ASHA American Speech-Language-Hearing Association

LSHSS

Research Article

Telepractice Treatment of Residual Rhotic Errors Using App-Based Biofeedback: A Pilot Study

Laura Peterson,[a] Christian Savarese,[b] Twylah Campbell,[c] Zhigong Ma,[c] Kenneth O. Simpson,[a] and Tara McAllister[c]

[a]Department of Speech-Language Pathology, Rocky Mountain University of Health Professions, Provo, UT [b]Department of Linguistics, Harvard University, Cambridge, MA [c]Department of Communicative Sciences and Disorders, New York University, NY

ARTICLE INFO

Article History:
Received May 27, 2021
Revision received August 6, 2021
Accepted September 30, 2021

Editor-in-Chief: Holly L. Storkel
Editor: Megann McGill

https://doi.org/10.1044/2021_LSHSS-21-00084

ABSTRACT

Purpose: Although mobile apps are used extensively by speech-language pathologists, evidence for app-based treatments remains limited in quantity and quality. This study investigated the efficacy of app-based visual–acoustic biofeedback relative to nonbiofeedback treatment using a single-case randomization design. Because of COVID-19, all intervention was delivered via telepractice.
Method: Participants were four children aged 9–10 years with residual errors affecting American English /ɹ/. Using a randomization design, individual sessions were randomly assigned to feature practice with or without biofeedback, all delivered using the speech app *Speech Therapist's App for /r/ Treatment*. Progress was assessed using blinded listener ratings of word probes administered at baseline, posttreatment, and immediately before and after each treatment session.
Results: All participants showed a clinically significant response to the overall treatment package, with effect sizes ranging from moderate to very large. One participant showed a significant advantage for biofeedback over nonbiofeedback treatment, although the order of treatment delivery poses a potential confound for interpretation in this case.
Conclusions: While larger scale studies are needed, these results suggest that app-based treatment for residual errors can be effective when delivered via telepractice. These results are compatible with previous findings in the motor learning literature regarding the importance of treatment dose and the timing of feedback conditions.
Supplemental Material: https://doi.org/10.23641/asha.18461576

Speech-language pathologists (SLPs) serving pediatric populations often struggle with large caseloads and heavy documentation requirements (Edgar & Rosa-Lugo, 2007; Kenny & Lincoln, 2012). As a consequence, many SLPs find themselves unable to achieve the frequency or dose of treatment recommended for the remediation of speech sound disorder (SSD) in children (Sugden et al., 2018). Treatment incorporating apps or computer programs could, in principle, help alleviate these challenges (McKechnie et al., 2018). For example, embedding speech practice in a gamified context could increase motivation and help clients complete a large number of trials per session. Apps and computer games could also be used to encourage home practice of speech targets, particularly if the software in question is able to provide accurate feedback on the user's productions (McKechnie, 2019). Given these potential advantages, as well as the decreasing cost and increasing prevalence of tablet technology, it is unsurprising that app-based and computer-based methods are on the rise in the treatment of pediatric SSD (Chen et al., 2016; Edwards & Dukhovny, 2017; Furlong et al., 2018; McKechnie et al., 2018).

Unfortunately, the high demand for technology-enhanced interventions for SSD has not been met with a

Correspondence to Tara McAllister: tkm214@nyu.edu. **Publisher Note:** This article is part of the Forum: Speech and Language Tele-Intervention: The Future Is Now. ***Disclosure:*** *Tara McAllister oversees the development of the staRt app described in this research. The staRt app is noncommercial at the present time but could transition to a commercial venture in the future. The other authors have declared that no other competing financial or nonfinancial interests existed at the time of publication.*

corresponding volume of evidence to support or refute the efficacy of these tools. For instance, in a systematic search of publicly available apps, Furlong et al. (2018) identified 132 unique apps intended for use in the treatment of pediatric SSD. Of these, only 19 apps (14%) were judged to have potential therapeutic benefit for children with SSD, suggesting a disconnect between software development and the rigor required for evidence-based clinical practice. There is a need for more research on apps designed for the treatment of SSD so that practicing SLPs can make evidence-based decisions about their use.

The long-term goal of the program of research described here is to document the relative efficacy of traditional versus biofeedback treatment for residual speech errors (RSEs) affecting American English /ɹ/ using app-based delivery of both treatment types. The app in question is *Speech Therapist's App for /r/ Treatment* (*staRt*), which is currently in development at New York University (NYU). The present pilot study reports the results of a single-case experimental study in which four participants received treatment using the staRt app. Although the study was originally planned to be conducted in person, this plan was disrupted by the COVID-19 pandemic. With social distancing requirements preventing in-person research activities for an unknown period of time, it was necessary to deliver staRt app treatment via telepractice. This change in study design was not ideal for the purpose of testing the efficacy of the staRt app, which was not developed for the telepractice context; this introduced some technical obstacles that are described in detail below. On the other hand, this modification allowed us to obtain evidence on the efficacy of telepractice delivery of traditional and biofeedback treatment for RSEs affecting English rhotics—a topic that has not, to our knowledge, been investigated in previous literature.

RSEs

The label RSEs is applied to describe a variety of speech production errors when they occur in older children and adults, typically affecting late-emerging sounds such as rhotics, sibilants, and laterals. These sounds have a protracted time course of emergence in the typical population, with some estimates suggesting that children with typical speech sound development as old as 8 years might still produce sounds such as American English /ɹ/ in a nonadultlike fashion (Smit et al., 1990; but see Crowe & McLeod, 2020). However, when these errors persist beyond 9 years of age, they are generally considered atypical (Shriberg et al., 1994). Children with RSEs pose a challenge for pediatric SLPs because many of them fail to respond to conventional forms of intervention even after months or years of treatment. A survey of SLPs in the school setting (Ruscello, 1995) indicated that many children with RSEs are discharged with their errors unresolved. Without successful treatment, however, errors may persist into adolescence or even adulthood; Culton (1986) estimated that 1%–2% of college-age adults present with RSEs. Although the impact of RSEs on overall intelligibility may be relatively minor, these errors are reported to affect peer perception (Crowe Hall, 1991) and socioemotional well-being (Hitchcock et al., 2015). Thus, there is a need for innovative solutions to increase the accessibility and effectiveness of treatment for RSEs.

While RSEs may affect a range of phonemes, this study is particularly concerned with errors affecting the rhotic /ɹ/ in North American English. Even though it is a high-frequency phoneme, /ɹ/ is one of the latest emerging sounds in typical development (Smit et al., 1990), and it is considered one of the most common treatment-resistant errors (Ruscello, 1995). Children's difficulties with /ɹ/ are, at least in part, attributable to its articulatory complexity. While most lingual sounds are produced with a single major point of constriction between the tongue and another structure, articulation of /ɹ/ involves two near-simultaneous lingual constrictions—one in the alveolar/palatal region and another in the pharynx (Boyce, 2015). In addition, the tongue constrictions for /ɹ/ can be difficult to cue because they are not externally visible, and they provide only weak tactile feedback due to the limited nature of tongue–palate contact. Finally, variability both within and across speakers in tongue shapes for /ɹ/ (Boyce, 2015; Mielke et al., 2016) can make it challenging for SLPs to determine appropriate articulatory cues to use in treatment. The acoustic cues for /ɹ/ are more stable: Its hallmark is a very low third-formant frequency (F3), yielding a small distance between the second and third formants (Espy-Wilson et al., 2000; Lehiste, 1964). In visual–acoustic biofeedback for /ɹ/, described in detail below, learners are encouraged to make articulatory adjustments to try to achieve the lowered F3 that characterizes /ɹ/.

Biofeedback Treatment for RSEs

Visual biofeedback utilizes technology to take measurements of certain physiological or behavioral components of a learner's performance and feed them back in a real-time visual display. A visual display makes explicit certain information that might otherwise be ambiguous or imperceptible to the learner (Davis & Drichta, 1980; Volin, 1998), which could help the learner exercise greater control over their own performance. Visual biofeedback has been used in speech-language pathology for several decades (e.g., Ruscello, 1995; Shawker & Sonies, 1985). One approach involves providing a display representing the position and movement of the articulators during speech, which can be compared with a visual model representing correct production of the target sound. Different technologies are

used to provide articulatory biofeedback for RSEs affecting /ɹ/. These include ultrasound imaging, which shows the shape and movements of the tongue (Bernhardt et al., 2005; Preston et al., 2019), and electropalatography, which displays areas of contact between the tongue and the palate (Hitchcock et al., 2017).

Visual–acoustic biofeedback is an alternative approach that implements the above-described principles using a real-time display of the acoustic signal of speech. This display may take the form of a real-time spectrogram (Shuster et al., 1992, 1995) or a linear predictive coding (LPC) spectrum (McAllister Byun, 2017; McAllister Byun & Campbell, 2016; McAllister Byun & Hitchcock, 2012). In either case, the learner can see the first three resonant frequencies of the vocal tract, formants F1, F2, and F3; these appear as bands of energy in a spectrogram and as peaks in an LPC spectrum. As in the articulatory case, biofeedback treatment involves comparing the real-time display to a model representing correct production. For /ɹ/, this involves calling attention to the lowered frequency of the third formant, F3, as described above. Although articulatory information is not part of the visual display, clinicians providing visual–acoustic biofeedback often incorporate articulator placement cues to help clients adjust their output to better match the model formant structure (McAllister Byun et al., 2016).

To date, evidence for the efficacy of visual biofeedback treatment for RSEs has typically taken the form of case studies and single-case experimental research. Some studies have investigated the effects of biofeedback as the sole treatment of interest, whereas others have compared biofeedback treatment against a traditional treatment comparison condition. (In traditional approaches to treatment of speech errors, such as that in Van Riper and Erickson [1996], the clinician typically provides auditory models of the target sound for the client to imitate, paired with verbal and/or visual cues for articulator placement.) For the purpose of this article, we limit our attention to the literature investigating visual–acoustic biofeedback, because this study involves a visual–acoustic biofeedback app. For an overview of the evidence base on the efficacy of ultrasound biofeedback, see the recent systematic review in Sugden et al. (2019).

A small but growing literature has suggested that visual–acoustic biofeedback can produce significant treatment gains in some children with RSEs affecting /ɹ/ who have not responded to previous forms of treatment. A quasi-experimental study in which children with RSEs received an initial period of traditional treatment followed by visual–acoustic biofeedback (McAllister Byun & Hitchcock, 2012) found that eight of 11 participants showed significant gains in perceptual and acoustic measures of /ɹ/ production after 10 weeks of treatment and that these gains occurred only after the transition to biofeedback for all but one participant. In a single-case randomization study (McAllister Byun, 2017) in which participants were provided with an equal number of biofeedback and traditional treatment sessions in a randomized order, three of seven participants showed significantly greater short-term gains in visual–acoustic biofeedback sessions than in traditional treatment sessions, whereas none showed a significant advantage in the opposite direction. Finally, a single-case experimental study of 11 children (McAllister Byun & Campbell, 2016) aimed to compare the effect of visual–acoustic biofeedback and traditional methods by dividing the treatment period into a biofeedback phase and a traditional phase, counterbalanced in order across participants. The results revealed a significant interaction between treatment condition and order, such that biofeedback was more effective than traditional methods in the first phase of treatment, but traditional treatment was more effective than biofeedback in the second phase. This is consistent with previous motor learning literature suggesting that knowledge of performance (KP) feedback—of which biofeedback is one type—is most effective when provided in the early stages of treatment (Maas et al., 2008). Overall, these results indicate that visual–acoustic biofeedback can produce significant gains in some children with RSEs, and they suggest that biofeedback may offer advantages over traditional methods, at least in the early stages of treatment targeting /ɹ/.

However, previous studies of biofeedback efficacy have also acknowledged several limitations. First, gains made in the treatment setting through biofeedback do not necessarily generalize to contexts in which enhanced feedback is not available; there is agreement on this point across studies of various biofeedback technologies (Gibbon & Paterson, 2006; McAllister Byun & Hitchcock, 2012; Sjolie et al., 2016). Second, treatment outcomes are variable across participants, with most studies reporting a mix of strong responders and nonresponders (e.g., McAllister Byun & Campbell, 2016; Preston et al., 2019). This heterogeneity in response makes it difficult to draw conclusions about the relative efficacy of biofeedback and traditional treatment, particularly in the small-*n* studies that form the current evidence base. Therefore, previous studies (e.g., McAllister Byun, 2017) have called for larger sample sizes to enable robust comparison of the efficacy of biofeedback and traditional treatment.

An additional limitation of biofeedback intervention pertains to the cost of the technologies used to generate the real-time visual display used in treatment. While actual costs vary depending on the technology used, systems used in biofeedback research are generally priced outside the operating budget of a typical SLP. The need for additional training to be able to use and interpret technologies for visualization of speech may also act as a barrier to wider adoption of biofeedback methods. McAllister Byun et al. (2017) suggested that these cost-related limitations could be addressed by making visual–acoustic biofeedback

available in the form of a mobile app. In addition, putting the app in the hands of a large number of practitioners could enable research–clinical partnerships that may help increase the scale of data collection and thereby provide stronger evidence on questions about the efficacy of biofeedback treatment.

The staRt App

staRt is an iOS app intended for use in a clinical setting for treatment of RSEs affecting /ɹ/. The staRt app provides visual–acoustic biofeedback in the form of a real-time LPC spectrum, with frequency on the *x*-axis and amplitude on the *y*-axis.[1] The LPC spectrum is styled as a wave in a "beach-themed environment with a cheerful gender-neutral palette" (McAllister Byun et al., 2017), shown in Figure 1; the peaks of the wave represent formants F1, F2, and so forth. An adjustable slider superimposed over the LPC spectrum serves as a visual target indicating approximately where the user's third formant (F3) should fall for a correct /ɹ/ sound. The default position of this target is determined from normative data based on the user's age and sex (Lee et al., 1999), which are entered during user profile creation, but the clinician can also customize the placement of the target. A self-paced tutorial module, designed for the child and the treating SLP to review together, introduces the LPC spectrum ("wave") and provides both text descriptions and visual examples of LPC spectra for correct and incorrect productions of /ɹ/.

The biofeedback display in its most basic form is presented in a prepractice module (termed "Free Play" in the app), where the learner and clinician can interact with the real-time LPC in an unstructured fashion. The app also features "Quiz" and "Quest" modules, intended for assessment and treatment, respectively. These modules are subdivided into syllable- and word-level options. Further detail on these components of the app is provided in the Method section, and complete word lists from both modes are included in the online supplement to this article at https://osf.io/6gyt4/.

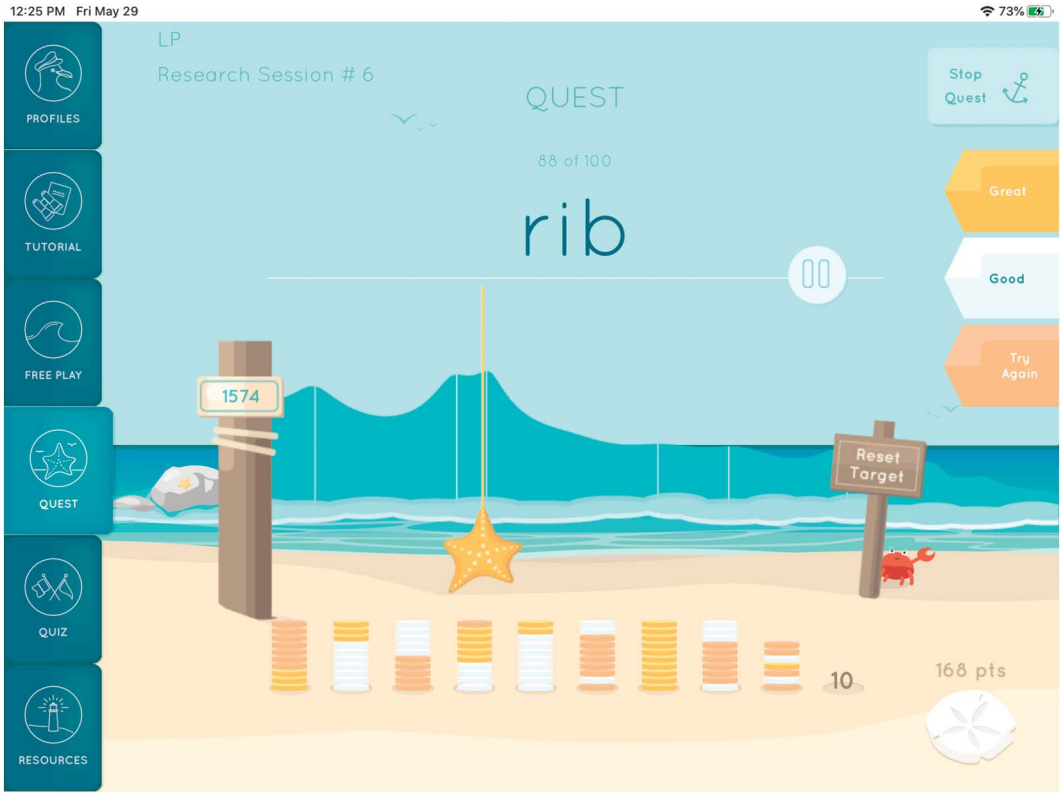
Previous research on the staRt app took the form of a proof-of-concept study examining the effects of the app with a single participant in the laboratory setting (McAllister Byun et al., 2017). The case study participant was a 12-year-old girl, pseudonym "Hannah," with RSEs affecting /ɹ/. As part of another research study at NYU, Hannah had previously received ten 30-min sessions of visual–acoustic biofeedback treatment using the commercially available software CSL Sona-Match (KayPENTAX), as well as ten 30-min sessions of traditional treatment. In that study, Hannah was judged to exhibit improved accuracy in /ɹ/ production within the treatment setting, but she showed minimal generalization of her gains to a context in which biofeedback was not available. The follow-up study was conducted to evaluate whether she would continue to show gains in /ɹ/ production accuracy in the treatment setting when biofeedback treatment was provided with the staRt app instead of Sona-Match and whether any gains would generalize beyond the treatment setting. With regard to generalization, the results of the follow-up study were disappointing: Acoustic measures showed that Hannah's /ɹ/ productions were slightly less accurate in a posttreatment maintenance probe than they had been at baseline. Within the treatment setting, however, both the treating clinician's ratings and acoustic measurements indicated that Hannah's accuracy while using the staRt app for biofeedback continued to increase at a rate comparable to that observed with CSL Sona-Match. Although the results in McAllister Byun et al. (2017) do not support the efficacy of biofeedback with regard to generalization, they do provide a proof-of-concept demonstration of a comparable within-treatment response using the staRt app versus CSL Sona-Match. Building off that finding, this study investigated the effects of treatment with the staRt app outside of the laboratory setting. While the original intent was to conduct a community-based in-person study, COVID-19 social distancing requirements made it necessary to adapt the study to the telepractice context.

Effectiveness of Speech Intervention Delivered via Telepractice

Telepractice delivery of clinical services for communication disorders was increasing even before the COVID-19 pandemic, and social distancing requirements in 2020 resulted in a dramatic increase in the number of SLPs using remote service delivery. Despite the widespread use of telepractice service delivery, there are limited data regarding the efficacy of speech intervention provided via telepractice; here, we focus on the evidence base pertaining to treatment of pediatric SSD. Wales et al. (2017) completed a systematic review of previous publications concerning the treatment of speech sounds via telepractice. They identified a total of seven qualifying studies that were judged to provide "limited but promising evidence" for the efficacy of telepractice service delivery in school-age children. Five of the studies investigated a combination of speech sound and language treatment via telehealth, whereas two studies (Grogan-Johnson et al., 2011, 2013) focused mainly on speech sound treatment and, thus, have the greatest relevance to the present investigation. In a comparison of 14 children with SSD from two different schools, one staffed in person and

[1]A real-time LPC speech processing algorithm is available in other software (e.g., KayPENTAX Computerized Speech Lab, MATLAB), but these are costly and can require considerable training to operate; they also do not operate on mobile devices.

Figure 1. Still frame from the staRt app. The real-time linear predictive coding spectrum used for biofeedback appears as the blue wave, with formants appearing as peaks in the wave; the starfish slider provides an adjustable target for F3. This screenshot shows an accurate production of /ɹ/ in which F2 and F3 are very close together.



one via telepractice, Grogan-Johnson et al. (2011) found no significant differences in participants' acquisition of target sounds. In a follow-up study (Grogan-Johnson et al., 2013), 14 children with SSD aged 6–10 years were randomly assigned to receive treatment in person or via telepractice. Both groups were found to improve significantly, with no significant between-groups difference.

Despite the promising nature of early studies, the existing body of evidence is insufficient for strong conclusions regarding the efficacy of telepractice treatment for pediatric SSD (Wales et al., 2017). Of the small number of published studies on the topic, most are limited by small sample sizes and/or suboptimal experimental design (e.g., using convenience samples or retrospective data collection). Information regarding the specific treatment methods used in telepractice is not always reported (e.g., Coufal et al., 2018). To the best of our knowledge, no previous published research has investigated the efficacy of visual–acoustic biofeedback as delivered via telepractice, nor have previous studies specifically focused on telepractice treatment in the population of children with RSEs. Thus, there is a need for further research to document the efficacy of specific interventions for SSD as delivered via telepractice.

Research Questions

This study contributes to this small but growing literature by measuring response to telepractice delivery of both traditional and visual–acoustic biofeedback treatment for residual rhotic errors. Although the sample size is small, the use of a single-case randomization design makes it possible to compare the two treatment types on a within-subject basis with high internal validity. As the preceding paragraphs indicate, there was limited precedent in the literature to guide our expectations for this study. However, it was clear at the outset that telepractice delivery of treatment for RSEs faces several technological hurdles. Videoconferencing technologies such as Zoom may introduce filtering or compression of the audio signal, which could affect the clinician's ability to classify a child's productions as accurate or distorted. Connectivity disruptions during a telepractice session may affect the clinician's ability to score productions or provide timely feedback. Finally, video transmission may create a temporal lag between the child's production and the response of the visual display in biofeedback treatment, which could make learning in this context more difficult. In addition to technical issues, we are familiar with anecdotal

suggestions that behavior management is more challenging in the telepractice setting than in the context of in-person interactions. Behavioral challenges could reflect the fact that the clinician has less direct influence over the client's behavior in a remote setting (e.g., the child can put the clinician on mute), as well as the possibility that some children were struggling with long hours of interacting via screens during the pandemic. Considering these factors, as well as the fact that most studies of intervention for RSEs yield a mix of responders and nonresponders, we hypothesized that most participants would make clinically significant gains in 16 sessions of traditional and biofeedback treatment delivered via telepractice. We further hypothesized that at least some participants would show an advantage for biofeedback over traditional treatment, as in McAllister Byun (2017). However, we tentatively predicted that effect sizes associated with biofeedback treatment of RSEs might be smaller in the telepractice context relative to previous studies involving in-person delivery of treatment.

Method

Recruitment

The NYU Institutional Review Board approved the investigation before implementation through a reliance agreement with Rocky Mountain University of Health Professions (Protocol No. FY2016–622). Participants, who could come from anywhere in the United States, were recruited through recruitment flyers, electronic mailing lists, and social media posts. Informed consent and assent were obtained electronically: The first author met with interested participants and their families via Zoom call to discuss requirements for participation and answer any questions. Signed consent and assent forms were uploaded to the study team via a link to a secure folder using Box software (Box, Inc.)

All treatment in this study was delivered by the first author, a certified SLP with 23 years of experience with pediatric clients. To prepare to deliver biofeedback treatment with the staRt app, the first author reviewed a series of training materials developed in connection with a randomized controlled trial comparing biofeedback and traditional treatment for /ɹ/ (McAllister et al., 2020), available online at https://osf.io/pg8sw/. She also reviewed a tutorial article describing recommended practices for traditional /ɹ/ intervention (Preston et al., 2020). The first and final authors jointly developed a master protocol for the study that specifies all the actions to be carried out at each step in the data collection process, from the administration of evaluation measures and baseline probes to the delivery of treatment. Finally, the same two authors jointly conducted a pilot treatment session, which allowed them to agree on a uniform standard for cueing and scoring children's rhotic productions.

Participants

The present pilot study involved four participants, who will be referred to by the pseudo-initials Y.L., O.R., R.T., and E.L. At the onset of the study, the participants were aged 10;3, 9;0, 9;2, and 10;2 (years;months), respectively. All participants spoke American English as their primary language and were exposed primarily to rhotic dialects of American English. They were required to exhibit difficulty producing /ɹ/, operationalized as a requirement to score at or below 40% correct on the initial administration of the Long Word Quiz instrument described below. Participants additionally completed an online administration of the Goldman-Fristoe Test of Articulation–Third Edition (GFTA-3; Goldman & Fristoe, 2015) and were required to score below the 8th percentile on that measure. No sounds other than /ɹ/ were scored as incorrect on the GFTA-3 for any of the participants in this study. Participants were also required to achieve a scaled score of 6 or higher on an online administration of the Word Classes and Recalling Sentences subtests of the Clinical Evaluation of Language Fundamentals–Fifth Edition (Wiig et al., 2013). A brief oral mechanism screening was administered via Zoom, and all participants were judged to exhibit no gross abnormalities in structure or function. Children who wore a palatal expander or other orthodontia that could affect tongue placement were excluded from the investigation. No hearing screening was administered due to the online nature of the study, but participants were required to have no history of hearing concerns, per parent report on a history questionnaire. Likewise, participants were required to have no major history of developmental disorder (e.g., Down syndrome and cerebral palsy) or neurobehavioral disorder (e.g., autism spectrum disorder and obsessive–compulsive disorder), per parent report. Because of high comorbidity of developmental speech and language disorder with attention-deficit/hyperactivity disorder (e.g., McGrath et al., 2008) and language-based learning disability (e.g., Hayiou-Thomas et al., 2017), these diagnoses were not treated as exclusionary. Full participant details are reported in Table 1.

Study Design

Single-case experimental studies are acknowledged as an important source of insight in clinical research for communication disorders (Byiers et al., 2012), where it is not always possible to conduct a well-powered group comparison due to factors such as the difficulty of recruiting clinical populations and the time-intensive nature of speech treatment. However, it can be challenging to compare

Table 1. Participant characteristics.

ID	Gender	Age at enrollment (years; months)	Length of previous treatment for rhotics	Former speech targets other than rhotics	CELF-5 Word Classes (scaled score)	CELF-5 Recalling Sentences (scaled score)	GFTA-3 (standard score, percentile)	Comorbid diagnoses
Y.C.	F	10;3	3.25 years	/ʃ, s/	10	9	40, < 0.1	Attention-deficit/hyperactivity disorder; receiving pragmatic language therapy
E.L.	M	10;2	6 years	/ʃ, tʃ/	15	17	52, 0.1	None
O.R.	F	9;0	None	None	11	11	40, < 0.1	None
R.T.	M	9;2	3 years	/l/ blends	9	9	42, < 0.1	Attention-deficit/hyperactivity disorder

Note. CELF-5 = Clinical Evaluation of Language Fundamentals–Fifth Edition; GFTA-3 = Goldman-Fristoe Test of Articulation–Third Edition; F = female; M= male.

multiple treatment conditions in a within-subject design because of the possibility of carryover effects (in which treatment in one condition influences performance in another condition) or order effects (in which participants' response differs depending on whether a particular treatment is introduced earlier or later in the course of the study). The single-case randomization design (Edgington & Onghena, 2007; Rvachew, 1988) has been proposed as an effective alternative to offset these challenges. In this design, each participant is exposed to two (or more) forms of treatment, with individual sessions randomly assigned to feature one type of treatment or the other. In addition to minimizing order effects, the randomization design makes it possible to use statistical hypothesis tests that are not applicable in the absence of randomization (Bulté & Onghena, 2008). In this study, individual treatment sessions were randomly assigned to feature either visual–acoustic biofeedback treatment or traditional treatment. Both treatment types were administered using the staRt app. The staRt app generated a randomized assignment for the type of intervention to be provided in each session before the implementation of the investigation. Randomization was constrained so that there was an equal number of sessions of each type, and no more than three sequential administrations of the same condition were implemented (Barlow & Hersen, 1984; Onghena & Edgington, 1994). Although the randomization design is considered a relatively good option for within-subject comparison of treatment methods, both carryover effects and order effects still represent threats to internal validity in this design, as we will discuss in more detail below.

Technology Setup

All evaluation and intervention sessions were completed via telepractice using a secure Zoom connection (Zoom Video Communications). The Zoom platform allowed the participant and investigator to interact through a webcam and speaker. A trial session was completed before the start of the study to familiarize the participant with the features of Zoom. Because the staRt app may be too resource intensive for easy use on an iPad with mirroring or screen-sharing technology, the investigators ran the app in the Xcode (Apple Inc.) iOS simulator on a MacBook Pro. The virtual audio mixer Soundflower (Ingalls, 2014) routed the participant's voice to the simulator to generate a visual–acoustic feedback display. The "Original sound" option was enabled in Zoom to minimize the influence of compression or echo suppression algorithms on audio transmission. Participants were required to use a wired connection to their home router to maximize bandwidth. Participants were provided with a standard set of headphones (Plantronics Blackwire C225, 20-Hz to 20-kHz stereo output range) for use in all study activities. Finally, while audio was shared over Zoom for treatment delivery, higher fidelity audio was desirable for the probes used to measure treatment progress. Therefore, for probes, participants and their parents were guided through the process of recording to their local device using Voice Memos on iOS and uploading the saved audio to the research team using a secure Box upload link. The built-in microphone in the Plantronics headset (100-Hz to 10-kHz frequency response) was used for this purpose.

Probes

Probe measures were administered to evaluate participants' progress in both long-term and short-term time frames over the course of the study. The Quiz function in the app, in which words are presented in random order with no prompts for feedback, was used to elicit all probes. The Long Word Quiz, used to assess changes in accuracy from the start to the end of the treatment period, was administered in three baseline sessions prior to the start of treatment, which meets What Works Clearinghouse (WWC) criteria for single-case experimental design "with reservations"; five probes per phase would be

required to meet WWC standards in full (Kratochwill & Levin, 2010). This probe consists of 50 words containing /ɹ/ in a balanced range of phonetic contexts. At the end of the study, the Long Word Quiz was readministered in three posttreatment sessions. In addition to these measures, a Short Word Quiz was administered at the start and end of each treatment session as a measure of short-term, within-session progress. The Short Word Quiz is a 25-word subset of the Long Word Quiz, with words randomly selected with stratification to preserve a balance of consonantal (i.e., onset) and vocalic (i.e., syllabic or postvocalic) variants of /ɹ/. This measure of within-session change is important for randomization tests comparing the incremental change associated with traditional and biofeedback treatment conditions. A stimulability probe modeled after Miccio (2002) was administered in baseline and maintenance sessions, but stimulability data were not evaluated as part of this study.

Treatment

Treatment in this study consisted of 16 individual sessions, with two to three sessions occurring per week. Prior to the start of structured treatment, two introductory sessions were conducted during the baseline phase. In the second baseline session, after administration of the Long Word Quiz, the clinician completed a script intended to familiarize the client with the articulatory requirements for producing perceptually accurate /ɹ/ sounds (similar to Preston et al., 2020). In the third baseline session, after probe administration, the clinician reviewed the staRt app tutorial with the participants, which introduced the visual–acoustic biofeedback display and familiarized the participant with the visual appearance of correct and incorrect /ɹ/. The clinician assessed each participant's comprehension of the visual–acoustic biofeedback display and provided 5 min of unstructured practice with the biofeedback display.

Each session began with a 10-min period of prepractice, a relatively unstructured interaction in which clinicians could provide various cues to try to elicit correct /ɹ/. This was followed by a period of structured practice eliciting /ɹ/ at the syllable or word level in 20 blocks of 10 trials (200 trials total). The determination to practice at the syllable or word level (Syllable Quest or Word Quest) was based on the treating clinician's ratings of the participant's accuracy on the Long Word Quiz, averaged across baseline sessions. Following McAllister et al. (2020), participants who exhibited at least 5% accuracy in words at baseline started at the word level, whereas participants with less than 5% accuracy started at the syllable level. The app allows the user to select the phonetic contexts in which /ɹ/ is targeted in a given session (categories: syllabic /ɹ/, prevocalic /ɹ/ with front vowels, prevocalic /ɹ/ with back vowels, postvocalic /ɹ/ with front vowels, and postvocalic /ɹ/ with back vowels). For this study, all five contexts were targeted in every session for all participants. The staRt app was used for stimulus presentation and response recording in both session types; the only difference was that the real-time LPC spectrum was visible in the biofeedback condition and suppressed in the traditional condition.[2] This allowed for all other parameters of treatment, such as the number of trials elicited, to be held constant.

During all sessions, the clinician wore Apple EarPods containing a built-in microphone. In each therapeutic exchange, the participant produced the syllable or word presented by the staRt app, and the clinician immediately scored each trial by tapping a button labeled with one of three possible ratings: *Great*, *Good*, or *Try again*; the meaning of these response choices was explained to the participant prior to the beginning of treatment. The clinician was instructed to assign the *Great* rating for adultlike rhotic productions, the *Good* rating for distortions with some rhotic quality, and the *Try again* rating for productions that were highly distorted or represented substitution of another sound. To increase participant motivation, the ratings were represented on screen as gold, silver, or bronze coins, and participants received messages indicating when they had achieved a new milestone such as a consecutive streak of *Great* ratings. The app encoded these ratings as numerical scores and used tallies of these scores to automatically adjust task difficulty in an adaptive fashion in the Word Quest mode only. Following principles from the challenge point framework (Guadagnoli & Lee, 2004; McAllister et al., 2021; Rvachew & Brosseau-Lapré, 2012), task difficulty was increased after a block in which the participant performed at least 80% correct, held constant after a block in which the participant scored between 50% and 80% correct, and decreased (unless already at the lowest level) after a block in which the participant performed below 50% correct. The within-session difficulty levels in staRt Word Quest mode are as follows: (1) one-syllable words, (2) two-syllable words, (3) one-syllable words with a competing phoneme (/l/ or /w/), (4) words from Levels 1–3 embedded in a standard carrier phrase ("Say [WORD] to me"), and (5) words from Levels 1–3 embedded in sentence frames randomly drawn from a set of 20 possibilities (e.g., "I made a label that said [WORD]"). In addition to these between-blocks modifications, participants could advance to higher complexity at the session level if they achieved at least 80% cumulative accuracy within a session. The between-sessions difficulty levels were (1) syllables presented in a blocked order (i.e., the same syllable presented in all 10 trials in a block), (2) syllables presented in a randomized order, (3) words presented in a blocked order, and (4) words presented in a randomized order.

[2]The staRt app has a "Show Wave" toggle switch in Quest mode that was used for this purpose.

At the end of each block, the app displayed a screen prompting the clinician to provide qualitative verbal feedback based on the participant's performance in the last 10 trials. In both biofeedback and traditional sessions, this KP feedback could pertain to either articulator placement or perceptual features of /ɹ/. In the biofeedback treatment condition only, qualitative feedback could also refer to the peaks in the LPC wave. In addition to this between-blocks feedback, the clinician provided KP feedback on two pseudorandomly selected trials within each block.

Measurement

The primary outcome variables used to assess treatment progress were the perceptually rated accuracy of /ɹ/ sounds in words on the Long Word Quiz measures administered in the baseline and maintenance phases, as well as the Short Word Quiz measures administered before and after each individual treatment session. Each quiz measure was scored in real time by the treating clinician,[3] but these unblinded ratings were judged to be subject to a potentially unacceptable degree of bias. Therefore, recordings of all Short and Long Word Quiz measures were rescored by blinded listeners via the Amazon Mechanical Turk (AMT) crowdsourcing platform. Crowdsourcing allows individuals from the general public to complete various types of online tasks, including rating speech productions (Nightingale et al., 2020; Sescleifer et al., 2018). Although speech ratings from untrained listeners using their own personal equipment may be subject to greater variability than trained listeners in a laboratory setting, validation studies have suggested that this variation can be offset by collecting a larger number of ratings than is customary for lab-based research. For instance, McAllister Byun et al. (2015) found that ratings aggregated across nine untrained listeners recruited through AMT were equivalent to ratings aggregated across three trained listeners.

Collection of ratings in this study was modeled after the suggestions from McAllister Byun et al. (2015). Target utterances by the participant in the audio recording of each probe measure were marked off in a TextGrid in Praat (Boersma & Weenink, 2019) and used to generate individual word-level files. These files were randomly ordered for presentation to listeners on AMT in blocks of 200 files. Each block was rated by nine unique listeners. Listeners rated the /ɹ/ sound in each word as correct or incorrect; they were provided with an orthographic representation of the target word but received no information about the identity of the speaker or the time point of elicitation. Participants on AMT were required to have IP addresses within the United States, report speaking English as their primary language, and report no history of speech or hearing difficulty. They were also required to pass an initial qualification task in which their ratings of 100 words containing /ɹ/ were compared against ratings assigned by trained listeners (McAllister Byun et al., 2015) as well as a listening test designed to confirm that the listener is wearing headphones (Woods et al., 2017). In addition, each block of files to be rated included 20 *catch trials* that were judged to be unequivocally correct or incorrect based on previous expert ratings. The blinded listeners needed to agree with the expert rater judgment on at least 80% of these tokens for their data to be included. A total of 23 raters contributed to the task, with each rater completing an average of 11.1 blocks (SD = 8.6, range: 1–24). Raters had a mean age of 43.7 years (SD = 11.5, range: 25–69).

Analyses

Two sets of analyses were carried out for this study. The first investigated the overall magnitude of participants' response to the combined treatment package without differentiating between traditional and biofeedback intervention types. This analysis drew on visual inspection of plots representing blinded listeners' ratings of tokens produced in baseline, intervention, and maintenance phases for each participant. Plots were generated in the R software environment (R Core Team, 2019) using the tidyverse family of packages (Wickham et al., 2019). In addition, effect sizes were calculated to quantify the magnitude of change from the pretreatment baseline to the postsession maintenance phase using Busk and Serlin's d_2 statistic (Beeson & Robey, 2006), which pools standard deviations across the baseline and maintenance phases to reduce instances where an effect size cannot be determined due to zero variance at baseline. Following Maas and Farinella (2012), this study used 1.0 as the minimum value to be accepted as clinically significant (i.e., the difference in pre- and posttreatment means must equal or exceed the pooled standard deviation). The second set of analyses addressed the relative efficacy of traditional versus biofeedback treatment. This involved computing within-session change for each session (difference in accuracy between the Short Word Quiz administered at the beginning vs. that administered at

[3]Because the clinician's scores were not used as an outcome measure, we judged the real-time scoring to be sufficient in most cases. The one exception to this generalization is the scoring of baseline Long Word Quizzes. Because scores on these probes were used to determine whether participants would start at the word or syllable level, it was judged to be important to rescore these probes from the high-fidelity audio instead of using the audio transmitted over Zoom. Scores differed by an average of 3 percentage points between the original and rescored probes. There were two instances (Y.C. and R.T.) where a participant who would have started at the word level based on the ratings of Zoom audio instead started at the syllable level based on the rescored probes. The two sets of scores are reported for comparison in Supplemental Material S1.

the end of the session) and comparing this value across treatment conditions. These values were used to compute the test statistic for a statistical comparison of conditions using the R package Single-Case Randomization Test (Bulté & Onghena, 2008).

Fidelity

Fidelity to the stated treatment protocol was evaluated in four of 16 treatment sessions per participant. Two sessions were selected from the first half of treatment and two from the second half, with one biofeedback and one traditional session in each half of treatment. Fidelity checks were conducted by the third author, who viewed the full video and audio record of each session and completed a checklist evaluating several parameters, including the number of trials elicited per block, whether and when verbal models were provided, and whether and when qualitative verbal feedback was provided. Qualitative information about the nature of verbal feedback was also noted as part of the fidelity check. One aspect of fidelity that was not included in the method adopted here was a second party's rescoring of the ratings given by the treating clinician during treatment. Because these within-session ratings influence participants' progress through the levels of the adaptive difficulty hierarchy, it would be desirable to include this aspect of fidelity in future studies.

The results of the fidelity check are as follows. The clinician provided qualitative feedback on two trials within the block for 97.8% of blocks. The target number of trials per block (10) was elicited in 83.4% of blocks. Blocks with a larger number of trials occasionally occurred if the participant rushed ahead of the clinician's prompting. For the 69/320 blocks that did not contain exactly 10 trials, 42 contained one extra trial, 14 contained two extra trials, and 13 contained three or more extra trials. No block elicited less than the target number of 10 trials. The clinician provided qualitative feedback after the block in 99.7% of blocks.

Results

Across-Sessions Change

Figure 2 provides a graphical depiction of blinded listeners' ratings of perceptual accuracy (percent correct) of Long Word Quiz measures elicited in all baseline and maintenance sessions as well as the Short Word Quiz measures administered immediately before and after each treatment session. Note that Figure 2 does not include a record of participants' performance on the 200 trials elicited during practice, which were rated only by the treating clinician and not by blinded listeners. The treating clinician's ratings are provided in Supplemental Material S1, along with a record of participants' progress through the adaptive difficulty hierarchy described previously. Figure 2 also does not provide any information regarding treatment condition; we return to condition-specific performance below. This figure does provide information about change in both long-term and short-term time frames: Long-term changes appear as trends in bar height over time, whereas short-term changes are reflected in the difference in height between the bars representing the presession and postsession probe for each session. Visual inspection of Figure 2 suggests that all participants increased their accuracy over the course of treatment, with the magnitude and timing of change varying across participants. This visual impression can be corroborated using the standardized effect sizes reported in Table 2. All four participants achieved effect sizes well above the threshold to be considered clinically significant (Maas & Farinella, 2012). Across participants, the average effect size was 27.64.

Within-Session Change

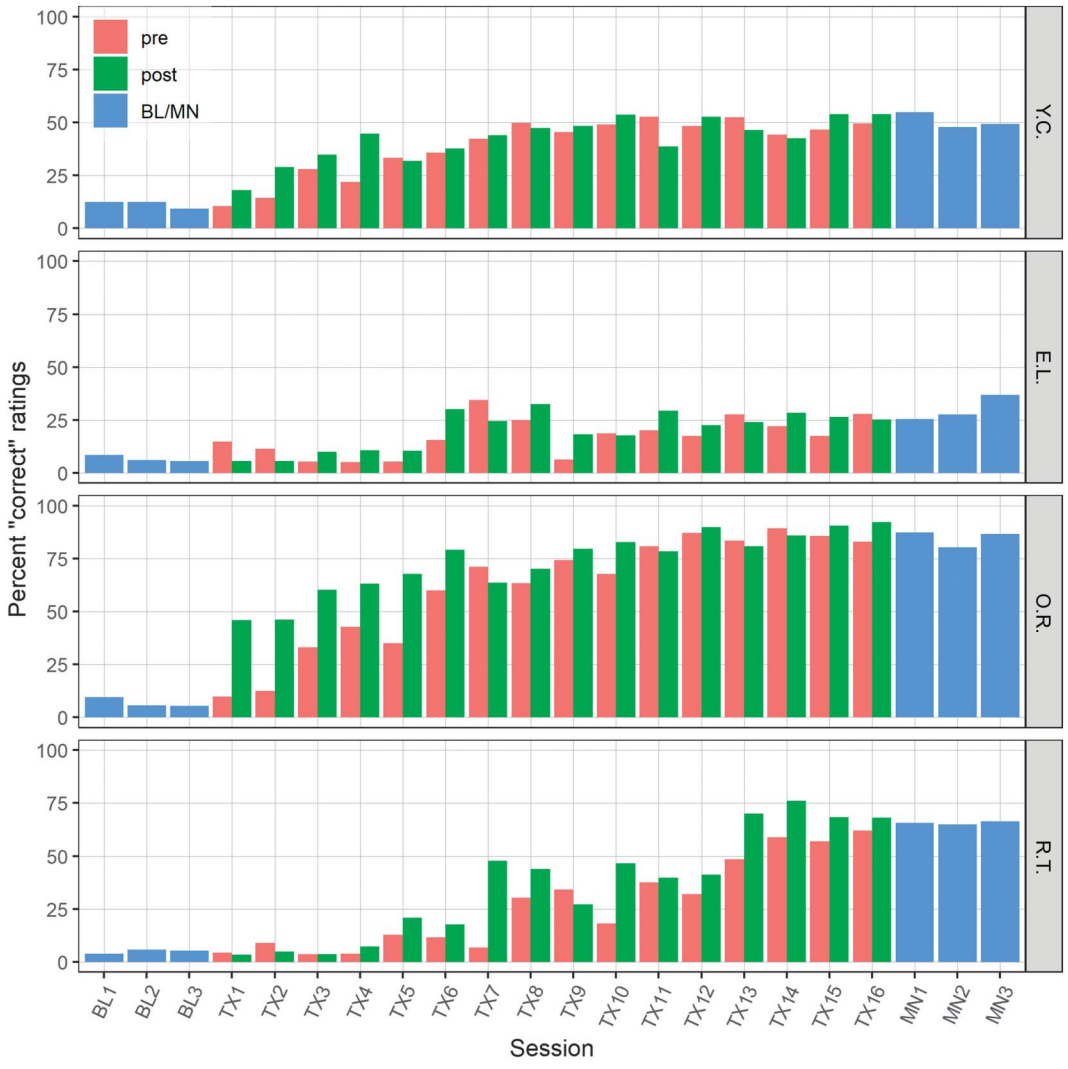
Figure 3 depicts within-session change over the course of the treatment period for each participant, with separate lines representing the biofeedback and traditional treatment conditions. In these plots, values above zero indicate improvement from the beginning to the end of each session (pre- to postsession probe). It is also possible to calculate the mean within-session change in each condition and ask whether it differs significantly across conditions. The results represented in Table 2 and Figures 2 and 3 are synthesized on a per-participant basis in the following section.

Qualitative Information on Treatment Response

Participant Y.C. began to respond to treatment almost immediately. She made her largest within-session gains in the first half of treatment, including a maximum improvement of 22.7 percentage points in Session 4, which featured traditional treatment. Her within-session gains leveled off in the second half of the study, but she maintained her earlier gains, resulting in a final d_2 of 13.27 (raw effect size of 39.24). She showed a similar magnitude of within-session change across traditional and biofeedback conditions, with a mean within-session change of 2.95 percentage points for biofeedback sessions and 3.69 percentage points for traditional sessions. The associated randomization test returned a p value of .86, which was not significant.

Participant E.L. showed minimal change in the first five treatment sessions, with a moderate increase in the sixth session (which represented the participant's fourth session featuring biofeedback). He continued to make small within-session gains for the rest of the treatment

Figure 2. Perceptually rated accuracy (percentage of "correct" ratings) in baseline (BL), treatment (TX; pre- and posttreatment probes), and maintenance (MN) sessions.



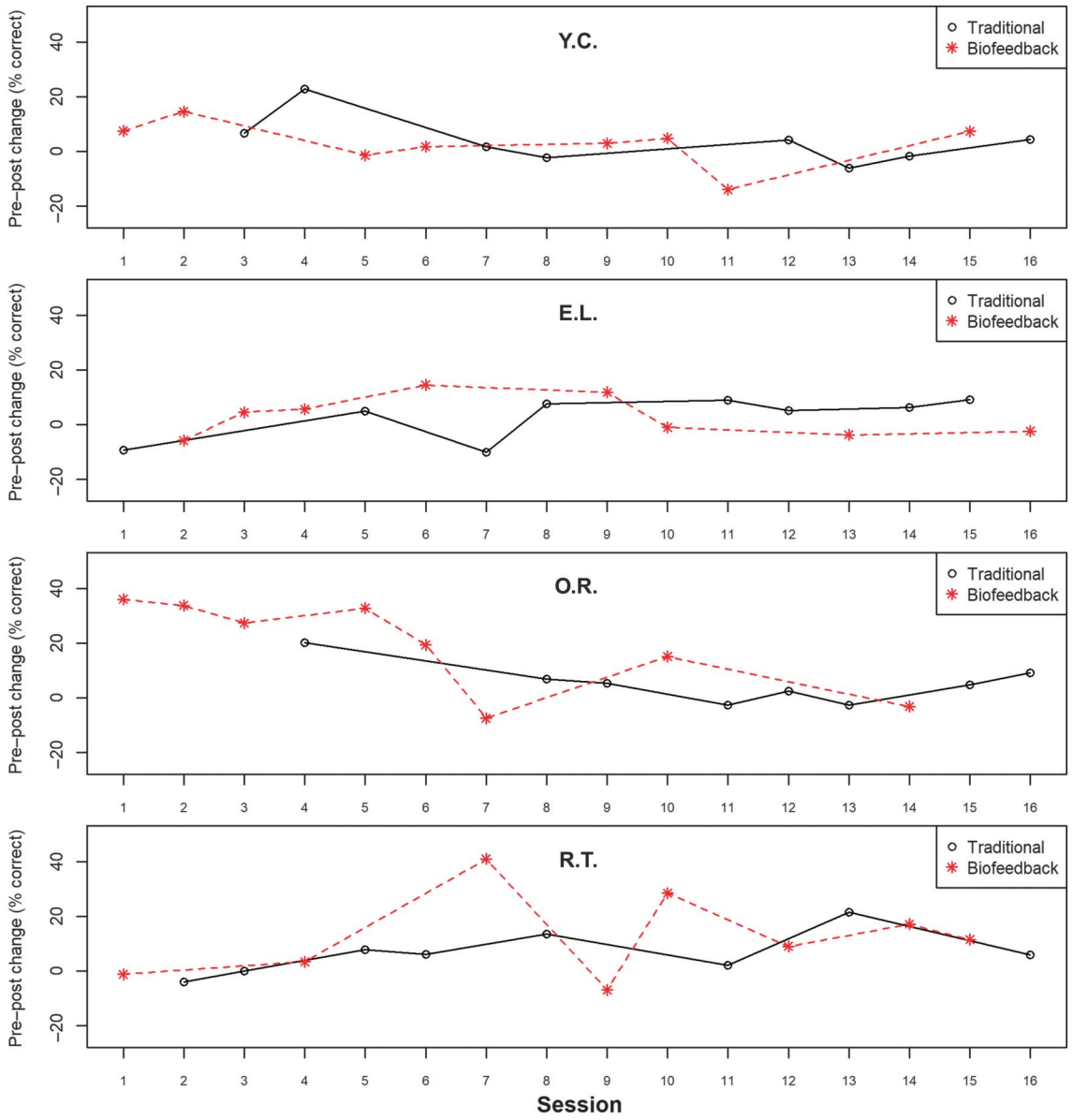
period, resulting in a final d_2 of 5.25 (raw effect size of 23.29). It was noted that in the first nine treatment sessions, the line representing within-session change in the biofeedback condition was consistently above the line representing within-session change in the traditional treatment condition (see Figure 3). After Session 9, this orientation flipped, and the traditional sessions produced larger gains than the biofeedback sessions. However, the magnitude of the difference between session types was minimal in both cases. The mean within-session change in percentage points was 2.95 for biofeedback sessions and 2.84 for traditional sessions. The associated randomization test returned a nonsignificant p value of .98.

Table 2. Means, standard deviations, and effect sizes reflecting blinded listeners' ratings (percentage of "correct" ratings) of participant performance before and after treatment.

Participant	Baseline mean (standard deviation)	Maintenance mean (standard deviation)	Raw effect size	d_2
Y.C.	11.47 (1.83)	50.71 (3.76)	39.24	13.27
E.L.	6.64 (1.55)	29.94 (6.08)	23.29	5.25
O.R.	6.83 (2.27)	84.74 (3.88)	77.91	24.48
R.T.	4.96 (1.03)	65.64 (0.75)	60.67	67.55

Participant O.R. had the second-largest d_2 among the four participants (24.48) and the largest raw effect size (39.24). She made her greatest gains in the first six sessions of treatment. Randomized assignment dictated that five of these first six sessions featured biofeedback. On average, her gains in biofeedback sessions were greater than her gains in traditional treatment sessions. The mean

Figure 3. Within-session change (pre- to postsession probe) in perceptually rated accuracy (percentage of "correct" ratings) in biofeedback and traditional treatment sessions.



within-session change in percentage points was 19.2 for biofeedback sessions and 5.45 for traditional sessions, and the associated randomization test returned a significant p value of .02. However, the fact that random assignment led to a clustering of biofeedback sessions in the first half of treatment poses an interpretive challenge because the investigators cannot determine if the observed differences can be attributed to treatment type, order, or a combination of both. This topic is further reviewed in the Discussion section.

Participant R.T. had the second-largest raw effect size among the four participants (60.67); because of low variability in baseline and maintenance probe scores, he had the largest d_2 (67.55). He showed minimal change in the first six sessions of treatment, followed by a considerable improvement (over 40 percentage points) in the seventh session, which featured biofeedback. R.T.'s gains were variable in subsequent sessions but were positive on average. While both of his largest within-session gains occurred on sessions in the biofeedback treatment condition, there is overlap between the lines representing traditional and biofeedback conditions in Figure 3. His mean within-session change in percentage points was 12.79 for biofeedback sessions and 6.6 for traditional sessions. The randomization test returned a nonsignificant p value of .35.

Discussion

Blinded listener ratings indicated that all participants in this investigation showed a strong response to the overall treatment package, which included biofeedback and traditional treatment. Effect sizes ranged from moderate (5.25) to very large (67.55), with a mean of 27.64. Visual inspection of plots supports the effect sizes in showing substantial gains in rhotic production accuracy over time. With respect to the difference in response during biofeedback and traditional sessions, two participants showed a negligible difference between conditions, whereas two participants showed a numerical advantage for biofeedback over traditional treatment. For one participant, O.R., the difference between biofeedback and traditional conditions returned a significant result in a randomization test ($p = .02$); however, this finding must be interpreted with caution because random assignment yielded a concentration of biofeedback in the early stages of treatment. The other participants showed no significant difference in within-session gains between biofeedback and traditional treatment conditions.

To the best of our knowledge, this is the first study investigating the efficacy of visual–acoustic biofeedback treatment for RSEs affecting /ɹ/ in the telepractice context. Although the small number of participants in this study limits the strength of the conclusions that can be drawn, the sizable gains made by all four participants in this study suggest that a treatment package combining traditional and visual–acoustic biofeedback treatment for residual /ɹ/ errors can be effective when delivered via telepractice.

Order of Biofeedback and Traditional Treatment Sessions

Participant O.R., a 9-year-old girl who had not received treatment for /ɹ/ prior to study enrollment, was the only individual in this investigation who exhibited a statistically significant difference in within-session progress between biofeedback and traditional treatment conditions. By random allocation, O.R. received biofeedback treatment in five of her first six sessions of treatment, the period when she made her greatest gains. As noted above, this creates ambiguity as to whether O.R. truly showed a stronger response to biofeedback than to traditional treatment or if she simply made the most progress in the first few sessions of treatment, with diminishing returns as her accuracy increased. A third possibility is that both the treatment condition and the time point in treatment were important for the large gains she made in the first six sessions. This is consistent with an idea that has been advanced in previous literature, namely, that biofeedback might have a more significant impact when provided early in the course of treatment. Before we briefly review this literature below, we note that data from participant E.L., who received biofeedback in four of his first six sessions, were also suggestive of an advantage for biofeedback in the early stages of treatment. E.L. showed a nonsignificant trend in which he made slightly larger gains in biofeedback than traditional sessions for the first half of treatment but then showed slightly greater gains in traditional than biofeedback sessions in the second half. On the other hand, participant Y.C. also received biofeedback in four of the first six sessions but did not show evidence of an advantage for biofeedback.

Previous literature on speech motor learning has proposed a theoretical basis for understanding how the order of biofeedback and traditional treatment conditions could influence learning outcomes. Because biofeedback provides detailed qualitative information about how to execute a motor plan,[4] it is considered most valuable in the initial stages of treatment, when the learner is still building an understanding of the motor target (Newell et al., 1990). Once a motor plan is internalized, ongoing KP feedback may actually inhibit generalization to novel targets (Ballard et al., 2012; Hodges & Franks, 2001; Maas et al., 2008). Therefore, prompt fading of KP feedback is recommended once a participant is familiar with the requirements for correct sound production (Maas et al., 2008; Newell et al., 1990). It may be optimal to provide biofeedback treatment in the initial acquisition phase, then use traditional treatment to stabilize the newly acquired motor plan and transfer it to additional contexts. Evidence in support of this theoretical framing was provided by McAllister Byun and Campbell (2016), who reported a significant interaction between treatment condition and order such that higher overall levels of accuracy were observed when biofeedback treatment was provided first, followed by traditional treatment. Preston et al. (2019) used a similar counterbalanced design to study the effects of ultrasound biofeedback relative to traditional treatment. Again, larger effect sizes were observed when biofeedback was provided in the first eight treatment sessions compared with the last eight treatment sessions. In sum, while it is not possible to fully disentangle the influence of order and treatment condition on the results in this investigation, both theoretical considerations and

[4]While this statement is unambiguously true for articulatory types of biofeedback such as ultrasound, there are differences of opinion as to whether visual–acoustic biofeedback also represents detailed qualitative information. The present authors view visual–acoustic biofeedback as providing information about how to execute a motor plan because we orient participants to the visual display and explain in what respect their production diverged from the target (e.g., "You need to move the third bump closer to the second bump," "Try to make the second bump stronger") and pair this feedback with verbal articulator placement cues. However, this represents an interesting point of distinction between articulatory and visual–acoustic types of biofeedback that should be investigated further. See Benway et al. (2021) for preliminary data directly comparing ultrasound and visual–acoustic biofeedback for RSEs affecting /ɹ/.

previous empirical research suggest that learning may be maximized when biofeedback is made available early in the course of treatment, as was the case for participant O.R.

Comparison With McAllister Byun (2017): Overall Treatment Response

The design of the present investigation is quite similar to the in-person investigation of biofeedback and traditional treatment for RSEs reported in McAllister Byun (2017). In both, participants received a combination of visual–acoustic biofeedback[5] and traditional treatment in a single-case randomization design, and outcomes were assessed using blinded listeners' perceptual ratings of /ɹ/ in pretreatment and posttreatment long word probes as well as presession and postsession short word probes. In the earlier investigation, participants completed 20 rather than 16 sessions, but the number of trials per session was lower than in the present investigation, as discussed below. The present investigation targeted both consonantal and vocalic variants of /ɹ/ for treatment; in contrast, McAllister Byun treated only vocalic /ɹ/. Both studies targeted children in the 9- to 15-year age range, with a slightly higher mean age for McAllister Byun than this study (12;3 vs. 9;6). Participants in the present investigation reported 0–6 years of previous treatment, whereas the McAllister Byun participants reported 0–11 years of treatment experience. Finally, both studies admitted participants with comorbid attention-deficit/hyperactivity disorder and/or language-based learning disability; two of four participants in the present investigation had such diagnoses, as did four of seven in McAllister Byun. Given the general similarities between the two investigations, here, we qualitatively compare and contrast their outcomes, although we recognize that this does not constitute a controlled comparison.

In the present investigation, response to the combined treatment package exceeded the minimum value to be considered clinically significant in all four participants, and the average standardized effect size (d_2) across participants was 27.64. In McAllister Byun (2017), response to the combined treatment package exceeded the minimum threshold to be considered clinically significant in four of seven participants. However, two participants demonstrated no gains in /ɹ/, and one showed a significant degree of change in the negative direction. The largest standardized effect size was 9.87, and the average d_2 across participants was 1.78.

In total, although the experimental design and participant histories were broadly comparable across the present investigation and McAllister Byun (2017), participant outcomes differed considerably across the two investigations. The present investigation yielded higher minimum, maximum, and average values for treatment response. This finding was unexpected, because we tentatively posited that technical challenges could reduce the magnitude of treatment gains in the telepractice context relative to in-person delivery. Here, we reflect further on factors that could account for the observed differences between the investigations.

The most plausible explanation for the discrepancy in outcomes is the existence of a difference in treatment dosage between the two investigations. Increased opportunities for practice are associated with positive gains in motor performance (Schmidt & Lee, 2005), and previous research has shown that outcomes for children with SSD are more favorable when treatment is delivered with a high cumulative intervention intensity. The evidence base for this generalization includes a systematic review of various interventions (Kaipa & Peterson, 2016) and a systematic review specifically focused on biofeedback treatment for RSEs (Hitchcock et al., 2019). Participants in McAllister Byun (2017) received 20 treatment sessions that were 30 min in duration, with each session eliciting 5 min of prepractice, followed by 60 trials in five blocks of 12 trials. In the current investigation, participants completed 16 sessions with a duration of 45–50 min, with each session featuring 10 min of prepractice, followed by 200 trials in 20 blocks of 10 trials. Thus, the cumulative dose of structured practice was 3,200 trials in the present investigation versus 1,200 trials in McAllister Byun. The actual difference in number of production attempts per session may be even larger because this study also featured a longer duration of prepractice. Thus, the higher cumulative dosage provided in this study provides the most likely explanation for the greater effect sizes observed relative to McAllister Byun.

As noted above, we hypothesized that technical challenges such as disrupted connectivity might result in reduced effect sizes in the telepractice setting relative to in-person service delivery. In practice, we observed few overt disruptions in the streaming signal; while we did not systematically track the occurrence of such issues, we estimate that we experienced something on the order of one to two minor interruptions per session. If a participant's production was cut off, the client would prompt a repetition after the interruption resolved. A more pervasive issue was the lag between participants' production of a word and the response of the staRt app display in the biofeedback treatment condition. While we were not able to quantify this lag, some

[5]In both studies, visual–acoustic biofeedback involved a real-time LPC spectrum representing formant frequencies; however, in McAllister Byun (2017), visual–acoustic biofeedback was delivered using the KayPENTAX Sona-Match software, whereas this study used the staRt app. In light of the other differences between the two studies discussed here, however, we will not attempt to speculate on any differences in performance between these two software options.

degree of temporal asynchrony is nearly always present. All participants were able to grasp the concept of how to interact with the biofeedback display despite this lag. However, we cannot rule out the possibility that the delayed nature of the response could reduce the efficacy of biofeedback delivered via telepractice, as we discuss in more detail below.

Comparison With McAllister Byun (2017): Difference Between Treatment Conditions

Participants in the present investigation varied in their relative degree of progress in biofeedback and traditional treatment conditions. Two participants (E.L. and Y.C.) showed no evidence of a difference between treatment types. R.T.'s average within-session gains were numerically higher in biofeedback than traditional treatment sessions, but this difference was not significant. Only O.R. showed a significant difference between treatment conditions, but the interpretation of this result is confounded by an order effect. By contrast, three of seven participants in the McAllister Byun (2017) single-case randomization investigation showed a significant advantage for biofeedback over traditional treatment. Thus, although the present investigation produced larger overall effect sizes, it yielded less evidence of a difference between biofeedback and traditional treatment conditions compared with the previous lab-based investigation. Many factors could account for this contrast between the investigations, including simple sampling error (i.e., the outcomes could change if different participants were randomly sampled from the population of children with RSEs). However, there is one possible explanation that accords with a hypothesis articulated earlier, namely, that delivery via telepractice does reduce the efficacy of biofeedback treatment, most likely because the biofeedback display is slower to respond and less sensitive to fine-grained differences when shared over videoconferencing software. Future research should include direct comparisons of visual–acoustic biofeedback treatment in in-person and telepractice settings in order to better understand potential differences in efficacy across service delivery contexts.

Limitations and Future Directions

The fact that no firm conclusions could be drawn from the comparison of traditional and visual–acoustic biofeedback treatment conditions represents a clear limitation of this study. One simple fix for the ambiguous outcome observed with participant O.R., in which the biofeedback treatment condition was overrepresented in the early stages of treatment, would be to place a lower cap on the number of consecutive sessions of a given type (e.g., no more than two consecutive sessions of the same type). However, McAllister Byun (2017) suggested that rapid alternation between treatment types may not be optimal for measuring differences between biofeedback and traditional treatment, because the short time elapsed between sessions could allow carryover between conditions. That article suggested that scheduling in which a single treatment is delivered over three or more consecutive sessions could allow for the effect of a single treatment to build up over time, which could make it possible to observe distinctions between interventions. However, such a design is also susceptible to order effects, as observed for O.R. in this study. An alternative possibility would be to use a single-case experimental design in which different treatment approaches are associated with different variants of /ɹ/ within individuals (e.g., consonantal variants are treated with traditional methods and vocalic variants with biofeedback). Unfortunately, comparisons across /ɹ/ variants are complicated by the fact that children often differ in accuracy across variants at baseline (e.g., Curtis & Hardy, 1959) and the fact that listeners show different rating behavior across variants (e.g., Klein et al., 2012). Ultimately, it appears that between-groups comparisons will be needed to provide strong evidence regarding the relative efficacy of treatments for RSEs.

An additional limitation of the present investigation is that the technological setup used here to deliver staRt app treatment via telepractice is not readily accessible to the average clinical SLP. The app's memory requirements make it difficult (but not impossible) to run on an iPad alongside mirroring or screen-sharing technology. The solution used here of running the app in the Xcode simulator software is freely available to anyone who owns a MacBook laptop (see instructions at https://osf.io/vxegp/), but setting up the Xcode simulator may require the assistance of a trained programmer. The staRt app development team is currently working to build a browser-based, platform-independent version of staRt that could be readily shared over video calls for use in telepractice. This version will also increase access to the staRt app for SLPs who do not have an iPad for use in their clinical practice. While these changes are still underway, clinicians may wish to note that participants in this study made gains in the traditional as well as the visual–acoustic biofeedback condition. Resources for providing traditional treatment for /ɹ/, developed for other studies by the authorial team and collaborators (McAllister et al., 2020), can be accessed at https://osf.io/pg8sw/ (see also Preston et al., 2020).

Future directions of investigation may further advance the knowledge base on visual–acoustic biofeedback and traditional treatment to remediate RSEs. Implementing this investigation in person with the staRt app is recommended to allow a qualitative comparison of results obtained through face-to-face intervention versus telepractice. Another alteration for future investigation involves manipulating the scheduling of sessions (intensive vs. distributed).

In addition to the difference in dosage described above, the present investigation differed from McAllister Byun (2017) in featuring a slightly more intensive session schedule (two to three sessions per week instead of two sessions per week). Hitchcock et al. (2019) suggested that participants who receive biofeedback treatment within a narrow time frame (i.e., high-intensity scheduling) may benefit more than individuals who receive the same dosage over a longer period. In a case series by Preston and Leece (2017), participants with RSEs affecting /ɹ/ received 14 hr of treatment, including ultrasound visual feedback, within 1 week. All four participants in that study exhibited statistically significant increases in rhotic production accuracy despite the short total duration of the intervention. It was reasoned that decreasing the time elapsed between sessions can lower participants' likelihood of reverting to incorrect motor production patterns for rhotics. Given these previous findings, future research could examine outcomes in a study with more intensive scheduling.

In the present investigation, all participants advanced to using rhotics in sentences through the adaptive difficulty framework described above (see Supplemental Material S1 for information about each participant's progress through the levels of the hierarchy). However, the generalization of /ɹ/ to sentences was not measured in the present investigation, partly because a sentence-level probe does not exist on the staRt app. A sentence probe would be a useful addition to the staRt app to assess the generalization of /ɹ/ in future investigations. Additional investigation, including longer term follow-up assessments (e.g., 1 month and 6 months posttreatment) and assessment of carryover of correct production to spontaneous conversation, will be important to fully understand the functional impact of the treatment investigated here. If limited generalization is observed, it might be beneficial to manipulate the treatment experience in ways intended to encourage increased self-monitoring, such as delaying knowledge of results feedback or asking participants to rate their own productions on some trials.

Conclusions and Clinical Implications

Practicing SLPs show considerable enthusiasm for mobile apps to enhance the delivery of treatment for SSD, but the evidence base supporting app-based interventions remains limited. This pilot study is part of a larger project to develop and evaluate the staRt app, which provides visual–acoustic biofeedback for residual errors affecting North American English /ɹ/. The first research question assessed effect sizes measuring the change in accuracy from the pretreatment baseline to the posttreatment maintenance phase, which represented participants' response to the overall treatment package delivered via telepractice. All four participants achieved effect sizes considered clinically significant, with a mean of 27.6, indicating strong overall positive response.

The second question used measures of short-term change to compare biofeedback versus traditional treatment conditions. The primary outcome variable was the change in percentage of perceptually correct /ɹ/ sounds on the Short Word Quiz administered before and after each treatment session. One participant demonstrated significantly larger gains in biofeedback than traditional treatment, but this result was confounded because her biofeedback sessions were concentrated at the start of the treatment period. No other participant showed a significant difference between visual–acoustic biofeedback and traditional treatment. Although the present investigation produced larger overall effect sizes than similar research in the laboratory setting, the difference between biofeedback and traditional treatment conditions was smaller than in previous research. One possibility is that delivery via telepractice may diminish the efficacy of biofeedback, which resulted in participants not showing an advantage for this treatment condition. A future investigation of staRt in an in-person setting is an important next step to determine the efficacy of biofeedback provided with the staRt app relative to traditional treatment for RSEs affecting /ɹ/.

The results of the present investigation, taken in combination with previous research, make some clear recommendations for clinical practice. First, the greater overall gains observed in this study relative to McAllister Byun (2017) speak to the importance of achieving as high a dose as possible when treating RSEs. Following Sugden et al. (2018), we recommend targeting a minimum of 100 trials per session. This study illustrates that, in the hands of an experienced clinician, 200 trials per hour can be achieved even in the telepractice context, which can be challenging from the standpoint of behavior management. Second, the present findings are consistent with previous studies (e.g., McAllister Byun & Campbell, 2016; Preston et al., 2019) in suggesting that biofeedback training may be most effective when presented early in the course of treatment and followed by traditional treatment to consolidate and generalize the learning acquired through biofeedback. Finally, these results provide a novel demonstration that both biofeedback and traditional articulatory treatment for RSEs affecting /ɹ/ can be effectively delivered over telepractice, suggesting that clinicians who work via telepractice can feel confident in including this population in their caseload.

Acknowledgments

Work on this project was partly funded by National Institute on Deafness and Other Communication Disorders Grants R41DC016778 (PI: Tara McAllister)

and R01DC017476 (PI: Tara McAllister) and by an ASHFoundation Clinical Research Grant awarded to Tara McAllister. The authors gratefully acknowledge Shari Sokol, Brandon McDevitt, Leah Black, Maria Hase, Kyung Hae Hwang, Martha Alman, Alison Trumpore, Thea Nihiser, and Amanda Spinogatti for their vital contributions as clinical partners in preliminary research leading up to this study. The authors also thank the members of the staRt app development team, including Tae Hong Park, Mario Svirsky, Will Haack, Sam Tarakajian, Helen Carey, and Jonathan Chin, and members of the Biofeedback Intervention Technology for Speech Lab who contributed to this project, especially Samantha Ayala, Samantha Beames, Kristina Doyle, Tabitha McCloud-James, Jelani White, and Graham Feeny. Finally, the second author extends his thanks to his thesis advisors, Julia Sturm and Diti Bhadra.

References

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

Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). Pergamon Press.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review, 16*(4), 161–169. https://doi.org/10.1007/s11065-006-9013-7

Benway, N. R., Hitchcock, E. R., McAllister, T., Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual /ɹ/ errors in American English: A single-case randomization design. *American Journal of Speech-Language Pathology, 30*(4), 1819–1845. https://doi.org/10.1044/2021_AJSLP-20-00216

Bernhardt, B., Gick, B., Bacsfalvi, P., & Adler-Bock, M. (2005). Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics & Phonetics, 19*(6–7), 605–617. https://doi.org/10.1080/02699200500114028

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer* (Version 6.1.06) [Computer software].

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

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*(2), 467–478. https://doi.org/10.3758/BRM.40.2.467

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. https://doi.org/10.1044/1058-0360(2012/11-0036)

Chen, Y.-P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A., Doube, W., & Morris, M. E. (2016). Systematic review of virtual speech therapists for speech disorders. *Computer Speech & Language, 37,* 98–128. https://doi.org/10.1016/j.csl.2015.08.005

Coufal, K., Parham, D., Jakubowitz, M., Howell, C., & Reyes, J. (2018). Comparing traditional service delivery and telepractice for speech sound production using a functional outcome measure. *American Journal of Speech-Language Pathology, 27*(1), 82–90. https://doi.org/10.1044/2017_AJSLP-16-0070

Crowe Hall, B. J. (1991). Attitudes of fourth and sixth graders toward peers with mild articulation disorders. *Language, Speech, and Hearing Services in Schools, 22*(1), 334–340. https://doi.org/10.1044/0161-1461.2201.334

Crowe, K., & McLeod, S. (2020). Children's English consonant acquisition in the United States: A review. *American Journal of Speech-Language Pathology, 29*(4), 2155–2169. https://doi.org/10.1044/2020_AJSLP-19-00168

Culton, G. L. (1986). Speech disorders among college freshmen. *Journal of Speech and Hearing Disorders, 51*(1), 3–7. https://doi.org/10.1044/jshd.5101.03

Curtis, J. F., & Hardy, J. C. (1959). A phonetic study of misarticulation of /r/. *Journal of Speech and Hearing Research, 2*(3), 244–257. https://doi.org/10.1044/jshr.0203.244

Davis, S. M., & Drichta, C. E. (1980). Biofeedback: Theory and application to speech pathology. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (pp. 283–286). Elsevier. https://doi.org/10.1016/B978-0-12-608603-4.50015-9

Edgar, D. L., & Rosa-Lugo, L. I. (2007). The critical shortage of speech-language pathologists in the public school setting: Features of the work environment that affect recruitment and retention. *Language, Speech, and Hearing Services in Schools, 38*(1), 31–46. https://doi.org/10.1044/0161-1461(2007/004)

Edgington, E., & Onghena, P. (2007). *Randomization tests*. CRC Press. https://doi.org/10.1201/9781420011814

Edwards, J., & Dukhovny, E. (2017). Technology training in speech-language pathology: A focus on tablets and apps. *Perspectives of the ASHA Special Interest Groups, 2*(10), 33–48. https://doi.org/10.1044/persp2.SIG10.33

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. https://doi.org/10.1121/1.429469

Furlong, L., Morris, M., Serry, T., & Erickson, S. (2018). Mobile apps for treatment of speech disorders in children: An evidence-based analysis of quality and efficacy. *PLOS ONE, 13*(8), Article e0201513. https://doi.org/10.1371/journal.pone.0201513

Gibbon, F. E., & Paterson, L. (2006). A survey of speech and language therapists' views on electropalatography therapy outcomes in Scotland. *Child Language Teaching and Therapy, 22*(3), 275–292. https://doi.org/10.1191/0265659006ct308xx

Goldman, R., & Fristoe, M. (2015). *Goldman-Fristoe Test of Articulation* (3rd ed.). Pearson.

Grogan-Johnson, S., Gabel, R. M., Taylor, J., Rowan, L. E., Alvares, R., & Schenker, J. (2011). A pilot exploration of speech sound disorder intervention delivered by telehealth to school-age children. *International Journal of Telerehabilitation, 3*(1), 31–42. https://doi.org/10.5195/ijt.2011.6064

Grogan-Johnson, S., Schmidt, A. M., Schenker, J., Alvares, R., Rowan, L. E., & Taylor, J. (2013). A comparison of speech sound intervention delivered by telepractice and side-by-side service delivery models. *Communication Disorders Quarterly, 34*(4), 210–220. https://doi.org/10.1177/1525740113484965

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

Hayiou-Thomas, M. E., Carroll, J. M., Leavett, R., Hulme, C., & Snowling, M. J. (2017). When does speech sound disorder

matter for literacy? The role of disordered speech errors, co-occurring language impairment and family risk of dyslexia. *The Journal of Child Psychology and Psychiatry, 58*(2), 197–205. https://doi.org/10.1111/jcpp.12648

Hitchcock, E. R., Harel, D., & McAllister Byun, T. (2015). Social, emotional, and academic impact of residual speech errors in school-aged children: A survey study. *Seminars in Speech and Language, 36*(4), 283–294. https://doi.org/10.1055/s-0035-1562911

Hitchcock, E. R., McAllister Byun, T., Swartz, M. T., & Lazarus, R. (2017). Efficacy of electropalatography for treating misarticulation of /r/. *American Journal of Speech-Language Pathology, 26*(4), 1141–1158. https://doi.org/10.1044/2017_AJSLP-16-0122

Hitchcock, E. R., Swartz, M. T., & Lopez, M. (2019). Speech sound disorder and visual biofeedback intervention: A preliminary investigation of treatment intensity. *Seminars in Speech and Language, 40*(2), 124–137. https://doi.org/10.1055/s-0039-1677763

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. https://doi.org/10.1080/02701367.2001.10608943

Ingalls, M. (2014). *Soundflower*. https://github.com/mattingalls/Soundflower/releases/tag/2.0b2

Kaipa, R., & Peterson, A. M. (2016). A systematic review of treatment intensity in speech disorders. *International Journal of Speech-Language Pathology, 18*(6), 507–520. https://doi.org/10.3109/17549507.2015.1126640

Kenny, B., & Lincoln, M. (2012). Sport, scales, or war? Metaphors speech-language pathologists use to describe caseload management. *International Journal of Speech-Language Pathology, 14*(3), 247–259. https://doi.org/10.3109/17549507.2012.651747

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. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*(2), 124–144. https://doi.org/10.1037/a0017736

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*(3), 1455–1468. https://doi.org/10.1121/1.426686

Lehiste, I. (1964). *Acoustical characteristics of selected English consonants*. Indiana University.

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., Hula, S. N. A., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology, 17*(3), 277–298. https://doi.org/10.1044/1058-0360(2008/025)

McAllister, T., Hitchcock, E. R., & Ortiz, J. A. (2021). Computer-assisted challenge point intervention for residual speech errors. *Perspectives of the ASHA Special Interest Groups, 6*(1), 214–229. https://doi.org/10.1044/2020_PERSP-20-00191

McAllister, T., Preston, J. L., Hitchcock, E. R., & Hill, J. (2020). Protocol for Correcting Residual Errors with Spectral, ULtrasound, Traditional Speech therapy Randomized Controlled Trial (C-RESULTS RCT). *BMC Pediatrics, 20*(1), Article 66. https://doi.org/10.1186/s12887-020-1941-5

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 visuall–acoustic biofeedback intervention for residual speech errors. *Frontiers in Human Neuroscience, 10*(567), 1–17. https://doi.org/10.3389/fnhum.2016.00567

McAllister Byun, T., Campbell, H., Carey, H., Liang, W., Park, T. H., & Svirsky, M. (2017). Enhancing intervention for residual rhotic errors via app-delivered biofeedback: A case study. *Journal of Speech, Language, and Hearing Research, 60*(6S), 1810–1817. https://doi.org/10.1044/2017_JSLHR-S-16-0248

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

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*(4), 384–401. https://doi.org/10.1111/1460-6984.12215

McGrath, L. M., Hutaff-Lee, C., Scott, A., Boada, R., Shriberg, L. D., & Pennington, B. F. (2008). Children with comorbid speech sound disorder and specific language impairment are at increased risk for attention-deficit/hyperactivity disorder. *Journal of Abnormal Child Psychology, 36*(2), 151–163. https://doi.org/10.1007/s10802-007-9166-8

McKechnie, J. (2019). *Exploring the use of technology for assessment and intensive treatment of childhood apraxia of speech* [Doctoral dissertation, The University of Sydney]. Sydney Digital Theses. http://hdl.handle.net/2123/20722

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech-Language Pathology, 20*(6), 583–598. https://doi.org/10.1080/17549507.2018.1477991

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

Mielke, J., Baker, A., & Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language, 92*(1), 101–140. https://doi.org/10.1353/lan.2016.0019

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

Nightingale, C., Swartz, M., Ramig, L. O., & McAllister, T. (2020). Using crowdsourced listeners' ratings to measure speech changes in hypokinetic dysarthria: A proof-of-concept study. *American Journal of Speech-Language Pathology, 29*(2), 873–882. https://doi.org/10.1044/2019_AJSLP-19-00162

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy, 32*(7), 783–786. https://doi.org/10.1016/0005-7967(94)90036-1

Preston, J. L., Benway, N. R., Leece, M. C., Hitchcock, E. R., & McAllister, T. (2020). Tutorial: Motor-based treatment strategies

for /r/ distortions. *Language, Speech, and Hearing Services in Schools, 51*(4), 966–980. https://doi.org/10.1044/2020_LSHSS-20-00012

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., McAllister, T., Phillips, E., Boyce, S., Tiede, M., Kim, J. S., & Whalen, D. H. (2019). Remediating residual rhotic errors with traditional and ultrasound-enhanced treatment: A single-case experimental study. *American Journal of Speech-Language Pathology, 28*(3), 1167–1183. https://doi.org/10.1044/2019_AJSLP-18-0261

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ruscello, D. M. (1995). Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders, 28*(4), 279–302. https://doi.org/10.1016/0021-9924(95)00058-X

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*. Plural.

Schmidt, R. A., & Lee, T. D. (2005). *Motor control and learning: A behavioral emphasis*. Human Kinetics.

Sescleifer, A. M., Francoisse, C. A., & Lin, A. Y. (2018). Systematic review: Online crowdsourcing to assess perceptual speech outcomes. *The Journal of Surgical Research, 232,* 351–364. https://doi.org/10.1016/j.jss.2018.06.032

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., Gruber, F. A., & Kwiatkowski, J. (1994). Developmental phonological disorders III: Long-term speech-sound normalization. *Journal of Speech and Hearing Research, 37*(5), 1151–1177. https://doi.org/10.1044/jshr.3705.1151

Shuster, L. I., Ruscello, D. M., & Smith, K. D. (1992). Evoking [r] using visual feedback. *American Journal of Speech-Language Pathology, 1*(3), 29–34. https://doi.org/10.1044/1058-0360.0103.29

Shuster, L. I., Ruscello, D. M., & Toth, A. R. (1995). The use of visual feedback to elicit correct /r/. *American Journal of Speech-Language Pathology, 4*(2), 37–44. https://doi.org/10.1044/1058-0360.0402.37

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. (1990). The Iowa Articulation Norms Project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55*(4), 779–798. https://doi.org/10.1044/jshd.5504.779

Sugden, E., Baker, E., Munro, N., Williams, A. L., & Trivette, C. M. (2018). Service delivery and intervention intensity for phonology-based speech sound disorders. *International Journal of Language & Communication Disorders, 53*(4), 718–734. https://doi.org/10.1111/1460-6984.12399

Sugden, E., Lloyd, S., Lam, J., & Cleland, J. (2019). Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders. *International Journal of Language & Communication Disorders, 54*(5), 705–728. https://doi.org/10.1111/1460-6984.12478

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

Volin, R. A. (1998). A relationship between stimulability and the efficacy of visual biofeedback in the training of a respiratory control task. *American Journal of Speech-Language Pathology, 7*(1), 81–90. https://doi.org/10.1044/1058-0360.0701.81

Wales, D., Skinner, L., & Hayman, M. (2017). The efficacy of telehealth-delivered speech and language intervention for primary school-age children: A systematic review. *International Journal of Telerehabilitation, 9*(1), 55–70. https://doi.org/10.5195/ijt.2017.6219

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wiig, E., Semel, E., & Secord, W. (2013). *Clinical Evaluation of Language Fundamentals–Fifth Edition (CELF-5)*. Pearson.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2