PERSPECTIVES SIG 19

Research Article

Computer-Assisted Challenge Point Intervention for Residual Speech Errors

Tara McAllister,[a] Elaine R. Hitchcock,[b] and José A. Ortiz[c]

Purpose: This preliminary case series investigated the effects of biofeedback intervention for residual rhotic errors delivered within a modified challenge point framework. In the challenge point framework, practice difficulty is adaptively adjusted with the goal of enhancing generalization learning. This study more specifically evaluated the feasibility of a computer-mediated implementation of challenge point treatment for rhotic errors using a custom open-source software, the Challenge Point Program.

Method: Participants were five native English speakers, ages 7;3–15;5 (years;months), who had established but not generalized correct rhotic production in previous treatment; overall treatment duration was flexible. Treatment incorporated either electropalatographic or visual-acoustic biofeedback and was structured by challenge point principles implemented using the Challenge Point Program software.

Results: Participants were highly variable in the magnitude of generalization gains attained. However, the median overall effect size was 4.24, suggesting that participants' response in treatment tended to exceed the minimum value considered clinically significant.

Conclusion: These findings provide preliminary evidence that computer-mediated implementation of the challenge point framework can be effective in producing generalization in some participants.

Supplemental Material: https://doi.org/10.23641/asha.13244501

The term *residual speech errors* (RSE) is used to refer to a subtype of speech sound disorder in which atypical speech patterns, typically distortions of late-developing sounds such as /ɹ/, /s/, or /l/, persist beyond 8–9 years of age. Although the impact of RSEs on overall intelligibility may be relatively minor, they can influence peer judgments (Crowe Hall, 1991) and negatively impact participation in academic and social–emotional contexts (Hitchcock et al., 2015). These errors are considered particularly challenging to remediate and often persist despite years of intervention, leading the clinical community to call on the research community to identify more effective treatment methods (Ruscello, 1995).

Much recent research on RSEs has investigated outcomes when real-time visual biofeedback is incorporated as a tool to understand and treat speech errors in children with RSE. Over the past decade, a number of Phase I clinical case studies have made the case that RSEs may be successfully eliminated when intervention is enhanced with some form of visual biofeedback (e.g., Adler-Bock et al., 2007; Shuster et al., 1992, 1995). More recently, higher quality evidence for the efficacy of biofeedback for residual errors has been derived from Phase II, single-subject experimental studies (e.g., Hitchcock et al., 2017; McAllister Byun & Campbell, 2016; McAllister Byun & Hitchcock, 2012; Preston et al., 2019).

Visual biofeedback can be delivered with various technologies, including ultrasound, in which an ultrasound probe held beneath the chin generates an image of the client's tongue during speech (e.g., McAllister Byun et al., 2014, Preston et al., 2013, 2019); electropalatography, which uses a pseudopalate to register and display areas of contact between the client's tongue and palate (e.g., Fletcher et al., 1991; Hitchcock et al., 2017); and electromagnetic articulography, where tracers affixed to the tongue are tracked in a magnetic field and displayed in three-dimensional space, potentially in connection with an animated tongue avatar that moves in real

[a]Department of Communicative Sciences and Disorders, New York University, NY
[b]Department of Communication Sciences and Disorders, Montclair State University, Bloomfield, NJ
[c]Department of Hearing and Speech Sciences, University of Maryland, College Park

Correspondence to Tara McAllister: tkm214@nyu.edu

Editor-in-Chief: Angela H. Ciccia
Editor: Alison Behrman

Received August 5, 2020
Revision received October 1, 2020
Accepted October 10, 2020
https://doi.org/10.1044/2020_PERSP-20-00191

Disclosures

Financial: Tara McAllister has no relevant financial interests to disclose. Elaine R. Hitchcock has no relevant financial interests to disclose. José A. Ortiz has no relevant financial interests to disclose.

Nonfinancial: Tara McAllister has no relevant nonfinancial interests to disclose. Elaine R. Hitchcock has no relevant nonfinancial interests to disclose. José A. Ortiz has no relevant nonfinancial interests to disclose.

time with the client's tongue (e.g., Katz et al., 1999). A final type of biofeedback involves a real-time visual display of the acoustic signal of speech, such as a real-time Linear Predictive Coding spectrum or spectrogram (e.g., Shuster et al., 1992; McAllister Byun, 2017; McAllister Byun & Hitchcock, 2012). Despite the diverse technologies involved, biofeedback intervention approaches are unified in that they all provide an *external visual display* as a new source of insight into the speech process. Using instrumentation to visualize articulatory or acoustic aspects of speech production in real time is intended to help the speaker exert more conscious control over a process that is typically largely unconscious (Gibbon & Paterson, 2006; Maas et al., 2008). Lastly, biofeedback intervention enables comparison of a speaker's production with a visual target, which can encourage self-monitoring.

Participants in the present case series were drawn from two larger studies of biofeedback treatment efficacy conducted by the first and second authors (Hitchcock et al., 2017; McAllister Byun, 2017). All participants exhibited rhotic misarticulation, one of the most common speech errors in American English. Participants were heterogeneous in other regards as discussed below. Although the remainder of this article will focus on the specific context of treatment for rhotic misarticulation, the challenge point principles under discussion here could also be applied equally to other speech sounds.

The North American English rhotic is one of the latest emerging sounds in typical development (Smit et al., 1990), and it occurs among the set of misarticulated sounds for an overwhelming majority of children with speech sound disorder. This difficulty acquiring rhotics is thought to be at least in part attributable to their articulatory complexity: while most sounds are produced with a single major constriction or narrowing of the vocal tract, /ɹ/ requires two near-simultaneous lingual constrictions, one near the palate and one in the pharyngeal space. When children misarticulate English rhotics, they tend to replace them with an articulatorily simpler sound, such as /w/ or /ə/. The English rhotic can appear as a consonant in syllable onset position in words like *red* and *free*; we term this subtype *consonantal /ɹ/*. It can also appear as the nucleus of a syllable, in which case it is transcribed /ɝ/ (as in *sir*) or /ɚ/ (as in *water*). Finally, it can occur in postvocalic position in words like *care* and *fear*. We will transcribe this postvocalic rhotic as the offglide of a rhotic diphthong (e.g., /kɛɚ/, /fɪɚ/), and we group it with syllabic rhotics in the category clinically termed *vocalic /ɚ/*.

Generalization in Biofeedback Treatment

Although previous studies have reported that biofeedback can succeed in eliciting perceptually correct rhotics from children who have not responded to other forms of intervention, such studies often report limited spontaneous generalization of these skills to a context in which biofeedback is not available (Fletcher et al., 1991; Gibbon & Paterson, 2006; McAllister Byun & Hitchcock, 2012; McAllister Byun et al., 2014). Gibbon and Paterson (2006) conducted a survey in which speech-language therapists were asked to describe the outcomes achieved by 60 children who had received electropalatographic (EPG) biofeedback intervention for speech impairment during the period from 1993 to 2003. Their survey results indicated that 87% of participants had made progress in their speech accuracy over the course of biofeedback treatment; however, the same proportion of participants were characterized as demonstrating at least some degree of difficulty in generalizing their newly acquired articulatory skills into the context of spontaneous speech.

The tendency for limited spontaneous carryover of gains acquired through biofeedback intervention finds a plausible explanation in the framework of principles of motor skill learning (e.g., Ballard et al., 2007). This framework makes a distinction between short-term enhancement of performance and longer term learning, also called retention or transfer, and it identifies certain conditions of practice as more facilitative of one type of learning than the other. Biofeedback is a detailed form of qualitative or knowledge of performance feedback. This type of feedback is known to help learners achieve a novel sensorimotor target, but once the target is well specified, continuing knowledge of performance feedback may lose its efficacy or even become detrimental (Hodges & Franks, 2002; Maas et al., 2008). Ongoing reliance on external feedback may inhibit generalization to naturalistic contexts in which this feedback is not available (Newell et al., 1990; Rvachew & Brosseau-Lapré, 2016). It has been suggested that detailed external feedback, including biofeedback, may be most effective when provided in early stages of training and withdrawn as the user increases his/her awareness of the target (Ballard et al., 2007; Fletcher et al., 1991; Newell et al., 1990). This study investigated the effects of biofeedback intervention when biofeedback was faded within the larger context of an adaptive *challenge point framework* (e.g., Rvachew & Brosseau-Lapré, 2016).

The Challenge Point Framework

The challenge point concept originates from motor learning research by Guadagnoli and Lee (2004) and was more recently applied to speech-motor learning by Rvachew and Brosseau-Lapré; it is also conceptually similar to Vygotsky's (1978) notion of the "zone of proximal development" in cognitive learning. Guadagnoli and Lee (2004) proposed the term to refer to the point within the range of functional task difficulty at which a learner will receive the optimal amount of feedback information to promote learning. If task difficulty is very low, the learner will demonstrate a high level of accuracy during practice, and he/she will not receive any error feedback. This is not considered optimal in a model of motor learning that treats errors as an opportunity to adjust parameter settings and thus improve future performance. On the other hand, if task difficulty is very high, the learner may not have resources available to encode feedback information for use in future motor performance.

Thus, Guadagnoli and Lee (2004) suggested that retention will be optimized when functional task difficulty is neither too high nor too low.

Adapting the challenge point concept to the specific context of speech-motor learning, Rvachew and Brosseau-Lapré (2016) identified a wide range of factors that contribute to the functional difficulty of a speech production task. Relevant parameters include the skill level of the learner, the complexity of the target, the schedule of practice, the amount of support or scaffolding offered by the clinician, and the nature of feedback provided. Hitchcock and McAllister Byun (2015) noted that some studies of biofeedback intervention for speech errors have held functional task difficulty fixed at a low level, for example, by making biofeedback available during all treatment trials. From a challenge point perspective, more generalization might have been expected in these studies if the practice context were gradually made less supportive, providing ongoing opportunities for speakers to learn from error feedback.

Applying the Challenge Point Framework in Biofeedback Intervention

Hitchcock and McAllister Byun (2015) conducted a case study of an 11-year-old girl with residual rhotic errors who showed strong within-session performance during a structured course of biofeedback treatment but did not generalize these gains. She was enrolled in a follow-up course of treatment representing a semistructured implementation of the challenge point framework. Successful generalization of correct rhotic production to the conversational level was observed and maintained at 1-month posttreatment. Of course, inferences about the general efficacy of challenge point treatment cannot be drawn without replication of these results.

Hitchcock and McAllister Byun (2015) noted one limitation of treatment in the challenge point framework: It involves making real-time adjustments in task difficulty in response to changes in participant performance, which can be challenging for anyone other than an experienced clinician. Their case study used a semistructured implementation of the challenge point framework, with a flowchart depicting a hierarchy of parameters to manipulate to adjust task complexity. The treating clinician used an Excel spreadsheet to enter scores for each item and calculate percent correct over each block of 10 trials. The clinician then applied predetermined criteria to advance, maintain, or move back a level in this complexity hierarchy. The successful outcome of the case study suggested that learners may still benefit when challenge point principles are applied in this more constrained fashion. However, even the semistructured implementation was found to require a high level of attention and effort from a skilled practitioner. Thus, the third author developed a computer program to automatically calculate accuracy and make adaptive adjustments in task difficulty.

The Challenge Point Program

The Challenge Point Program (CPP) is a PC-based software that encodes a structured version of the challenge point hierarchy. It is a free and open-source project (available at http://blog.umd.edu/cpp/download/) that was created with the goal of making it easier for clinicians with a wide range of experience levels to incorporate challenge point principles into their intervention for speech sound disorders. To facilitate adaptive adjustment of difficulty over the course of speech intervention, the CPP presents targets for production, tallies accuracy, and makes adjustments to increase or decrease task difficulty. While the parameters of the program can be customized by defining a new "study type" (see Figure 1), this article will highlight the settings used in the current study, which also correspond with the default settings of the program.

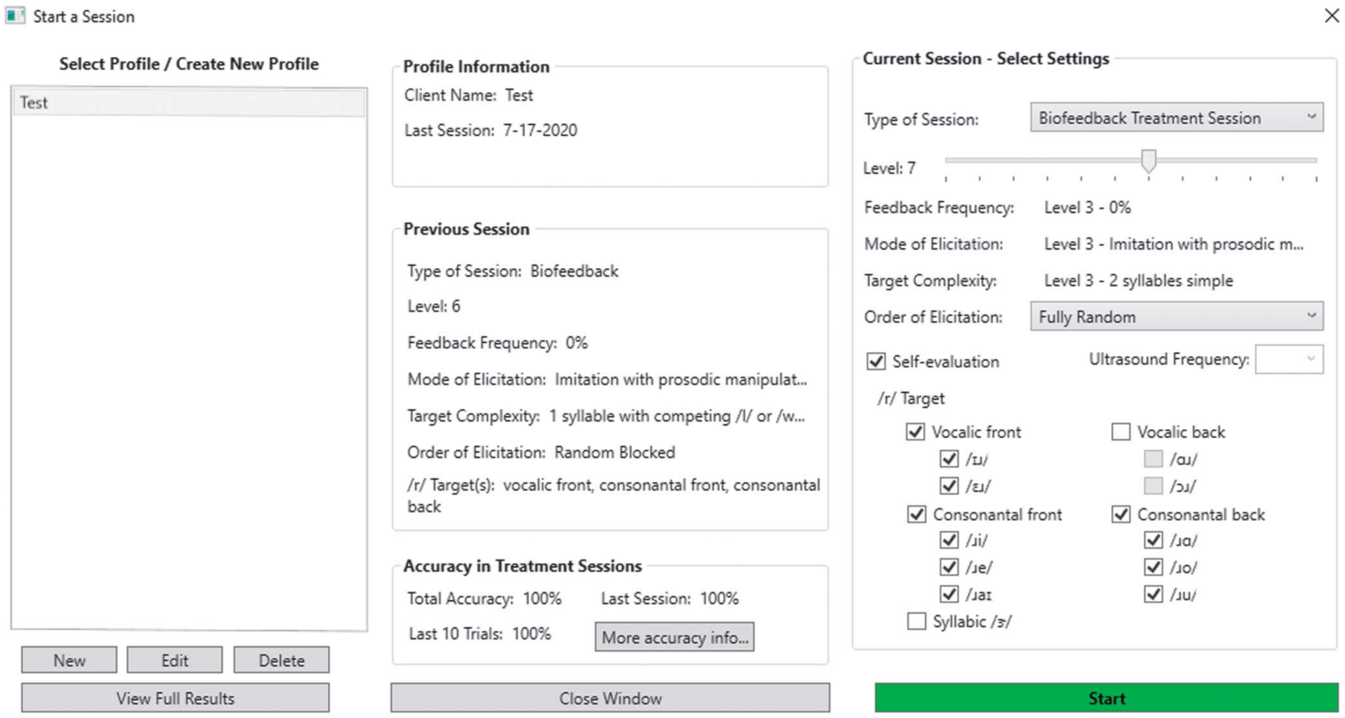
In a basic therapy exchange, the CPP presents an orthographic representation of a target utterance for the client to produce. The treating clinician scores the production via keypress as either correct (1) or incorrect (0), and this accuracy score is registered by the program. After every 10 trials, the program tallies the score and determines whether any adjustments in difficulty are indicated. If the participant's accuracy in the most recent block of 10 trials fell between 50% and 80%, it is judged that the task represents roughly the correct level of difficulty for their current ability, and no adjustments are made. If accuracy was 80% or higher, the program increases the difficulty of one parameter to make practice more challenging. If accuracy was 50% or lower, the program decreases the difficulty of one parameter. The functional task difficulty is determined by three parameters, which are adjusted on a rotating schedule: the frequency with which biofeedback is made available, the mode of elicitation, and the complexity of the syllables or words presented. Detailed information about these parameters and their levels are provided in the description of study methods that follows.

The CPP additionally makes it possible to save a profile for each participant (see Figure 1). Profiles can be used to review participant progress over time, including within-session performance and performance on probe measures administered before or after treatment. In addition, when a user selects a saved profile, the CPP automatically returns to the point in the hierarchy reflecting the level attained by that participant at the end of the previous session.

Research Objectives

This study aimed to replicate the case study from Hitchcock and McAllister Byun (2015) with a larger number of participants. Specifically, this study aimed to document learning outcomes when five participants with RSE, having begun to establish correct production of their rhotic targets in a previous course of biofeedback treatment, completed a follow-up course of intervention structured according to challenge point principles. In addition, the CPP was

Figure 1. Challenge Point Program window with user profile and previous session settings.



Start a Session

Select Profile / Create New Profile

Test

Profile Information

Client Name: Test

Last Session: 7-17-2020

Previous Session

Type of Session: Biofeedback

Level: 6

Feedback Frequency: 0%

Mode of Elicitation: Imitation with prosodic manipulat...

Target Complexity: 1 syllable with competing /l/ or /w...

Order of Elicitation: Random Blocked

/r/ Target(s): vocalic front, consonantal front, consonantal back

Accuracy in Treatment Sessions

Total Accuracy: 100%

Last Session: 100%

Last 10 Trials: 100%

Current Session - Select Settings

Type of Session: Biofeedback Treatment Session

Level: 7

Feedback Frequency: Level 3 - 0%

Mode of Elicitation: Level 3 - Imitation with prosodic m...

Target Complexity: Level 3 - 2 syllables simple

Order of Elicitation: Fully Random

☑ Self-evaluation

Ultrasound Frequency:

/r/ Target

- ☑ Vocalic front
  - ☑ /ɪɹ/
  - ☑ /ɛɹ/
- ☑ Consonantal front
  - ☑ /ɹi/
  - ☑ /ɹe/
  - ☑ /ɹaɪ
- □ Syllabic /ɝ/
- □ Vocalic back
  - □ /ɑɹ/
  - □ /ɔɹ/
- ☑ Consonantal back
  - ☑ /ɹɑ/
  - ☑ /ɹo/
  - ☑ /ɹu/

used instead of the semistructured implementation of challenge point treatment from the original case study.

Method

Study Design

Procedures for this study were approved by the institutional review boards at New York University and Montclair State University. This study utilized a structured case series design. Participants began in a baseline phase that was randomly assigned to have a duration of three, four, or five sessions. In each baseline session, participants were recorded producing a 50-item word probe, a five-item sentence probe, and a 30-item stimulability probe. In the treatment phase of the study, a 25-item subset of the word probe was administered at the beginning of every session, along with all five sentences. To assess maintenance of any gains made in therapy, the full 50-word probe, sentence probe, and stimulability probe were re-administered in three sessions after the end of all treatment. No feedback was provided during probe administration. A complete list of the words and sentences used is provided in Online Supplement A.

The duration of the study was flexible. Participants signed a consent form agreeing to complete up to 20 sessions; sessions were scheduled to occur once per week. However, participants could be discharged sooner if they reached the highest level of the challenge point hierarchy; this was the case for one participant. After 20 sessions, participants who had not yet reached ceiling-level performance were given the option to extend their participation. Two participants elected to extend their participation until they completed all levels of the challenge point hierarchy, with one reaching this goal in 38 sessions and the other in 51 sessions. Two additional participants, who had made limited progress in their 20 treatment sessions, opted to discontinue intervention.

Participants

The study was completed by five native speakers of American English ranging in age from 7;3 to 15;5 (years; months; M_{age} = 11;0).[1] Three of the five participants were male. Participants were enrolled in this study as a follow-up to their participation in one of two larger studies of biofeedback treatment for rhotic misarticulation conducted by the first and second authors. Three participants had previously completed 20 sessions of traditional and visual-acoustic biofeedback intervention over 10 weeks (McAllister Byun, 2017), and two had received 16 sessions of EPG biofeedback over 8–10 weeks (Hitchcock et al., 2017). All participants who had demonstrated some positive response to treatment over the course of the initial study were invited to continue in this follow-up intervention program; the participants evaluated

[1] Although this study focuses on the treatment of residual speech errors, the two youngest participants (ages 7;2 and 7;6) fell below the age range customarily associated with RSE and might thus be regarded simply as children with developmental rhotic errors. Because we do not draw comparisons across subjects in this study, this heterogeneity is considered unproblematic.

here are those who elected to pursue this option. A sixth participant enrolled in follow-up therapy but did not complete the study due to scheduling conflicts.

For enrollment in the original studies, participants were required to score within 1 *SD* of their age mean on the "Auditory Comprehension" subtest of the Test of Auditory Processing Skills–Third Edition. They were also required to pass a pure-tone hearing test (1000, 2000, and 4000 Hz at 20 dB HL) and to exhibit no gross structural or functional abnormality in an oral mechanism screening evaluation. To exclude individuals with more global speech deficits, participants were required to demonstrate no more than two sounds in error other than rhotics.

As noted above, all participants who enrolled in this follow-up study had shown some improvement in the ability to produce perceptually accurate rhotics in the treatment setting. However, they exhibited varying degrees of difficulty in generalizing correct production to a context in which biofeedback was not available. Table 1 reports the effect size that was calculated for each individual in connection with their participation in one of two previous biofeedback intervention studies. All measurements reflect the standardized effect size d_2 (Beeson & Robey, 2006), which is discussed in more detail in the Analyses section.

Treatment Protocol

Structure of Sessions

In this case series, treatment took place in two different locations and involved two different biofeedback technologies: visual-acoustic biofeedback (provided using the Sona-Match Module of the Kay-Pentax Computerized Speech Lab) and EPG biofeedback (provided using a Complete-Speech Palatometer V1.0). The goal of this article is not to compare the efficacies of these technologies; rather, it aims to investigate the feasibility of the CPP as a means to promote generalization of gains made in biofeedback intervention of either type. Further detail on the nature of visual and verbal cues provided during visual-acoustic biofeedback intervention can be found in McAllister Byun (2017), while the nature of the EPG biofeedback treatment was detailed in Hitchcock et al. (2017).

Although the technology used to provide biofeedback differed across the EPG and visual-acoustic arms of this study, all other parameters were held constant. Treatment was provided in weekly 45-min individual sessions eliciting 60 trials of utterances containing rhotic targets. Each session began with a 5-min period of free-play during which participants could try various manipulations in an effort to match the visual biofeedback target. The free-play period was also used to familiarize participants with any new words or prosodic patterns that were likely to be encountered within a session. The remainder of the session was spent completing the 60 trials in blocks of five. Before each block of five, the clinician provided a verbal cue (e.g. "Try to make your wave match the line" for visual-acoustic biofeedback; "Try to make the blue dots turn green" for EPG biofeedback). In some levels, the clinician also provided a model of the target utterance. The client then produced the target, which the clinician scored by pressing 1 (correct), 0 (incorrect), or 5 (intermediate/distorted) on the keyboard. After five trials, the CPP would display a screen reading "Pausing for feedback," and the clinician would provide verbal summary feedback indicating which of the five trials she perceived to be most accurate. The program would then advance to the next block.

CPP Within-Session Parameters

As described above, if a participant's accuracy over two blocks (10 trials) was 80% or better, the CPP would adjust one parameter to increase difficulty in the next block. If accuracy again reached or exceeded 80%, another manipulation was added to further increase difficulty. If accuracy fell at or below 50%, these manipulations were withdrawn in reverse order of application to reduce difficulty. Written prompts were displayed to cue the clinician to make any necessary changes. All participants started at the most basic level of the CPP hierarchy, independent of baseline accuracy.[2] At the beginning of all subsequent sessions, the participant's starting point in the challenge point hierarchy was based on performance in the previous session. Within-session adjustments in difficulty affected three categories on a rotating basis. Table 2 lays out all possible combinations of within-session parameter settings and assigns a number (1 through 12) to each. Bold text is used to highlight the individual parameter that was changed in each transition between levels.

The first category of parameter pertained to the frequency with which visual biofeedback was made available, with levels of 100%, 50%, and 0%. At the 50% biofeedback level, feedback was provided according to a one-block-on, one-block-off schedule. Feedback reduction was achieved by minimizing or covering the biofeedback display.

The second category pertained to mode of elicitation. The four levels of this category were imitation of the clinician's model, independent reading, imitation of the clinician's model of modified prosody (interrogative and exclamatory intonation contours), and independent reading with modified prosody (Preston et al., 2017). Prosodic manipulations were randomly selected by the program and were specified by adding punctuation (?, !, or .). Prosodic manipulations were initially applied at the block level, for example, all five words in a block produced with question intonation.

The third category affected the word shape and linguistic context in which rhotic targets were elicited, with seven levels: one-syllable words, one-syllable words with a competing speech sound (/l/ or /w/), two-syllable words, two-syllable words with a competing speech sound, words in a carrier phrase, words in sentences, and words in sentences with multiple rhotic words. At the carrier phrase and sentence levels, words were randomly selected from any of the first

[2] This was for the purpose of experimental uniformity, with the expectation that participants with a higher level of baseline accuracy would quickly ascend to a level appropriate to their ability.

Table 1. Participant background data, including standardized effect size (d_2) measured in connection with an immediately preceding period of biofeedback treatment.

Biofeedback type	Pseudonym	Gender	Age at enrollment	Speech treatment history	Effect size (d_2) in previous biofeedback treatment
Visual-acoustic	Ian	M	15;5	11 years (private); 10 sessions visual-acoustic biofeedback and 10 sessions traditional treatment (McAllister Byun, 2017)	1.5 (McAllister Byun, 2017)
Visual-acoustic	Garrett	M	14;1	2.5 years (school); 10 sessions visual-acoustic biofeedback and 10 sessions traditional treatment (McAllister Byun, 2017)	9.87 (McAllister Byun, 2017)
Visual-acoustic	Clara	F	10;10	2.5 years (school and private); 10 sessions visual-acoustic biofeedback and 10 sessions traditional treatment (McAllister Byun, 2017)	2.13 (McAllister Byun, 2017)
EPG	Ethan	M	7;6	2 years; 16 sessions EPG biofeedback (Hitchcock et al., 2017)	could not be computed due to zero variance; estimated to be 0.0 (Hitchcock et al., 2017)
EPG	Jenna	F	7;3	6 months; 16 sessions EPG biofeedback (Hitchcock et al., 2017)	–0.05 (Hitchcock et al., 2017)

Note. M = male; F = female; EPG = electropalatographic.

four levels and displayed in a single carrier phrase ("Say *WORD* again") or in one of a set of five sentence frames. At the highest level of complexity, the same five sentence frames were minimally modified to contain a second rhotic sound in addition to the target word. The complete list of words used in treatment, which does not overlap with the words used in probe measures, is provided in Online Supplement B.

CPP Across-Session Parameters

Two additional parameters were adjusted based on the participant's cumulative accuracy across all 60 trials in the most recent treatment session. Like the within-session parameters, these across-session parameters were adjusted when cumulative accuracy was at or above 80% or at or below 50%. The first across-session parameter was order of target elicitation. The lowest level of complexity was a fully blocked order in which each target word was elicited in a block of five consecutive trials. At the next level, random-blocked, each block of five trials targeted a single rhotic variant (see following paragraph), but different words could be used to elicit the same variant within a block (e.g., *barn, hard, star*). The highest level of complexity was a fully random order, in which different words and rhotic variants could co-occur within a single block of five.

The other parameter that was manipulated on an across-session basis was the rhotic variant(s) targeted in practice. Rhotic targets were elicited in different subsets of phonetic contexts that were individually selected for each participant. Within the major "vocalic" and "consonantal" categories, targets were subcategorized by the front or back place of the neighboring vowel. Syllabic rhotics, which have a central place of articulation, were included in both vocalic categories. To promote rapid progress, treatment began by targeting what was judged to be the most facilitative

Table 2. Within-session levels in the Challenge Point Program software.

Level	Biofeedback frequency	Mode of elicitation	Stimulus complexity
1	100%	Imitate clinician's model	1 syllable simple (e.g., *rope*)
2	**50%**	Imitate clinician's model	1 syllable simple
3	50%	**Read independently**	1 syllable simple
4	50%	Read independently	**1 syllable with competing /l/ or /w/** (e.g., *role*)
5	**0%**	Read independently	1 syllable with competing /l/ or /w/
6	0%	**Imitation with prosodic manipulation**	1 syllable with competing /l/ or /w/
7	0%	Imitation with prosodic manipulation	**2 syllables simple** (e.g., *rotate*)
8	0%	**Independent reading with prosodic manipulation**	2 syllables simple
9	0%	Independent reading with prosodic manipulation	**2 syllables with competing /l/ or /w/** (e.g., *rolling*)
10	0%	Independent reading with prosodic manipulation	**Words in carrier phrases** (e.g., "*Say 'rope' again"*)
11	0%	Independent reading with prosodic manipulation	**Words in sentences** (e.g., *"I put 'rope' at the top of my list")*
12	0%	Independent reading with prosodic manipulation	**Sentences with multiple /r/ targets** (e.g., "*When he said 'rope' she got really mad at him"*)

Note. Parameters (represented in columns) change on a rotating basis between levels; the parameter that was changed in a given level is in bold.

context for a given speaker, based on performance in the baseline period. The same target variant was maintained as the participant moved through the levels of the CPP hierarchy. Once a participant reached the top of the hierarchy of within-session parameters (Level 12) for a given variant, the decision whether or not to advance to a new variant was based on the treating clinician's ratings of accuracy in the generalization probe measures administered at the start of each session. If the participant's accuracy in the most recent word-level probe was at least 80% for the current target, a new variant was selected, again favoring the most facilitative variant based on previous probe measures. Otherwise, practice on the current target would continue until the participant maintained (or returned to) Level 12 while also demonstrating at least 80% accuracy for that target in the word-level probe.

Criterion for Discharge

Because gains can generalize from one phonetic context to another, a criterion was established whereby participants could be discharged without completing the treatment hierarchy for all categories. Whenever a participant reached the top level of the CPP hierarchy for a particular rhotic variant, the clinician reviewed that participant's performance on generalization probes (both word and sentence) administered at the start of the last two sessions. If the participant showed at least 90% accuracy at the word level and at least 80% accuracy at the sentence level across both of the two preceding sessions, the participant was judged to have met the criterion for discharge, and the study proceeded to the maintenance probe period.

Measurement

To examine participants' progress over the course of treatment, we used the binary accuracy ratings (correct/incorrect) assigned by the treating clinician in an online fashion over the course of treatment. We recognize that the treating clinician, as an unblinded listener, is not able to provide ratings free from bias. Despite this limitation, we report these ratings because they formed the basis for each participant's movement through the CPP hierarchy. We did use blinded listeners to measure probes elicited before and after treatment, as described below. We also acknowledge that this study included clinicians at two sites, introducing the potential for discrepancies in the criteria used to classify within-session productions as correct or incorrect. However, both clinicians (one at each site) had a similar duration of experience in providing biofeedback treatment and were employed in their respective research labs for approximately 1 year prior to the current study, ensuring familiarity with study protocols and scoring conventions. The clinicians were individually trained by the first and second authors, with training activities including joint ratings of sample sound files.

Words elicited in baseline and maintenance probe sessions were extracted, pooled, and presented in random order for rating by naïve listeners who were blinded to treatment condition and speaker identity. Listeners saw the orthographic target for each word and were instructed to assign a binary "correct" or "incorrect" rating to the rhotic sound in each word. These raters were recruited through Amazon Mechanical Turk (AMT), an online crowdsourcing platform that is widely used to solicit large pools of participants for experimental tasks in fields such as behavioral psychology and linguistics. Results obtained online in this crowdsourced fashion are known to be more variable than lab-based data, but with a sufficiently large number of raters, crowdsourced ratings have been found to converge with responses obtained from trained experts (Ipeirotis et al., 2014). McAllister Byun et al. (2015) investigated the validity of crowdsourced listeners' ratings in the specific context of children's rhotic sounds. They found that binary ratings aggregated over 205 naïve listeners recruited through AMT were strongly correlated with both expert listener ratings and an acoustic measure of rhoticity.

In this study, binary ratings of accuracy were collected following the protocol laid out in McAllister Byun et al. (2015). The Institutional Review Board at New York University approved this use of AMT to obtain speech ratings, and participants and their parents gave permission for recordings to be shared anonymously with outside listeners for rating purposes. For each speech token, ratings were collected from at least nine unique AMT listeners. Raters were required to have U.S.-based Internet protocol (IP) addresses and, per self-report, to be native speakers of English with no history of speech or hearing impairment. A total of 513 unique attempts to complete the task were logged, of which 63 were excluded for chance-level performance on attentional catch trials. Successful completions of the task originated from 189 unique IP addresses (where the number of unique IP addresses represents a close but not perfect proxy for the number of unique raters). Demographically, the raters who successfully completed the task ranged in self-reported age from 20 to 68 years; 43 of the 50 U.S. states were represented in their self-report of state of residence. To assess interrater reliability, the proportion of raters who agreed with the modal rating (i.e., the most frequently selected rating) was calculated for each token. Across the full set of included responses, this value was calculated to be 84.3%. A one-proportion z test indicated that this value significantly exceeded the level of agreement expected by chance (50%), $\chi^2 = 13687$, $p < .0001$. For further analysis, responses were aggregated using $\hat{p}_{correct}$, defined as the proportion of listeners who classified each token as a "correct r sound" in the binary forced-choice task.

Analyses

To evaluate each participant's within-session performance, we rely primarily on visual inspection of plots representing the treating clinician's ratings of accuracy for each set of 10 trials, as well as the current level of difficulty in the CPP hierarchy. In addition, effect sizes were calculated for each individual by comparing the mean value of $\hat{p}_{correct}$ across all baseline probes versus the mean across all maintenance probes. Effect sizes were standardized using Busk and

Serlin's d_2 statistic (Beeson & Robey, 2006), which pools standard deviations across baseline and maintenance periods in order to minimize the number of cases in which effect size cannot be calculated due to zero variance at baseline. Following Maas and Farinella (2012), this study adopted 1.0 as the minimum value of d_2 that would be regarded as clinically meaningful (i.e., the difference between pre- and posttreatment means must exceed the pooled standard deviation). However, d_2 estimates can be inflated when variance is low (Howard et al., 2015). Therefore, unstandardized effect sizes were also considered to determine whether an individual's pattern of performance was consistent with a meaningful effect of treatment.

Results

Effect Size

Effect sizes representing change in $\hat{p}_{correct}$ are reported for all participants in Table 3. The first column represents the number of treatment sessions completed by each participant. The second column shows participants' mean $\hat{p}_{correct}$ in the baseline period and the corresponding standard deviation, and the third column shows the equivalent mean and standard deviation across the three maintenance sessions. The final columns report the standardized effect size, d_2. Table 3 reflects a wide range of variability in overall response to treatment across individuals, with a minimum of −1.3 (Ian) and a maximum of 22.49 (Clara). The median effect size across all participants was 4.24, suggesting that participants' overall response to the course of treatment tended to be positive and to exceed the minimum threshold to be considered clinically significant. Individual patterns of response will be examined in detail below.

Visual Inspection and Comparison to Effect Size

The plots in Figures 2–6 depict participants' accuracy and progress through levels of the CPP hierarchy over the course of treatment. Each figure is made up of facets that represent individual sessions of treatment. Within each facet, the x-axis represents blocks of 10 trials (six blocks per session). The y-axis represents the percent of trials that were scored as correct in a given block of 10. Each point is plotted with a numeral that corresponds with the CPP level (between 1 and 12, as seen in Table 2) that was represented in that block of practice. For instance, if a block is represented with the numeral "3," this indicates that child was independently reading one-syllable words with biofeedback available in 50% of trials. Finally, the color of the numeral and the type of the associated line indicate which rhotic variant was targeted in the block in question.

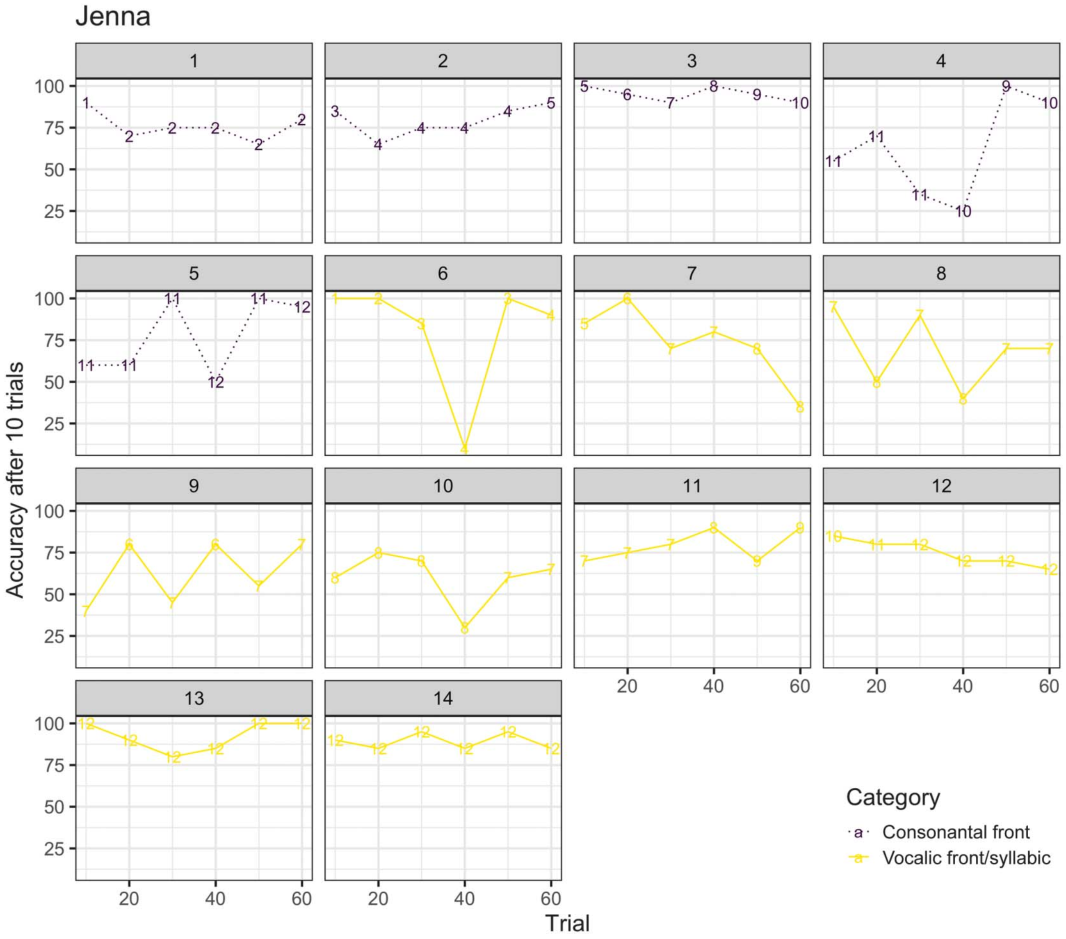
Jenna, age 7;3, made the most rapid progress out of all participants in this study, illustrated in Figure 2. Her first target variant was consonantal /ɹ/ in a front vowel context. By the end of the second session of practice, she reached Level 5, at which biofeedback is reduced to 0%. Her within-session accuracy fell considerably on the introduction of sentence-length stimuli (Level 11) in Session 4, causing her to drop back to lower levels. By the end of the fifth session, though, she reached Level 12, which involves producing sentence-level utterances with multiple rhotic words. Following the criteria described above, Jenna's accuracy on word-level probes administered at the start of each session was examined, and because her accuracy on the current variant exceeded 80%, a new target, vocalic /ɚ/ in a front vowel context, was introduced in Session 6. Jenna was mostly accurate in producing this target at the word level, although there was a brief but dramatic decrease in accuracy when a competing speech sound was introduced (Level 4). She reached 0% biofeedback frequency (Level 5) in the seventh session, and over the next four sessions, she made variable progress through levels involving imitation and reading of stimuli with prosodic manipulation. In the 12th session, she reached sentence-length stimuli (Level 11) and moved on to sentences containing multiple rhotic sounds. For the final two sessions, Jenna practiced at the top level of the CPP hierarchy (Level 12). After the 14th session, Jenna was judged to exceed 80% accuracy in producing the current target sound in the within-treatment word probe, which would qualify her to move on to another target. However, she also showed spontaneous generalization of treatment gains to untreated targets. Based on the treating clinician's ratings, she exhibited over 90% total accuracy on the word probe and over 80% accuracy on the sentence probe measure in two consecutive sessions by Session 14, thus meeting criteria to be discharged from treatment. Her cumulative within-session accuracy was calculated to be 76.6%, which was the highest of all participants in this study (see Table 3).

Blinded listeners' ratings of Jenna's accuracy on probe measures are also consistent with rapid progress. The overall effect size, pooled across vocalic and consonantal variants, was large (d_2 = 8.4). The effect size for vocalic targets was particularly strong (d_2 = 11.7), corroborating the clinician's ratings showing a steady rise in perceptually rated accuracy over the duration of the study. However, for consonantal /ɹ/,

Table 3. Measures of accuracy ($\hat{p}_{correct}$) and effect size for all participants.

Subject	No. of sessions	Baseline *M* (*SD*)	Maintenance *M* (*SD*)	d_2
Jenna	14	46.23 (3.55)	77.27 (3.88)	8.36
Ethan	20	10.90 (0.88)	19.44 (2.99)	4.24
Clara	50	28.54 (3.85)	94.49 (1.53)	22.49
Garrett	36	59.65 (6.53)	66.84 (2.16)	1.37
Ian	20	9.56 (3.26)	5.97 (1.26)	–1.30

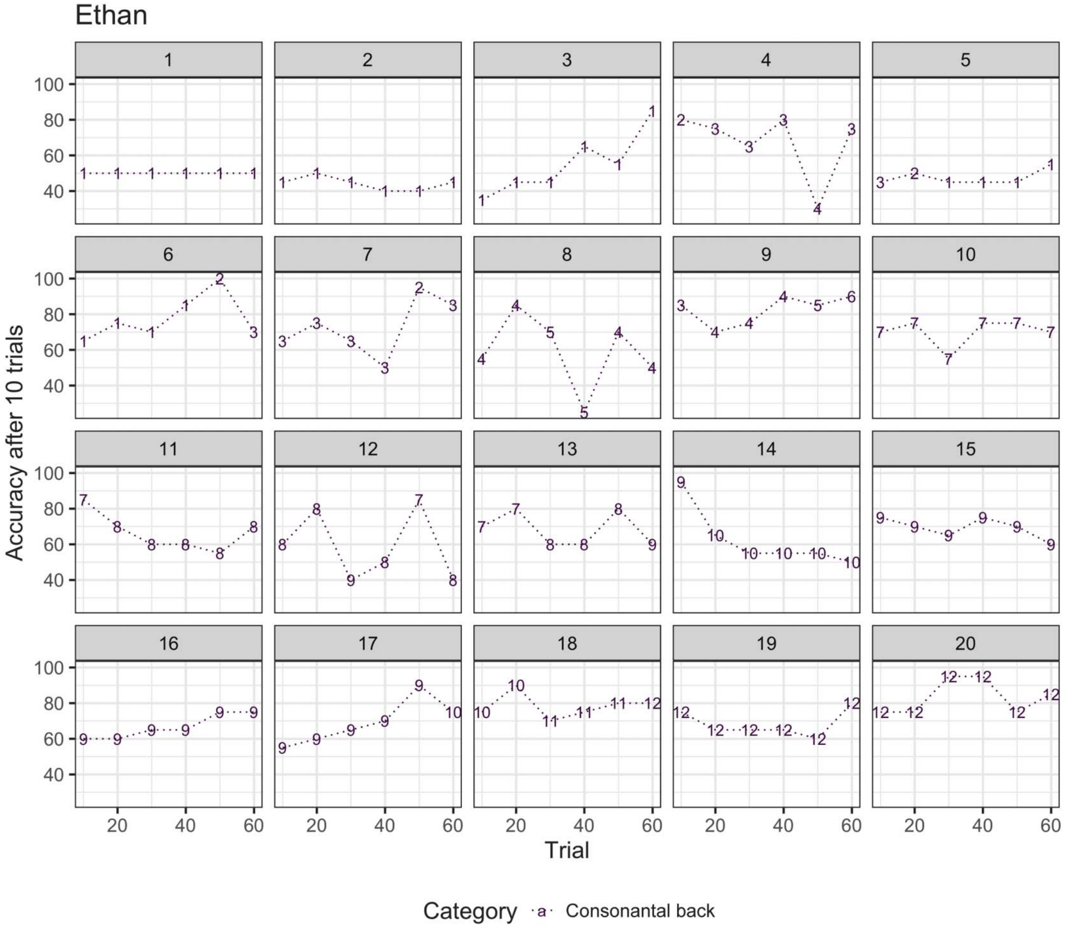
Figure 2. Plot of within-session accuracy for participant Jenna. Facets represent treatment sessions; *x*-axis represents blocks of 10 trials; *y*-axis represents accuracy in that block based on treating clinician's ratings; plotted numeral represents current Challenge Point Program level (1 through 12); color and line type represent rhotic variant targeted.



there was divergence in the clinician's record of progress in practice and blinded listeners' ratings of generalization gains. The blinded listeners perceived Jenna's productions of consonantal /ɹ/ as quite accurate at baseline (mean $\hat{p}_{correct}$ of 78.1%, *SD* of 9.0). In the maintenance period, accuracy was still fairly high but fell below the baseline level (mean $\hat{p}_{correct}$ of 60.3%, *SD* of 10.0). This suggests that Jenna's generalization learning was less robust than it seemed based on her performance during treatment and on within-session probes. We return to this issue in the Discussion section.

Ethan, age 7;6, made gains that were both slower to emerge and smaller in magnitude than those exhibited by Jenna; his within-treatment progress is illustrated in Figure 3. His initial target was consonantal /ɹ/ in a back vowel context. His accuracy was low for the first two sessions but rose in the third session, leading him to achieve Level 2 (practice with reduced feedback frequency) at the start of the fourth session. Ethan struggled notably when a competing speech sound (Level 4) was introduced in Session 4, and poor overall performance in Session 5 brought him back down to Level 1. He regained ground in Sessions 6–7, reaching Level 5 (biofeedback frequency reduced to 0%) in Session 8. However, withdrawal of biofeedback was associated with an immediate drop in accuracy, resulting in a return to Level 4 and the reintroduction of biofeedback, which was not consistently withdrawn until Session 10. In Sessions 10–13, Ethan worked through Levels 6–8, involving reduced feedback and prosodic manipulations. When he reached Level 9 (two-syllable words with competing speech sounds) in Session 12, he again showed a dramatic drop in accuracy, and he continued to struggle to produce /ɹ/ in the context of a competing speech sound through Session 17. Ethan reached the highest level of the CPP hierarchy for the selected target by the end of Session 18. However, he did not meet the criterion of 80% accuracy in word probes administered at the start of each session and therefore did not move on to a new rhotic variant. Ethan and his family elected not to continue his participation after the standard 20 sessions. His cumulative within-session accuracy was calculated to be 65.9% (see Table 3), the lowest of all participants in this study.

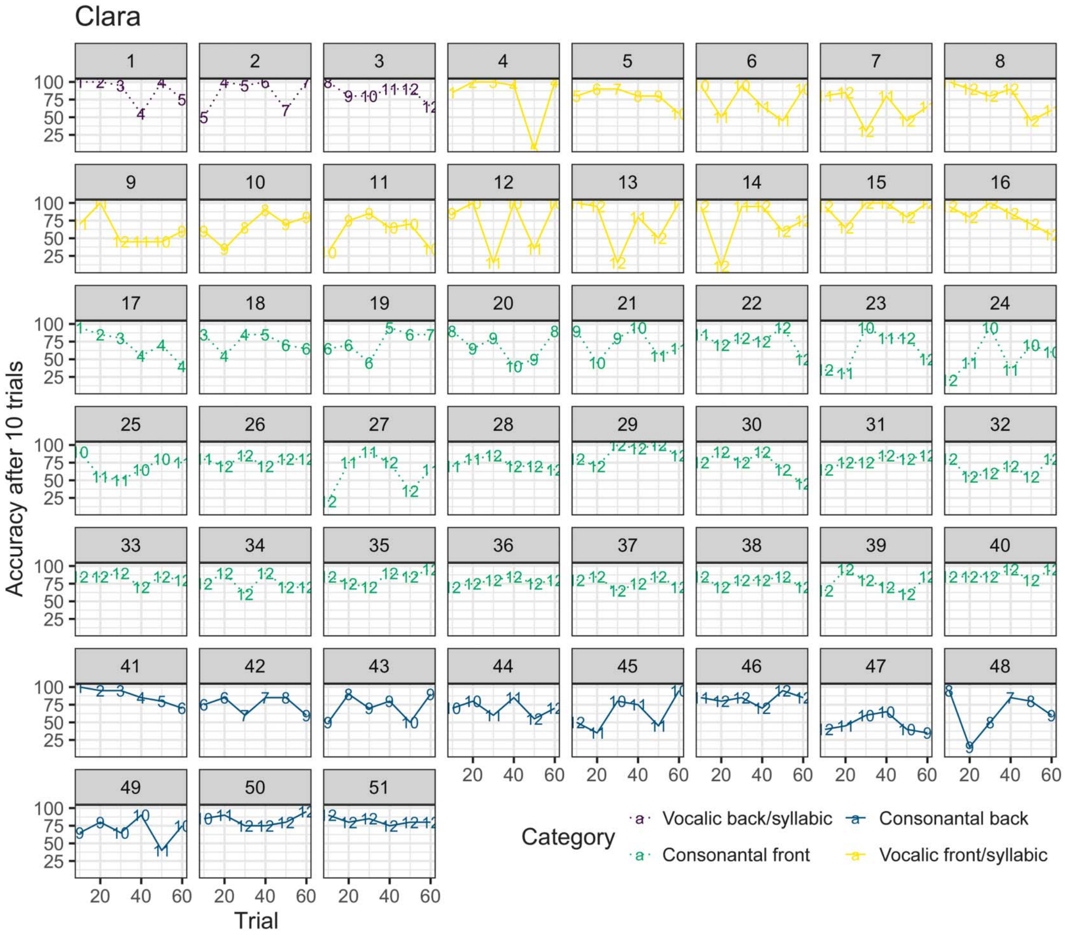
Figure 3. Plot of within-session accuracy for participant Ethan. Facets represent treatment sessions; *x*-axis represents blocks of 10 trials; *y*-axis represents accuracy in that block based on treating clinician's ratings; plotted numeral represents current Challenge Point Program level (1 through 12); color and line type represent rhotic variant targeted.



Blinded listeners' ratings of generalization probe measures show that Ethan's accuracy in the maintenance interval exceeded his baseline accuracy for both vocalic and consonantal rhotic variants. Although standardized effect sizes were suggestive of moderate gains for both vocalic ($d_2 = 4.1$) and consonantal ($d_2 = 2.7$) targets, the raw magnitude of change in $\hat{p}_{correct}$ from baseline to maintenance was less than 10 percentage points. Thus, Ethan represents a case where standardized effect sizes are inflated by low variability; despite his gains within the treatment setting, his generalization learning was modest in magnitude.

Clara, age 10;10, was the most successful among the three participants who received CPP treatment incorporating visual-acoustic biofeedback; her within-session progress can be seen in Figure 4. She made rapid progress on her initial target, vocalic /ɚ/ in a back vowel context, and reached the top of the CPP hierarchy for that target by the end of the third session. A new target, vocalic /ɚ/ in a front vowel context, was introduced in the fourth session, and she progressed to 0% frequency of biofeedback (Level 5) by the end of that session and then reached sentence-level stimuli (Level 10) by the end of the subsequent session. Her main difficulty was in stabilizing correct production in sentences containing multiple rhotic targets; practice at and around this level, with various fluctuations in accuracy, continued from the seventh session until the target was finally completed in the 16th session. A new untreated target, consonantal /ɹ/ in a front vowel context, was introduced in the 17th session. She reached phrase-level stimuli (Level 9) for this target by the 20th session. Clara then opted to extend her participation in order to complete all levels of the CPP hierarchy for this target. As with previous targets, Clara spent most of her time in practice at Level 12, sentences containing multiple rhotic targets. Although she maintained a high level of performance in practice, her accuracy in within-treatment word probes hovered below the 80% accuracy criterion until it was finally achieved in the 40th session. At this point, Clara again opted to continue her participation in order to complete the CPP hierarchy for all targets. Her gains for her final target, consonantal /ɹ/ in a back vowel context, were more rapid. She reached the top of the CPP hierarchy in the 44th session and met criteria for discharge in her 51st session.

Figure 4. Plot of within-session accuracy for participant Clara. Facets represent treatment sessions; *x*-axis represents blocks of 10 trials; *y*-axis represents accuracy in that block based on treating clinician's ratings; plotted numeral represents current Challenge Point Program level (1 through 12); color and line type represent rhotic variant targeted.

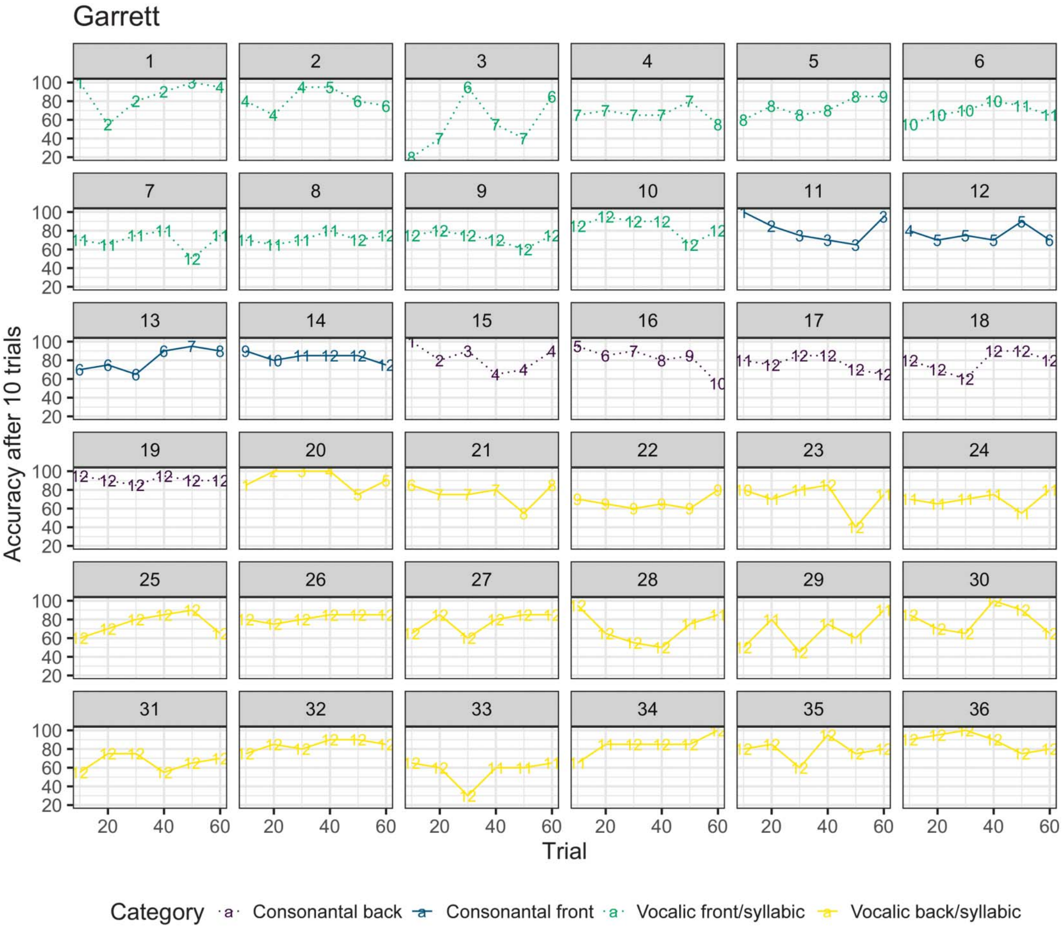


Her cumulative within-session accuracy was calculated to be 74.1% (see Table 3). Blinded listeners' ratings indicated that she sustained ceiling-level accuracy for both consonantal and vocalic /r/ at the word level in maintenance probes, yielding a final pooled effect size of 22.5. The effect size of her gains was greater for consonantal /ɹ/ (d_2 = 43.2), which started with a lower baseline accuracy, than for vocalic /ɚ/ (d_2 = 18.1).

Garrett, aged 14;1, had shown the greatest gains in the preceding biofeedback treatment study (McAllister Byun, 2017). He was judged by blinded listeners to produce rhotic targets with roughly 60% accuracy in the baseline phase. Vocalic /ɚ/ in a front vowel context was selected as the most facilitative variant and was thus targeted first in treatment. Figure 5 shows that Garrett moved steadily through successive levels of complexity in the CPP hierarchy. Biofeedback frequency was reduced to 0% (Level 5) by the second session of treatment. Sessions 3 through 5 featured practice at Levels 6–8, involving imitation and independent reading of words with prosodic manipulation. In Sessions 6 through 8, he moved through phrase and sentence levels, and Sessions 9–10 involved practice in sentences with multiple rhotic words (Level 12). In the 11th session, Garrett's accuracy for this target also exceeded 80% in word probes administered at the start of treatment, and so a new target, consonantal /ɹ/ in a front vowel context, was selected. Progress for this target was similarly rapid, and another new target, consonantal /ɹ/ in a back vowel context, was introduced in the 16th session. Garrett opted to continue his participation past 20 sessions in order to complete the CPP hierarchy for all targets. His final target, vocalic /ɚ/ in a back vowel context, was introduced in the 20th session. He again ascended rapidly through the lower levels of the CPP hierarchy, but in this case, he was slow to stabilize accurate performance at the highest level. Moreover, the treating clinician's ratings indicated that he did not achieve the 80% accuracy criterion in word probes until his 36th session, at which point treatment was discontinued. Garrett's cumulative within-session accuracy was 76.3% (see Table 3).

The extended duration of time required for Garrett to achieve 80% accuracy in word-level probes as rated by

Figure 5. Plot of within-session accuracy for participant Garrett. Facets represent treatment sessions; *x*-axis represents blocks of 10 trials; *y*-axis represents accuracy in that block based on treating clinician's ratings; plotted numeral represents current Challenge Point Program level (1 through 12); color and line type represent rhotic variant targeted.

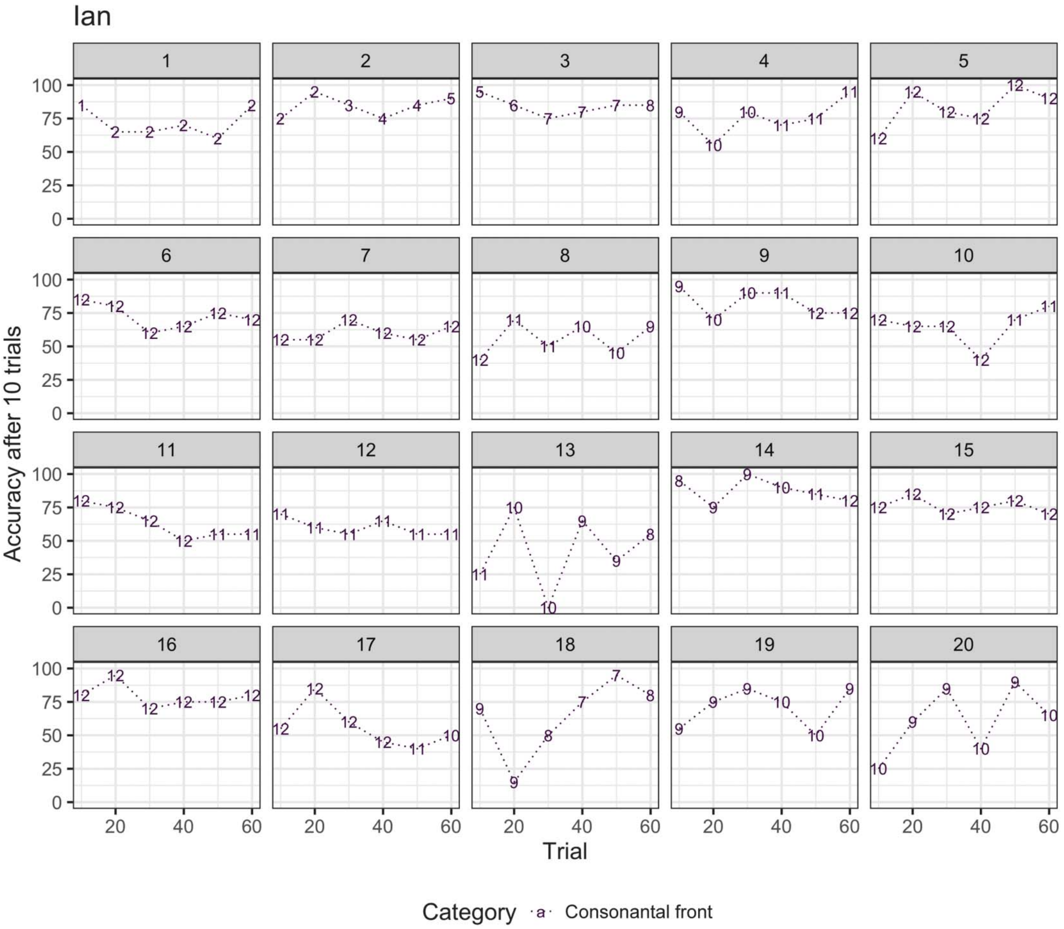


the treating clinician, even though he was practicing in sentences with multiple rhotic targets with a high level of accuracy, points to a disconnect between his performance within versus outside the treatment setting. Blinded listeners' ratings provided confirmation of this dissociation: Although his rhotic production accuracy in word-level generalization probes was higher during the maintenance phase than at baseline, the magnitude of this change was small for vocalic targets (d_2 = 1.8) and null for consonantal targets (d_2 = 0.1).

Ian, aged 15;5, was also noteworthy for the large discrepancy observed between his accuracy within the treatment setting and his accuracy on generalization probes. Both consonantal and vocalic rhotic variants were produced with a similarly low level of accuracy at baseline; consonantal /ɹ/, which had not been explicitly targeted in his previous biofeedback treatment, was selected for treatment with the CPP. Within the treatment setting (see Figure 6), Ian made immediate progress, achieving a reduction to 0% biofeedback in the second session and an increase in stimulus complexity to the sentence level in the fourth session. However, he struggled to sustain accurate production at the sentence level, particularly for sentences with multiple rhotic targets (Level 12). He initially achieved this level in Session 5 and maintained it in Sessions 6 and 7 before dropping down to lower levels due to low accuracy in Session 8. This variable performance continued for Sessions 9 through 14. He regained the top of the hierarchy in Session 14 and maintained a high level of accuracy throughout Sessions 15 and 16. However, despite this high level of performance within the practice setting, the treating clinician's ratings indicated that accurate production still was not carrying over to word probes administered at the start of treatment. Accordingly, no new target was introduced. In the final three practice sessions, Ian showed signs of frustration, dropping below Level 10 (words in carrier phrases) for the first time since Session 9. He opted not to continue participation in this study beyond the standard 20 sessions. His cumulative within-session accuracy was calculated to be 69.8% (see Table 3).

Blinded listeners' ratings of word probes elicited before and after Ian's treatment corroborate the treating clinician's

Figure 6. Plot of within-session accuracy for participant Ian. Facets represent treatment sessions; *x*-axis represents blocks of 10 trials; *y*-axis represents accuracy in that block based on treating clinician's ratings; plotted numeral represents current Challenge Point Program level (1 through 12); color and line type represent rhotic variant targeted.



impression of a lack of generalization. In fact, he showed small decreases in accuracy from pre- to posttreatment. This regression was more pronounced for vocalic targets, which had shown some progress as the targets of treatment in the previous biofeedback study but were not targeted as part of Ian's CPP treatment. Overall, despite his consistently high performance within the treatment setting, Ian's final effect size of −1.3 reflected a small but meaningful net loss of accuracy over his 20 weeks of treatment.

Discussion

As a preliminary case series, this research does not allow for strong conclusions about the efficacy of challenge point-based practice in producing generalization. In particular, given the absence of a control group or condition, we can make no claim about the relative efficacy of generalization practice with and without a challenge point structure. However, we argue that this research can furnish several valuable observations in spite of its preliminary nature.

Differing Individual Profiles of Generalization Learning

All participants in the study were observed to advance to higher levels of complexity in the challenge point framework, indicating that they were perceived by the treating clinician to produce more accurate rhotics during practice sessions. However, in keeping with previous biofeedback research (e.g., McAllister Byun & Hitchcock, 2012; McAllister Byun et al., 2014; Preston, Brick, & Landi, 2013; Preston, Leece, & Maas, 2016), individuals were highly variable in the rate at which they improved in treatment, and even more variable in the extent of gains they exhibited on generalization probes administered with no feedback. Two participants showed large effect sizes on generalization measures, two showed a small to moderate degree of response, and one showed a negative overall effect size.

Starting with the two strong responders, Clara and Jenna, we argue that these cases show that it is possible for children with RSE to make significant generalization gains when engaging in biofeedback practice following a computer-based implementation of challenge point principles. However, we lack information about how these participants would have responded to intervention structured in a different way; it is possible that they would have shown a similar magnitude of response even if treatment did not follow challenge point principles. Controlled comparisons along these lines represent an important next step.

It is also potentially informative to contrast these two cases, since they differed strikingly in the duration of treatment. Both Jenna and Clara met discharge criteria by demonstrating a high level of accuracy across all rhotic variants, but Jenna reached this point after only 14 sessions, versus a total of 51 sessions for Clara. In Clara's case, blinded listeners' ratings of untreated probe words corroborated the treating clinician's judgment in showing a large magnitude of improvement for both vocalic and consonantal rhotic variants. For Jenna, blinded listeners' ratings of maintenance probes showed robust gains for vocalic variants, but for consonantal /ɹ/, which began with a high level of accuracy, there was a decrease in accuracy to slightly below baseline levels.[3] While no conclusive interpretation can be attached to these findings, they raise the possibility that the higher cumulative dose of practice completed by Clara was more conducive to long-term generalization than the lower dose delivered to Jenna (Hitchcock et al., 2019). Thus, the criteria for discontinuing intervention might need to be made more stringent in future implementations of challenge point practice. However, controlled comparisons are needed before strong conclusions can be drawn.

Data from other participants in the case series provide a strong indication that the cumulative dose of practice cannot be the only factor influencing generalization learning. Participants Ian and, to a lesser extent, Garrett, demonstrate that it is possible to produce perceptually accurate rhotics in the treatment setting over an extended duration of practice and yet still fail to show gains on generalization probes administered with no feedback. Ian provides a particularly striking example. Because he showed increased accuracy within the treatment setting, he advanced to higher levels of the CPP hierarchy, and from the second session on, he completed most of his practice at a level of complexity that involved no biofeedback. Despite sustaining a high level of accuracy in this no-biofeedback practice, he showed no improvement on generalization probes administered at the start of each treatment session. That is, even after a session of producing rhotics with a high level of accuracy without biofeedback, he would return to baseline accuracy on the word probe administered at the start of the next session.

One possible interpretation for this pattern of behavior would posit that Ian started each session unable to access a motor plan to produce a perceptually accurate rhotic without additional support. He then reestablished an accurate motor plan over the course of within-session practice, possibly during the period of free-play with biofeedback that was made available at the start of each session (even after he had advanced to no-biofeedback levels of the challenge point hierarchy). Upon leaving the practice setting, though, he would return to reinforcing his habitual, inaccurate motor plan, and by the start of the next session, he would have to repeat the process of establishment all over again. This supports the idea that not only the dose but also the scheduling of practice may play a critical role in generalization learning. It has been proposed that increasing the intensity of practice sessions, for example, condensing a larger number of trials and sessions into a shorter duration of time, with fewer opportunities to reestablish the old motor pattern between correct trials, might be facilitative of generalization learning (Preston, Leece, & Maas, 2016). Controlled studies comparing an equal dose of intervention delivered on an intensive schedule, in contrast with the distributed schedule of this study, will be essential to test this hypothesis (Hitchcock et al., 2019).

It is possible that individual characteristics of study participants contributed to their differential response to treatment. For example, both of the strong responders in this study were female, and the moderate to minimal responders were male. However, with such a small sample size, it is impossible to know if gender does serve as a significant predictor of response to treatment. Age is another factor that needs to be investigated in larger correlational studies predicting response to treatment. Finally, it is known that auditory-perceptual sensitivity plays an important role in determining the accuracy with which a speaker produces speech sound contrasts (see discussion in Rvachew & Brosseau-Lapré, 2016). Perceptual acuity may be particularly relevant in biofeedback treatment, since an individual who has difficulty making an auditory distinction between correct and incorrect rhotics will also be less able to monitor the accuracy of their own productions when biofeedback is unavailable. This generates the prediction that individuals with strong auditory perception will, on average, show greater retention and generalization of gains from the treatment setting than comparable individuals with poorer perception. Although an appropriately fine-grained perceptual measure for rhotics was not available at the time this study was carried out, such a measure has since been developed for use in ongoing research (McAllister Byun & Tiede, 2017).

Advantages of Computer-Based Treatment

These preliminary findings do not provide evidence that practice using the CPP software is more effective in inducing generalization than practice following a more traditional approach. Nevertheless, we see value in disseminating these early results, in part because we believe that certain properties of the CPP software can be clinically beneficial independent of the question of any advantage over traditionally structured practice. First, clinicians may wish to

[3] In a 1-month follow-up probe, Jenna's accuracy for consonantal /ɹ/ as rated by blinded listeners increased to 81.8%, suggesting that the decrease in accuracy during the maintenance phase may have been a temporary fluctuation.

use the free software because it facilitates the process of tracking an individual's progress in treatment over time. Each participant's progress is saved in a separate profile. The program automatically tallies the scores entered by the clinician and reports the percent correct in each successive probe measure, broken down by rhotic variant. The clinician's scores of accuracy within each session are not displayed, but they are logged, and it is relatively straightforward to extract these data and generate a plot tracking a client's accuracy over time and across different rhotic variants.

A second advantage is that the frequently changing, goal-driven structure of practice using CPP may increase motivation for young clients. The fact that the task alters slightly from block to block, as well as the fact that practice targets a level of difficulty that is neither too hard nor too easy, may help learners remain engaged. Moreover, because the client is aware that advancing to the next step in the hierarchy is contingent on accuracy in the current level, they may show enhanced motivation to "beat the level." The power of these milestones is best illustrated by Garrett and Clara, who both continued their treatment well beyond the minimum scheduled duration of 20 sessions (36 sessions for Garrett and 51 for Clara). Both indicated that they extended their participation because they wanted to reach the top of the within-treatment hierarchy for all targets. Participants and their families informally reported that they enjoyed the sense of achievement associated with reaching a higher level.

Conclusions

This case series investigated the effects of biofeedback intervention for residual rhotic errors delivered within a challenge point framework, where practice difficulty is adaptively adjusted with the goal of enhancing generalization learning. The study more specifically examined the effects of computer-mediated challenge point treatment using a custom software, the CPP. Five children who had previously established but not generalized correct rhotic production received individual treatment that incorporated either EPG or visual-acoustic biofeedback and was structured by challenge point principles implemented in the CPP. The total duration of treatment was flexible. Participants were highly variable in the magnitude of generalization gains attained, although the median overall effect size of 4.24 suggested that participants tended to have a positive response of clinically significant magnitude. These mixed results suggest that challenge point treatment can be effective in producing generalization in some participants, but controlled comparison of treatment with and without challenge point principles will be necessary before strong conclusions can be drawn. The free and open-source CPP software may offer benefits in other areas, such as client motivation and systematic record-keeping, which recommend it independent of its theoretical grounding in the challenge point framework.

Acknowledgments

This project was funded by National Institute on Deafness and Other Communication Disorders Awards R01DC013668 (D. H. Whalen, PI) and R01DC017476 (T. McAllister, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Adler-Bock, M., Bernhardt, B. M., Gick, B., & Bacsfalvi, P. (2007). The use of ultrasound in remediation of North American English /r/ in 2 adolescents. *American Journal of Speech-Language Pathology, 16*(2), 128–139. https://doi.org/10.1044/1058-0360(2007/017)

Ballard, K. J., Maas, E., & Robin, D. A. (2007). Treating control of voicing in apraxia of speech with variable practice. *Aphasiology, 21*(12), 1195–1217. https://doi.org/10.1080/02687030601047858

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

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

Fletcher, S. G., Dagenais, P. A., & Critz-Crosby, P. (1991). Teaching consonants to profoundly hearing-impaired speakers using palatometry. *Journal of Speech and Hearing Research, 34*(4), 929–943. https://doi.org/10.1044/jshr.3404.929

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

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior, 36*(2), 212–224. https://doi.org/10.3200/JMBR.36.2.212-224

Hitchcock, E. R., 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. (2015). Enhancing generalization in biofeedback intervention using the challenge point framework: A case study. *Clinical Linguistics & Phonetics, 29*(1), 59–75. https://doi.org/10.3109/02699206.2014.956232

Hitchcock, E. R., McAllister Byun, T., Swartz, M., & 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. (2002). Learning as a function of coordination bias: Building upon pre-practice behaviours. *Human Movement Science, 21*(2), 231–258. https://doi.org/10.1016/S0167-9457(02)00101-X

Howard, D., Best, W., & Nickels, L. (2015). Optimising the design of intervention studies: Critiques and ways forward. *Aphasiology, 29*(5), 526–562. https://doi.org/10.1080/02687038.2014.985884

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

Knowledge Discovery, 28(2), 402–441. https://doi.org/10.1007/s10618-013-0306-1

Katz, W. F., Bharadwaj, S. V., & Carstens, B. (1999). Electromagnetic articulography treatment for an adult with Broca's aphasia and apraxia of speech. *Journal of Speech, Language, and Hearing Research, 42*(6), 1355–1366. https://doi.org/10.1044/jslhr.4206.1355

Maas, E., & Farinella, K. A. (2012). Random versus blocked practice in treatment for childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research, 55*(2), 561–578. https://doi.org/10.1044/1092-4388(2011/11-0120)

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

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

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

McAllister Byun, T., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders, 53,* 70–83. https://doi.org/10.1016/j.jcomdis.2014.11.003

McAllister Byun, T., & 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., Hitchcock, E. R., & Swartz, M. T. (2014). Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention. *Journal of Speech, Language, and Hearing Research, 57*(6), 2116–2130. https://doi.org/10.1044/2014_JSLHR-S-14-0034

McAllister Byun, T., & Tiede, M. (2017). Perception-production relations in later development of American English rhotics. *PLOS ONE, 12*(2), e0172022. https://doi.org/10.1371/journal.pone.0172022

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

Preston, J. L., Brick, N., & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology, 22*(4), 627–643. https://doi.org/10.1044/1058-0360(2013/12-0139)

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

Preston, J. L., Leece, M. C., McNamara, K., & Maas, E. (2017). Variable practice to enhance speech learning in ultrasound biofeedback treatment for childhood apraxia of speech: A single case experimental study. *American Journal of Speech-Language Pathology, 26*(3), 840–852. https://doi.org/10.1044/2017_AJSLP-16-0155

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

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., & Brosseau-Lapré, F. (2016). *Developmental phonological disorders: Foundations of clinical practice*. Plural.

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

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

Vygotsky, L. S. (1978). Interaction between learning and development. *Readings on the Development of Children, 23*(3), 34–41.