size over the course of the study. One participant, Felix, had a final effect size of −0.99. Visual inspection showed very little variation in accuracy over the course of the study, consistent with an interpretation of no meaningful change in either direction. The second participant, Lucas, had a similar effect size (−0.96) but showed much greater variability in performance within treatment sessions. This case will be examined in greater detail in the following in connection with the discussion of relative response to biofeedback versus traditional treatment. The last participant, Evan, showed a small but continuous decline in accuracy over the course of treatment. This could reflect a loss of motivation: Evan made gains in the first two sessions but showed little progress thereafter, and by the 9th or 10th treatment session, his scores slowly began slipping below baseline level. An alternative is that he could have developed a maladaptive strategy that led him to produce rhotics in a way that blinded listeners judged to be slightly less accurate than his baseline productions. The overall magnitude of this decrease was less than 10 percentage points, but the standardized effect size was quite large (−5.3). This is partly a reflection of unusually low variance in Evan's accuracy in baseline and maintenance sessions: His pooled standard deviation in $\hat{p}_{correct}$ was 1.41, whereas the standard deviation averaged across the group was 3.2 (SD = 1.2).

Individual Results: Randomization Tests

Separate randomization tests were conducted for each individual. Recall that a significant result in a randomization test suggests that one treatment played a disproportionate role in driving the participant's overall pattern of change over time. The results of the randomization test were significant in three out of seven participants: Clara ($mean_{BF} - mean_{TRAD} = 8.59$, $p = .03$), Piper ($mean_{BF} - mean_{TRAD} = 4.76$, $p < .001$), and Lucas ($mean_{BF} - mean_{TRAD} = 13.96$, $p < 0.001$). In all three cases, the results were indicative of larger gains during biofeedback treatment sessions than traditional treatment sessions. Figure 5 presents randomization plots from these three cases; participants who did not demonstrate a significant result in the randomization test are omitted for brevity. These plots present all 20 treatment sessions in sequential order.

For the first participant, Clara, the line representing change within biofeedback sessions remains above the line representing change within traditional sessions throughout the great majority of the treatment period; the lines only cross in the final three sessions. The plot for the second participant, Piper, has a different trajectory because she reached ceiling-level accuracy after five sessions and thus ceased to show any meaningful within-session change in either treatment condition. However, the line representing her gains within the biofeedback condition is well separated from the line representing traditional treatment during those first five sessions.

Figure 5. Change in $\hat{p}_{correct}$ from preprobe to postprobe across sessions, plotted only for those participants with a significant difference on the randomization test. The *y*-axis represents the magnitude of change in $\hat{p}_{correct}$ from pre- to posttreatment probe within a given session. Biofeedback (BF) sessions are represented with a solid line and circles and traditional (Trad) treatment with a dashed line and stars.



The final participant, Lucas, is noteworthy because he showed a significant difference between biofeedback and traditional treatment sessions, even though his overall trajectory was flat, indicating no meaningful long-term change over the course of the study. In this case, the significant difference between treatment conditions is driven by sizable negative changes in traditional treatment sessions in the first half of the study—that is, blinded listeners judged Lucas's speech to be less accurate after the first five sessions of traditional treatment than at baseline, whereas his accuracy remained generally unchanged after the first five biofeedback treatment sessions.[9] A likely interpretation is that Lucas initially assumed a maladaptive tongue posture in response to one or more of the articulator placement cues provided during traditional treatment. No equivalent maladaptive response appears to have been present during biofeedback treatment sessions.

Discussion

Effects of Treatment Package

Recall that all of the measures reported here reflect participants' performance on generalization probes (i.e., untreated words elicited without feedback) representing the treated category of vocalic rhotics. The combined package of biofeedback and traditional treatment had a meaningful positive effect on rhotic production accuracy for four out of seven participants in the present study. Standardized

[9]Note that there is a missing data point in connection with Session 13, a biofeedback treatment session: The pre- to posttest change could not be computed because the recording of the posttest probe was lost.

effect sizes were large (greater than 5.0) for two participants and were more modest (between 1.5 and 2.5) for two other participants. In the latter two cases, visual inspection supported the impression that these individuals showed improvement in response to treatment. Two participants showed no change over the course of the study, and one demonstrated a small but consistent decline in accuracy. The presence of nonresponders in the treated group is not particularly unusual; most studies of intervention for residual speech errors, including biofeedback treatment, report one or more participants who do not demonstrate a response to intervention within the duration of the study (e.g., McAllister Byun, Hitchcock, & Swartz, 2014; Preston et al., 2014). However, the proportion of participants who did not respond to treatment is sizable. Additional discussion of nonresponders is offered in the following.

Effects of Traditional Versus Biofeedback Treatment

Individual randomization tests revealed a significant difference between biofeedback and traditional treatment in three out of seven individuals. In all three cases, these tests indicated that biofeedback sessions were associated with larger gains than traditional sessions. However, there were two other participants who exhibited a meaningful overall response to treatment but showed no significant difference between traditional and biofeedback sessions. Thus, these results should not be interpreted as evidence that traditional treatment is ineffective. Rather, they suggest that the inclusion of biofeedback treatment can, in some cases, enhance the magnitude of gains beyond what might be observed through traditional intervention alone.

The design of the present study made it possible to analyze both short- and long-term changes in participants' rhotic production accuracy, revealing several cases in which short-term changes (from pretest to posttest within a treatment session) appeared to dissociate from the long-term learning trajectory (from baseline to maintenance). Two of these cases, Clara and Ian, showed relatively large short-term gains paired with a modest overall change. This pattern, in which gains from the treatment setting do not generalize to a wider context or are not sustained over time, is a widely documented problem that has received particular attention in the literature on biofeedback; we return to this issue in the following. A more unusual pattern in the present study was that exhibited by Lucas, who showed sizable short-term decreases in accuracy in the first five traditional treatment sessions, paired with no significant overall change from baseline to maintenance. Because no decrease in accuracy was observed following early biofeedback sessions, it was strongly suspected that Lucas was adopting a maladaptive strategy in response to one or more of the articulator placement cues provided as part of traditional treatment. It is known that typical speakers differ in the tongue postures they adopt to produce rhotics (e.g., Delattre & Freeman, 1968; Zhou et al., 2008), and previous research has advocated for flexible treatment approaches that allow the child to find the rhotic tongue posture that best fits his or her individual vocal tract (McAllister Byun, Hitchcock, & Swartz, 2014). One possible advantage of visual–acoustic biofeedback is that it allows the clinician to provide a clear target and specific feedback without committing to a particular tongue shape target (McAllister Byun, Swartz, et al., 2016). In Lucas's case, fortunately, the maladaptive strategy seems to have been eliminated by the midpoint of treatment; on the other hand, he did not succeed in replacing his preexisting articulatory pattern with one that produced perceptually accurate rhotics.

Limitations and Future Directions

The widely varying nature of the observed responses to treatment, including several nonresponders, suggests that some individuals may be better suited to benefit from biofeedback intervention than others. The participant sample observed here was quite heterogeneous, but there were no clear relationships between demographic variables and treatment outcomes. For example, this study included a large proportion of children (four of seven) with comorbid ADHD and language-based learning disability, but of the four children with comorbidities, two showed a null or negative response, and two were judged to demonstrate a clinically significant response to treatment. In addition, there was no significant correlation between age and magnitude of response to treatment ($r = .2$, $p = .61$). Note that the small sample size of a single-subject experimental design, such as this one, provides little opportunity to draw confident conclusions about any demographic predictors of response to treatment. Larger-scale studies, potentially including meta-analyses over multiple small studies, are needed to identify factors that can reliably predict an individual's likelihood of responding to biofeedback treatment.

For the four participants who did demonstrate a significant response to treatment, changes tended to be relatively small in magnitude, particularly when assessing how gains were sustained across three posttreatment maintenance sessions. Furthermore, the present study exclusively examined participants' progress on the treated category, vocalic rhotics; generalization to consonantal rhotics was not measured. It is not unusual to observe limited generalization gains after a relatively brief duration of intervention for residual rhotic errors (Gibbon & Paterson, 2006; McAllister Byun & Hitchcock, 2012). On the other hand, the present study incorporated challenge point principles that were specifically intended to maximize generalization to words produced without feedback (Hitchcock & McAllister Byun, 2014). In light of this, it is disappointing that multiple participants who made progress within treatment still showed limited generalization. This points to a need for further research investigating the influence of the challenge point structure on the magnitude of generalization gains in biofeedback intervention.

One important modification that should be incorporated into future studies is an increase in dose frequency (i.e., the number of trials elicited per session). The number of trials per session in this study was determined on the

basis of previous intervention research (e.g., McAllister Byun & Hitchcock, 2012). However, children in those studies were younger, on average, than the present participants. Older participants can easily handle significantly larger numbers of trials, and this may result in larger effect sizes (e.g., Edeal & Gildersleeve-Neumann, 2011).

Another direction for future investigation is raised by the observation that the differences between biofeedback and traditional treatment, shown in Figure 5, tend to be most prominent in the early phases of treatment. (In Piper's case, this was attributed to a ceiling effect; however, Clara and Lucas also showed diminished separation between conditions toward the end of the treatment period.) This is reminiscent of an existing literature describing the manner in which motor learning is influenced by qualitative or knowledge of performance feedback, of which biofeedback is one type (Maas et al., 2008; Volin, 1998). Previous research suggests that when a movement target is novel or unfamiliar, learning can be enhanced through detailed qualitative feedback; however, knowledge of performance feedback has been shown to lose its advantage or even have a detrimental effect when the target is already well specified (Hodges & Franks, 2001; Maas et al., 2008). Further investigation should assess the possibility that improved outcomes might be possible in connection with a treatment package consisting of an early phase of biofeedback treatment, followed by a period of traditional articulatory intervention to encourage generalization.

A final possibility to consider in future research is that the rapid alternation between biofeedback and traditional treatment used in the present study may not have been optimal for observing differences between the two conditions. Increasing the number of consecutive sessions of a given treatment type could allow the effects of that treatment to build up before switching, which might, in turn, make it possible to observe larger differences between conditions. In the long term, a randomized controlled trial is likely to represent the best way to systematically measure the difference in efficacy of biofeedback versus traditional interventions for residual errors.

Conclusions

This study addressed two experimental objectives. The first was to collect systematic evidence on the efficacy of a combined treatment package, including traditional articulatory treatment and visual–acoustic biofeedback intervention for residual rhotic errors. Four out of seven participants showed a clinically significant effect size for the treated category that was supported by visual inspection of the data over time, while three showed a null or negative response. A second objective was to compare the relative magnitude of short-term gains made in biofeedback versus traditional treatment sessions. Individual randomization tests revealed significant differences for three participants. In these three cases, gains associated with biofeedback intervention exceeded gains in traditional treatment; there were no cases in which gains in traditional treatment significantly exceeded gains in biofeedback. Taken together, these results support the effectiveness of visual–acoustic biofeedback intervention for some individuals with residual rhotic errors. However, the presence of nonresponders and participants who exhibited limited generalization make it clear that visual–acoustic biofeedback is not a silver bullet solution. Larger-scale research is needed to make a confident judgment about the relative efficacy of traditional versus biofeedback intervention and to identify predictors that indicate which candidates are most likely to demonstrate a successful response to biofeedback treatment.

Acknowledgments

The author acknowledges support for this research by the National Institutes of Health Grant NIH R03DC 012883 and also by a travel fellowship to attend the Institute of Education Sciences Single-Case Design and Analysis Institute 2014. The author gratefully acknowledges the contributions of the following individuals: for treatment delivery, Heather Campbell; for developing the Challenge-R software, José Ortiz and Elaine Hitchcock; for programming support, Daniel Szeredi; for statistical consultation, Daphna Harel; for data collection and management, numerous student assistants at New York University, notably Laine Cialdella, Tala Ginsberg, Deanna Kawitzky, and Christopher Nightingale. Many thanks also to all participants and their families for their cooperation throughout the study.

References

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

Bacsfalvi, P. (2010). Établissement des composantes linguales du son/r/à l'aide d'ultrasons chez trois adolescents avec un implant cochléaire. [Attaining the lingual components of /r/ with ultrasound for three adolescents with cochlear implants]. *Revue canadienne d'orthophonie et d'audiologie, 34*(3), 206–217.

Becker, M., & Levine, J. (2010). *Experigen—an online experiment platform*. Retrieved from https://github.com/tlozoot/experigen

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

Bernhardt, B. M., & Stemberger, J. P. (1998). *Handbook of phonological development from the perspective of constraint-based nonlinear phonology*. San Diego, CA: Academic Press.

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

Boyce, S., & Espy-Wilson, C. (1997). Coarticulatory stability in American English /r/. *The Journal of the Acoustical Society of America, 101,* 3741–3753.

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

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

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental

behavioral research. *PLoS One, 8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

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

Davis, S. M., & Drichta, C. E. (1980). Biofeedback theory and application in allied health. *Biofeedback and Self-Regulation, 5,* 159–174.

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

Edeal, D. M., & Gildersleeve-Neumann, C. E. (2011). The importance of production frequency in therapy for childhood apraxia of speech. *American Journal of Speech-Language Pathology, 20,* 95–110.

Edgington, E. S. (1987). *Randomization tests* (2nd ed.). New York, NY: Marcel Dekker.

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 153–184). Washington, DC: American Psychological Association.

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

Flipsen, P., Jr., Shriberg, L. D., Weismer, G., Karlsson, H. B., & McSweeny, J. L. (2001). Acoustic phenotypes for speech-genetics studies: Reference data for residual /ɝ/ distortions. *Clinical Linguistics & Phonetics, 15,* 603–630.

Gibbon, F., & Paterson, L. (2006). A survey of speech and language therapists' views on electropalatography therapy outcomes in Scotland. *Child Language Teaching and Therapy, 22,* 275–292.

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

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,* 212–224.

Hagiwara, R. (1995). *Acoustic realizations of American /r/ as produced by women and men* (Unpublished doctoral dissertation). University of California, Los Angeles.

Hesketh, C. A., Nightingale, C., & Hall, R. A. (2000). Phonological awareness therapy and articulatory training approaches for children with phonological disorders: A comparative outcome study. *International Journal of Language & Communication Disorders, 35,* 337–354.

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

Hitchcock, E. R., & McAllister Byun, T. (2014). Enhancing generalization in biofeedback intervention using the challenge point framework: A case study. *Clinical Linguistics & Phonetics, 29,* 59–75.

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

Howard, D., Best, W., & Nickels, L. (2015). Optimising the design of intervention studies: Critiques and ways forward. *Aphasiology, 29,* 526–562.

Ipeirotis, P. G., Provost, F., Sheng, V. S., & Wang, J. (2014). Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery, 28,* 402–441.

Kaderavek, J. N., & Justice, L. M. (2010). Fidelity: An essential component of evidence-based practice in speech-language pathology. *American Journal of Speech-Language Pathology, 19,* 369–379.

Katz, W., McNeil, M., & Garst, D. (2010). Treating apraxia of speech (AOS) with EMA-supplied visual augmented feedback. *Aphasiology, 24,* 826–837.

Klein, E. S. (1996). Phonological/traditional approaches to articulation therapy: A retrospective group comparison. *Language, Speech, and Hearing Services in Schools, 27,* 314–323.

Klein, H. B., McAllister Byun, T., Davidson, L., & Grigos, M. I. (2013). A multidimensional investigation of children's /r/ productions: Perceptual, ultrasound, and acoustic measures. *American Journal of Speech-Language Pathology, 22,* 540.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34,* 26–38.

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

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 91–126). Washington, DC: American Psychological Association.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105,* 1455–1468.

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,* 613–622.

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,* 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,* 277–298.

Magloughlin, L. (2016). Accounting for variability in North American English /ɹ/: Evidence from children's articulation. *Journal of Phonetics, 54,* 51–67.

Martin, N. A., & Brownell, R. (2005). *Test of Auditory Processing Skills–Third Edition*. Novato, CA: Academy Therapy Publications.

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

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., & 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,* 207–221.

McAllister Byun, T., Hitchcock, E. R., & Ortiz, J. (2014). *Challenge-R: Computerized challenge point treatment for /r/ misarticulation*. Talk presented at American Speech-Language-Hearing Association, Orlando, FL.

McAllister Byun, T., Hitchcock, E. R., & Swartz, M. T. (2014). Retroflex versus bunched in treatment for /r/ misarticulation: Evidence from ultrasound biofeedback intervention. *Journal of Speech, Language, and Hearing Research, 57,* 2116–2130.

McAllister Byun, T., Swartz, M. T., Halpin, P. F., Szeredi, D., & Maas, E. (2016). Direction of attentional focus in biofeedback treatment for /r/ misarticulation. *International Journal of Language & Communication Disorders, 51,* 384–401.

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,* 871–884.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411–419.

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

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,* 2102–2115.

Psychology Software Tools, Inc. (2012). E-Prime 2.0 [Computer software]. Retrieved from http://www.pstnet.com

Robey, R. R. (2004). A five-phase model for clinical-outcome research. *Journal of Communication Disorders, 37,* 401–411.

Ruscello, D. M. (1995). Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders, 28*(4), 279–302.

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

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

Shipley, K., & McAfee, J. (2008). *Assessment in speech-language pathology: A resource manual*. Clifton Park, NY: Cengage Learning.

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

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

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

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

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

Sjölander, K., & Beskow, J. (2006). *Wavesurfer*. Retrieved from http://www.speech.kth.se/wavesurfer/

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,* 779–798.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43,* 155–167.

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,* 2633–2634.

Van Riper, C., & Erickson, R. L. (1996). *Speech correction: An introduction to speech pathology and audiology* (9th ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., & Choe, A. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *The Journal of the Acoustical Society of America, 123,* 4466–4481.

Appendix A

Rhotic Probe Words

Word	Consonantal or Vocalic	Included in which probes
rose	Consonantal	Baseline–maintenance, within treatment (pre and post)
rude	Consonantal	Baseline–maintenance, within treatment (pre and post)
raw	Consonantal	Baseline–maintenance, within treatment (pre and post)
rob	Consonantal	Baseline–maintenance, within treatment (pre and post)
rug	Consonantal	Baseline–maintenance, within treatment (pre and post)
race	Consonantal	Baseline–maintenance, within treatment (pre and post)
red	Consonantal	Baseline–maintenance, within treatment (pre and post)
ride	Consonantal	Baseline–maintenance, within treatment (pre and post)
reach	Consonantal	Baseline–maintenance, within treatment (pre and post)
raft	Consonantal	Baseline–maintenance, within treatment (pre and post)
scarf	Vocalic	Baseline–maintenance, within treatment (pre and post)
dark	Vocalic	Baseline–maintenance, within treatment (pre and post)
farm	Vocalic	Baseline–maintenance, within treatment (pre and post)
chair	Vocalic	Baseline–maintenance, within treatment (pre and post)
stare	Vocalic	Baseline–maintenance, within treatment (pre and post)
weird	Vocalic	Baseline–maintenance, within treatment (pre and post)
clear	Vocalic	Baseline–maintenance, within treatment (pre and post)
beard	Vocalic	Baseline–maintenance, within treatment (pre and post)
fork	Vocalic	Baseline–maintenance, within treatment (pre and post)
sword	Vocalic	Baseline–maintenance, within treatment (pre and post)
turn	Vocalic	Baseline–maintenance, within treatment (pre and post)
nurse	Vocalic	Baseline–maintenance, within treatment (pre and post)
worm	Vocalic	Baseline–maintenance, within treatment (pre and post)
ladder	Vocalic	Baseline–maintenance, within treatment (pre and post)
hammer	Vocalic	Baseline–maintenance, within treatment (pre and post)
rock	Consonantal	Baseline–maintenance only
wrong	Consonantal	Baseline–maintenance only
robe	Consonantal	Baseline–maintenance only
run	Consonantal	Baseline–maintenance only
rules	Consonantal	Baselinevmaintenance only
read	Consonantal	Baseline–maintenance only
ring	Consonantal	Baseline–maintenance only
rake	Consonantal	Baseline–maintenance only
wrap	Consonantal	Baseline–maintenance only
rip	Consonantal	Baseline–maintenance only
barn	Vocalic	Baseline–maintenance only
star	Vocalic	Baseline–maintenance only
board	Vocalic	Baseline–maintenance only
door	Vocalic	Baseline–maintenance only
floor	Vocalic	Baseline–maintenance only
scare	Vocalic	Baseline–maintenance only
share	Vocalic	Baseline–maintenance only
tear	Vocalic	Baseline–maintenance only
cheer	Vocalic	Baseline–maintenance only
year	Vocalic	Baseline–maintenance only
sir	Vocalic	Baseline–maintenance only
stir	Vocalic	Baseline–maintenance only
butter	Vocalic	Baseline–maintenance only
flower	Vocalic	Baseline–maintenance only
mother	Vocalic	Baseline–maintenance only

Appendix B

Details of Protocol for Online Collection of Perceptual Ratings

Binary perceptual ratings of rhotic sounds elicited in probe measures were collected from nonspecialist listeners recruited through the online crowdsourcing platform Amazon Mechanical Turk and directed to a task hosted on the Experigen online experiment presentation platform (Becker & Levine, 2010). Following McAllister Byun, Halpin, and Szeredi (2015), each token was initially presented to nine unique listeners for rating. Due to 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 cleanup blocks to collect additional ratings. If at least eight ratings had not been collected after three cleanup rounds, items were discarded.

Upon initiating the task, raters were informed that they would hear words containing r sounds produced by children of varying ages and that their job was to rate each r sound as correct or incorrect. In each trial, participants saw the target word in standard orthography and heard the child's production of the word, which they could listen to up to three times. Prior to rating any experimental stimuli, raters were required to complete 20 training trials in which they received feedback on the accuracy of their responses. Raters were then required to complete a 100-item eligibility test, in which their responses were evaluated by using a criterion that combines acoustic measures and experienced listener ratings, as described in McAllister Byun, Halpin, and Szeredi (2015). Only raters who passed this criterion were eligible to proceed to rate experimental trials for pay.

Files were presented in random order in blocks of 220, of which 150 were experimental trials, 50 were filler trials, and 20 were catch trials. Half of the filler files were collected from typically developing speakers, and half were elicited from children receiving treatment in a previous study. Fillers were included so that the sample of words raters heard would not be overly skewed toward incorrect productions and so that raters would not become overly familiar with the voices of the seven participants in the present study. The 20 catch trials were items hand selected to represent unambiguously correct or incorrect production. They were used to monitor a rater's attention to the task: If a participant did not score above chance on the catch trials in a block, the rater received a warning, and results from that block were discarded. Raters could complete multiple blocks, but after five blocks, they were required to pass the eligibility test again to continue. A total of 307 unique raters completed at least one block.