AJSLP

Research Article

Comparing Biofeedback Types for Children With Residual /ɹ/ Errors in American English: A Single-Case Randomization Design

Nina R. Benway,[a] Elaine R. Hitchcock,[b] Tara McAllister,[c] Graham Tomkins Feeny,[c] Jennifer Hill,[d] and Jonathan L. Preston[a]

Purpose: Research comparing different biofeedback types could lead to individualized treatments for those with residual speech errors. This study examines within-treatment response to ultrasound and visual-acoustic biofeedback, as well as generalization to untrained words, for errors affecting the American English rhotic /ɹ/. We investigated whether some children demonstrated greater improvement in /ɹ/ during ultrasound or visual-acoustic biofeedback. Each participant received both biofeedback types. Individual predictors of treatment response (i.e., age, auditory-perceptual skill, oral somatosensory skill, and growth mindset) were also explored.
Method: Seven children ages 9–16 years with residual rhotic errors participated in 10 treatment visits. Each visit consisted of two conditions: 45 min of ultrasound biofeedback and 45 min of visual-acoustic biofeedback. The order of biofeedback conditions was randomized within a single-case experimental design. Acquisition of /ɹ/ was evaluated through acoustic measurements (normalized F3–F2 difference) of selected nonbiofeedback productions during practice. Generalization of /ɹ/ was evaluated through acoustic measurements and perceptual ratings of pretreatment/posttreatment probes.
Results: Five participants demonstrated acquisition of practiced words during the combined treatment package. Three participants demonstrated a clinically significant degree of generalization to untreated words on posttreatment probes. Randomization tests indicated one participant demonstrated a significant advantage for visual-acoustic over ultrasound biofeedback. Participants' auditory-perceptual acuity on an /ɹ/–/w/ identification task was identified as a possible correlate of generalization following treatment.
Conclusions: Most participants did not demonstrate a statistically significant difference in acoustic productions between the ultrasound and visual-acoustic conditions, but one participant showed greater improvement in /ɹ/ during visual-acoustic biofeedback.
Supplemental Material: https://doi.org/10.23641/asha.14881101

Speech sound errors continue through adolescence for an estimated 1%–2% of the population (Flipsen, 2015). The American English rhotic /ɹ/ is reported to be the most commonly occurring residual speech error and is considered to be difficult to treat (Ruscello, 1995). Recent evidence demonstrates that many children with residual speech errors benefit from motor-based treatment including the use of visual biofeedback (Preston et al., 2014, 2017), but not all children demonstrate generalization of learning following treatment (McAllister Byun, 2017; Preston et al., 2018). Therefore, continued validation of treatments for residual errors, as well as investigation of the factors that contribute to treatment response, remains a clinically important priority.

It is possible that differences in client response to treatment are influenced by client-specific factors, which could interact with the biofeedback modalities used. Although there are various forms of visual biofeedback available for treating speech sound disorders, there is a lack of literature exploring whether children respond to one biofeedback

[a]Department of Communication Sciences & Disorders, Syracuse University, NY
[b]Department of Communication Sciences and Disorders, Montclair State University, NJ
[c]Department of Communicative Sciences and Disorders, New York University, NY
[d]Department of Applied Statistics, Social Science, and Humanities, New York University, NY
Correspondence to Nina R. Benway: nrbenway@syr.edu

Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Stacy Betz
Received July 17, 2020
Revision received November 13, 2020
Accepted March 27, 2021
https://doi.org/10.1044/2021_AJSLP-20-00216

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

treatment differently than another. Understanding if some children demonstrate greater improvement during one biofeedback type is an important step in individualizing treatments for individuals who do not respond to treatment. This study, therefore, compared magnitude of improvement in rhotic production during ultrasound and visual-acoustic biofeedback, when both modalities were alternated within a course of motor-based speech sound treatment in children and adolescents with residual rhotic distortions. This study also begins to explore if observed treatment response correlates with age, auditory-perceptual skill, oral somatosensory skill, or an individual's growth mindset.

Motor Learning Treatment Parameters

Recent studies (e.g., Hitchcock et al., 2017; McAllister Byun, 2017; Preston et al., 2018) have refined a series of theoretically motivated and empirically tested principles to enhance the *acquisition* and *generalization* of speech sounds. These principles are based on the premise that speech motor control may be governed by principles similar to those that influence nonspeech motor control (Maas et al., 2008). Specifically, within the schema-based motor learning framework, cognitive schemas are hypothesized to include representations of motor commands and associated contextual variables, sensory feedback, and knowledge of movement outcomes. For speech sounds such as /ɹ/, this information may include what a correct speech movement feels like, what a correct speech movement sounds like, and the articulatory targets or velocities for those movements (Guenther, 2016). The goal of motor-based speech sound intervention, then, is to update an individual's schema for the speech movements through practice and feedback. Disordered motor schemas are thought to be remediated most effectively through the systematic manipulation of *principles of motor learning*, which include prepractice variables, practice variables, and feedback variables (Maas et al., 2008). Prepractice variables address motivation and readiness for learning, frustration tolerance, and understanding of the task (i.e., what constitutes a correct production); practice variables dictate the amount, distribution, variability, attentional focus, and complexity of practice trials; and feedback variables control the type, frequency, and timing of information about the success of the movement.

The schema-based motor learning framework makes an important distinction between the concepts of performance during practice and generalized learning: *Acquisition* concerns the *performance* of a practiced skill on trained words within the treatment setting, while *learning* involves the *generalization* of the practiced motor plan to untrained words. Paradoxically, the clinical factors that facilitate acquisition of a motor plan (i.e., speech sound) do not necessarily enhance learning of the movement (Maas et al., 2008). When the goal of the current stage of intervention is the acquisition of the speech motor skill (e.g., teaching a rhotic sound to a child who is minimally accurate in syllable-level production), treatment is believed to be most effective when simple, consistent targets are presented within linguistically homogenous treatment blocks. Feedback facilitating acquisition should be frequent and immediate, focusing on establishing detailed knowledge of the articulatory movements required for correct production. On the other hand, transfer of the newly acquired speech sound to novel contexts is hypothesized to be enhanced when the speech sound is practiced in complex linguistic utterances with random variation of word position and prosodic context between trials. Feedback that promotes learning is reduced in frequency and focuses on perceptual acceptability of the production.

Visual Biofeedback as Targeted Knowledge of Performance

Knowledge of performance (KP) feedback—detailed feedback about the execution of a movement—is one feedback parameter believed to enhance motor acquisition in early stages of speech sound intervention (Maas et al., 2008). Recent evidence suggests that KP may contribute to motor learning as well (McKechnie et al., 2020; Preston et al., 2019). KP can be delivered in real time to a client through biofeedback interventions, which utilize a real-time display to provide information during a client's speech sound production. The nature of the information delivered by KP differs across biofeedback modalities, and there is currently a lack of empirical evidence comparing the relative efficacy of biofeedback modalities.

Ultrasound Biofeedback

Ultrasound biofeedback provides visual information about the shape and movements of the tongue. In an ultrasound image, the dorsal surface of the tongue appears as a hyperechoic boundary whose movements are reproduced on a computer screen in nearly real time. This tongue image is utilized clinically as a visual cue, allowing the client to compare their articulatory movements against a model template representing correct production of a speech target. For correct productions of /ɹ/, for example, ultrasound images may reveal elevation of the tongue tip or blade, lowering of the posterior tongue dorsum, and tongue root retraction (see Preston et al., 2020, for a full discussion). For /ɹ/ distortions, ultrasound images may reveal lowering of the tip or blade, elevation of the dorsum, and/or limited tongue root retraction. There is a growing body of single-case experiments and group studies showing that intervention incorporating ultrasound biofeedback can facilitate greater gains in the perceived accuracy of speech than nonbiofeedback treatment for individuals with residual speech errors, including in those whose errors continue following traditional treatment (Preston et al., 2019; Sugden et al., 2019).

Visual-Acoustic Biofeedback

Visual-acoustic biofeedback displays the moving acoustic signal accompanying a given vocal tract configuration for a speech sound. Because of the salient formant structure for /ɹ/, visual-acoustic biofeedback for /ɹ/ can be presented either in the form of a linear predictive coding spectral envelope (e.g., McAllister Byun & Hitchcock, 2012)

or a spectrogram (e.g., Shuster et al., 1995).[1] The visual-acoustic biofeedback display also serves as a visual clinical cue; throughout intervention, the client is instructed to adjust their speech sound production in a way that makes the display more accurately match a visual template representing the hallmark spectral characteristics of the targeted sound. For /ɹ/, this template would emphasize the small difference between the third formant (F3) and the second formant (F2) that characterizes accurate productions (Espy-Wilson et al., 2000). Previous single-case experiments have indicated that visual-acoustic biofeedback may enhance intervention outcomes for some children, even after nonresponse to traditional treatment, and these advantages may be increased when presented early in treatment (McAllister Byun, 2017; McAllister Byun & Campbell, 2016; McAllister Byun & Hitchcock, 2012).

Visual Biofeedback Supplements Internal Auditory and Somatosensory Feedback

Auditory and somatosensory feedback pathways—which convey information about what the speaker's own speech production sounds like and feels like—are hypothesized to play a large role in speech sound learning (Guenther, 2016). In typical speakers, acuity in discerning auditory information and somatosensory information may impact the auditory and somatosensory targets that are formed for a given sound, which contributes to differences in speech sound production (Ghosh et al., 2010; McAllister Byun & Tiede, 2017). Furthermore, there is evidence that children with speech sound disorder differ in their ability to form accurate representations of what a production should sound like (i.e., auditory targets; see, e.g., Cialdella et al., 2020) or what a production should feel like (i.e., somatosensory targets; see, e.g., McNutt, 1977). Auditory and somatosensory targets may also be influenced by individual differences in the weighting of the feedback pathways. For example, some adults with typical speech development have been shown to prioritize the information they receive from auditory versus somatosensory feedback, which Lametti et al. (2012) described as a "sensory preference." When participants in that study were simultaneously exposed to different (conflicting) combinations of somatosensory perturbation and auditory perturbation during real-time feedback, the authors observed individual differences in how participants changed their feedforward motor plan due to the feedback received. Some participants corrected their speech in response to the information contained in the somatosensory perturbation, and some corrected their speech in response to the information contained in the auditory perturbation. This suggests that there are individual differences in how speakers prioritize information conveyed through either the somatosensory or the acoustic feedback channel.

[1]Different acoustic representations may be more appropriate for other speech sounds (e.g., discrete Fourier transform spectrum for fricatives).

Individual Factors May Contribute to Speech Motor Learning

It has been thought for some time that children with residual errors may benefit from individualized treatments (Shuster et al., 1992). This conclusion continues to be supported by recent research: In noncomparative studies utilizing treatments that emphasize motor schema generalization with visual biofeedback, for every two children who demonstrate generalization following 14–16 sessions of motor-based treatment, one child does not (McAllister Byun, 2017; Preston et al., 2018). It is possible that individuals with speech sound disorders may also prioritize auditory or somatosensory feedback pathways. This raises the question of whether it is possible that certain forms of KP feedback might result in greater speech sound acquisition and learning for certain individuals because of the feedback pathway overtly emphasized by the biofeedback as well as the individual's own sensory preference. For example, the two biofeedback methods introduced above can be considered to differ with respect to the KP information that is directly available via the visual display. Li et al. (2019) posited that ultrasound biofeedback can be thought of as providing KP that supplements the client's naturally occurring somatosensory feedback, because it provides overt information about articulator placement that is ordinarily received only through internal somatosensory channels while not directly displaying detailed feedback about the acoustic properties of the speech signal. Conversely, Li et al. posited that visual-acoustic biofeedback provides KP that supplements the client's naturally occurring auditory feedback, as it displays overt, real-time tracking of auditory-acoustic properties of the speech signal while not directly displaying information about the location or direction of articulator movements.

If children do prioritize either somatosentory or auditory information, the pathway that is prioritized might be related to the individual's specific profile of sensory strengths and weaknesses. That is, it is possible that children who have relatively weaker somatosensory representations than acoustic representations might respond to the biofeedback modalities in a different manner than children who have weaker acoustic representations than somatosensory representations. It is additionally unknown if, in individuals with asymmetric sensory profiles (i.e., typical sensory profiles in one domain and atypical sensory profiles in another domain), greater therapeutic benefit would be derived from intervention that bolsters a pathway of relative perceptual strength or from intervention that reinforces the weaker pathway. For instance, individuals with poor auditory-perceptual acuity might benefit more from the well-defined auditory target provided by visual-acoustic biofeedback, or they might benefit more from leveraging their robust somatosensory target during ultrasound biofeedback.

Factors Possibly Related to Treatment Nonresponse

Factors contributing to treatment nonresponse are currently unclear, and exploration of these factors remains a high clinical priority given that some children are discharged

from clinical caseloads without improvement (Ruscello, 1995). In the following experiment, we explore if auditory perceptual skills, somatosensory skills, a child's age, or an element of global learning processes might be related to treatment response (either a differential response to biofeedback modalities or generalization following the combined treatment program).

It is well established that speech development involves the refinement of auditory and somatosensory targets for speech, and individual differences in these domains may affect speech sound learning (Cabbage et al., 2016; Hoffman et al., 1985; Ohde & Sharf, 1988; Rvachew & Jamieson, 1989). Recently, Preston et al. (2020) reported that auditory-perceptual acuity on a /ɹ/–/w/ categorical perception task was significantly correlated with amount of change following 14 sessions of ultrasound biofeedback treatment for children with residual rhotic errors. Specifically, children with better auditory-perceptual acuity made greater improvement in /ɹ/ production accuracy. Cialdella et al. (2020) likewise observed this same relationship in females with higher auditory-perceptual acuity. However, it is not yet evident whether oral somatosensory skills for speech-related tasks might also predict treatment response.

One additional unexplored factor regarding response to speech sound intervention is an individual's ability to detect errors and allocate attention to mistakes. An individual's *implicit theory of intelligence* and associated *mindset* (Dweck & Leggett, 1988) have been used to index the amount of attention that a child or adult directs to mistakes when processing stimuli, as well as their improvement on future attempts at a difficult task. A *growth mindset* is associated with the belief that effort leads to positive outcomes, that challenges are opportunities to learn, and that learning occurs from mistakes (Schroder et al., 2017). In contrast, a *fixed mindset* is associated with avoidance of challenge, low persistence, and negative self-concept of ability (Schroder et al., 2017). Children with a growth mindset are believed to direct more attention to their mistakes as a means to improve future performance than children with a fixed mindset. Moreover, following moments of failure, children with a growth mindset engage in solution-oriented instruction, self-monitoring, self-motivation, and utilization of more strategies than children with a fixed mindset. Accordingly, children with a growth mindset are reported to be more likely to show task improvement following failure, while children with a fixed mindset respond to failure with reduced performance (Diener & Dweck, 1978, 1980). Despite these differences, mindsets are not believed to be related to overall ability level because children with both fixed and growth mindsets experience the same rate of success up until the moment of failure (Dweck & Leggett, 1988; Schroder et al., 2017). Adults and children with a growth mindset show differences in neurocognitive profiles compared to those with a fixed mindset, which have been interpreted to reflect differences in attentional allocation to mistakes and processing of information related to their errors (Moser et al., 2011; Schroder et al., 2014, 2017). To date, however, there has not been exploration of this construct with respect to speech motor learning in children with speech sound disorders.

Research Questions and Hypotheses

We address our research questions within the context of an intervention study. We start by quantifying acquisition (Research Question 1) and then generalization (Research Question 2) in response to a combined treatment package, then present our primary research question (Research Question 3) as a comparison of biofeedback conditions. Specifically, we first examine if the combined treatment program with ultrasound and visual acoustic biofeedback resulted in /ɹ/ improvement in trained words during session practice. This was addressed by our Research Question 1: Is there evidence of acoustic improvement in /ɹ/ during practiced words (i.e., acquisition) for some participants during the combined biofeedback treatment package? We hypothesized, based on previous noncomparative studies of the two biofeedback types, that participants would demonstrate acquisition of /ɹ/ in response to the treatment.

We additionally wished to observe if the combined biofeedback treatment program resulted in improvements on untrained words at the posttreatment time point. This aim was addressed in Research Question 2, a pre-to-post treatment within-subject comparison: Do 10 visits of the combined biofeedback treatment program result in significant generalization of correct /ɹ/ to untrained words for some participants? Because previous studies (e.g., Preston et al., 2019) have shown that generalization may be enhanced with biofeedback relative to nonbiofeedback treatment, we hypothesized that the combined biofeedback treatment program might result in generalization to untreated words for some individuals.

The primary aim of this study was to see if some children demonstrated greater improvement in /ɹ/ during one biofeedback treatment or the other, as a means of individualizing clinical intervention (Shuster, Ruscello, & Smith, 1992). This aim underlies Research Question 3, a between-series, within-subject comparison: Do some participants with rhotic distortions show greater acoustic improvement in /ɹ/ during acquisition with the use of either ultrasound or visual-acoustic biofeedback? We hypothesized that at least some participants would show greater improvement in /ɹ/ during one type of biofeedback relative to the other. This hypothesis arises from previous clinical intervention research reporting that both types of biofeedback, taken individually, tend to yield a mix of responders and nonresponders (McAllister Byun, 2017; Preston et al., 2018). The hypothesis is also grounded in previous evidence demonstrating that individuals are heterogeneous in sensory acuity and may prioritize the information received by one sensory feedback channel over the other (e.g., Ghosh et al., 2010; Lametti et al., 2012). Because the two biofeedback modalities studied here provide differing KP feedback about the sensorimotor act of speech production, they may produce different effects based on the learner's sensory acuity or preference. Ultrasound biofeedback is hypothesized

to provide supplementary KP about articulatory targets for /ɹ/, which are typically experienced through internal somatosensory feedback, while visual-acoustic biofeedback is hypothesized to provide supplementary KP about auditory targets for /ɹ/, which are typically experienced through internal auditory feedback.

Finally, we conducted exploratory analyses to examine factors related to greater improvement in /ɹ/ during one biofeedback treatment or another (i.e., Research Question 3), as well as factors related to generalization on posttreatment probes in response to the combined treatment program (i.e., Research Question 2). The identification of factors that influence response to treatment may inform clinical treatment planning. We considered these investigations to be exploratory for several reasons. First, we had limited power at the present sample size. Second, the methods used to quantify participant perceptual factors are still undergoing validation. Third, we intended for these results to inform future, well-powered investigations.

Regarding Research Question 3, we posited if individuals who exhibit a different magnitude of improvement between the two biofeedback types during acquisition might exhibit an asymmetry in performance across auditory versus somatosensory acuity. Regarding Research Question 2, we sought to replicate the finding of Preston et al. (2020): Auditory-perceptual acuity predicts generalization of speech sounds to untrained words following the combined treatment program. We also explored whether oral somatosensory skills, age, and growth mindset related to generalization, expecting that individuals with higher somatosensory skills or higher growth mindset may show greater generalization. We did not believe that the age of participants would influence generalization.

Method

This single-case experiment was designed with reference to the What Works Clearinghouse standards for single-case design (Kratochwill et al., 2010) and is reported herein with reference to SCRIBE (Single-Case Reporting guidelines In BEhavioral interventions; Tate et al., 2016) and American Psychological Association guidelines (Appelbaum et al., 2018). We examined Research Questions 1 and 2 using within-subject pretreatment to posttreatment comparisons. Research Question 3 was examined using a between-series, within-subject randomized block single-case design. Nonparametric correlations were used to explore factors that predict individual difference in treatment response. Further details about the study design appear following the description of the treatment package. This project is an arm of a larger multisite clinical trial: Correcting Residual Errors with Spectral, Ultrasound, and Traditional Speech Therapy (McAllister et al., 2020). The three participating sites—New York University, Syracuse University, and Montclair State University—have received institutional review board approval from the Biomedical Research Alliance of New York. Informed consent for participation was obtained from parents/guardians, while informed assent was obtained from participants. All treatment visits were delivered by licensed speech-language pathologists who had undergone training in the specific treatment methodology. All assessment and treatment materials, as well as the modules used for clinician training, are freely available through Open Science Framework (https://osf.io/3qf2m/).

Participants and Recruitment

Eight children and adolescents (four boys and four girls) who demonstrated difficulty producing the Mainstream American English rhotic /ɹ/ were initially recruited from university clinic waiting lists and local speech-language pathologists for potential enrollment in the study. Five children were recruited at Syracuse University, and three children were recruited at Montclair State University. (New York University was involved only in data processing.)

Enrolled participants were required to have begun learning American English by the age of 3;0 (years;months). English was required to be the child's dominant language, or balanced with another language in cases of bilingualism. Participants were also required to hear a rhotic dialect of American English in the home. Exclusionary criteria included a diagnosis of a neurobehavioral disorder (e.g., autism spectrum disorder, obsessive–compulsive disorder, Tourette's) or permanent hearing loss. To rule out childhood apraxia of speech, participants were required to score in the "not apraxic" range on the Syllable Repetition Task (Shriberg et al., 2009), with sound additions on < 20% of nonwords. Participants were also required to demonstrate consistent productions (i.e., no segmental or prosodic differences) for at least 10 of 12 items on the Inconsistency subtest of the LinguiSystems Articulation Test–Normative Update (Bowers & Huisingh, 2018). All participants passed the Clinical Evaluation of Language Fundamentals–Fifth Edition Screening Test (Wiig et al., 2013); a hearing screening at 20 dB HL at 500, 1000, 2000, and 4000 Hz bilaterally; and a brief oral structural exam demonstrating lingual movement within functional limits for speech sounds including /ɹ/. Participants were required to demonstrate scores not lower than 1.3 *SD*s below the mean on the Matrix Reasoning subtest of the Wechsler Abbreviated Scales of Intelligence–Second Edition (Wechsler, 2011). Participants were required to score below the 8th percentile on the Goldman-Fristoe Test of Articulation–Third Edition (Goldman & Fristoe, 2015) and demonstrate less than 30% accuracy on the initial /ɹ/ word probe as judged by the speech-language pathologist. Probes assessing stimulability for /ɹ/ at the syllable level were also collected but were not exclusionary. Finally, phonological awareness ability was measured using the Phonological Awareness Composite and the Nonword Repetition subtest of the Comprehensive Test of Phonological Processing–Second Edition (Wagner et al., 2013); these scores were also not exclusionary.

Enrolled participant characteristics are summarized in Table 1. One of the eight recruited participants was excluded for not passing the childhood apraxia of speech screening. The other seven participants met all inclusionary criteria and

Table 1. Participant characteristics.

Participant	Age	Sex	Baseline Word Probe 1 percent correct	Baseline Word Probe 2 percent correct	Baseline Word Probe 3 percent correct	Average baseline stimulability probe percent correct	WASI-II T score	GFTA-3 standard score	SRT PCC	SRT additions	LAT inconsistency score	CTOPP-2 PA composite score	CTOPP-2 NWR scaled score
3101	9;9	Female	0	0	0.1	1	40	56	88	1	0	103	9
3102	11;10	Male	0.14	0.12	0.2	6.33	37	51	84	0	0	86	12
3104	9;9	Female	0	0	0	0	55	57	100	1	0	100	14
6102	15;8	Female	0	0	0.06	0	65	40	88	0	2	94	7
6103	14;11	Male	0.04	0	0	0	42	40	100	0	0	105	12
6104	9;5	Female	0	0	0	0.33	52	40	96	0	1	94	11
6108	14;6	Male	0	0	0.02	0	44	40	100	1	2	96	5

Note. WASI-II = Wechsler Abbreviated Scales of Intelligence–Second Edition Matrix Reasoning subtest (Wechsler, 2011); GFTA-3 = Goldman-Fristoe Test of Articulation–Third Edition (Goldman & Fristoe, 2015); SRT = Syllable Repetition Task (Shriberg et al., 2009); PCC = percent consonants correct; LAT = LinguiSystems Articulation Test–Normative Update (Bowers & Huisingh, 2018); CTOPP-2 = Comprehensive Test of Phonological Processing–Second Edition (Wagner et al., 2013).

passed all screening measures, were admitted to the study, and completed treatment. All were White monolingual speakers of American English. This article reports on these three boys and four girls, who ranged in age from 9;5 to 15;8 (M = 12;3, SD = 2;9). These participants demonstrated variation with regard to individual speech sound disorder histories and previous speech intervention. Descriptions of participant speech sound histories are provided in the supplemental data set available at Open Science Framework (https://osf.io/3qf2m/).

Descriptive Measures

Participants meeting inclusionary criteria completed two additional baseline visits in which /ɹ/ probes were readministered. These baseline visits also included measures of oral somatosensory skill, auditory-perceptual ability, and the administration of a Speech Mindset Scale. These measures are available through the study's Open Science Framework.

Oral somatosensory skill was measured through a researcher-developed articulatory awareness task that required participants to reflect on tongue position and movement of the tongue within the oral cavity during speech sound production (see Appendix A). Following training on the task, participants were required to imitate speech sounds and indicate the relative position of the tongue for each sound. The task consisted of one module assessing consonant production and three modules assessing vowel production. For the consonant modules, participants were asked to identify which part of their tongue (front, back) was used following the participants' articulation of nine speech sounds (/t/, /d/, /k/, /g/, /s/, /l/, /n/, /z/, and /ŋ/). For vowel modules, participants were prompted to produce pairs of contrasting vowels and then were asked to identify which vowel in each pair had a more back tongue position (Task 2; two training items, seven trials), a lower tongue position (Task 3; one training item, eight trials), or a higher tongue position (Task 4; nine trials). The vowel tasks included only targets unaffected by dialect differences (such as /æ/ tensing seen in many dialects of American English). Presentation of the task, including directions and auditory speech sound prompts, was standardized through PsychoPy v. 3.0.7 (Peirce, 2007). The main outcome measure was the percent of correct responses overall.

Auditory-perceptual ability was measured through an /ɹ/–/w/ identification task presented over Sennheiser HD 280 Pro headphones to participants at a comfortable listening level. The auditory identification task utilized a two-alternative forced choice paradigm, in a manner similar to that which is described by McAllister Byun and Tiede (2017). Typical productions of /ɹ/ and /w/ were extracted from exemplars of "rake" and "wake," and a continuum of synthetic consonants was created in which the formant structure of the phone incremented evenly from /ɹ/ to /w/ using the Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram algorithm (Kawahara et al., 2013). This resulted in a stimulus set in which some tokens consistently sound like "rake," some consistently sound like "wake," and some are ambiguous. Participants were trained to identify the stimulus as "rake" or "wake" using the maximally distinct continuum endpoints, while more ambiguous tokens from the interior of the continuum were presented repeatedly during the main task trials. The percent of "rake" responses for each step of the continuum was plotted and fitted to a logistic function. Following previous studies (e.g., Benway et al., 2021; McAllister Byun & Tiede, 2017; Preston et al., 2020), the width of the fitted function from the 25th to the 75th percentile of probability was treated as an index of auditory-perceptual acuity. Specifically, the ability to consistently assign the same ambiguous token to the same perceptual category results in a narrower boundary between the perceptual categories for "rake" and "wake" and is suggestive of higher auditory-perceptual acuity. Administration of this task was standardized using a custom software (Ortiz, 2017).

As discussed previously, a child's propensity for a growth or fixed mindset may be related to learning performance. A nine-question, researcher-developed survey (Speech Mindset Scale) was adapted from previous explorations of mindset in children (Park et al., 2017; Schroder et al., 2014). Participants were introduced to two hypothetical siblings, Skyler and Peyton. Within each prompt, Skyler represents a fixed mindset (e.g., "In speech class, I like to practice words that are very easy so I can get a lot right"), while Peyton represents a growth mindset (e.g., "In speech class, I like to practice words that are very hard so I can learn more about making speech sounds."). The participant selected how often they feel the same way as either sibling (Always like Skyler, Sometimes like Skyler, Equal to both, Sometimes like Peyton, Always like Peyton; e.g., Mellor & Moore, 2014). The prompts of the Speech Mindset Scale focus on the evaluation of effort, seeking feedback after mistakes, and seeking challenge within the context of school, leisure, and speech production. The task was presented using REDCap software (Harris et al., 2009). Participant responses were totaled to represent a sum of endorsed items ranging from 0 to 36, with 0 representing consistent endorsement of "Always like Skyler" (most fixed mindset) and 36 representing consistent endorsement of "Always like Peyton" (most growth mindset). The Speech Mindset Scale is included as Appendix B.

Baseline Probes, Dynamic Assessment, and Biofeedback Orientation

Three pretreatment baseline probes were obtained in which /ɹ/ was elicited in untreated words (with no clinician model) and in a stimulability task in which the participant imitated a clinician model of /ɹ/ in syllables representing 16 different phonetic contexts. Participants completed the same /ɹ/ word probes in three visits after the end of the full treatment program. The probe list is available through the study's Open Science Framework page (https://osf.io/3qf2m/).

After the three baseline visits, participants engaged in a 90-min dynamic assessment visit that introduced basic tongue anatomy and tongue shapes required for /ɹ/ (see Preston et al., 2020). Participants were introduced to a

variety of magnetic resonance images and tracings representing sagittal and coronal views of the tongue while learning that perceptually accurate /ɹ/ productions generally have a constriction in the anterior aspect of the oral cavity as well as tongue root retraction into the oropharynx. This scripted instruction was followed by a relatively unstructured period of articulatory cueing and shaping aimed at eliciting perceptually accurate /ɹ/; these two elements of prepractice were completed in roughly 45 min. The remaining 45 min of the dynamic assessment visit elicited intensive motor-based practice in the manner described below for the main treatment program, with the important exception that no biofeedback was used during the dynamic assessment visit (see Supplemental Material S1 for a video example of a structured practice during the dynamic assessment visit). The goal of dynamic assessment was to introduce the articulatory requirements for /ɹ/, thus providing some familiarity with the articulatory cues that would be used in subsequent visits before the real-time biofeedback displays and associated clinical feedback were introduced.

Following the dynamic assessment visit, participants completed an orientation visit consisting of two 48-min segments introducing each of the biofeedback conditions. The order of orientation of biofeedback types was randomized for each participant. Participants were guided through a period of scripted instruction explaining how to interpret the visual display of each technology. This was followed by a period of relatively unstructured elicitation that gave participants the chance to observe how articulator movements impact the real-time ultrasound or visual-acoustic display. Articulatory cueing, introduced during the previous dynamic assessment visit, was provided to help participants better approximate /ɹ/ while viewing the visual feedback display. No structured practice occurred during this visit.

Biofeedback Treatment Program

All treatment materials are available through the study's Open Science Framework page, and video examples of biofeedback treatment are included in Supplemental Materials S2 and S3. The treatment was designed to include 10 visits, with each visit allowing for a comparison of acquisition (within-session accuracy on practiced items) during two treatment conditions. These 10 visits occurred twice per week, with each visit including both ultrasound and visual-acoustic biofeedback conditions. Therefore, participants completed 20 biofeedback conditions (10 each for ultrasound and visual-acoustic biofeedback) in the 10 visits. Each condition was 48 min in length, with each visit totaling 101 min (comprising two conditions plus breaks). In all, the combined treatment program consisted of 15 hr of treatment, exclusive of transitional breaks. Stimulus lists were developed that featured /ɹ/ in five phonetic contexts: nucleic /ɝ/, prevocalic /ɹ/ before front vowels and before back vowels, and postvocalic /ɹ/ following front vowels and following back vowels. The lists contained a balanced representation of words and nonwords. Two different, yet comparable, stimulus lists (e.g., List A contained *raid* and *serve,* while List B contained *rate* and *curve*) were also randomly assigned to the two conditions within each visit in order to minimize the potential for treatment effects to carry over between the two conditions occurring on the same day. Because there is evidence that speech motor plans are stored at the monosyllable level (Guenther, 2016), using different monosyllable exemplars in the different conditions theoretically minimizes carryover between biofeedback types. Two targets, either words or nonwords, representing each phonetic context were randomly selected from the overall list for practice in each condition. Each biofeedback condition began with prepractice (e.g., Maas et al., 2008; Preston et al., 2020) designed to provide orientation to the target movement and detailed clinical feedback about correct productions. Clinical feedback was provided through the use of traditional visual and auditory placement cues paired with constant reference to the real-time biofeedback display. Prepractice focused on direct imitation of a subset of the exemplar word lists randomized for each biofeedback condition. Prepractice ended after 15 min or once there were three correct productions for each of the rhotic targets.

Following a 3-min break, participants engaged in structured practice for the remaining 30 min of the biofeedback condition or until 200 practice trials were completed. Treatment was standardized across clinicians and sites using customized, open-source software for stimulus presentation and response recording in the context of speech treatment (McAllister, Hitlock, et al., 2020). Stimuli included monosyllabic words that did not adapt in difficulty. These targets were presented by the software in a blocked fashion, with each exemplar selected for the condition elicited in two consecutive blocks of 10 trials. For each trial, participants read the stimulus prompt presented by the software. Simultaneous biofeedback was made available to participants for the first eight of 10 trials within each block (the individual modalities are described in more detail below). There were two trials per block without biofeedback, which served two purposes. It allowed for a nonbiofeedback trial to be used for acoustic measurement, as described below, and aimed to prevent participants from becoming entirely dependent on biofeedback. After each production, the clinician recorded their own perceptual judgment and the software prompted the clinician to deliver a specific type of feedback (i.e., KP, knowledge of results [KR], or no feedback). For the first five blocks (50 trials), KP was randomly assigned to occur on five of 10 trials and KR was randomly assigned to two of 10 trials. For the sixth block and beyond, KP was faded to three of 10 trials while KR was maintained at two of 10 trials. Non-KP/non-KR trials were designated as "no clinician feedback" trials. Feedback containing KP was designed to refer to the position of the articulators (e.g., Preston et al., 2020), reference the visual biofeedback display, and provide a correct model of the target. For example, when delivering KP for a visual-acoustic trial, the clinician might say "not quite, lift your tongue tip high in order to make the third bump move to the left like this: *reen.*" Clinicians provided holistic perceptual impressions for KR feedback (i.e., "correct!" or "not quite"). Participants rated their own productions on two trials per block to practice speech sound error detection.

At the end of each block of 10 trials, participants received general feedback emphasizing the importance of effort and challenge, framed according to the principles of growth mindset (Dweck & Leggett, 1988; Paunesku et al., 2015).

Throughout this practice, the stimulus presentation software was visible on the upper one third of the computer screen and the biofeedback display was visible on the lower two thirds of the screen (see Supplemental Materials S2 and S3). Ultrasound biofeedback was provided using either a MicroUS or Echoblaster 128 probe set to a depth of 90 mm, stabilized with a custom chest-strapped stabilizer (Hitchcock & Reda, 2020). Ultrasound biofeedback was displayed at both treatment sites using Echo Wave II software (TELEMED Medical Systems, 2019). Visual-acoustic biofeedback was displayed using Computerized Speech Lab 4500b, Sona-Match Module (PENTAX Medical, 2019), which was connected to the treatment computer via a Lynx E44 sound card. Magnetic resonance images illustrating target tongue shapes for /ɹ/ (Boyce, 2015), as well as condition-specific illustrations, were also available to the participant and clinician throughout prepractice and treatment. Examples of correct and incorrect productions as seen for each biofeedback display are shown in Figure 1.

Treatment Fidelity and Achieved Treatment Intensity

Treatment fidelity was measured to determine the extent to which individual trials were delivered as specified in the study protocol. A total of four videos per participant were pseudorandomly selected to assess treatment fidelity, stratified to provide coverage for each biofeedback condition and the beginning, middle, and end of treatment. The aspects of treatment reviewed for fidelity were those theorized to be clinically potent within a biofeedback treatment program built upon the principles of motor learning. These aspects included the availability of biofeedback on 80% of trials, an auditory model before each treatment block, a prompt for self-evaluation, KR on indicated trials, and KP on indicated trials. For KP specifically, we quantified whether the clinician's feedback appropriately referenced the position of the articulators, the biofeedback display, and an acoustic model. We counted two general patterns of clinician behavior as deviations: failing to provide the specified feedback type on a trial (e.g., articulators not mentioned on a KP trial) and providing feedback that was not specified on a trial (e.g., KR provided on a no feedback trial). Table 2 reveals that study clinicians had very high levels of fidelity in providing biofeedback and auditory models, cueing self-evaluation, and providing KR feedback. Fidelity was lower, but still greater than 80%, for the three KP feedback components as operationalized for this study (i.e., display reference, articulator reference, and acoustic model). We found the most common KP feedback error to be a clinician reference to the biofeedback display on a "no verbal feedback" trial, which contributed to the lower treatment fidelity for KP than KR. Fidelity to the treatment protocol, overall, was judged to be acceptable.

Treatment during the 30 min of structured practice per biofeedback condition revealed average dose per participant ranged from 96.7 trials (SD = 11.5; Participant 3101) to 197.2 trials (SD = 5.7; Participant 6108), with 137.4 trials (SD = 41.6) representing the group-level average. The group-average cumulative intervention intensity was 2748 (1934–3943) trials.

Study Design and Randomization

We examined the within-subject data in two complementary ways. For Research Questions 1 and 2, we only

Figure 1. Examples of correct and distorted /ɹ/ during biofeedback. Panels A and B show the linear predictive coding spectrum as seen during treatment. Each image has a template in red representing a "good" /ɹ/ for that individual's age and gender. Panel A shows a perceptually correct /ɹ/, while Panel B shows a distorted /ɹ/ (third formant is too high). Panels C and D show the ultrasound display as seen during treatment, with the white line showing the dorsal surface of the tongue. The tongue shape template is not visible in these images. Panel C shows a perceptually correct "bunched" /ɹ/ (a high tongue blade, a low tongue body [dorsum], and tongue root retraction). Panel D shows a distorted /ɹ/ with a high tongue body (dorsum).

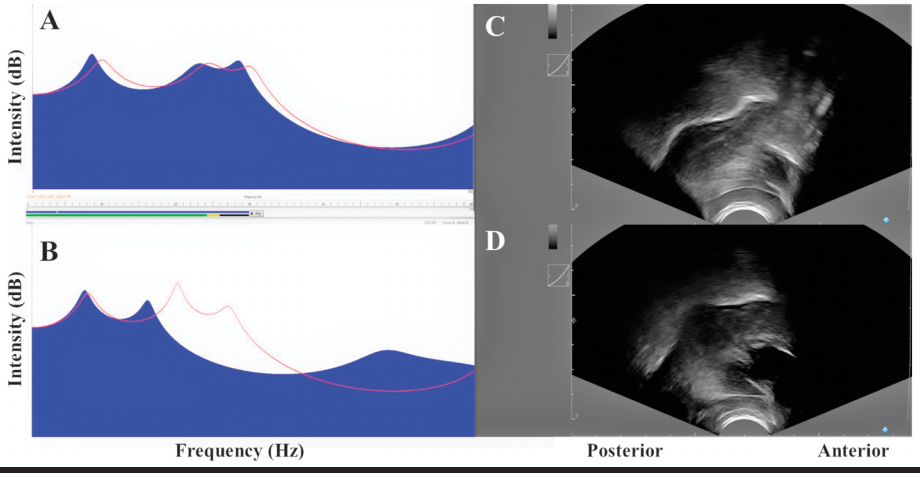


Table 2. Treatment fidelity.

Dose and fidelity	Biofeedback as indicated	Preblock auditory model	Prompt for self-evaluation	KR as indicated	KP – reference position of articulators	KP – reference biofeedback display	KP – provide acoustic model
Expected dose (trials per block)	8/10	1/10	2/10	2/10	5/10, fading to 3/10 after 50 trials	5/10, fading to 3/10 after 50 trials	5/10, fading to 3/10 after 50 trials
Observed fidelity to expected dose	99.9%	97.9%	99.7%	97.9%	82.7%	80.4%	85.2%

Note. Fidelity was calculated on a per-trial basis. KR = knowledge of results; KP = knowledge of performance.

considered the effects of the combined treatment program. Research Question 1 examined if the rate of change during acquisition for each participant was different than zero, and concerns the same within-treatment productions analyzed in our primary research question. Research Question 2 examined if there was quantifiable change in rhotic production from the mean of three pretreatment time points to the mean of three posttreatment time points. Acoustic measures and listener ratings for this analysis were obtained during the production of untrained words before treatment began and following the culmination of treatment (i.e., motor skill generalization).

The primary research hypothesis (accompanying Research Question 3) was addressed through between-series comparison of acoustic measures obtained during the production of trained words (i.e., motor skill acquisition) during ultrasound biofeedback or visual-acoustic biofeedback conditions, within blocks of treatment visits, for each participant. This experimental design, a randomized block design (see Kratochwill & Levin, 2010; Rvachew & Matthews, 2017), alternates study conditions within time-sensitive statistical blocks (in this case, within a treatment visit). Statistical blocking by treatment visit, specifically, is important to control for effects that may happen within larger timeframes, such as maturation or consolidation of learning. If an overall time trend is present, defining the statistical block as the treatment visit will still allow for the comparison of condition mean difference within the block because the time trend will affect both observations in a given day of the study to a similar extent. Additionally, the incorporation of randomization in the design legitimizes the use of inferential statistics to test the null hypothesis for Research Question 3: For a given participant, there is no difference in acoustic production of /ɹ/ between the two biofeedback conditions. Randomization represents a methodological and statistical improvement over traditional single-case replication (Kratochwill & Levin, 2010) and can be considered an "N-of-1 randomized controlled trial" (e.g., Rvachew & Matthews, 2017). For this investigation, the order of the two biofeedback conditions within each visit was randomly determined a priori by the data management team at New York University. The randomization sequence for each participant is included as Supplemental Material S4. No design or procedural changes occurred after the start of the study. Participants and parents were formally interviewed after the midpoint of treatment and at follow-up to monitor for adverse effects of the treatment. A schematic of the study is included in Figure 2.

Analysis Methodology

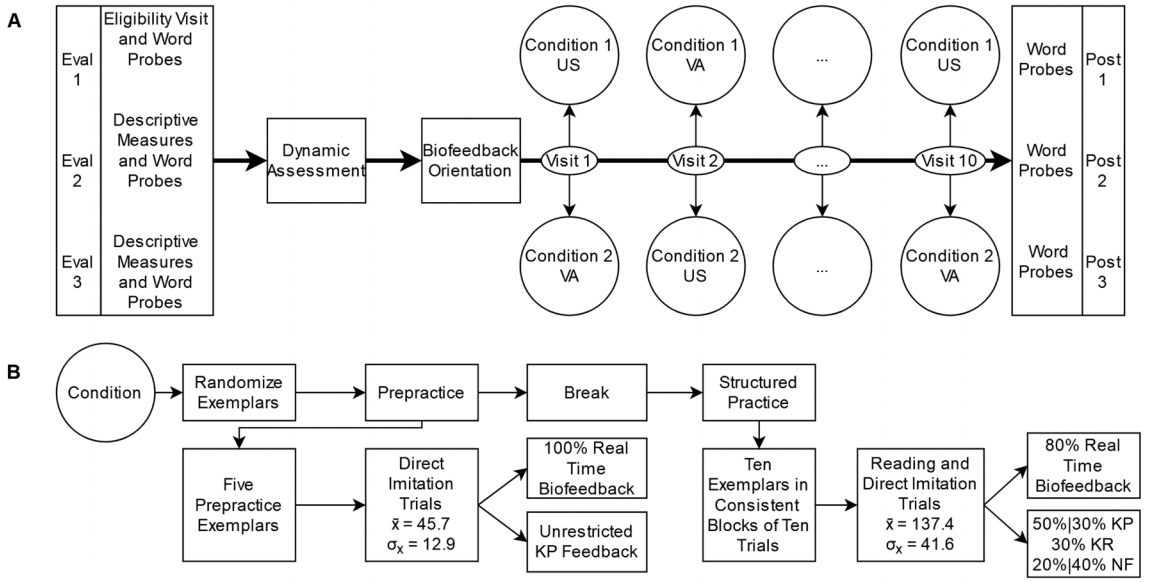
Audio Capture

Because this study employed acoustic analysis, audio capture was standardized across both treatment sites. All participants wore headset condenser microphones with a cardioid polar pattern (AKG C520) with standardized mic-to-mouth distance set by template at the start of each visit. The audio signal was split at a Behringer UMC 404HD Audio Interface; from this unit, a digitized signal was sent to the computer for screen capture purposes and analog signals were sent to an external Marantz PMD 661 MKIII and a Pentax CSL 4500b connected to the treatment computer via a Lynx E44 sound card. Because of this setup, one headset microphone simultaneously fed the main recording source, the backup recording source, and the visual-acoustic biofeedback display. All audio was recorded at 44.1 kHz and 16-bit resolution resulting in lossless 16-bit PCM audio inside WAV containers.

Acoustic Analysis

Acoustic measures were used to examine all three hypotheses. In American English, rhotic distortions are characterized acoustically by a large distance between the F2 and F3, while standard rhotic productions are characterized by an F3 that approaches the frequency of F2 (Espy-Wilson et al., 2000). Formant measurements were made from /ɹ/ sounds produced in the ninth of 10 trials in every treatment block (which was always produced with no biofeedback) to assess acquisition for Research Questions 1 and 3. Formant measurements were made from every production in the baseline and posttreatment word probes to assess generalization for Research Question 2. Because the age-normalized difference between F3 and F2 has been found to best correlate with perceptual ratings of /ɹ/ production accuracy (Campbell et al., 2018), it was used as the primary acoustic outcome for this study. Age and gender norms for normalization were obtained from Lee et al. (1999). Normalized formant values were generated by taking the difference between the observed

Figure 2. Study design and treatment methodology. Panel A shows the rapidly alternating single-case randomized block design and a hypothetical condition order for illustrative purposes. For each participant, the order of biofeedback presentation was randomized to the two treatment conditions within each of the 10 visits (i.e., the statistical blocking unit). Research Question 1 concerned the between-series, within-subject comparison of performance on trained words to measure if some subjects demonstrated greater motor acquisition to one biofeedback condition or the other. Research Question 2 compared performance on untrained words during the three pretreatment evaluation probes to the three posttreatment probes to measure generalization following the combined treatment program. Panel B shows the structure of the treatment program, along with group-level averages of achieved dosage. US = ultrasound; VA = visual-acoustic; KP = knowledge of performance; KR = knowledge of results; NF = No feedback.



/ɹ/ formant value for each production and the normative /ɹ/ formant value for the appropriate age (in years) and gender, and then the difference was divided by the standard deviation observed for that age and gender in the normative sample. In this way, the normalized age-and-gender formant values are *z* scores.[2] Analysis was completed by research assistants who were blinded to information about participant age, treatment condition, and visit number.

This study employed a hybrid approach to acoustic analysis that included manual verification of automated data extraction. Word-level tokens designated for analysis were manually segmented from audio recordings using Praat (Boersma & Weenink, 2019). Phoneme boundaries were automatically placed within these word tokens using the Montreal Forced Aligner v 1.0.1 (McAuliffe et al., 2017) wrapper to the Kaldi Speech Recognition Toolkit (Povey et al., 2011). Default English acoustic models were utilized by the forced aligner during alignment. The alignment dictionary, CMUdict (Carnegie Mellon Speech Group, 2014), was expanded to include grapheme-to-phoneme mappings of the nonwords employed during the present study. The automatically aligned phoneme boundaries generated by the Montreal Forced Aligner were manually confirmed by trained research assistants, who adjusted boundaries to exclude areas of inconsistent formant tracking as estimated by Praat. Research assistants also flagged tokens for manual review if no regions of valid automated formant tracking could be found. Formant extraction (Burg method) for F2 and F3 was then automated using a 50-ms Gaussian window surrounding the midpoint of each identified /ɹ/ interval (Lennes, 2003). Formant estimation settings that most accurately tracked with the visible areas of energy concentration on the spectrogram were selected for each speaker, as suggested by Derdemezis et al. (2016). These estimation settings were reached by consensus among three trained raters before the forced-aligned phoneme boundaries were verified and the formants were extracted. The automated formant extraction values were screened for plausibility, manually reviewed, and—if needed—manually remeasured in a manner similar to Hillenbrand et al. (1995). Screening of automated formant values and the procedure for manual remeasurement is described in Supplemental Material S5.

Both interrater and intrarater reliability were calculated for a randomly selected 20% of files to examine the average intercorrelation of the formant measurements. The intraclass correlation coefficient (ICC) for intrarater reliability, with raters entered as a random effect, was good for F2 (ICC = .88, 95% confidence interval [CI] [.85, .90]) and excellent for F3 (ICC = .93, 95% CI [.91, .94]). The intraclass correlation coefficient for interrater reliability was good for both F2 (ICC = .83, 95% CI [.80, .85]) and F3 (ICC = .85, 95% CI [.83, .88]). All ICC calculations

[2]Because the F3–F2 distance is so great in disordered rhotics, the age- and gender-normalized F3–F2 distance may also be larger than typically encountered *z* scores (see Campbell et al., 2018).

were completed using the psych package (Revelle, 2019) in R Version 3.6.2 (R Core Team, 2017).

Perceptual Analysis

Research Question 2 examined generalization of treatment gains to untrained words following the completion of the entire treatment program, so we supplemented acoustic measures with perceptual ratings that would indicate whether any transfer to unpracticed words was apparent to blinded listeners. We did not use perceptual ratings to address acquisition (Research Questions 1 and 3) because perceptual ratings may not be gradient enough to capture incremental speech sound changes in response to acquisition.

Binary perceptual ratings (correct/incorrect) were obtained for untrained words from the three pretreatment and three posttreatment probes following the protocol described in Preston and Leece (2017). Each probe token was extracted from the audio of the treatment and saved to individual WAV files. These files were randomized to listening modules presented by a Praat script in groups of 100 tokens representing audio from all study participants at a given treatment site. Raters provided correct/incorrect decisions after hearing the audio recording. No visual information about the speech productions was available to raters. The raters could replay a token freely before submitting a judgment, with no time limit. The ratings were made independently by three experienced clinicians from the alternate treatment site, who were blinded to the subject and time point at which the recording was collected. Gwet's chance-corrected agreement coefficient (Gwet, 2014) was selected to quantify reliability across multiple raters. This measure has been found to be more robust to the paradoxes of kappa in the case of high or low rating prevalence (Wongpakaran et al., 2013), a salient consideration for correct/incorrect ratings in cases of possible treatment (non)response. The calculations were performed in R using the irrCAC package (Gwet, 2019). Benchmarking of the strength of the coefficient was completed in the same package, comparing the agreement coefficient and its standard error relative to the coefficient benchmarks of Altman (i.e., poor, fair, moderate, good, very good; Altman, 1990). Rater reliability, overall, was calculated to be "very good" for recordings of participants from one treatment site ($\gamma = .85$, $SE = .01$, (95% CI [0.83, 0.87]) and "good" for recordings of participants from the other treatment site ($\gamma = .79$, $SE = .02$, (95% CI [0.75, 0.82]).

Visual Inspection and Statistical Analyses

The analysis of single-case intervention research is strongest when it employs both visual inspection and quantitative comparison methods that include effect size calculation and hypothesis testing (Kratochwill & Levin, 2010). Visual inspection of trend was used for Research Questions 1 and 2. Trend was defined as systematic increase or decrease in outcome values that is either linear or nonlinear (e.g., Manolov & Onghena, 2018). Visual inspection of trend for Research Question 1 was considered if, for individual participants, normalized F3–F2 distance lowered (i.e., showed acoustic improvement) during the time course of treatment visits. Visual inspection of trend for Research Question 2 was considered if, for individual participants, normalized F3–F2 distance was lower (i.e., showed acoustic improvement) during posttreatment probes than pretreatment probes. Visual inspection of overlap was used for Research Question 3. Overlap was deemed to occur if the lines connecting points within one condition crossed with the other series (e.g., Manolov & Onghena, 2018). Research Question 3 was examined if, for individual participants, plotted points representing normalized F3–F2 distance were consistently lower (i.e., showed greater acoustic improvement) in one biofeedback condition versus the other.

We used three different types of statistical analyses to examine our research questions: multilevel modeling (e.g., Harel & McAlister, 2019), effect size (e.g., Beeson & Robey, 2006), and randomization tests (e.g., Rvachew & Matthews, 2017). Multilevel modeling was used for Research Questions 1 and 2. These questions, broadly, examined if the rate of change (i.e., slope) during or following treatment was significantly different from zero for each participant. Specifically, Research Question 1 examined trends related to acquisition of /ɹ/ during the combined treatment package and considers 1,894 tokens. Research Question 2 asked whether the participants exhibited a clinically significant degree of generalization to untrained words after the combined treatment package, compared to their pretreatment performance, and considered a separate set of 1,794 tokens. In the context of multilevel models, it is possible to examine the significance of individual slopes without adjustment for multiple comparisons (Gelman et al., 2012). Both models used restricted maximum likelihood estimation and variance components covariance structure, and were completed in PROC MIXED (SAS Institute Inc, 2018).

For Research Question 1, a multilevel model was fit to quantify individual trajectories of age- and gender-normalized F3–F2 during the course of treatment. These models contained no fixed effects, only random intercepts for subject with a random slope on time (Visit 1–Visit 10). These random intercepts and slopes allow for quantification of baseline performance level and rate of change over the course of treatment for each participant.

Two additional multilevel models were fit for Research Question 2: one for acoustic data and one for perceptual data. These models (again, without fixed effects) were used to estimate participant-specific random intercepts and random slopes. These intercepts and slopes quantified individual trajectories of age- and gender-normalized F3–F2 distance, as well as listener perceptual rating, from pretreatment to posttreatment.

For Research Question 2, standardized effect sizes were also calculated to quantify the amount of change in the acoustic or perceptual measure for each participant following treatment. The difference between pooled baseline and posttreatment means was divided by the standard deviation pooled across baseline and maintenance visits to arrive at the standardized effect size, Busk and Serlin's d_2 (Beeson & Robey, 2006). We selected an effect size that measured *pooled* variance in order to reduce the number of

cases for which a valid effect size cannot be calculated due to zero or near-zero variance in the baseline phase. An effect size of ±1.0 was used as the threshold for a clinically significant response to the overall treatment program, following from previous speech sound treatment research reasoning that a change in mean accuracy must be at least as large as the standard deviation of baseline variability in order to have a meaningful impact on a client's functioning (Maas & Farinella, 2012).

Randomization tests (e.g., Rvachew & Matthews, 2017) were the third statistical method employed. Recall that Research Question 3 concerned the rapid alternation of biofeedback conditions and asked whether participants demonstrated greater improvement in /ɹ/ during one of the two biofeedback treatment conditions, as indexed by acoustic measures of within-treatment performance. Randomization tests were used to evaluate the null hypothesis that there would be no difference in normalized F3–F2 distance between biofeedback conditions for any given participant. No correction of p values for multiple comparisons was needed because independent data sets were used in each randomization test. The test statistic for each participant was the difference between the mean normalized F3–F2 in all ultrasound conditions and the mean normalized F3–F2 in all visual-acoustic conditions ($mean_{US} - mean_{VA}$). The randomization test quantifies how extreme the observed between-condition difference is for a participant, versus a distribution of 2^{10} possible pseudostatistics in which the condition labels "ultrasound" and "visual acoustic" were randomly assigned to all possible options within the statistical block for that participant.[3] The null hypothesis that there is no difference in acoustic improvement in /ɹ/ between biofeedback conditions for a given participant is rejected when the observed test statistic is more extreme (i.e., lower mean normalized F3–F2 in one condition vs. another) than 95% of pseudostatistics generated. Rejection of the null hypothesis would indicate that the difference observed is better described by the actual treatment condition assigned to each session (e.g., ultrasound or visual-acoustic) than randomly assigned condition labels. Randomization tests were completed using the SCRT: Single-Case Randomization Tests package (Bulte & Onghena, 2008) in R.

Finally, we explored how participant factors correlated with response to treatment. Our first exploration examined the relationship between participant factors and response to the biofeedback conditions. We examined how participants' auditory-perceptual acuity and articulatory awareness were associated with the magnitude of the randomization test statistic. Next, we explored the relationship between participant factors (mindset, auditory-perceptual acuity, and articulatory awareness) and generalization. A calculation in the program G* Power (Faul et al., 2013) suggested that—with an alpha of .05, a power of .8, and a sample size of 7—these explorations were powered to detect significance in correlations with an absolute value greater than .74. We used nonparametric correlations due to the small sample size.

[3]As discussed later, data loss occurred for one visit. That participant, therefore, had 2^9 possible pseudostatistics.

Results

Participant performance on baseline measures and treatment outcomes are summarized in Table 3 and are expanded upon in a supplementary data set available at Open Science Framework. Auditory-perceptual acuity scores and articulatory awareness scores were centered and standardized relative to a sample of 48 typically developing children (aged 9–16 years) who have participated in the normative comparison arm of the current project. Total scores on the Speech Mindset Scale are also presented but were not standardized in this exploration, as children who participated in speech therapy were not included in the normative comparison group.

Research Question 1: Is There Evidence of Acoustic Improvement in /ɹ/ During Practiced Words (i.e., Acquisition) for Some Participants During the Combined Biofeedback Treatment Package?

We visually inspected overall visit performance, regardless of biofeedback condition, to observe whether participants appeared to demonstrate acquisition of /ɹ/ over the course of the combined biofeedback treatment. In the time series line graphs (see Figure 3), acquisition manifests as a downward slope over time, because perceptually accurate /ɹ/ is associated with lower values of F3–F2 distance. The time series line graphs in Figure 3 show individual patterns of within-condition accuracy, based on average normalized F3–F2 distance as measured during nonbiofeedback trials for each treatment condition over time. Audio failure resulted in data loss for one participant for one of the 20 biofeedback conditions (3104, visit 9), so the entire visit was excluded from the randomization analysis. Visual inspection of the visit-by-visit trends present in Figure 3 panes suggests that five subjects made nonlinear progress within each visit over the course of treatment, regardless of biofeedback condition: Participants 3102, 6102, 6103, 6104, and 6108. Three nonlinear trend patterns (e.g., Manolov & Onghena, 2018) were observed: flattened trends (most notably, 6102 and 6103), trends initiated by stable data (most notably, 3102), and alternations between larger and smaller acoustic values (most notably, 6104 and 6108). All trends that were observed occurred in the expected direction (lowering), indicating a therapeutic (rather than maladaptive) acquisition of a more perceptually accurate /ɹ/.

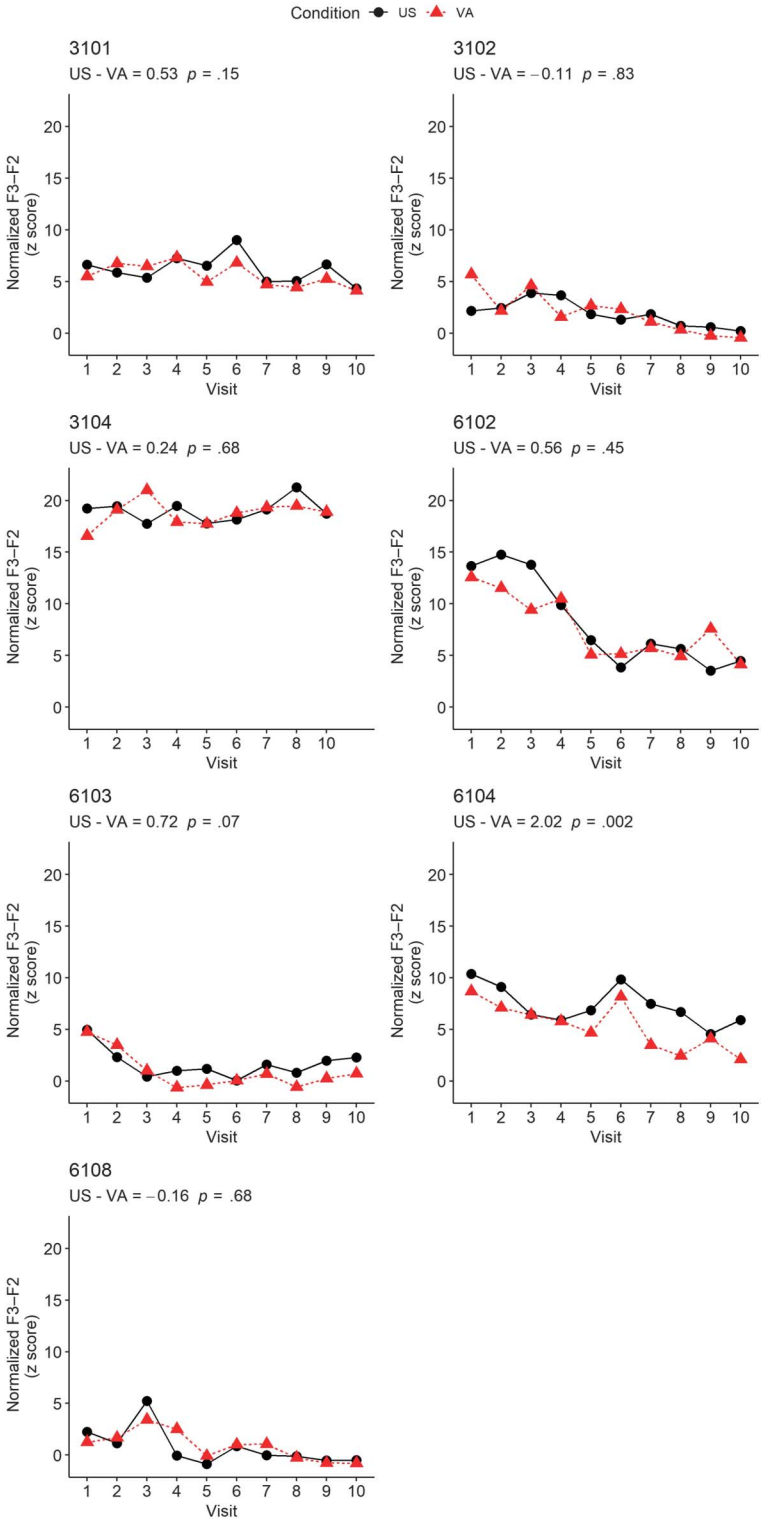
As a more formal counterpart to this visual inspection (e.g., Manolov & Moeyaert, 2017), we used restricted maximum likelihood estimation to model subject-specific intercepts in a multilevel model that allowed for individual variation in intercept (baseline ability) and slope (rate of change across visits), as described above. One thousand eight hundred ninety-four productions informed this model. Slope estimates from a multilevel model predicting normalized F3–F2 distance as a function of visit provide support for six participants demonstrating an overall trend in the expected direction that is significantly different from zero: Participant 3101 ($\hat{\gamma} = -.18$, $SE = .07$, $p = .011$), 3102 ($\hat{\gamma} = -.45$, $SE = .07$, $p < .0001$), 6102 ($\gamma = -1.0$, $SE = .07$, $p < .0001$),

Table 3. Descriptive measures and results.

Participant	Speech Mindset Scale	Auditory-perpetual acuity	Articulatory awareness	Randomization Test Statistic (US - VA)	*p*	Pretreatment F3–F2 *M* (*SD*)	Posttreatment F3–F2 *M* (*SD*)	Standardized acoustic effect size	Pretreatment perceptual mean rating (*SD*)	Posttreatment perceptual mean rating (*SD*)	Standardized perceptual effect size
3101	26	–0.43	–0.44	0.53	0.15	9.37 (4.01)	4.98 (3.46)	–1.17	0.09 (0.19)	0.4 (0.39)	1.01
3102	20	–2.36	0.75	–0.11	0.83	4.6 (2.16)	4.59 (2.55)	0	0.18 (0.27)	0.39 (0.38)	0.64
3104	28	–0.97	–0.44	0.24	0.68	17.11 (3.6)	16.74 (3.98)	–0.1	0.01 (0.05)	0.01 (0.07)	0
6102	20	–4.3	–0.44	0.56	0.45	13.03 (5)	13.14 (6.49)	0.02	0.03 (0.12)	0.04 (0.11)	0.09
6103	27	–0.22	–2.11	0.72	0.07	6.34 (1.91)	1.29 (2.56)	–2.24	0.05 (0.13)	0.79 (0.31)	3.11
6104	16	–3.4	–2.11	2.02	0.002	11.68 (2.88)	11.12 (2.65)	–0.2	0.01 (0.05)	0.01 (0.06)	0
6108	23	0.12	–1.87	–0.16	0.68	7.77 (3.69)	1.14 (2.41)	–2.13	0.02 (0.1)	0.46 (0.35)	1.71

Note. Auditory-perceptual acuity and articulatory awareness are presented as *z* scores, with auditory-perceptual acuity reverse coded such that higher scores represent better acuity. Acoustic improvement (F3–F2 distance) is associated with effect sizes that are more strongly negative. Improvement in perceptual judgment (0 = incorrect, 1 = correct) is associated with effect sizes that are more strongly positive. Acoustic effect sizes < –1 and perceptual effect sizes > 1 were considered to be clinically significant (e.g., Maas & Farinella, 2012). US = ultrasound; VA = visual-acoustic.

Figure 3. Time series line graphs comparing the normalized within-condition F3–F2 distance for each subject. Perceptually correct /ɹ/ productions have lower F3–F2 values. F3–F2 distance is measured in Hertz, but as a *z*-standardized score, the *y*-axis for age- and gender-normalized F3–F2 is unitless. Test statistics and randomization test *p* values are provided for each subject. US = ultrasound; VA = visual-acoustic.



6103 ($\hat{\gamma}$ = − .17, *SE* = .06, *p* = .0038), 6104 ($\hat{\gamma}$ = − .50, *SE* = .06, *p* < .0001), and 6108 ($\hat{\gamma}$ = − .35, *SE* = .05, *p* < .0001). Participant 3104 ($\hat{\gamma}$ =.067, *SE* = .07, *p* = .35), however, did not demonstrate a statistically significant trend during visits of the combined treatment program. Together, both visual inspection and quantitative analysis support acquisition of /ɹ/ for five out of the seven subjects. One participant (3101) exhibited improvement according to quantitative measures that were not salient during visual inspection. One participant (3104) did not demonstrate acquisition of /ɹ/ according to either analysis.

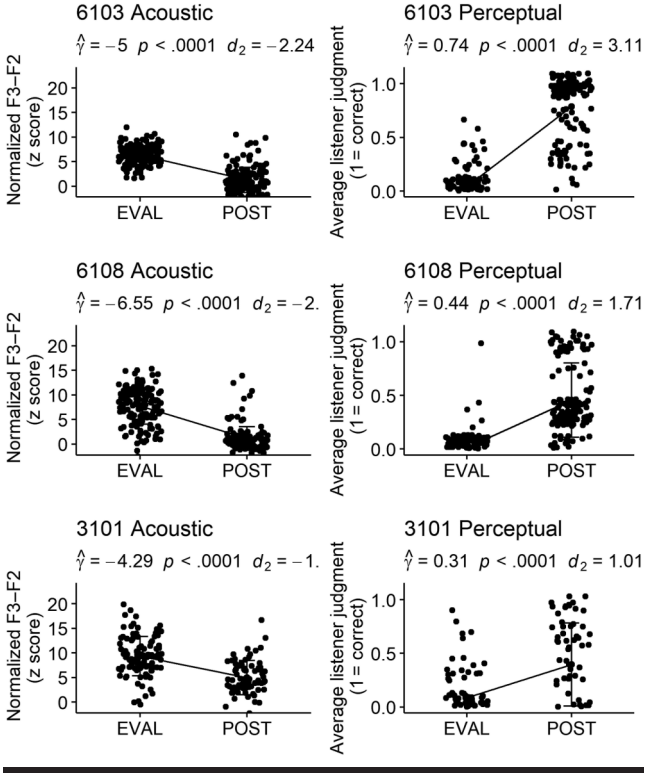
Research Question 2: Do 10 Visits of the Combined Biofeedback Treatment Program Result in Significant Generalization of Correct /ɹ/ to Untrained Words for Some Participants?

Acoustic and perceptual measures of 1,794 /ɹ/ productions were compared for untrained words elicited in probes at pretreatment and posttreatment time points using a multilevel model that allowed for estimates of this pre–post difference to vary across the seven participants. The associated acoustic and perceptual effect sizes, reported in Table 3 and alongside each plot in Figures 4 and 5, show a wide range of variability in overall response to treatment across individuals. Statistically significant *acoustic* change occurred from pretreatment to posttreatment for three participants, with these three participants also demonstrating clinically significant standardized effect sizes: Participant 3101 ($\hat{\gamma}$ = −4.29, *SE* = .54, *p* < .0001, d_2 = −1.17), 6103 ($\hat{\gamma}$ = −5.00, *SE* = .41, *p* < .0001, d_2 = −2.24), and 6108 ($\hat{\gamma}$ = −6.55, *SE* = .41, *p* < .0001, d_2 = −2.13). A second model, with corresponding effect sizes, supported significant *perceptual* changes following treatment for the same three participants: 3101 ($\hat{\gamma}$ = .31, *SE* = .03, *p* < .0001, d_2 = 1.01), 6103 ($\hat{\gamma}$ = .74, *SE* = .02, *p* < .0001, d_2 = 3.11), and 6108 ($\hat{\gamma}$ = .44, *SE* = .02, *p* < .0001, d_2 = 1.71). Acoustic and perceptually rated changes for these three participants are shown in Figure 4. Participant 3102 demonstrated a statistically significant increase in average perceptual rating ($\hat{\gamma}$ = .21, *SE* = .03, *p* < .0001). This effect size, however, was below the threshold for change Maas and Farinella (2012) considered to be clinically significant (d_2 = 0.64). This participant did not show a statistically significant acoustic change. The remaining participants, shown alongside participant 3102 in Figure 5, did not demonstrate a statistically significant acoustic or perceptual change on generalization probes following the combined treatment program.

Research Question 3: Do Some Participants With Rhotic Distortions Show Greater Acoustic Improvement in /ɹ/ During Acquisition With the Use of Either Ultrasound or Visual-Acoustic Biofeedback?

A preliminary mixed model was run to examine the effect of condition number, biofeedback condition, and order of biofeedback condition on the normalized acoustic

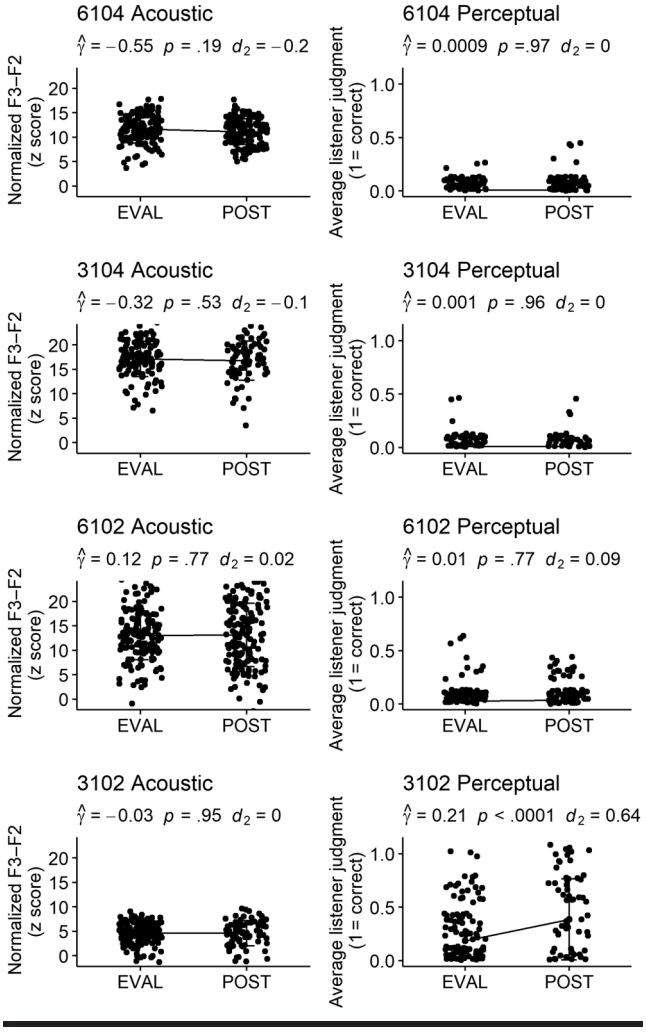
Figure 4. Participants demonstrating significant generalization from pretreatment to posttreatment. Graphs representing change from baseline to posttreatment on acoustic and perceptual measures for three participants judged to demonstrate generalization to untrained words. The coefficient and significance of participant-level random slopes is provided, as well as effect size measures of pre-to-post change. For acoustic response (left side), lower plotted values represent a more adultlike production. F3–F2 distance is measured in Hertz, but as a *z*-standardized score, the *y*-axis for age- and gender-normalized F3–F2 is unitless. The perceptual ratings (right side) reflect the average of three listeners' perceptual judgment: 1 = unanimously correct, 0 = unanimously incorrect. Bars represent standard deviations. Perceptual points have been jittered to prevent overlap obscuring data visualization.



measure. This model did not support a significant effect of condition order ($\hat{\beta}$ = − .024, *SE* = .14, *p* = .87). That is, our data do not provide evidence that participants did significantly better in the second condition of each day, which would be seen if what was learned in the first condition each day transferred to the second condition. We interpret this as evidence that our design successfully minimized carryover within each day of treatment.

The same 1,894 rhotic productions across seven subjects that were used to address to Research Question 1 were also used for Research Question 3. Visual inspection of the time series line graphs in Figure 3 shows the range of variability in participants' mean acoustic response during the ultrasound biofeedback condition (black circles/solid line) or the visual-acoustic biofeedback condition (red triangles/dotted line) presented on the same day. For Participant 6104, lower values of the dotted line connecting the visual-acoustic conditions relative to the solid line

Figure 5. Participants not demonstrating generalization from pretreatment to posttreatment. Graphs representing change from baseline to posttreatment on acoustic and perceptual measures for four participants judged to show no generalization to untrained words. The coefficient and significance of participant-level random slopes is provided, as well as effect size measures of pre-to-post change. For acoustic response (left side), lower plotted values represent a more adultlike production. F3–F2 distance is measured in Hertz, but as a *z*-standardized score, the *y*-axis for age- and gender-normalized F3–F2 is unitless. The perceptual ratings (right side) reflect the average of three listeners' perceptual judgment: 1 = unanimously correct, 0 = unanimously incorrect. Bars represent standard deviations. Perceptual points have been jittered to prevent overlap obscuring data visualization.



connecting ultrasound conditions indicate greater improvement in /ɹ/ during visual-acoustic conditions, as the lines connecting the two conditions do not cross each other. A similar pattern was observed, to a lesser extent, for Participant 6103, whose data demonstrated no instances of overlap within the last seven visits. A different pattern is seen for Participant 6102. For this participant, the conditions are well separated at the beginning of treatment until overlap occurs at Visit 4 and then frequently thereafter.

Randomization test statistics and significance levels for each participant are presented in Table 3 and Figure 3. Of the seven participants, one randomization test reached significance (Participant 6104: $mean_{US} - mean_{VA} = 2.02$, $p = .002$). In examining 6104's response by visit (the statistical blocking unit), a lower normalized F3–F2 distance during visual-acoustic biofeedback conditions was replicated across all 10 visits. The magnitude of the test statistic for Participant 6104 indicated that their age- and gender-normalized F3–F2 distance was, on average, 2.02 standardized units closer to the mean in visual acoustic conditions than in ultrasound conditions. Greater improvement in /ɹ/ during visual-acoustic conditions over ultrasound conditions for this participant was seen regardless of the order of conditions during a visit.

Explorations of Age, Auditory-Perceptual Acuity, Articulatory Awareness, and Growth Mindset Related to Biofeedback Response and Generalization

We explored whether greater improvement in /ɹ/ during the different biofeedback types could be influenced by relative acuity in different sensory domains. Data visualizations are available in Supplemental Materials S6 and S7.

Because ultrasound biofeedback and visual-acoustic biofeedback are hypothesized to target different sensory domains, an asymmetry in sensory skill might be expected for participants with a differential pattern of treatment response. Participant 6104, who did demonstrate a statistically significant /ɹ/ improvement in visual-acoustic sessions versus ultrasound sessions, did not demonstrate the predicted asymmetry between the auditory-perceptual acuity scores and articulatory awareness scores. This participant was notable, however, for demonstrating the poorest global performance, scoring lower than 2 *SD*s below the mean for both the auditory and articulatory measures.

Factors associated with the evidence of generalization after the combined treatment program were also explored. We examined nonparametric correlations between participant-level factors and magnitude of acoustic improvement, indexed by the effect size of the change in acoustically measured accuracy from pre- to posttreatment. Correlations indicate that participant age was not related to generalization (Spearman $\rho = 0.0$, $p = 1$). The correlation of Speech Mindset Score with generalization was moderate in magnitude but not statistically significant in this small sample (Spearman $\rho = -.41$, $p = .36$). Articulatory awareness also had a moderately strong correlation with generalization but was not statistically significant (Spearman $\rho = .69$, $p = .085$). Auditory-perceptual acuity, however, was strongly and significantly related to generalization (Spearman $\rho = -.86$, $p = .024$) in this small *n* sample. Individuals with better performance on the current measure of speech perception—a more consistent ability to assign repeated presentations of ambiguous tokens to the same phonemic category—demonstrated a greater degree of generalization to untrained words following the combined treatment program. Individuals with poorer auditory perceptual acuity along the /ɹ/–/w/ continuum showed poorer

generalization of motor skill to untrained words in response to the combined treatment program.

Discussion

This study primarily sought to compare the relative magnitude of acoustic improvement in /ɹ/ during ultrasound and visual-acoustic biofeedback in children who received both types of treatment, using a single-case randomized block design. Separate aims were to determine if the combined biofeedback treatment program resulted in significant acquisition of /ɹ/ in practiced words and posttreatment generalization to untrained words. We also explored several individual factors as potential correlates of treatment response and nonresponse. While our current observations are preliminary in nature due to the small sample size involved, this study motivates research with larger samples that, in the long term, may inform the individualization of treatment for children with residual speech errors (e.g., Shuster et al., 1992) through the identification of the most effective biofeedback technique considering the client's pattern of strengths and weaknesses in sensory domains and elsewhere.

Our first hypothesis, that participants would demonstrate *acquisition* of /ɹ/ in response to the treatment, was evident in five of the seven participants. For these participants, the acoustic structure of rhotic productions improved within the treatment setting over the course of 10 visits during the combined treatment program. Visual inspection of trend (e.g., Manolov & Onghena, 2018) indicates that acquisition for several participants was nonlinear. Some participants demonstrated a downward (improved) pattern following a stable period. Clinical intuition suggests that these are the children who require a few sessions before treatment benefit is seen. Other participants in this study, however, demonstrated an immediate downward (improved) pattern and then plateaued. Clinical intuition suggests that these are the children who respond quickly to treatment and would benefit most from an approach that adapts in difficulty from the beginning of treatment.

Regarding Research Question 2, only three of the seven participants showed *generalization* to untrained words following the combined treatment program. These three participants provide support for our initial hypothesis and replicate the generalization seen in previous biofeedback studies (e.g., Hitchcock et al., 2017; McAllister Byun, 2017; Preston et al., 2019). Importantly, the present gains were observed following a treatment with substantial *acquisition* focus—that is, one with simple speech targets (i.e., syllables and words), blocked practice during treatment, and frequent KP feedback. While previous studies report a generalization rate of approximately 2/3, the rate of generalization seen in this study was somewhat lower (3/7). Both treatment- and subject-related reasons may contribute to this discrepancy. For example, McAllister Byun (2017) and Preston et al. (2018) utilized treatment paradigms that adapted in difficulty in response to participant performance in order to maintain stimulus presentation around a participant's *challenge point* (e.g., Guadagnoli & Lee, 2004; Hitchcock & McAllister Byun, 2015). This is believed to support motor learning. Stimulus presentation in this study was nonadaptive, however, in order to facilitate direct comparisons between biofeedback conditions. This decision may have contributed to the present generalization rate being lower than that which has been previously reported.

In this combined treatment program, participants experienced two biofeedback conditions each day. Regarding Research Question 3, visual inspection of data overlap indicates that one of the seven participants (Participant 6104) showed greater accuracy in /ɹ/ production (measured by normalized F3–F2 distance) during the visual-acoustic biofeedback relative to the ultrasound biofeedback condition during all 10 treatment visits; this observation was corroborated by the statistical analysis. This finding provides some limited statistical support for our hypothesis that some participants may show dissimilar acoustic improvement in /ɹ/ during different biofeedback conditions. Because this study used a single-case design, however, follow-up research with a larger sample size will be needed to estimate the prevalence of a differential response to biofeedback in a representative sample of children with residual rhotic errors.

Furthermore, we examined whether individuals who demonstrated greater improvement in /ɹ/ during one biofeedback modality might exhibit an asymmetry in auditory and somatosensory acuity. This was grounded in the observation that different biofeedback modalities can be thought of as enhancing different sensory domains (Li et al., 2019) and that typical adults have been observed to exhibit a "sensory preference" between these domains (Lametti et al., 2012). In the present sample, the participant who exhibited a differential biofeedback response scored lower than 2 *SD*s below the mean of a comparison sample of typically developing children on measures of *both* articulatory awareness and auditory-perceptual acuity. Future development of additional auditory and somatosensory measures may help elucidate the relationship between sensory skills and greater improvement in /ɹ/ with a particular modality of biofeedback.

We also sought to replicate previous findings that auditory-perceptual acuity was a predictor of generalization following the combined treatment program. Generalization was indeed predicted by a subject's auditory-perceptual acuity in our small sample: Greater generalization was associated with better performance on an auditory-perceptual measure in which participants identified tokens on a synthetic speech continuum from /ɹ/ to /w/ (e.g., McAllister Byun & Tiede, 2017). The size of this correlation was large and supports previous research suggesting that individuals with relatively better auditory-perceptual acuity are more likely to generalize speech motor improvement to untrained words (Cialdella et al., 2020; Preston et al., 2020). We also explored whether articulatory awareness, age of participants, and growth mindset related to generalization. In this small sample, articulatory awareness was not significantly related to pre-to-post acoustic effect size. Interestingly, however, the correlation between the two was large and, in our sample, the children who had stronger auditory-perceptual

acuity demonstrated a range of below-average articulatory skills. Additional research on larger samples can examine if this is salient in light of our premise that some children do demonstrate differing sensory abilities. Furthermore, age of participants was not related to amount of generalization in the current sample. Another predictor of generalization that was investigated—growth mindset—yielded a moderate correlation in the predicted direction, but this correlation did not reach significance in the present sample size. Thus, additional research with larger sample sizes will be needed to determine whether pretreatment mindset is associated with speech intervention outcomes over and above one's auditory-perceptual acuity.

Although sample size for this current study was relatively small, there are several factors that serve to mitigate the inherent limitations of a small sample. The rapidly alternating treatment, with both conditions focusing on the same speech sound target in the same day, allowed for comparison of the two biofeedback conditions within the same stage of motor learning. Using the same speech sound target alleviates concerns with assumptions that two different speech sounds have the same level of motoric difficulty or are equally easy to remediate (a salient consideration for the complex vocal tract configuration of /ɹ/). Additionally, care was taken to minimize condition-to-condition learning effects through the randomization of distinct word lists to conditions. We also measured response to treatment for both research questions only from productions in which no biofeedback was available to the participants, addressing the concern that correct production might be possible in the immediate context of biofeedback but not once biofeedback is faded. Generalization response was measured on untreated utterances that were not produced in direct imitation. Fidelity results additionally indicate that, overall, the treatment was delivered as designed and the aspects of treatment believed to be clinically potent were provided in the intended manner. Acoustically measured findings for the generalization research question were corroborated by blinded listeners' perceptual ratings. Finally, our acoustic methodology allowed for accurate processing of a high volume of tokens. This process of extracting and measuring formant values, specifically, is a useful method for the acoustic analysis of /ɹ/ in part because the /ɹ/ sound exhibits clear formant structure in a spectrogram. However, this method may be less useful for the acoustic analysis of other phones with less clear formant structure; other acoustic methods and measures can be utilized in those cases.

Clinical Implications

Several findings of this study have potential implications for clinical decision making in the treatment of children with residual rhotic errors. This study demonstrates that biofeedback can lead to acquisition in some children with /ɹ/ errors. A subset of children might demonstrate generalization of the motor plan for /ɹ/ to unpracticed words. The scripts used to introduce the /ɹ/ sound and biofeedback modalities within this study are freely available to clinicians who wish to use these aspects of the study within their practice (https://osf.io/3qf2m/). The study results corroborate other small-scale studies that have found biofeedback, broadly, may facilitate motor plan acquisition and generalization in children with speech sound disorder (e.g., Sugden et al., 2019), including in those who still demonstrate difficulty following previous speech sound treatment. Furthermore, this study expands on previous studies by showing that acquisition and generalization can occur as a result of a treatment program that combines two biofeedback types.

These previous biofeedback studies, however, do not directly compare biofeedback modalities. The current investigation provides evidence that some children—the majority in the current sample—may not demonstrate a statistically significant difference in performance to ultrasound or visual-acoustic biofeedback. However, this study provides the first evidence that some children—only one in the current sample of seven—may be better suited for one biofeedback modality over the other. This suggests that either modality could be adopted in an evidence-based practice setting, to be combined with the clinician's internal evidence and the client's considerations during treatment planning. This evidence also suggests that, if the clinician deems that insufficient progress is being made with one biofeedback tool, evaluating the client's response to another biofeedback tool may be warranted. Within this study, we found that the biofeedback modality can be switched and children can continue to show progress in acquisition and generalization.

It is also noteworthy that we found no evidence of an association between age and magnitude of treatment response in this study: Participants at every age showed evidence of generalization. This lends credence to previous studies (i.e., Preston & Leece, 2017) demonstrating that even older adolescents can benefit from high-quality, empirically motivated treatment paradigms. Taken together, the current findings do not support discharging students with rhotic errors from caseload based on age or lack of treatment response, and it is suggested that treatments integrating alternative (biofeedback) modalities and principles of motor learning be considered for inclusion in treatment planning for individuals who have not (yet) demonstrated a treatment response.

Limitations and Future Directions

As a small-scale study utilizing single-case experimental design, this investigation was limited in the range of analyses that could be conducted and questions that could be answered. A large-scale investigation is needed to determine the absolute efficacy of ultrasound biofeedback versus visual-acoustic biofeedback and the client factors (such as sensory skills) that moderate responses to these treatments. Clinically, it would be of interest to examine participant responses if this study provided treatment over a longer period of time, or with different treatment parameters. For example, two individuals did not meet our threshold for across-visit acquisition of the target sound. It is unknown if these children would have made gains in a

treatment paradigm with a higher average cumulative intervention intensity and/or a more intense schedule (e.g., Hitchcock et al., 2019; Preston & Leece, 2017). Future studies can examine how manipulating these parameters might support motor skill acquisition for such children.

Regarding Research Question 2, four individuals did not demonstrate evidence of generalization. However, it is important to note that the treatment paradigm used here was designed to focus on the principles of *acquisition*. It is unknown how many of the participants would have demonstrated generalization to untrained words in a paradigm that explicitly targeted motor *learning* (e.g., McAllister et al., 2020), which would involve (a) increasing the complexity of linguistic utterances and (b) increasing the variation between phonological and prosodic context, while (c) decreasing KP feedback. It is also unknown how many participants would have demonstrated generalization if a longer treatment duration was used. Future studies can investigate the impact of such paradigms on response to biofeedback treatment.

Our primary research question, Research Question 3, was measured with a between-series randomized block single-case design. There are advantages to single-case research, as these designs are better able to reflect upon a given individual's response to treatment, as might be seen when working with individuals in a clinical setting. This is particularly important given the need for individualized treatments for children with residual speech errors. With the rapid alternation needed to make between-series, within-subject comparisons, however, there is the potential for carryover from one condition to the other (Kratochwill & Levin, 2010), which could underestimate the difference between conditions. Knowing this possibility, the present investigation included a priori attempts to minimize the carryover of learning from one condition to the other and to understand if the daily order of treatments was related to accuracy. Our analysis showed that the outcome measure in each condition was not related to whether that condition occurred first or second on each day. One participant did demonstrate an acoustic difference in rhotics between treatment conditions, an effect that was replicated within that subject 10 times. It may be noted, however, that this study did not meet Horner's (2005) "conceptual norm" (Kratochwill & Levin, 2010, p. 127) that an effect is demonstrated after three across-subject replications. The nature of the replications required, however, depends on the study design. For the within-subject randomization design, it is key to demonstrate replication within, rather than across, subjects. The randomization test makes it possible to compare the observed within-subject replication relative to an alpha level, which Kratochwill and Levin (2010) characterize as a stronger way to account for threats to internal validity than replication across subjects. Because it is possible, however, that the true difference between treatment conditions for individual participants is underestimated in the current design, it motivates a larger, between-group investigation of the role of individual differences in response to biofeedback conditions.

Furthermore, it would be useful, clinically, to know why some children do not respond to biofeedback (or other treatments). This study was not adequately powered to investigate a priori hypotheses regarding factors predicting response to biofeedback, but did begin to explore these questions in order to inform larger studies. These future studies can better examine the relationships between treatment response, individual predictors, and different parameters of treatment in order to identify evidence-based treatments for those who do not respond to biofeedback.

Conclusions

This study sought to investigate whether some children and adolescents with residual speech errors demonstrated a greater acoustic response to either ultrasound or visual-acoustic biofeedback. The study also examined the extent to which the combined treatment program resulted in posttreatment generalization to untrained words for some individuals. Five out of seven participants demonstrated acquisition during the course of treatment according to visual and statistical analysis. Our primary research question revealed that one of seven participants demonstrated significantly improved /ɹ/ formant structure in the visual-acoustic condition relative to the ultrasound condition. Three of seven participants demonstrated a clinically significant degree of posttreatment generalization to untrained words in response to the combined treatment program. Another relevant finding showed that individuals with better performance on a measure of auditory-perceptual acuity exhibited a greater magnitude of generalization learning from biofeedback treatment, a finding consistent with previous research. This study adds to the evidence base that biofeedback can be an effective part of intervention for children and adolescents with residual rhotic speech errors.

Acknowledgments

This project is funded by the National Institute on Deafness and Other Communication Disorders (NIH R01DC017476). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank the participants and families for their time and participation in the study. The authors also gratefully acknowledge Megan Leece, Michelle Turner Swartz, and Laura Ochs for their ongoing contributions to project oversight and treatment, as well as the following individuals involved in data processing and analysis: Sam Ayala, Lauren Bergman, Twylah Campbell, Kelly Garcia, Olesia Gritsyk, Lynne Harris, Benny Herbst, Heather Kabakoff, Maya Kumar, Robbie Lazarus, Haley Rankin-Bauer, and members of NYU BITS Lab. We also appreciate the contributions of study consultants Frank Guenther, Doug Shiller, and Jose Ortiz.

References

Altman, D. G. (1990). *Practical statistics for medical research*. CRC Press. https://doi.org/10.1201/9780429258589

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). "Journal article reporting standards

for quantitative research in psychology: The APA Publications and Communications Board task force Report": Correction to Appelbaum et al. (2018). *American Psychologist, 73*(7), 947. https://doi.org/10.1037/amp0000389

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., Garcia, K., Hitchcock, E., McAllister, T., Leece, M., Wang, Q., & Preston, J. L. (2021). Associations between speech perception, vocabulary, and phonological awareness skill in school-aged children with speech sound disorders. *Journal of Speech, Language, and Hearing Research, 64*(2), 452–463. https://doi.org/10.1044/2020_JSLHR-20-00356

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer* [Computer program] (Version 6.0.43). http://www.praat.org/

Bowers, L., & Huisingh, R. (2018). *LAT-NU: LinguiSystems Articulation Test–Normative Update*. Pro-Ed.

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

Bulte, 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

Cabbage, K. L., Hogan, T. P., & Carrell, T. D. (2016). Speech perception differences in children with dyslexia and persistent speech delay. *Speech Communication, 82,* 14–25. https://doi.org/10.1016/j.specom.2016.05.002

Campbell, H., Harel, D., Hitchcock, E., & McAllister Byun, T. (2018). Selecting an acoustic correlate for automated measurement of American English rhotic production in children. *International Journal of Speech-Language Pathology, 20*(6), 635–643. https://doi.org/10.1080/17549507.2017.1359334

Carnegie Mellon Speech Group. (2014). *Carnegie Mellon Pronouncing Dictionary (CMUdict) Version cmudict-0.7b)*.

Cialdella, L., Kabakoff, H., Preston, J. L., Dugan, S., Spencer, C., Boyce, S., Tiede, M., Whalen, D. H., & McAllister, T. (2020). Auditory-perceptual acuity in rhotic misarticulation: Baseline characteristics and treatment response. *Clinical Linguistics & Phonetics, 35*(1), 19–42. https://doi.org/10.1080/02699206.2020.1739749

Derdemezis, E., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., & Bolt, D. M. (2016). Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *American Journal of Speech-Language Pathology, 25*(3), 335–354. https://doi.org/10.1044/2015_AJSLP-15-0020

Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. *Journal of Personality and Social Psychology, 36*(5), 451–462. https://doi.org/10.1037/0022-3514.36.5.451

Diener, C. I., & Dweck, C. S. (1980). An analysis of learned helplessness: II. The processing of success. *Journal of Personality and Social Psychology, 39*(5), 940–952. https://doi.org/10.1037/0022-3514.39.5.940

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*(2), 256–273. https://doi.org/10.1037/0033-295x.95.2.256

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2013). *G* Power Version 3.1.7* [Computer software]. Uiversität Kiel.

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

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Ghosh, S. S., Matthies, M. L., Maas, E., Hanson, A., Tiede, M., Ménard, L., Guenther, F. H., Lane, H., & Perkell, J. S. (2010). An investigation of the relation between sibilant production and somatosensory and auditory acuity. *The Journal of the Acoustical Society of America, 128*(5), 3079–3087. https://doi.org/10.1121/1.3493430

Goldman, R., & Fristoe, M. (2015). *Goldman-Fristoe Test of Articulation–Third Edition*. Pearson.

Guadagnoli, M., & Lee, T. (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

Guenther, F. H. (2016). *Neural control of speech*. MIT Press.https://doi.org/10.7551/mitpress/10471.001.0001

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics.

Gwet, K. L. (2019). *irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC)*. R package version 1.0. https://cran.r-project.org/web/packages/irrCAC/vignettes/overview.html

Harel, D., & McAllister, T. (2019). Multilevel Models for Communication Sciences and Disorders. *Journal of Speech, Language, and Hearing Research, 62*(4), 783–801. https://doi.org/10.1044/2018_JSLHR-S-18-0075

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872

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

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., & Reda, J. (2020, October 21-24). *A new ultrasound probe stabilizer for speech therapy* [Conference presentation]. Ultrafest IX, Virtual Conference.

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*(02), 124–137. https://doi.org/10.1055/s-0039-1677763

Hoffman, P. R., Daniloff, R. G., Bengoa, D., & Schuckers, G. H. (1985). Misarticulating and normally articulating children's identification and discrimination of synthetic [r] and [w]. *Journal of Speech and Hearing Disorders, 50*(1), 46–53. https://doi.org/10.1044/jshd.5001.46

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional children*, *71*(2), 165–179.

Kawahara, H., Morise, M., Banno, H., & Skuk, V. G. (2013). Temporally variable multi-aspect N-way morphing based on interference-free speech representations. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (pp. 1–10). IEEE. https://doi.org/10.1109/APSIPA.2013.6694355

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S., Rindskopf, D. M., & Shadish, W. R. (2010). *What Works Clearinghouse: Single-case design technical documentation*. http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

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.

Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *Journal of Neuroscience, 32*(27), 9351–9358. https://doi.org/10.1523/JNEUROSCI.0404-12.2012

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

Lennes, M. (2003). *Collect_formant_data_from_files. praat* [Praat script].

Li, J. J., Ayala, S., Harel, D., Shiller, D. M., & McAllister, T. (2019). Individual predictors of response to biofeedback training for second-language production. *The Journal of the Acoustical Society of America, 146*(6), 4625–4643. https://doi.org/10.1121/1.5139423

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)

Manolov, R., & Moeyaert, M. (2017). How can single-case data be analyzed? Software resources, tutorial, and reflections on analysis. *Behavior Modification, 41*(2), 179–228. https://doi.org/10.1177/0145445516664307

Manolov, R., & Onghena, P. (2018). Analyzing data from single-case alternating treatments designs. *Psychological Methods, 23*(3), 480–504. https://doi.org/10.1037/met0000133

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

McAllister, T., Hitchcock, E. R., & Ortiz, J. (2020). Computer-assisted challenge point intervention for residual speech errors. *SIG 19 Perspectives on Speech Science, 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), 66. https://doi.org/10.1186/s12887-020-1941-5

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech 2017,* 498–502. https://doi.org/110.21437/Interspeech.2017-1386

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Murray, E., McCabe, P., & Ballard, K. J. (2020). The influence of type of feedback during tablet-based delivery of intensive treatment for childhood apraxia of speech. *Journal of Communication Disorders, 87,* 106026. https://doi.org/10.1016/j.jcomdis.2020.106026

McNutt, J. C. (1977). Oral sensory and motor behaviors of children with /s/ or /r/ misarticulations. *Journal of Speech and Hearing Research, 20*(4), 694–703. https://doi.org/10.1044/jshr.2004.694

Mellor, D., & Moore, K. A. (2014). The use of Likert scales with children. *Journal of Pediatric Psychology, 39*(3), 369–379. https://doi.org/10.1093/jpepsy/jst079

Moser, J. S., Schroder, H. S., Heeter, C., Moran, T. P., & Lee, Y.-H. (2011). Mind your errors: Evidence for a neural mechanism linking growth mind-set to adaptive posterror adjustments. *Psychological Science, 22*(12), 1484–1489. https://doi.org/10.1177/0956797611419520

Ohde, R. N., & Sharf, D. J. (1988). Perceptual categorization and consistency of synthesized /r-w/ continua by adults, normal children and /r/−misarticulating children. *Journal of Speech and Hearing Research, 31*(4), 556–568. https://doi.org/10.1044/jshr.3104.556

Ortiz, J. A. (2017). *Perceptual task* [Computer software]. WordPress. http://joseaortiz.com/software

Park, H. W., Rosenberg-Kima, R., Rosenberg, M., Gordon, G., & Breazeal, C. (2017). Growing growth mindset with a social robot peer. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 137−145). IEEE. https://doi.org/10.1145/2909824.3020213

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science, 26*(6), 784–793. https://doi.org/10.1177/0956797615571017

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

PENTAX Medical. (2019). *Sona-Match* [software]. https://www.pentaxmedical.com/

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., &Schwarz, P. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

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., Leece, M. C., & Maas, E. (2017). Motor-based treatment with and without ultrasound feedback for residual speech-sound errors. *International Journal of Language & Communication Disorders, 52*(1), 80–94. https://doi.org/10.1111/1460-6984.12259

Preston, J. L., McAllister, T., Phillips, E., Boyce, S., Tiede, M., Kim, J. S., & Whalen, D. H. (2018). Treatment for residual rhotic errors with high- and low-frequency ultrasound visual feedback: A single-case experimental design. *Journal of Speech, Language, and Hearing Ressearch, 61*(8), 1875–1892. https://doi.org/10.1044/2018_JSLHR-S-17-0441

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

Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., & Maas, E. (2014). Ultrasound visual feedback treatment and practice variability for residual speech sound errors. *Journal of Speech, Language, and Hearing Research, 57*(6), 2102–2115. https://doi.org/10.1044/2014_JSLHR-S-14-0031

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Revelle, W. (2019). *psych: Procedures for personality and psychological research, Northwestern University, Evanston, Illinois, USA*. R package version 1.9.12. http://CRAN.R-project.org/package=psych

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., & Jamieson, D. G. (1989). Perception of voiceless fricatives by children with a functional articulation disorder. *Journal of Speech and Hearing Disorders, 54*(2), 193–208. https://doi.org/10.1044/jshd.5402.193

Rvachew, S., & Matthews, T. (2017). Demonstrating treatment efficacy using the single subject randomization design: A tutorial and demonstration. *Journal of Communication Disorder, 67,* 1–13. https://doi.org/10.1016/j.jcomdis.2017.04.003

SAS Institute Inc. (2018). *Statistical Analysis Software (SAS)*. Version 9.04.01.

Schroder, H. S., Fisher, M. E., Lin, Y., Lo, S. L., Danovitch, J. H., & Moser, J. S. (2017). Neural evidence for enhanced attention to mistakes among school-aged children with a growth mindset. *Developmental Cognitive Neuroscience, 24,* 42–50. https://doi.org/10.1016/j.dcn.2017.01.004

Schroder, H. S., Moran, T. P., Donnellan, M. B., & Moser, J. S. (2014). Mindset induction effects on cognitive control: A neurobehavioral investigation. *Biological Psychology, 103,* 27–37. https://doi.org/10.1016/j.biopsycho.2014.08.004

Shriberg, L. D., Lohmeier, H. L., Campbell, T. F., Dollaghan, C. A., Green, J. R., & Moore, C. A. (2009). A nonword repetition task for speakers with misarticulations: The Syllable Repetition Task (SRT). *Journal of Speech, Language, and Hearing Research, 52*(5), 1189–1212. https://doi.org/10.1044/1092-4388(2009/08-0047)

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.

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

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

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., Horner, R., Kazdin, A., Kratochwill, T., McDonald, S., Sampson, M., Shamseer, L., Togher, L., Albin, R., Backman, C., Douglas, J., Evans, J. J., Gast, D., Manolov, R., . . . Wilson, B. (2016). The Single-Case Reporting guideline In BEhavioural interventions (SCRIBE) 2016 statement. *Journal of School Psychology, 56,* 133–142. https://doi.org/https://doi.org/10.1016/j.jsp.2016.04.001

TELEMED Medical Systems. (2019). *Echo Wave II. Version 4.0*. https://www.telemedultrasound.com/

Wagner, R., Torgesen, J., Rashotte, C., & Pearson, N. A. (2013). *CTOPP-2: Comprehensive Test of Phonological Processing* (2nd ed.). Pro-Ed. https://doi.org/10.1037/t52630-000

Wechsler, D. (2011). *WASI-II: Wechsler Abbreviated Scale of Intelligence–Second Edition*. APA PsycTests. https://doi.org/10.1037/t15171-000

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

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology, 13*(1), 61. https://doi.org/10.1186/1471-2288-13-61

Appendix A (p. 1 of 3)

Protocol for Articulatory Awareness Task

For an automated version of this task, visit the Open Science Framework (https://osf.io/3qf2m/).
Part A. Do you use the front or the back of your tongue?

Scoring: The sum of correct responses on all tasks.

Instructions for administrator: After reading the instructions to the participant, elicit the remaining items within the frame listed in the "prompt" column. After the child demonstrates the correct production, ask, "Did you use the front or the back of your tongue?" Record and correct incorrect productions. Mark correct responses on the form provided.

Instructions for participant: In English, some sounds are made with the front of the tongue and some are made with the back of the tongue. I am going to ask you to repeat some sounds and tell me if you feel you are using the front or the back of your tongue.

The first sound is 'th', like in 'think.' I want you to say 'th' three times. /θʌ θʌ θʌ/. [Pause.]

While you say it, think about whether feel like you are using the front or the back of your tongue.

Did you use the front or the back of your tongue? [Provide feedback as indicated.]

[Continue to Item 2.]

	Prompt	Correct	Incorrect Production	Comments
PRAC 1	Say /θ/ like in think. /θʌ θʌ θʌ/	f		"Good. You use the front of the tongue to say /θ/." OR "Not quite. You use the front of the tongue to say /θ/." If child seems to be focused on the vowel sound, say "Pay attention to the first sound, the consonant sound, when you make your decision."
2.	Say /z/ like in zero. /zʌ zʌ zʌ/	f		
3.	Say /t/ like in tea. /tʌ tʌ tʌ/	f		
4.	Say /k/ like in king. /kʌ kʌ kʌ/	b		
5.	Say /d/ like in dig /dʌ dʌ dʌ/	f		
6.	Say /s/ like in see /sʌ sʌ sʌ/	f		
7.	Say /l/ like in lead /lʌ lʌ lʌ/	f		
8.	Say /g/ like in go /gʌ gʌ gʌ/	b		
9.	Say /n/ like in new /nʌ nʌ nʌ/	f		

Part B. Which vowel is further back?

Instructions for administrator: After reading the instructions to the participant, elicit the remaining items within the frame listed in the "prompt" column. After the child demonstrates the correct production, ask, "Which one feels further back?" Record and correct incorrect productions. Mark correct responses.

Instructions for participant: We move our tongue in different directions when we make vowel sounds. Sometimes we move the tongue up, down, forward or backward. When we say these sounds, our lips often move, too, but I want you to just pay attention to the direction you feel your tongue moving in. We'll start by focusing on whether your tongue feels forward in your mouth, near your front teeth, or further back. I'm going to ask you to repeat two vowel sounds after me, then say them together three times. Then I want you to tell me if your tongue feels further back for the first or second vowel.
Let's try one. [PRAC 1] Say 'ah' like in 'hot'. [Pause] Say 'ey' like in 'hate'. [Pause] /ɑ eɪ ɑ eɪ ɑ eɪ/ [Pause]
Which one feels further back?

Appendix A (p. 2 of 3)

Protocol for Articulatory Awareness Task

[Continue to PRAC 2]

	Prompt	Correct	Incorrect Production	Comments
PRAC	Say ‘ah’ like in ‘hot’	1		“Good. For /ɑ/ the tongue is further back in the mouth than for /e/.”
1.	Say ‘ey’ like in ‘hate’ /ɑ eɪ ɑ eɪ ɑ eɪ/			OR “Not quite. For /e/ the tongue is forward in the mouth. It’s further back for /ɑ/.”
PRAC	Say ‘ooh’ like in ‘hoot’	1		“Good. For /u/ the tongue is further back in the mouth than for /i/.”
2.	Say ‘ee’ like in ‘heat’ /u i u i u i/			OR “Not quite. For /i/ the tongue is forward in the mouth. It’s further back for /u/.”
3.	Say ‘ey’ like in ‘hate.’ Say ‘oh’ like in ‘hoed’ /eɪ oʊ eɪ oʊ eɪ oʊ/	2		
4.	Say ‘ih’ like in ‘hit’ Say ‘ooh’ like in ‘hoot’ /ɪ u ɪ u ɪ u/	2		
5.	Say ‘ah’ like in ‘hot’ Say ‘ae’ like in ‘hat’ /ɑ æ ɑ æ ɑ æ/	1		
6.	Say ‘ih’ like in ‘pit’ Say ‘oo’ like in ‘put’ /ɪ ʊ ɪ ʊ ɪ ʊ/	2		
7.	Say ‘eh’ like in ‘pen’ Say ‘aw’ like in ‘pawn’ /ɛ ɔ ɛ ɔ ɛ ɔ/	2		
8.	Say ‘ah’ like in ‘pot’ Say ‘eh’ like in ‘pet’ /ɑ ɛ ɑ ɛ ɑ ɛ/	1		
9.	Say ‘ey’ like in ‘fate’ Say ‘oo’ like in ‘foot’ /eɪ ʊ eɪ ʊ eɪ ʊ/	2		

Part C. Which vowel is lower?

Instructions for administrator: After reading the instructions to the participant, elicit the remaining items within the frame listed in the “prompt” column. After the child demonstrates the correct production, ask, “Which one feels lower?” Record and correct incorrect productions. Mark correct responses.

Instructions for participant: Now we’re going to focus on whether your tongue is higher in your mouth, or lower. Remember to focus on your tongue, not your lips or your jaw. We’ll do the same thing as before: repeat each vowel after me, then say them together three times. Then I want you to tell me if your tongue feels lower for the first or second vowel.

[Continue to PRAC 1]

	Prompt	Correct	Incorrect Production	Comments
PRAC	Say ‘ah’ like in ‘hot’	1		“Good. For /ɑ/ the tongue is lower in the mouth than for /i/.”
1.	Say ‘ee’ like in ‘heat’ /ɑ i ɑ i ɑ i/			OR “Not quite. For /i/ the tongue is higher in the mouth. It’s lower for /ɑ/.”
2.	Say ‘ooh’ like in ‘loon’ Say ‘aw’ like in ‘lawn’ /u ɔ u ɔ u ɔ/	2		
3.	Say ‘eh’ like in ‘bet’ Say ‘ee’ like in ‘beet’ /ɛ i ɛ i ɛ i/	1		

(*table continues*)

Appendix A (p. 3 of 3)

Protocol for Articulatory Awareness Task

	Prompt	Correct	Incorrect Production	Comments
4.	Say ‘ooh’ like in ‘hoot’ Say ‘uh’ like in ‘hut’ /u ʊ u ʊ u ʊ/	2		
5.	Say ‘ah’ like in ‘hop’ Say ‘oh’ like in ‘hope’ /ɑ oʊ ɑ oʊ ɑ oʊ/	1		
6.	Say ‘oo’ like in ‘foot’ Say ‘aw’ like in ‘fought’ /ʊ ɔ ʊ ɔ ʊ ɔ/	2		
7.	Say ‘ey’ like in ‘hate’ Say ‘ee’ like in ‘heat’ /eɪ i eɪ i eɪ i/	1		
8.	Say ‘ah’ like in ‘hot’ Say ‘ooh’ like in ‘hoot’ /ɑ u ɑ u ɑ u/	1		

Part D. Which vowel is higher?

Instructions for administrator: After reading the instructions to the participant, elicit the remaining items within the frame listed in the “prompt” column. After the child demonstrates the correct production, ask, “Which one feels higher?” Record and correct incorrect productions. Mark correct responses.

Instructions for participant: Now we’re going to do the same thing, but this time we’re going to flip it around so you tell me which vowel feels higher in your mouth. Remember to focus on your tongue, not your lips or your jaw.

[Continue to Item 1.]

		Correct	Incorrect Production	Comments
1.	Say ‘aw’ like in ‘lawn’ Say ‘ooh’ like in ‘loon’ /ɔ u ɔ u ɔ u/	2		
2.	Say ‘ee’ like in ‘beet’ Say ‘eh’ like in ‘bet’ /i ɛ i ɛ i ɛ/	1		
3.	Say ‘uh’ like in ‘hut’ Say ‘ooh’ like in ‘hoot’ /ʊ u ʊ u ʊ u/	2		
4.	Say ‘ah’ like in ‘hot’ Say ‘ee’ like in ‘heat’ /ɑ i ɑ i ɑ i/	2		
5.	Say ‘oh’ like in ‘hope’ Say ‘ah’ like in ‘hop’ /oʊ ɑ oʊ ɑ oʊ ɑ/	1		
6.	Say ‘aw’ like in ‘fought’ Say ‘oo’ like in ‘foot’ /ɔ ʊ ɔ ʊ ɔ ʊ/	2		
7.	Say ‘ooh’ like in ‘hoot’ Say ‘ah’ like in ‘hot’ /u ɑ u ɑ u ɑ/	1		
8.	Say ‘ee’ like in ‘heat’ Say ‘ey’ like in ‘hate’ /i eɪ i eɪ i eɪ/	1		

Appendix B

Speech Mindset Scale

The Speech Mindset Scale, originally administered via REDCap, is reproduced in text form below.

Skyler and Peyton are siblings. They are going to describe what they like or don't like about school. I want you to think about whether you are more like Skyler or more like Peyton, based on what they said. Think about how often you feel the same way as each sibling, and then click the circle that goes with the answer that best describes your feelings for that one statement. Here are the choices:

- Always like Skyler
- Sometimes like Skyler
- Equal to both
- Sometimes like Peyton
- Always like Peyton

Try the first question, below.

Skyler says, "I like going to gym class." Peyton says, "I do not like going to gym class." Are you more like Skyler or Peyton?

Do you have any questions about the first item? If so, ask the researcher. If not, you can continue with the items below.

1. Skyler says, "I think a person has a certain amount of intelligence and that stays pretty much the same." Peyton says, "I think a person can get smarter and smarter all of the time." Are you more like Skyler or Peyton?
2. Skyler says, "In math class, I like to do math problems that are very easy so I can get a lot right." Peyton says, "In math class, I like to do math problems that are very hard even if I get some wrong, so I can learn more about math." Are you more like Skyler or Peyton?
3. Skyler says, "In speech class, I like to practice words that are very easy so I can get a lot right." Peyton says, "In speech class, I like to practice words that are very hard so I can learn more about making speech sounds." Are you more like Skyler or Peyton?
4. Skyler says, "I like to play games on the easiest levels so I can win every time." Peyton says, "I like to play games on the hardest levels so I can get better at playing the game." Are you more like Skyler or Peyton?
5. Skyler says, "I think some things in school are too difficult, no matter how hard I try." Peyton says, "I think if I work really hard in school, I can do difficult things." Are you more like Skyler or Peyton?
6. Skyler says, "After I make a mistake saying the "r" sound in speech class, I try again the same way." Peyton says, "After I make a mistake saying the "r" sound in speech class, I think about what I can do differently next time." Are you more like Skyler or Peyton?
7. Skyler says, "I like it best when things are really easy to do, and I don't have to try too hard." Peyton says, "I like it best when things are hard, and I have to figure them out." Are you more like Skyler or Peyton?
8. Skyler says, "If I ask questions my teacher will think I'm not smart." Peyton says, "If I ask questions I can learn new things." Are you more like Skyler or Peyton?
9. Skyler says, "If my friends are better than me at something, they will always stay better." Peyton says, "If my friends are better than me at something, I can still get better than them if I work hard." Are you more like Skyler or Peyton?