ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Baseline Stimulability Predicts Patterns of Response to Traditional and Ultrasound Biofeedback Treatment for Residual Speech Sound Disorder

Tara McAllister,[a] Amanda Eads,[a] Heather Kabakoff,[b] Marc Scott,[c] Suzanne Boyce,[d,e] D. H. Whalen,[d,f] and Jonathan L. Preston[d,g]

[a] Department of Communicative Sciences and Disorders, New York University, NY [b] Department of Neurology, Grossman School of Medicine, New York University, NY [c] Department of Applied Statistics, Social Science, and Humanities, New York University, NY [d] Haskins Laboratories, New Haven, CT [e] Department of Communication Sciences and Disorders, University of Cincinnati, OH [f] Program in Speech-Language-Hearing Sciences, Graduate School and University Center, City University of New York, NY [g] Department of Communication Sciences and Disorders, Syracuse University, NY

ARTICLE INFO

Article History:
Received March 13, 2022
Revision received May 10, 2022
Accepted May 10, 2022

Editor-in-Chief: Cara E. Stepp
Editor: Raymond D. Kent

https://doi.org/10.1044/2022_JSLHR-22-00161

ABSTRACT

Purpose: This study aimed to identify predictors of response to treatment for residual speech sound disorder (RSSD) affecting English rhotics. Progress was tracked during an initial phase of traditional motor-based treatment and a longer phase of treatment incorporating ultrasound biofeedback. Based on previous literature, we focused on baseline stimulability and sensory acuity as predictors of interest.

Method: Thirty-three individuals aged 9–15 years with residual distortions of /ɹ/ received a course of individual intervention comprising 1 week of intensive traditional treatment and 9 weeks of ultrasound biofeedback treatment. Stimulability for /ɹ/ was probed prior to treatment, after the traditional treatment phase, and after the end of all treatment. Accuracy of /ɹ/ production in each probe was assessed with an acoustic measure: normalized third formant (F3)–second formant (F2) distance. Model-based clustering analysis was applied to these acoustic measures to identify different average trajectories of progress over the course of treatment. The resulting clusters were compared with respect to acuity in auditory and somatosensory domains.

Results: All but four individuals were judged to exhibit a clinically significant response to the combined course of treatment. Two major clusters were identified. The "low stimulability" cluster was characterized by very low accuracy at baseline, minimal response to traditional treatment, and strong response to ultrasound biofeedback. The "high stimulability" group was more accurate at baseline and made significant gains in both traditional and ultrasound biofeedback phases of treatment. The clusters did not differ with respect to sensory acuity.

Conclusions: This research accords with clinical intuition in finding that individuals who are more stimulable at baseline are more likely to respond to traditional intervention, whereas less stimulable individuals may derive greater relative benefit from biofeedback.

Supplemental Material: https://doi.org/10.23641/asha.20422236

Speech sound disorder (SSD) is a broad label that characterizes individuals whose speech production patterns do not converge with normative expectations for speakers of their language and dialect. Many individuals who present with SSD in early childhood do converge on more typical speech patterns by late school age, either through treatment or through maturation (To et al., 2022). Approximately 25% of children with SSD continue to exhibit atypical speech patterns past the age of 6 years (Shriberg et al., 1999). When these patterns persist past 8–9 years of age, a child might be identified as having a residual SSD (RSSD). RSSD is considered particularly challenging to

Correspondence to Tara McAllister: tkm214@nyu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

treat, and some children are dismissed from clinician caseloads prior to full remediation; in an estimated 1%–2% of the overall population, RSSD continues through adolescence and into adulthood (Culton, 1986; Flipsen, 2015). Deviations in RSSD tend to cluster on late-emerging sounds such as /ɹ/ and /s/ (Ruscello, 1995); our focus here is on /ɹ/, which is often characterized as the most common and most challenging residual deviation for speakers of North American English. This may be due to the articulatory complexity required to produce perceptually accurate /ɹ/. While most sounds are produced with a single major lingual constriction, /ɹ/ requires near-simultaneous dual constrictions: one near the palate/alveolar ridge and one in the pharynx. In addition, there is extensive variation in the tongue shapes used for /ɹ/, both within and across speakers (Boyce, 2015; Delattre & Freeman, 1968; Westbury et al., 1998); in this context, speech-language pathologists (SLPs) may find it challenging to identify successful cueing strategies for a given child. Although the impacts of RSSD on intelligibility may be relatively minor, speech differences have been linked to negative perception by peers (e.g., Crowe Hall, 1991), with potential ramifications for socioemotional well-being (Hitchcock et al., 2015). Developing effective treatments and determining which children show the strongest or weakest response to those treatments is a high priority for clinical research on RSSD.

Traditional motor-based treatment for /ɹ/ can lead to improved speech sound production in some children with RSSD (e.g., McAllister Byun & Hitchcock, 2012; Preston et al., 2017; Ruscello & Shelton, 1979; Shriberg, 1975). In motor-based treatment, the SLP typically provides auditory models, verbal cues, and shaping strategies to encourage appropriate articulator placement for /ɹ/ (e.g., Boyce, 2015; Preston, Benway, et al., 2020; Van Riper & Erickson, 1996). Treatment usually involves repetitive drilling, beginning in simple contexts and proceeding to stimuli of increasing complexity. However, motor-based treatments are not always effective for RSSD, and clinicians have called for improved treatment approaches for this population (Ruscello, 1995).

Ultrasound Biofeedback for /ɹ/ in Children With RSSD

In recent years, considerable attention has been devoted to the possibility that RSSD might be remediated more effectively or efficiently through treatment incorporating visual biofeedback. Biofeedback uses instrumentation to create a real-time visual display of the current function of some physiological process (Davis & Drichta, 1980; Volin, 1998), adding a new modality to the typical sensory experience of producing speech. This is intended to help learners gain more conscious control over processes that are typically highly automatized. It may also be beneficial for individuals with weak perceptual ability in auditory and/or somatosensory domains, which are generally understood to provide the primary targets of speech, as discussed subsequently.

Several biofeedback technologies have been used for treating RSSD, including visual–acoustic biofeedback, which displays the resonant frequencies of the vocal tract (e.g., McAllister Byun & Hitchcock, 2012), and electropalatographic biofeedback, which displays an image of regions of contact between the tongue and the palate (e.g., Hitchcock et al., 2017). Possibly the most thoroughly studied type of biofeedback is ultrasound biofeedback, in which an ultrasound probe placed beneath the chin is used to provide a real-time view of the surface of the tongue. In a systematic review of 28 ultrasound biofeedback treatment studies encompassing over 100 participants with various forms of developmental SSD, Sugden et al. (2019) found that ultrasound biofeedback was associated with positive outcomes in many, but not all, children. There is increasingly high-quality evidence in support of the use of ultrasound biofeedback for RSSD specifically. In a single-case experimental study, Preston et al. (2019) found greater improvements in /ɹ/ production for six children who received eight sessions of ultrasound biofeedback treatment versus six children who received eight sessions of traditional articulation treatment. Several other case studies and single-case experimental studies have also reported that children with RSSD affecting /ɹ/ show benefit from ultrasound biofeedback treatment (Adler-Bock et al., 2007; McAllister Byun et al., 2014; Preston & Leece, 2017; Preston et al., 2014, 2017, 2018).

Despite reporting positive outcomes on average, the existing literature on biofeedback intervention has highlighted heterogeneity in participants' treatment response and has noted that it should be a priority to understand these differences. Virtually all studies to date have reported a mix of strong responders and nonresponders, where response is generally measured based on change in listener ratings of the target sound in untreated word contexts. Lack of response to treatment is commonly reported in roughly one quarter to one third of individuals receiving biofeedback treatment (McAllister Byun & Campbell, 2016; Preston et al., 2019). Research in the "personalized learning framework" suggests that learning can be optimized by identifying profiles of baseline ability that might be linked to different profiles of treatment response (Perrachione et al., 2011; Wong et al., 2017). For instance, Perrachione et al. (2011) found that the optimal paradigm to train participants in a pitch-learning task differed for individuals with high versus low performance in a pitch identification task at baseline. Applying this framework to the study of treatment methods such as biofeedback could make it possible to pair individuals with an approach

expected to be most effective for their profile of strengths and limitations.

Potential Predictors of Treatment Response

Stimulability

To apply the personalized learning framework in the context of ultrasound biofeedback intervention, it is necessary to identify plausible predictors of treatment response. Volin (1998), who measured performance in a nonlinguistic respiratory control task with biofeedback, put forward a clear candidate in the form of baseline stimulability. Stimulability has been defined as "a generalized measure of a child's ability to correct errors in an imitative context" (Miccio et al., 1999, citing Milisen, 1954); it generally involves prompting the learner to imitate the clinician's model of a target sound in isolation or in a simple context such as a nonword syllable. Volin (1998) used participants' performance on the experimental task after a brief period of initial training as the index of stimulability. Participants (n = 36) were randomly assigned to receive further training in a biofeedback condition or a condition in which only verbal knowledge of performance (KP) feedback was provided. Participants with low baseline stimulability tended to make gains in the biofeedback condition but not in the verbal feedback conditions; participants with moderate baseline stimulability showed comparable gains in accuracy in both conditions. However, participants with high stimulability at baseline showed a large decrement in performance when assigned to receive biofeedback training, versus negligible change in the KP feedback condition. Volin hypothesized that biofeedback may interfere with performance in highly stimulable learners by forcing them to recalibrate their motor plan to match an external specification, even though they already possessed an internal representation that was largely successful. These results suggested that biofeedback is most helpful for the least stimulable learners and may be contraindicated for the most stimulable. For moderate stimulability, biofeedback may be beneficial but may not outperform traditional verbal cueing.

The findings from Volin (1998) are consistent with general principles of motor learning (e.g., Maas et al., 2008). Detailed qualitative or KP feedback, of which biofeedback is one type, is thought to be most effective in early stages of learning when the learner is working to understand the parameters of the movement (Newell et al., 1990). As a learner's skill increases, detailed feedback has been associated with diminishing returns and may even inhibit generalization to novel contexts (Ballard et al., 2012; Hodges & Franks, 2001). However, there is relatively little research examining the connection between stimulability and response to biofeedback in the specific context of speech–motor learning. Previous non–biofeedback studies have found a significant relationship between stimulability and generalization during treatment for SSDs (Miccio, 1995; Powell et al., 1991). Higher stimulability is generally associated with greater gains in the absence of treatment (Powell & Miccio, 1996; To et al., 2022), as well as greater progress in the treatment setting (Carter & Buck, 1958; Irwin et al., 1966; but cf. Sommers et al., 1967). In light of this predictive power, stimulability testing is widely used in clinical practice; in a 2019 survey of school-based SLPs, over 80% of respondents indicated that they assessed stimulability as part of a typical evaluation for SSD (Farquharson & Tambyraja, 2019). In the specific context of biofeedback, multiple studies have argued that technologically enhanced feedback may have its greatest impact in the early stages of intervention, when learners are presumed to be at their least stimulable (for ultrasound biofeedback, see Preston et al., 2019; for visual–acoustic biofeedback, see McAllister Byun & Campbell, 2016, and Peterson et al., 2022). However, these studies did not specifically examine individual differences in baseline stimulability as predictors of treatment response.

Sensory Acuity

Since speech is a sensorimotor skill, efforts to predict treatment response should also focus on measuring aspects of sensorimotor control. Following models such as DIVA (Guenther, 2016), HSFC (Hickok, 2012), and FACTS (Parrell et al., 2019), we understand speech as a process of acquiring and refining motor plans to achieve goals in auditory and somatosensory space. These motor plans must be adjusted in response to auditory and somatosensory feedback. Sensorimotor models of speech production predict that speakers who represent a given speech sound with a narrower region in sensory space should also be more precise in their phonetic realization of that sound. Empirical research in recent decades (e.g., Brunner et al., 2011; Ghosh et al., 2010; Perkell et al., 2004; Villacorta et al., 2007) has accumulated a body of evidence that individual variation in speech production does indeed correlate with individual differences in auditory and/or somatosensory sensitivity.

Auditory-Perceptual Acuity

Previous research has found significant associations between individual differences in auditory-perceptual acuity for speech sounds and individual variation in speech production in typical adults (Newman, 2003; Perkell et al., 2004; Villacorta et al., 2007) and children/adolescents (McAllister Byun & Tiede, 2017). In the clinical literature, a 2019 meta-analysis showed that, on average, children

with SSD exhibit lower performance on speech perception tasks than typically developing peers, although perceptual deficits are not universal in SSD (Hearnshaw et al., 2019). Children who misarticulate a sound may have specific difficulty distinguishing correct versus incorrect versions of that target (e.g., Rvachew & Jamieson, 1989); in the specific context of RSSD, Shuster (1998) found that children with RSSD affecting /ɹ/ tended to perceive both correct and misarticulated /ɹ/ as acceptable variants. Recent research has also suggested that auditory-perceptual acuity may act as a significant predictor of response to biofeedback treatment for RSSD affecting /ɹ/. In a randomized controlled trial with 38 children ages 8–16 years receiving ultrasound biofeedback therapy for /ɹ/ misarticulation, Preston, Hitchcock, and Leece (2020) found that pretreatment auditory-perceptual acuity was positively correlated with treatment response for children who participated in ultrasound biofeedback treatment. Better auditory-perceptual acuity was associated with better /ɹ/ production outcomes regardless of whether or not children received speech perception training along with biofeedback treatment. Cialdella et al. (2021) extended this finding in a retrospective study of 59 participants ages 9–15 years who received treatment for RSSD affecting /ɹ/ using various biofeedback technologies. That study found that the children with RSSD showed significantly poorer auditory-perceptual acuity (wider boundary region in an auditory identification task with stimuli along a synthetic continuum from *rake* to *wake*) than 48 age-matched peers with typically developing speech. The study also found a positive correlation between pretreatment auditory-perceptual acuity and treatment response that was significant in female but not in male participants. Finally, in a single-case experimental study of seven participants with RSSD affecting /ɹ/ who received both visual–acoustic and ultrasound biofeedback, Benway et al. (2021) found that individuals with more acute auditory perception at baseline exhibited a larger magnitude of change over the course of treatment. Despite the small sample size, this association was strong and significant (Spearman rho = .86, $p = .024$).

Somatosensory Acuity

Oral somatosensory feedback helps refine sensory targets during speech development. Somatosensory feedback may include information about the position and movements of body structures (proprioception and kinesthesia), as well as the tactile sense of touch and pressure as body structures come into contact with other surfaces (Guenther, 2016). A number of studies have examined oral tactile acuity in connection with speech production. Oral tactile awareness can be measured through stereognosis tasks, in which participants are asked to identify shapes or letters presented in the oral cavity (Attanasio, 1987), as well as through tasks measuring thresholds for the detection of touch or vibration. Children and young adults with RSSD have been observed to show reduced somatosensory acuity relative to typically developing peers on tasks such as two-point discrimination and oral stereognosis (Fucci, 1972; Fucci & Robertson, 1971; McNutt, 1977; Ringel et al., 1968), although other studies have failed to find an association between oral somatosensory acuity and articulation, particularly in younger speakers (Lonegan, 1974; Madison & Fucci, 1971). In addition, Moreau and Lass (1974) found that performance on an oral form discrimination task was significantly correlated with stimulability in a sample of children with SSD.

Purpose and Hypotheses

Following the personalized learning framework, this study aimed to identify individual baseline characteristics that might be associated with distinct profiles of response to different types of treatment for RSSD affecting /ɹ/. All participants received an initial phase of traditional treatment followed by a longer period of ultrasound biofeedback treatment. To evaluate the relationship between stimulability and response to different types of treatment, we administered the same stimulability probe in the pretreatment baseline phase (pretreatment time point), after the initial phase of traditional treatment (posttraditional time point), and again after ultrasound biofeedback treatment (posttreatment time point). We used model-based clustering (Banfield & Raftery, 1993) to identify participants with similar profiles of performance over the course of treatment. We hypothesized that this process would identify at least two clusters reflecting differences in baseline stimulability, relative response to traditional and biofeedback treatment, and/or overall magnitude of response. Based on Volin (1998) and other literature reviewed previously, we predicted that greater stimulability at baseline would be associated with a smaller relative magnitude of response to biofeedback treatment compared to traditional treatment. We then compared the resulting clusters with respect to age, auditory-perceptual acuity, and somatosensory acuity. We hypothesized that the clusters identified on the basis of differences in treatment response would show differences with respect to at least one of our sensory measures, which were selected for their relevance to speech–motor control.

Finally, we obtained blinded listeners' perceptual ratings of /ɹ/ sounds in word probes elicited at the beginning and end of treatment. This allowed us to examine participants' average magnitude of treatment response in a fashion more comparable to prior studies (e.g., Preston et al., 2019), which typically focus on generalization of perceptually accurate production to untrained words spoken

Table 1. Characteristics of 33 children with residual speech sound disorder included in treatment.

Variable	*M*	*SD*	Minimum	Maximum
Age (years;months)	10;7	17.7 mo	9;0	14;7
GFTA-2 standard score	< 73.3	11.2	< 40	91
PPVT-4 standard score	115	14.4	86	152
CELF-4 recalling sentences scaled score	12.1	3.2	7	18
CTOPP-2 phonological awareness composite score	98.4	12.2	65	127

Note. GFTA-2 = Goldman–Fristoe Test of Articulation–Second Edition; PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition; CELF-4 = Clinical Evaluation of Language Fundamentals–Fourth Edition; CTOPP-2 = Comprehensive Test of Phonological Processing–Second Edition.

nonimitatively. We additionally examined whether this overall response magnitude differed across clusters.

Method

This was a single-group study designed to evaluate changes from pre- to posttreatment. Data were collected from two participating sites: Haskins Laboratories and New York University. Institutional review board approval was obtained from the respective institutions. Three certified SLPs completed assessments and treatment at each site: two at New York University and one at Haskins Laboratories.

Participants

Participants between ages 9 and 15 years were recruited through flyers posted throughout the community and through referrals from local SLPs. During an initial screening appointment, parents completed a case history and demographic form. A total of 33 participants met all criteria for inclusion in the study, including 18 at New York University and 15 at Haskins Laboratories. Four additional participants were evaluated but not included in the final study because they did not meet all inclusionary criteria.[1] Participants completed a written consent and assent process.

Inclusionary Criteria

All participants included in the study were native speakers of rhotic dialects of American English, per parent report. Participants were required to demonstrate /ɹ/ distortions in conversational speech and on the Goldman-Fristoe Test of Articulation–Second Edition (GFTA-2; Goldman & Fristoe, 2000), as determined by a certified SLP. To establish a relatively uniform level of severity of RSSD affecting /ɹ/, participants were required to fall below 30% accuracy on a standard list of 50 words assessing /ɹ/ accuracy as scored by the SLP. Participants were required to achieve a minimum score of 80 on the Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007) and to pass a pure-tone hearing screening at 20 dB bilaterally at 500, 1000, 2000, and 4000 Hz.

Speech motor function was assessed using a maximum performance task (see Thoonen et al., 1999). This examination evaluated presence of possible dysarthria and apraxia by evaluating measured duration of sustained phonemes including /ɑ/, /f/ /s/, and /z/; speed and accuracy in repeated production of /pɑ/, /tɑ/, and /kɑ/; and accurate production of /pɑtɑkɑ/ at the quickest rate possible. Participants' performance then fell into one of three categories: A score of 0 was representative of no apparent signs of dysarthria or apraxia, 1 was indicative of an uncertain diagnosis of dysarthria or apraxia, and a score of 2 meant likely dysarthria or apraxia. Participants were required to receive a score below 2 for inclusion in the study.

Once eligibility was confirmed, a second pretreatment assessment visit was conducted. The Recalling Sentences subtest of the Clinical Evaluation of Language Fundamentals–Fourth Edition (Semel et al., 2003) and the Comprehensive Test of Phonological Processing–Second Edition (Wagner et al., 2013) were administered for descriptive purposes. The tasks used to predict treatment response, outlined subsequently, were also administered. Table 1 shows descriptive characteristics for included participants.

Assessment of Auditory-Perceptual Acuity

Auditory-perceptual acuity for /ɹ/ was assessed using the categorical perception task described by McAllister Byun and Tiede (2017). A synthetic nine-step continuum between the sounds /ɹ/ and /w/ was presented in a customized computerized task. Stimuli were generated from natural productions of the words *rake* and *wake* spoken by a 10-year-old female participant. (The endpoints of *rake* and *wake* were chosen because misarticulated productions of

[1]Due to experimenter error, one participant was enrolled in the study despite not meeting all inclusionary criteria (baseline accuracy in the standard /ɹ/ word probe exceeded the maximum cutoff of 30% correct). This participant was identified at the data processing stage, and their scores appeared as extreme outliers relative to other participants. Their data were thus excluded from the analyses and results reported here.

syllable-initial /ɹ/ are often perceived as /w/ or a similar sound such as the labiodental glide /ʋ/, although other distortions also occur.) The steps along the continuum were synthesized from the weighted average of the linear predictive coding coefficients, the gain, and the associated residuals computed from the voiced region of the aligned natural *rake* and *wake* utterances. During the task, participants listened to and categorized each step of the continuum 8 times, presented in random order, for a total of 72 trials. In a two-alternative forced-choice task, each token was presented twice with a 200-ms interstimulus interval, and then participants selected *rake* or *wake* on the computer screen via mouse click. This task was completed with headphones in a soundproof booth or a sound-attenuated room. For each participant, auditory-perceptual acuity was measured by plotting the proportion of presentations of each continuum step that were identified as *rake*. These data points were then fitted to logistic functions via maximum-likelihood estimation. Auditory-perceptual acuity for /ɹ/ was defined as the width of the fitted function from the 25th to the 75th percentile of probability. In this task, a narrower boundary region between *rake* and *wake* is indicative of a higher degree of response consistency, which has been interpreted as evidence of more acute auditory perception (Hazan & Barrett, 2000; McAllister Byun & Tiede, 2017).

Assessment of Somatosensory Acuity

An oral stereognosis task was administered to assess the tactile aspect of somatosensory acuity.[2] This task, which was administered per the specifications in Steele et al. (2014), uses Teflon strips with raised embossed letters ranging in size from 2.5 to 8 mm. The clinician presented each strip to the participant with the top of the letter oriented toward the back of their mouth. Participants, who wore dark glasses to obscure visual access to the strips, were instructed to insert each plastic strip into their mouth and use their tongue tip to identify each letter. Participants received an initial training explaining how the strips would be oriented and describing key characteristics of the stimuli (e.g., letters were capital and could repeat). After comprehension of this information was confirmed, the child completed one practice trial and then proceeded to the main task. Letters were presented in an adaptive staircase fashion in which size increased following an incorrect response and decreased following a correct response. No accuracy feedback was provided. The task was terminated after eight reversals in direction or 28 total trials, whichever occurred first. Following Steele et al. (2014), the score was calculated as the average size of letters in correct trials only. As such, higher scores are indicative of lower somatosensory acuity. Because the protocol and materials for this task were still being prepared at the time this study was initiated, 11 participants completed this task at pretreatment, 14 participants completed the task at posttreatment, and the remaining eight participants were unable to complete the task because their participation ended prior to the availability of task materials. Results pertaining to somatosensory acuity will be interpreted with caution in light of these limitations.

[2]Oral stereognosis can be defined as the ability to recognize shapes or symbols using the sense of touch with oral structures (Fucci & Robertson, 1971). It assesses oral somatosensory function in a more complex and functional way than measures involving the detection of tactile stimuli, such as vibrotactile threshold or two-point discrimination tasks (Jacobs et al., 1998). Stereognosis tasks also permit active manipulation of the object to be identified (e.g., searching the object with the tongue), in contrast with tactile detection tasks, which do not allow movement of the body part being assessed (Lukasewycz & Mennella, 2012). Because oral stereognosis was the only oral somatosensory task administered in this study, we use the general term *somatosensory acuity* in connection with performance on this task.

Treatment

Treatment occurred in one phase of traditional treatment and a phase of ultrasound biofeedback treatment that included a brief period of intensive treatment and a longer period of less intensive treatment focused on generalization. Target selection was based on participant performance on the stimulability probe described below (see details in Supplemental Material S1). For each participant, treatment targets included syllabic /ɝ/ plus the three least accurate contexts, based on the stimulability probe, selected from the following: postvocalic /ɹ/ following a front vowel, postvocalic /ɹ/ following a back vowel, onset /ɹ/ before a front vowel, or onset /ɹ/ before a back vowel. Target contexts remained stable across phases of treatment.

Phase I: Intensive Traditional Treatment

Phase I consisted of intensive non–biofeedback motor-based (henceforth, "traditional") treatment in which participants were provided three 1.5-hr sessions of motor-based speech therapy (with no ultrasound use) within roughly 1 week. Each treatment session consisted of prepractice and structured practice portions. During prepractice, participants received visual and verbal cues aimed at eliciting correct /ɹ/, including descriptions of tongue positioning, phonetic placement cues, and shaping strategies (for examples of such cues, see Preston, Benway, et al., 2020). A document listing suggested cues was provided to the clinicians as guidance; additional strategies were allowed at the clinician's discretion. Targets in prepractice were selected to feature two exemplars for each of the child's four contexts (e.g., /mɝ/, /ɝg/; /ɹi/, /ɹeɪ/; /ɹu/, /ɛɹ/; /ɑɹ/, /ɔɹ/). The structured practice component of each Phase I session began after all target syllables were produced correctly at least 3 times (24 correct productions in total) or

after 50 min, whichever came first. In Phase I, structured practice was guided by the Challenge Point Program (CPP; McAllister et al., 2021). The software presented 96 trials in blocks of six, with targets equally divided across the four chosen contexts for the child. At the start of each block, the clinician provided one verbal model. For each production, the clinician entered a score of correct (1) or incorrect (0) into the software. Based on the clinician's rating, the software displayed knowledge of results feedback on the screen (i.e., "That's right!" or "Not quite") following Trials 1, 3, and 5 in each block, and the clinician was prompted to provide detailed KP feedback at the end of the block. This first phase of treatment was nonadaptive, meaning that all participants received the same practice structure regardless of their accuracy.

Phase II-a: Intensive Ultrasound Treatment

During treatment Phase II-a, participants received three 1.5-hr sessions of intensive ultrasound therapy within a 1-week period. Intensive ultrasound sessions began with instruction about the function of the ultrasound and how to interpret ultrasound images. A Siemens Acuson X300 ultrasound machine with C6–2 transducer was used for biofeedback at the Haskins Site; at the NYU site, the same device was used with a C8–5 transducer. During prepractice, magnetic resonance imaging (MRI) images of the vocal tracts of 22 speakers producing /ɹ/ were displayed (Boyce, 2015; Tiede et al., 2004), and example tongue images were selected for participants to use as visual models. Sample tongue shapes were selected based on the clinician's judgment of which aspects of practiced shapes were facilitative during prepractice. The ultrasound could be used in both sagittal and coronal views. In sagittal view, participants were instructed to focus on elevating the anterior tongue (tip, blade, or anterior dorsum), lower the posterior tongue dorsum, or retract the tongue root to achieve an accurate /ɹ/. In coronal view, the focus was on elevating the sides of the tongue and forming a groove in the midline of the posterior tongue dorsum. No head stabilization was used, but the probe was stabilized in a microphone stand and repositioned as needed by the clinician or participant. Each Phase II-a session began with up to 50 min of prepractice followed by 30 min of structured practice. As in Phase I, participants could advance to structured practice early if they produced 24 perceptually correct utterances. Structured practice in Phase II-a was identical to Phase I in terms of the number of trials and the nature and timing of clinician feedback; the only difference was the inclusion of real-time ultrasound biofeedback.

Phase II-b: Ultrasound Treatment for Generalization

In Phase II-b of treatment, participants received two sessions per week for 8 weeks with ultrasound biofeedback. Sessions were 45 min to 1 hr in duration. The syllables/words targeted in each session were randomly selected by the CPP program from its built-in lists of words/syllables, with three items initially selected to represent each of the participant's four target contexts. Each Phase II-b session began with prepractice that was identical to Phase II-a, except that the duration was limited to 15 min or 24 correct productions, whichever came first. Phase II-b structured practice consisted of 216 /ɹ/ trials, elicited in blocks of six trials, or a cumulative session duration of 1 hr. Phase II-b treatment differed from the previous phases in that structured practice was adaptive; that is, the conditions of practice were adjusted to be more or less challenging based on the child's recent performance. This was accomplished using the CPP software, which was designed to facilitate adaptive practice embodying the principles of the challenge point framework for motor learning (Guadagnoli & Lee, 2004; Matthews et al., 2021). If accuracy in a block of six trials exceeded 80%, the next block would feature an increase in complexity; if accuracy in a block fell below 50%, complexity in the next block would decrease. Within-session changes in complexity could manipulate the frequency of verbal feedback from the clinician (from four to two trials per block) and the complexity of the practiced items (from syllables to words in sentences). Other adaptive changes were made between sessions based on cumulative accuracy in the previous session; these included changes in the order of presentation of practice trials (fully blocked, random blocked, or fully random), and the frequency at which ultrasound feedback was made available (from 80% to 30% to 0% of blocks). See McAllister et al. (2020) for additional details of the CPP software and the complexity hierarchy. CPP settings were carried over from one session to the next so that the complexity at the start of a session was the same as that achieved at the end of the previous session.

Treatment Fidelity

Fidelity in treatment was assessed by blinded trained students who reviewed recordings of 102 pseudorandomly selected treatment sessions, representing roughly 14% of all sessions completed. Pseudorandom selection was used to ensure a balanced representation of participants and phases of treatment across fidelity checks. In each check, the student reviewed a video or audio recording of the session[3] while also inspecting CPP records indicating what stimuli were expected to be presented and what cues were expected to be provided in each block of six trials. They followed a checklist to answer the following questions for

[3]In the ultrasound treatment condition, video was obtained as part of the ultrasound setup process. In the traditional treatment condition, no video was obtained; hence, the fidelity check was performed on the basis of the session audio recording.

each block. (a) Was a verbal model provided at the start of the block? (b) For biofeedback sessions only, was a visual target (MRI image or custom trace) made available? (c) Was the correct number of trials completed? (d) Was biofeedback presented or withheld as indicated by the CPP? (e) Was qualitative feedback provided as indicated by the CPP? Results of the fidelity check are reported in Table 2. All parameters had at least 90% fidelity, indicating an appropriate level of adherence to the stated experimental protocol.

Measurement

Stimulability Probes

A standard stimulability probe eliciting /ɹ/ was administered at the pretreatment, posttraditional, and posttreatment time points. The probe, which is provided in Supplemental Material S1, consisted of 15 syllables that were elicited 3 times each. Nine targets elicited vocalic /ɹ/ (three syllabic /ɝ/; three postvocalic with front vowels, and three postvocalic with back vowels), and six elicited consonantal /ɹ/ in consonant–vowel syllables (three with front and three with back vowels). Syllables elicited in the stimulability probe were excluded from lists of targets used during treatment. In stimulability probe administration, the treating clinician provided a visual and auditory model of each syllable and cued the child to repeat with their "best /ɹ/ sound."

Using Praat software for acoustic analysis (Boersma & Weenink, 2019), the third author and a trained graduate research assistant measured the formant frequencies of the target /ɹ/ intervals. An optimal filter order (e.g., five formants in 5000 Hz, following Praat conventions) was selected for each participant (Burris et al., 2014) by comparing different settings and selecting the one that yielded the best match between automated tracks and visible areas of energy concentration. At that selected formant setting, the researcher selected a time point during which the first three formants of the rhotic interval were judged to be stable and representative of that production, particularly with respect to the third formant (F3). A Praat script (Lennes, 2003) was used to extract measurements of the first three formant frequencies from a 50-ms Gaussian window generated around the selected point. For each production, the difference between the second and third formants (F3–F2) was calculated in Hertz and then converted to *z* scores in order to account for expected differences in formant frequencies related to age and gender. This normalization process was carried out using published means and standard deviations (Flipsen et al., 2001; S. Lee et al., 1999). Normalized F3–F2 distance was selected based on previous research suggesting that it is the acoustic index of rhoticity that corresponds most closely with expert ratings of perceptual accuracy (Campbell et al., 2018; Dugan et al., 2019). To assess reliability, the second author followed the same procedure for 20/99 files (20%) representing a range of participants and time points. Intraclass correlation with single random raters was used to assess the reliability of F3–F2 measurements between the original and remeasured files. The ICC was computed to be 0.95, indicating strong agreement between formant measurements carried out by different individuals (Koo & Li, 2016).

Word Probes

While performance on stimulability probe measures was the primary focus of this study, participants were also recorded producing untreated words containing /ɹ/ in a nonimitative fashion. A fixed list of words (see Supplemental Material S1) was elicited at the beginning and end of the study to assess the overall magnitude of generalization learning over the course of all treatment, including both traditional and ultrasound biofeedback phases. No word probe was administered at the posttraditional time point because minimal generalization learning was expected after such a brief period of treatment. To further increase the real-world relevance of this generalization measure, instead of acoustic measurement, word probes were rated for perceptual accuracy by blinded untrained listeners recruited using the Amazon Mechanical Turk (AMT) crowdsourcing platform. Previous literature has explored online crowdsourcing as a means to obtain ratings of speech productions from members of the general public (Nightingale et al., 2020; Sescleifer et al., 2018). Listeners recruited in this way do not have specific training in speech rating, and they use different equipment to complete the task, introducing an additional source of noise into the rating data. Nevertheless, empirical studies have shown that when responses are averaged over a sufficient number of listeners,

Table 2. Treatment fidelity indicating the percentage of treatment blocks (groups of six trials) meeting the criteria for each treatment site.

Fidelity domain	NYU site (out of 1,100 blocks)	Haskins site (out of 1,638 blocks)
Provided model at start of each block	99.8%	99.1%
MRI or custom tongue image available for reference in the block (biofeedback only)	90.7%	100%
Correct number of trials completed in the block	99.5%	100%

Note. NYU = New York University; MRI = magnetic resonance imaging.

crowdsourced ratings are comparable to trained listener ratings obtained in the laboratory setting. Specifically, McAllister Byun et al. (2015) found that ratings from nine untrained AMT listeners converged with ratings derived from three trained listeners.

Following McAllister Byun et al. (2015), target utterances from audio recordings of each probe session were labeled with a textgrid file in Praat (Boersma & Weenink, 2019) and used to generate individual word-level sound files. These sound files were randomized and presented to listeners on AMT in blocks of 200 words. Each token was independently judged by nine unique listeners, who rated the /ɹ/ sound as correct or incorrect. Listeners were blinded to the time point of elicitation and the identity of the speaker, but they were provided with an orthographic representation of the target word. Listeners were required to have IP addresses within the United States, report speaking English as their primary language, and report no history of speech or hearing difficulty. In addition, they were required to pass a qualification task in which their ratings of 100 /ɹ/ words were compared against ratings from expert listeners (see details in McAllister Byun et al., 2015). Finally, each block of 200 files included 20 catch trials that were judged to be unequivocally correct or incorrect based on previous expert ratings. The listeners on AMT needed to agree with the expert rater judgment on at least 80% of these tokens for their data to be included. A total of 26 raters (17 females and 9 males) contributed to the task, with each rater completing an average of 11.2 blocks (SD = 8.5, range: 1–28). Raters had a mean age of 40.8 years (SD = 9.2 years, range: 27–60 years).

Analyses

The data set used for the clustering analysis consisted of three measurements for each participant: mean normalized F3–F2 distance for the stimulability probes administered at pretreatment, posttraditional, and posttreatment time points. These three raw measures were used as input to a principal components analysis (PCA) in R (R Core Team, 2019). PCA identifies orthogonal components characterized by sequentially maximal variance derived from a set of measures. While PCA is commonly used for dimensionality reduction, due to the low dimensionality of our raw data (only three observations per participant), PCA was used to rotate the data with no change in dimensionality. The principal component (PC) scores were then used as input to unsupervised model-based clustering analyses using the "mclust" package (Scrucca et al., 2016). The optimal number of clusters was selected based on Bayesian information criterion (BIC) values (Schwarz, 1978) as suggested by Fraley and Raftery (1998) and supported by visual inspection of classification plots. The resulting clusters were qualitatively examined to characterize their respective patterns in terms of baseline stimulability and changes over the course of treatment. They were also compared with respect to age, auditory acuity, and somatosensory acuity using a single multivariate logistic regression model and visual inspection of boxplots. A final set of analyses examined participants' generalization learning as indicated by blinded listeners' perceptual ratings of probes composed of untreated words. Scores at pre- and posttreatment were compared across clusters using a multivariate t test (Hotelling's T^2), and magnitude of change from pre- to posttreatment was compared across clusters using an independent-samples t test. Complete code and deidentified data to reproduce the analyses reported here can be found on the Open Science Framework at https://osf.io/fe3pr/.

Results

Stimulability Probes

Figure 1 shows each participant's acoustically measured accuracy (normalized F3–F2 distance) for stimulability probes at the pretreatment, posttraditional, and posttreatment time points. The median F3–F2 distance is represented with a horizontal bar at each time point. Recall that because F3–F2 distance is smaller in perceptually accurate /ɹ/ sounds, this measure decreases as accuracy increases. In Figure 1, the group median decreases slightly in the brief initial phase of intensive traditional treatment, followed by a larger reduction in the longer phase of treatment featuring ultrasound biofeedback. At all time points, the plot shows considerable heterogeneity across participants.

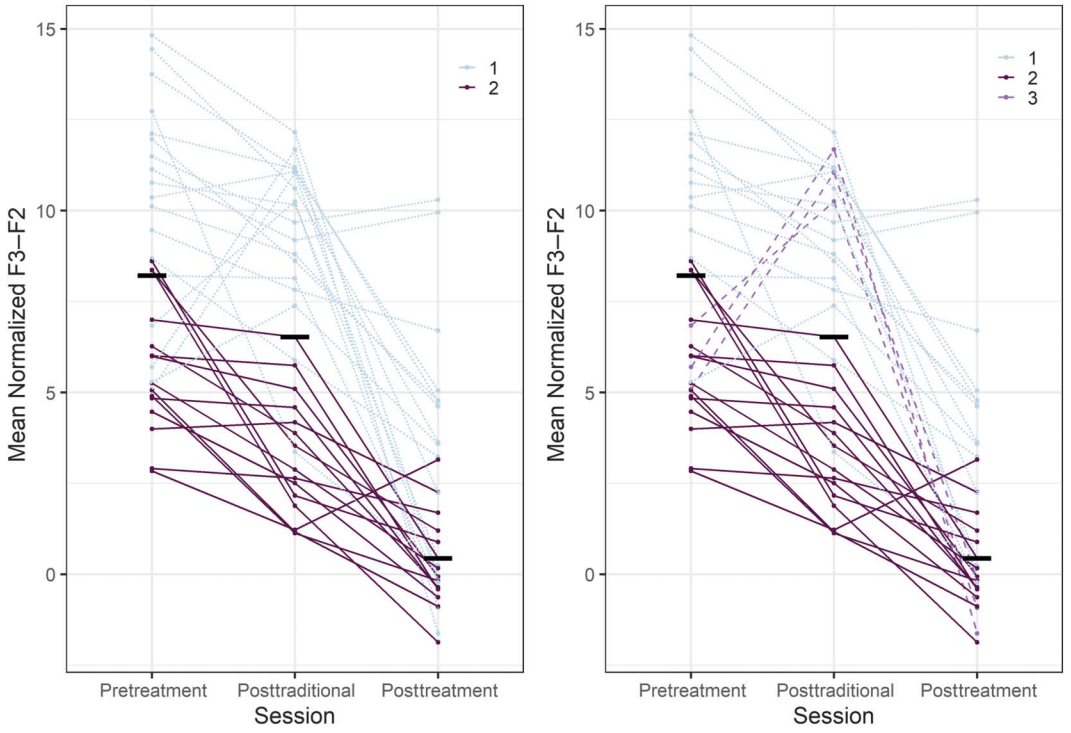
PCA

PCA was applied to the acoustic measures (mean normalized F3–F2 distance for each participant at each of the three time points) to rotate the space to a coordinate system where variance is maximized along uncorrelated axes. The resulting PCs are linear combinations of the original features. All three time points were found to load relatively equally onto PC 1. Posttreatment acoustically measured accuracy loaded most strongly on PC 2 (where it contrasted with posttraditional accuracy). Pretreatment accuracy was not found to load on PC 2 but was the strongest factor for PC 3. The complete table of loadings of original variables on PCA components is reported in Supplemental Material S2.

Cluster Analysis

Scores resulting from the previously described PCA were used as input to a model-based clustering algorithm.

Figure 1. Individual participants' acoustically measured accuracy in stimulability probes at pretreatment, posttraditional (after Phase I), and posttreatment (after Phase II-b) time points. Horizontal bars mark the median at each time point. Line color and type reflect classification in a two-cluster (left panel) and a three-cluster (right panel) model.

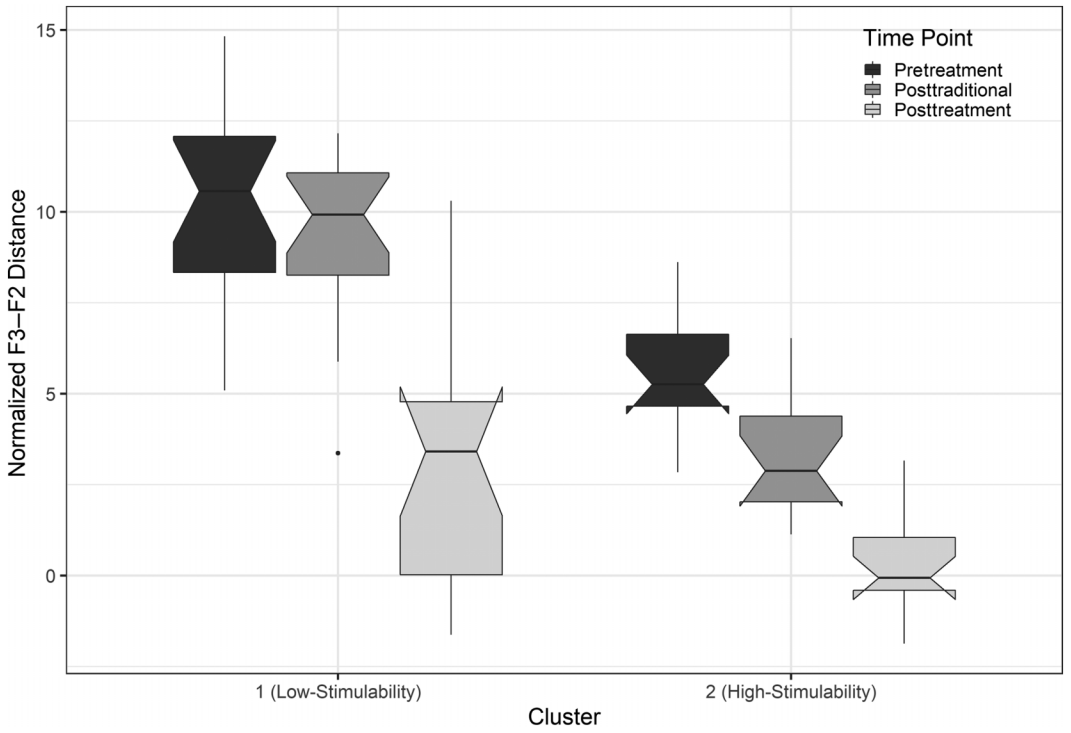


The "mclust" package (Scrucca et al., 2016) examines a range of choices for the number of groups under different covariance structures and selects the best solution using the BIC. The optimal solution identified three clusters with VVI covariance structure (diagonal, variable volume, and shape); however, a very similar BIC score was obtained by a two-cluster solution with VII covariance structure (spherical, variable volume, and equal shape). These solutions were further evaluated through visual inspection of plots of each participant's accuracy over time, with trajectories colored by classification (see Figure 1). Figure 1 shows that individuals in the third cluster do pattern differently from the others, but the cluster includes only three participants. Therefore, our primary analyses will focus on the two-cluster solution, with a brief exploratory consideration of the three-cluster option.

Figure 2 shows boxplots of accuracy for each cluster at each time point in the two-cluster solution. The first cluster (n = 18) is characterized by very low initial accuracy; the median normalized F3–F2 distance is 10.6, indicating that participants tended to produce /ɹ/ with substantially greater separation of F2 and F3 than their age-matched peers. After the intensive phase of traditional treatment, the median normalized F3–F2 distance decreased to 9.9, indicating a very small improvement in accuracy. Finally, after ultrasound treatment, normalized F3–F2 distance dropped substantially. However, the median z score is still 3.4, indicating that age-typical levels of acoustic accuracy in /ɹ/ production were not attained. The second cluster (n = 15) is characterized by a lower median F3–F2 distance of 5.3 at baseline, indicating that these participants started with a higher level of accuracy than those in the first cluster. This group of participants made a larger magnitude of improvement in the traditional phase of treatment than the first group; the median decreased to 2.9 and the interquartile ranges of the boxplots for the pretreatment and posttraditional time points do not overlap. After ultrasound treatment, this group of participants made further progress, finishing with a median z score very close to zero. Thus, at the end of the treatment period, these participants demonstrated virtually typical acoustics of /ɹ/ in the stimulability probe.

For our analyses moving forward, it will be helpful to attach descriptive labels to the clusters of participants. The first cluster, which began with low acoustically measured accuracy and remained well below normative levels of accuracy by posttreatment, will be termed the *low-stimulability group*. The second cluster, which began with higher acoustically measured accuracy and achieved age-appropriate median accuracy by posttreatment, will be called the *high-stimulability group*. Although these descriptive labels characterize the clusters in terms of their initial

Figure 2. Boxplots showing acoustically measured accuracy in stimulability probes at pretreatment, posttraditional, and posttreatment time points for participants in both clusters (two-cluster solution).



presentation (i.e., exhibiting high versus low stimulability at baseline), it is important to keep in mind that scores at all three time points were taken into consideration by the clustering algorithm.

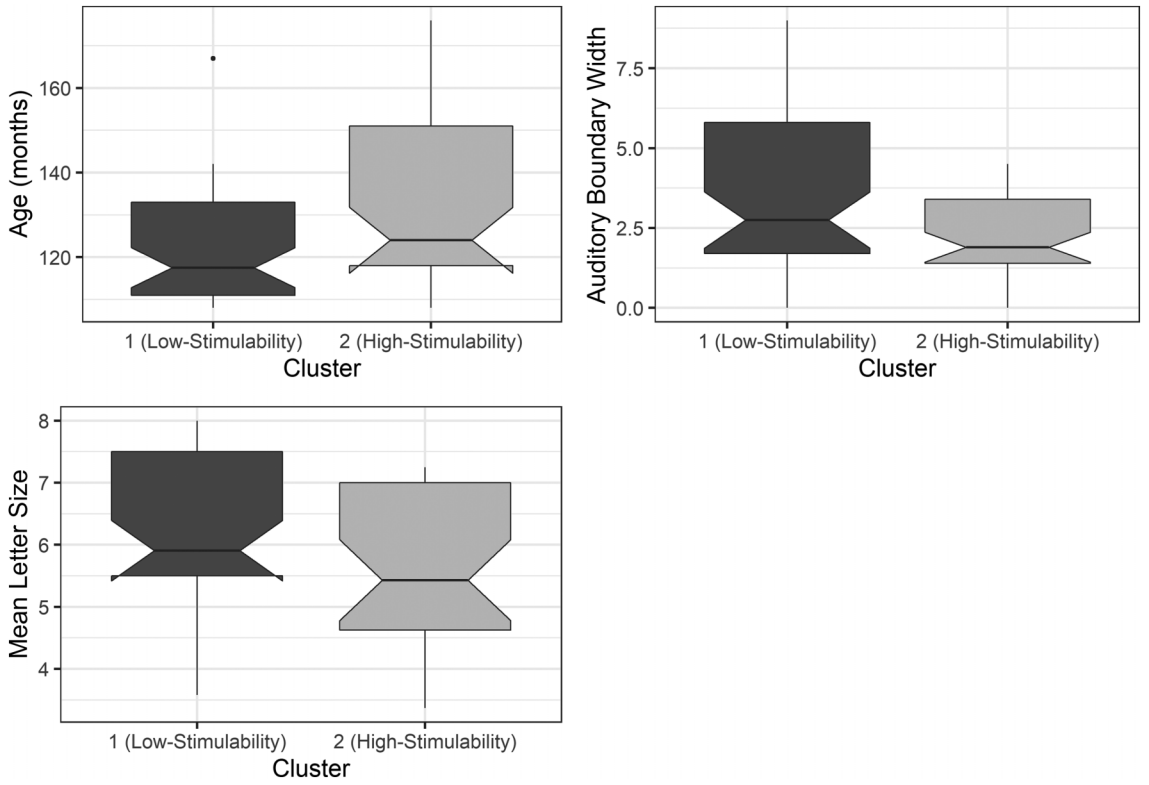
It is of interest to qualitatively compare these two clusters with regard to their relative magnitude of progress in each type of treatment. The low-stimulability cluster showed minimal change in the brief initial phase of traditional treatment, making the majority of their gains in the longer phase of ultrasound biofeedback treatment that followed. By contrast, the high-stimulability cluster made roughly comparable gains in the traditional and biofeedback phases, even though the traditional phase of treatment was much shorter. Because the group median for the high-stimulability cluster reached a typical level of accuracy by posttreatment, it is possible that these participants' magnitude of improvement in the biofeedback treatment condition was limited by a ceiling effect; we return to this point in the Discussion section.

Age and Sensory Characteristics of the Selected Clusters (Two-Cluster Solution)

Figure 3 allows qualitative comparison of age and sensory characteristics across clusters. The low-stimulability cluster had a slightly lower median age than the high-stimulability cluster. With respect to auditory-perceptual acuity, the low-stimulability cluster had a slightly higher median boundary width than the high-stimulability cluster, indicating less acute perception. Likewise, for the stereognosis task, the low-stimulability cluster showed a slightly higher median letter size, indicating lower somatosensory acuity. As previously noted, eight out of the 33 participants were missing data for the stereognosis task, which was still being developed at the start of the study. While these missing data limit statistical power, the comparison is still considered valid because the data are missing completely at random (i.e., the probability of exclusion was the same for all individuals), with the consequence that any comparison of means or likelihood-based models will yield unbiased parameter estimates (Ibrahim & Molenberghs, 2009). Moreover, participants with missing data were evenly distributed across the two clusters, with four in each cluster.

A single multivariate logistic regression was used to test the significance of differences between the two clusters with respect to the characteristics shown in Figure 3. This analysis indicated that the clusters did not differ significantly with respect to age, auditory-perceptual acuity, or somatosensory acuity. Because previous literature suggested an interaction between sex and auditory-perceptual acuity in predicting treatment response (Cialdella et al., 2021), we examined an alternative model with this interaction included; however, that model also yielded no significant predictors. Complete results of both regressions are available in Supplemental Material S3. While it is

Figure 3. Boxplots comparing clusters (two-cluster solution) with respect to age, auditory boundary width in the identification task, and mean letter size in the stereognosis task.



certain that the power of our model to detect significant effects was limited by the large number of predictors relative to the small number of datapoints, visual inspection of the boxplots in Figure 3 corroborates the statistical findings in showing no dramatic differences between clusters with respect to any of the characteristics measured.

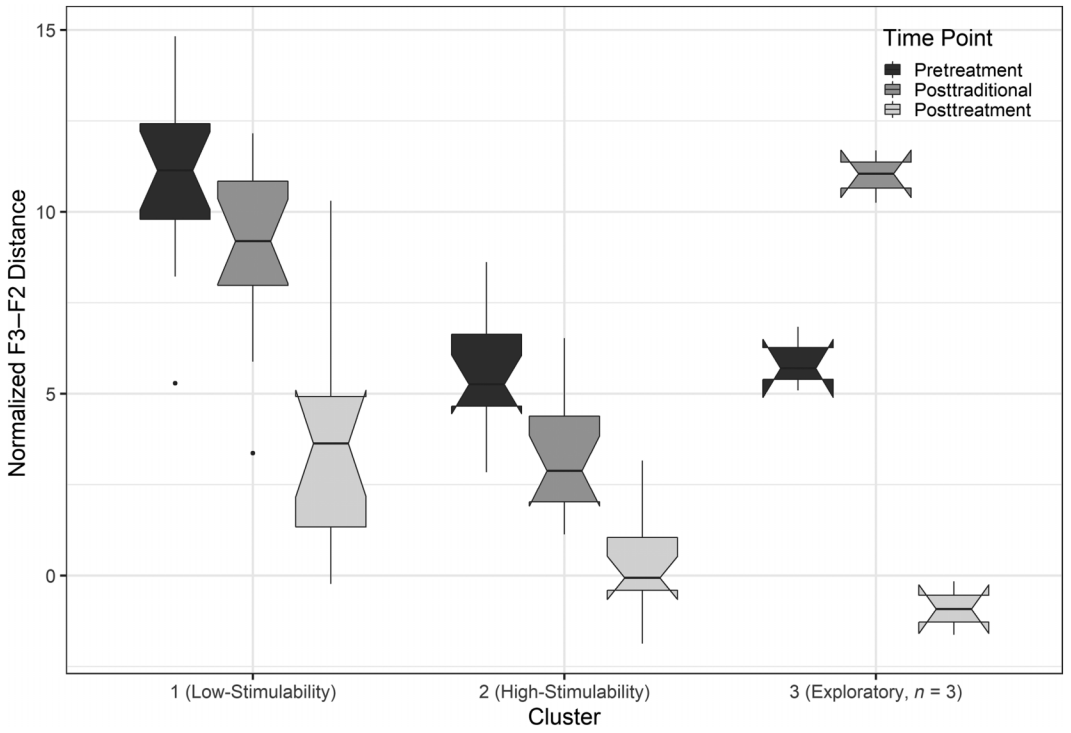
Exploratory Examination of Three-Cluster Solution

As previously indicated, the three-cluster solution must be interpreted with caution because the third cluster contained only three participants; however, we briefly discuss it here as an exploratory direction to inform future research. Figure 4 provides boxplots of acoustically measured accuracy for all time points in the three-cluster solution. At the pretreatment time point, the participants in the third cluster show accuracy similar to the participants in the high-stimulability cluster. However, they contrast with both the low- and high-stimulability clusters in their response to traditional treatment: Instead of making small to moderate gains, these three participants exhibited a maladaptive response, showing a substantially lower level of accuracy at the posttraditional time point relative to the pretreatment time point. During ultrasound treatment, all participants in the third cluster changed course and made dramatic gains, ending with a slightly negative normalized F3–F2 distance (indicating that their /ɹ/ sounds were somewhat hyperarticulated relative to typically developing peers). In the two-cluster solution, the three participants making up the third cluster are merged into the low-stimulability cluster, as shown in Figure 1; no other cluster affiliations change between the two solutions. Because of the small size of the third cluster, we will avoid statistical comparisons between groups in the three-cluster solution. We do note qualitatively that the participants in the third cluster had a younger median age (9;3 [years; months]) than participants in either of the other two clusters (10;0 and 10;4, respectively). However, the children making up this cluster were in fact heterogeneous with respect to age (9;0, 9;3, and 13;11). The third cluster also had the highest median stereognosis score (mean letter size of 6.0 vs. 5.81 for Cluster 1 and 5.43 for Cluster 2).

Word Probes

Although the primary focus of this study was on acoustically measured accuracy of stimulability probes, perceptual ratings of word probes administered at the pre- and posttreatment time points were also obtained for comparability with previous treatment studies. Across

Figure 4. Boxplots showing acoustically measured accuracy in stimulability probes at pretreatment, posttraditional, and posttreatment time points for participants in all clusters (exploratory three-cluster solution).



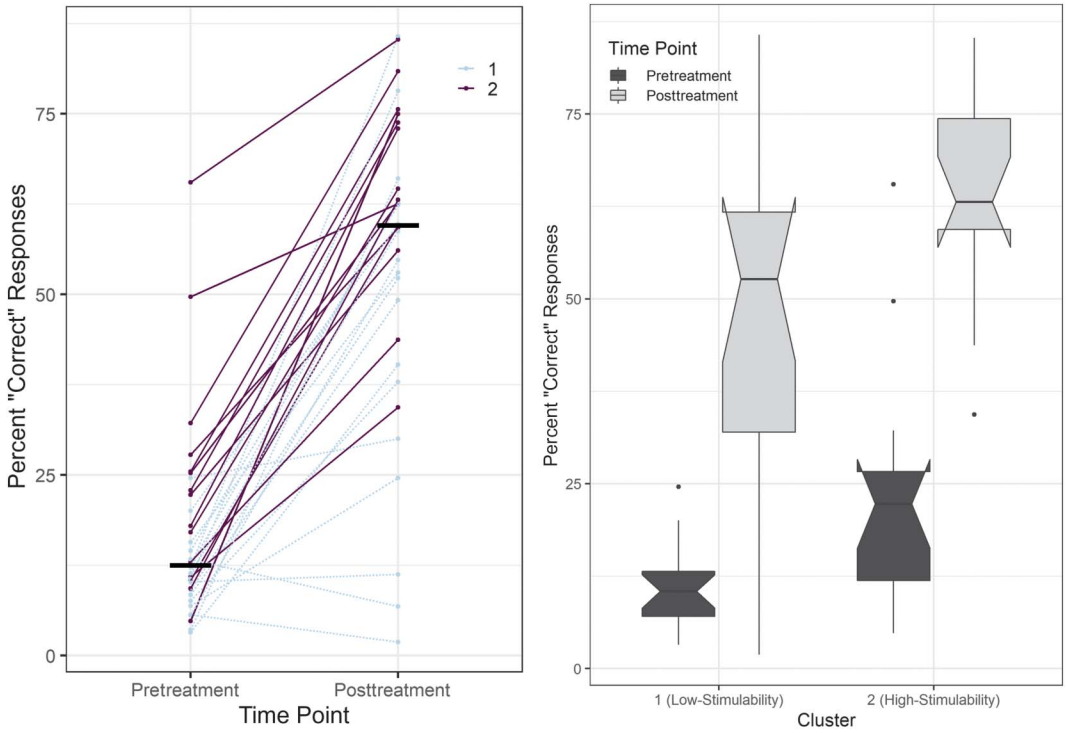
participants, the median perceptually rated accuracy (i.e., percentage of "correct" ratings out of total ratings) was 12.5% at the pretreatment time point, whereas after the end of all treatment, the median accuracy increased to 59.6%. The average raw effect size (change in perceptually rated accuracy from pre- to posttreatment on the generalization probe) was 38.0 percentage points. However, inspection of individual-level data (see Figure 5, left panel) revealed considerable variation, including individuals who began with much higher accuracy than the median across participants,[4] and participants whose accuracy remained low at posttreatment. The two participants with very high perceptually rated accuracy at baseline still made at least some progress over the course of treatment. Three other individuals began with low accuracy and made little or no change over the course of treatment, appearing as low outliers in the posttreatment plot. One additional participant showed a baseline accuracy of roughly 25% and made very little change from pre- to posttreatment. Based on these generalization probe measures, it appears that four of the 33 participants were nonresponders to treatment, whereas the remaining 29 participants made at least some gains over the course of the study.

Associations Between Clusters and Word Probes

We also asked whether the two clusters discussed previously, which were determined based on acoustic measures of stimulability probes, differed with respect to generalization learning as measured by the change in perceptually rated accuracy of words produced at pre- and posttreatment. Figure 5 shows plots of perceptually rated word probe data (shown individually in the left panel and pooled across participants in the right panel) with partitioning by cluster. The patterns shown in Figure 5 are generally compatible with the results from acoustically measured stimulability data. The participants in the low-stimulability cluster started and ended with lower median accuracy than those in the high-stimulability cluster. While some participants in the low-stimulability cluster made large to very large gains, all four nonresponders were part of this group. Participants in the high-stimulability cluster started with higher accuracy and generally did make progress. No participant reached ceiling-level performance, although it is possible that progress in

[4]Two participants show baseline accuracy of 50% or greater, which seems to conflict with our requirement that participants show < 30% accuracy on the baseline word probe. However, inclusion was based on perceptual ratings by a study SLP, whereas these plots reflect ratings by naïve listeners on AMT. Although ratings aggregated across samples of AMT listeners do, on average, converge with trained listener consensus ratings (McAllister Byun et al., 2015), differences in rating behavior between trained and untrained listeners have also been documented, particularly for tokens with a distorted or intermediate quality (e.g., Klein et al., 2012).

Figure 5. Individual and pooled measures of perceptually rated accuracy (mean percent of "correct" responses by blinded listeners) in word probes before and after treatment, partitioned by cluster.



treatment could slow down as the room for growth diminishes at higher levels of accuracy. A multivariate t test indicated that the two clusters differed significantly in their distribution of scores at pre- and posttreatment (Hotelling's $T^2(2, 30) = 13.9$, $p = .004$). However, they did not differ significantly in the magnitude of change from pre- to posttreatment (univariate $t[29.4] = -0.78$, $p = .44$); generalization learning over the course of all treatment was comparable across groups.

Discussion

This study examined patterns of treatment response in 33 children with RSSD affecting /ɹ/ who received 10 weeks of intervention incorporating both traditional and ultrasound biofeedback treatment methods. On average, participants responded positively to the course of treatment; the average raw effect size (change in perceptually rated accuracy in word probes from pre- to posttreatment) was 38 percentage points, and only four participants were judged to exhibit no meaningful response to treatment. The primary goal of this study was to investigate factors that might moderate response to traditional and ultrasound biofeedback treatments. Based on previous research, notably Volin (1998), we were particularly interested in the idea that different trajectories of response to treatment would show an association with differences in baseline stimulability. A stimulability probe was administered prior to any treatment, after a brief initial phase of traditional treatment, and again at the end of the study, following a longer phase of ultrasound biofeedback treatment. Acoustically measured accuracy on these probes was submitted first to PCA and then to a model-based clustering analysis.

Two major clusters were identified. The first cluster was described as the "low-stimulability" cluster and was characterized by very low acoustically measured accuracy at baseline and minimal response to traditional treatment. Most participants in this cluster did make progress over the longer period of ultrasound biofeedback treatment, although their median acoustically measured accuracy at posttreatment still differed from expectations for typically developing children. The second "high-stimulability" cluster was characterized by higher baseline accuracy and a larger magnitude of response to the brief initial phase of traditional treatment. Further progress was observed during ultrasound biofeedback treatment, and median acoustically measured accuracy on stimulability probes at posttreatment was consistent with expectations for typically developing children. The clusters were not found to differ significantly with respect to

age, auditory-perceptual acuity, or somatosensory acuity (with the caveat that missing data for the stereognosis measure used resulted in a low-powered comparison). They also did not differ with respect to the overall magnitude of change in perceptually rated accuracy on word probes administered before and after all treatment as a measure of generalization learning. A third cluster with only three participants was examined on an exploratory basis. Participants in this cluster showed a maladaptive response to traditional treatment but improved in response to ultrasound biofeedback treatment. However, additional data collection will be needed before this pattern of response can be regarded as robust.

Overall, we interpret the results of this study as compatible with predictions based on Volin's (1998) study of response to biofeedback in a nonspeech task. Volin suggested that learners who already possessed an "effective internal representation of the task" (p. 89) did not benefit from biofeedback, in contrast with learners with lower stimulability. This is consistent with general principles of motor learning suggesting that detailed qualitative feedback is most effective for learners who are still working to explore the parameters of a movement task (Newell et al., 1990). In this study, while the high- and low-stimulability groups were not found to differ with respect to sensory acuity, they did differ in their ability to approximate the acoustic target for /ɹ/. At baseline, individuals in the high-stimulability group possessed a motor plan that yielded a partial approximation of the acoustics of /ɹ/; they then tended to show some ability to refine this motor plan in a subsequent period of traditional treatment. Individuals in the low-stimulability group produced /ɹ/ sounds that were acoustically much further from the norm for their age, suggesting that they were still exploring the articulatory parameters needed to achieve successful production of /ɹ/. Consistent with expectations from previous literature, it was this group of learners who showed a stronger response to biofeedback than traditional treatment.

Our results do diverge from those of Volin (1998) in some respects: The previous study documented a large decrement in performance among high-stimulability individuals in the biofeedback condition and suggested that biofeedback may be contraindicated for high-stimulability learners. No comparable negative effect of biofeedback treatment was observed among high-stimulability learners in this study.[5] Rather, the asymmetry in relative response to traditional versus biofeedback types of treatment in this study was driven primarily by the fact that high-stimulability participants tended to respond to both treatments to a similar degree, whereas low-stimulability participants tended to make the majority of their gains in the ultrasound biofeedback condition. This difference may reflect the fact that participants were required to produce /ɹ/ with low accuracy at baseline to be included in this study, whereas Volin (1998) did not have any inclusionary criteria relevant to the experimental measure of interest.

This study additionally tested whether the clusters representing different profiles of treatment response also differed with respect to other factors, with the goal of identifying characteristics that could be used to predict treatment response or make a recommendation for one type of treatment or another. However, none of the measures we investigated (age, auditory-perceptual acuity, and somatosensory acuity) differed significantly across clusters. This was an unexpected finding since previous research using the same identification task administered here has found that auditory-perceptual acuity at baseline is predictive of the magnitude of response to treatment (Benway et al., 2021; Preston, Hitchcock, & Leece, 2020). It is important to note that this study did not directly investigate the relationship between auditory-perceptual acuity and overall treatment effect size[6]: We compared the two clusters, and as indicated previously, they did not differ significantly in overall magnitude of change in accuracy over the course of treatment. Still, we hypothesized that different profiles of response might be associated with differences in auditory-perceptual acuity, yet that hypothesis was not supported. The inclusion of an interaction with participant sex, as suggested by Cialdella et al. (2021), did not change this outcome. Likewise, there were no differences in somatosensory acuity between the clusters identified in this study. It is possible that these null results may reflect limitations of the sensory measures we used, as well as the small sample size. We discuss these and other limitations subsequently.

Caveats and Limitations

We interpreted our findings as suggestive of a difference in relative response to traditional and ultrasound biofeedback treatment methods in children with high versus low stimulability at baseline. However, the two treatments were delivered in a fixed order, and traditional treatment was provided only in a brief intensive phase, whereas ultrasound biofeedback was provided over a longer duration. As a consequence, there is a possible alternative interpretation whereby the high- and low-stimulability groups differed simply in the amount of time needed to

[5]There is one exception to this generalization: One participant in the high-stimulability group had a mean normalized F3–F2 distance of 2.8 at pretreatment, showed a decrease (i.e., improvement) to 1.2 at the posttraditional time point, and then increased to 3.2 at the post-treatment time point, suggesting a maladaptive response to the ultrasound phase of treatment.

[6]This correlation was investigated post hoc and was not significant, Pearson $r(31) = -.20$, $p = .25$.

respond to treatment. The participants in the high-stimulability group responded almost immediately, whereas the participants in the low-stimulability group made progress only at a later date when the effects of treatment had had more time to accumulate. If traditional treatment had been provided over a longer duration, some of these participants might have shown a more favorable response.

The idea that children with lower baseline stimulability simply make progress at a slower pace is not entirely consistent with the observation that the participants in the low-stimulability group showed a larger magnitude of progress than their high-stimulability counterparts in the same duration of ultrasound treatment. On the other hand, the magnitude of change during ultrasound treatment may have been limited by ceiling effects on the stimulability probe for the high-stimulability participants. That is, the participants in the high-stimulability group did tend to respond to biofeedback treatment, and some participants might, in principle, have continued to make progress in this condition if they had not already reached the target pronunciation. However, this does not alter our finding that the high- and low-stimulability groups differed in their patterns of relative response because the groups also differed robustly in their magnitude of response to traditional treatment, and that difference cannot be explained by a ceiling effect.

In short, because of the possibility of both order effects and ceiling effects, we cannot rule out the possibility that the two clusters identified here differ primarily in how fast they respond to treatment of any type, rather than a difference in the relative benefit they derive from two different types of treatment. However, we think that this is the less likely interpretation for several reasons. First, our proposed interpretation is consistent with the theoretically motivated account of the relationship between stimulability and response to biofeedback put forward by Volin (1998). Second, some participants exhibited a response to traditional treatment that was not just neutral but actively negative, followed by a strong positive response during ultrasound treatment. For these participants, at least, there appears to be a qualitative difference between the two treatment conditions. Finally, many of the participants in this study, including those in the low-stimulability group, had received multiple years of traditional treatment without success prior to their involvement in this study. This suggests that simply providing a longer duration of traditional treatment, as opposed to a change in the type of intervention, would be unlikely to produce the large gains observed in Phase II of this study. For a robust investigation of the relationship between stimulability and response to different types of treatment, future studies should consider using a counterbalanced order of treatment delivery or a between-subjects design in which individuals are randomized to receive traditional or biofeedback treatment, potentially with stratification by baseline stimulability.

The lack of significant differences between clusters with respect to sensory acuity may reflect limitations of the measures used here. Many previous studies showing relationships between speech production and auditory-perceptual acuity have used tasks involving discrimination rather than identification of speech sounds (e.g., Perkell et al., 2004; Villacorta et al., 2007). Thus, it is possible that differences between the two clusters would emerge if we used a discrimination task to examine auditory-perceptual acuity. However, previous research using the exact auditory identification task adopted here has in fact shown associations with patterns of treatment response (e.g., Benway et al., 2021; Preston, Hitchcock, & Leece, 2020), which limits the plausibility of this explanation. More robust concerns can be raised regarding the task used here to assess oral somatosensory acuity. First, the stereognosis task functions primarily as an assessment of tactile acuity of the tongue tip. It might thus be expected to predict precision in production of apical consonants such as /t/ or /s/. However, /ɹ/ is produced with limited overall tongue–palate contact and lacks apical contact, with the result that this specific measure of somatosensory acuity is poorly matched to the speech production task of interest. Letter recognition tasks have also been criticized on the grounds that some letters are easier to recognize than others and that differences in cognitive ability or literacy experience could confound the measurement of oral somatosensory acuity (Appiani et al., 2020; J. Lee et al., 2022). The latter factor is of particular concern in the SSD context since children with SSD are at elevated risk for reading difficulties (Tambyraja et al., 2020). Finally, successful performance of the task also requires mental rotation of letters (which are presented with the top of the letter facing the back of the participant's oral cavity), with the consequence that results may show a confounding effect of visuospatial ability. In future research, we will collect additional somatosensory measures that may be more appropriate for our population and speech sounds of interest; see discussion in Gritsyk et al. (2021). Finally, as acknowledged throughout this article, our power to assess associations with oral somatosensory acuity was low due to the fact that the stereognosis measure was not obtained from eight of the 33 participants.

An additional limitation of this study pertains to the small size of the sample (n = 33) available as input to the clustering analysis. As previously discussed, the model-based clustering solution with the absolute lowest BIC score actually featured three rather than two clusters, but the small size of the third cluster prevented us from treating this solution as robust. Visual inspection of the third cluster found that the three participants included in this cluster did show a distinct and interesting pattern of

response; that is, they showed a distinct decrement in accuracy after the brief initial phase of traditional treatment, followed by a very strong response to ultrasound biofeedback treatment. This pattern of response raises the possibility that there may be some participants for whom traditional treatment methods are contraindicated. However, additional data collection on a larger scale will be needed to attach a strong interpretation to this finding and to identify any other factors (e.g., differences in sensory acuity) that could be used to identify individuals as likely to exhibit this pattern of response. We also note that the small size of our sample appears to have limited our ability to discern other subgroups representing distinct profiles of response. For instance, there were four participants who did not respond to either type of intervention over the duration of the study. These individuals were grouped with the low-stimulability cluster in the present analysis, but with a larger sample, nonresponders might emerge as their own cluster. It would be valuable to conduct a large enough study to isolate this cluster, which would then allow us to ask whether other factors such as sensory acuity can be used to predict response or nonresponse to treatment.

A final limitation of this study pertains to the interpretation of our findings regarding the overall magnitude of treatment response. As previously noted, the mean change in perceptually rated accuracy of untreated words over the course of treatment was large (38 percentage points), and the rate of nonresponse (4/33 participants, or 12.1%) was low relative to previous single-case studies of ultrasound biofeedback, which have commonly reported 25%–33% of participants to be nonresponders (e.g., Preston et al., 2019). However, our measure of effect size must be interpreted with some caution because we have only one observation at each time point. Our previous single-case experimental studies have probed accuracy at multiple time points both before and after treatment and calculated standardized effect sizes that take both magnitude of change and session-to-session variability into consideration. The present measure may thus overestimate participants' progress because some of the gains observed might not be stable over time. In addition, longer term generalization measures, as well as measures of accuracy at higher levels of linguistic complexity (e.g., sentences or connected speech), will be necessary to have confidence in the generalizability of these results to real-world clinical contexts.

Clinical Implications and Conclusions

This study yielded several null findings; in particular, the two groups identified by model-based clustering did not differ with respect to our measures of auditory-perceptual acuity or somatosensory acuity. However, the study did return one clear result: Individuals with low stimulability at baseline tended to show limited response in an initial phase of traditional treatment, whereas individuals with high stimulability at baseline tended to make progress in this initial phase, in spite of its brief duration. This finding translates to a straightforward recommendation for clinical practitioners to assess baseline stimulability and take it into consideration when choosing a treatment approach for clients with RSSD. For individuals with high stimulability, it is reasonable for treatment to begin with traditional methods, but individuals with lower stimulability should be preferentially allocated to biofeedback treatment when it is available. In this study, all participants in the high-stimulability group had a normalized mean F3–F2 distance of 8.6 or less, whereas all but four participants in the low-stimulability group fell above that cutoff. Since acoustic measurements are not always obtained in customary clinical practice, we initially sought to report these cutoffs in terms of the treating clinician's perceptually rated accuracy as well. However, we discovered that although study participants varied widely in acoustically measured accuracy, the overwhelming majority were judged by the treating clinician to produce 0% fully correct productions on the stimulability probe at baseline. This may be a case where acoustic measurements could be of true clinical utility, differentiating more versus less stimulable participants when binary ratings of perceptual accuracy are limited by a floor effect. Alternatively, the use of a continuous rather than a binary rating scale (e.g., Meyer & Munson, 2021) could allow more and less stimulable participants to be differentiated on the basis of perceptual judgment.

Of course, it is an unfortunate fact that relatively few clinicians have access to ultrasound technology for use in biofeedback intervention. Clinicians who lack access to ultrasound may wish to try a different form of biofeedback that is more readily available. Visual–acoustic biofeedback, which provides a real-time display of the acoustic signal of speech that can be compared to a model representing a desired pronunciation, can be accessed through free or low-cost mobile apps such as the Speech Therapist's App for /ɹ/ Treatment, or staRt (McAllister Byun et al., 2017; Peterson et al., 2022). However, it is not guaranteed that the present findings regarding the relationship between stimulability and relative response to traditional versus ultrasound biofeedback treatment will also hold for visual–acoustic biofeedback. Future studies should investigate this question directly, as well as systematically compare the efficacy of different types of biofeedback. (For an initial step in this direction, see Benway et al., 2021.) Readers are referred to the ASHA Practice Portal for SSD (American Speech-Language-Hearing

Association, n.d.) for further information about a range of intervention approaches, including biofeedback.

Overall, the present research endorsed a generalization that accords with a common clinical intuition, namely, that specialized treatment approaches like biofeedback may have greater utility for learners with low baseline stimulability relative to those with higher accuracy at baseline. Despite the unsurprising nature of the findings, we believe that it is useful to show concordance between empirical research and clinical intuition in this way. The present findings are also generally in agreement with previous research on biofeedback learning in a nonspeech context (Volin, 1998) although unlike that study, we did not find evidence of a detrimental effect of biofeedback training in high-stimulability learners. Although the design of the study prevents us from drawing strong conclusions, this study also supported the efficacy of treatment incorporating ultrasound biofeedback, with an average gain of 38 percentage points for participants with RSSD in this study. There is ample justification for further research investigating the efficacy of biofeedback intervention for RSSD, including between-subjects comparisons that will support conclusions regarding the relative efficacy of biofeedback versus traditional treatment approaches.

Acknowledgments

Research reported in this publication was supported by National Institute on Deafness and Other Communication Disorders Awards R01DC013668 (D. H. Whalen, PI), F31DC018197 (Heather Kabakoff, PI), and R01DC017476 (Tara McAllister, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank Siemens Corporation, which provided a temporary loan of the Acuson ultrasound scanner and probes for research purposes at no cost to the authors. Additional thanks to Emily Phillips and Erin Doty for treatment delivery, Graham Tomkins Feeny for data management, Laine Cialdella for formant measurement, and José Ortiz for custom software development.

References

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

American Speech-Language-Hearing Association. (n.d.). *Speech sound disorders: Articulation and phonology* [Practice portal]. https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/

Appiani, M., Rabitti, N. S., Methven, L., Cattaneo, C., & Laureati, M. (2020). Assessment of lingual tactile sensitivity in children and adults: Methodological suitability and challenges. *Food, 9*(11), 1594. https://doi.org/10.3390/foods9111594

Attanasio, J. S. (1987). Relationships between oral sensory feedback skills and adaptation to delayed auditory feedback. *Journal of Communication Disorders, 20*(5), 391–402. https://doi.org/10.1016/0021-9924(87)90027-X

Ballard, K. J., Djaja, D., Arciuli, J., James, D. G., & van Doorn, J. (2012). Developmental trajectory for production of prosody: Lexical stress contrastivity in children ages 3 to 7 years and in adults. *Journal of Speech, Language, and Hearing Research, 55*(6), 1822–1835. https://doi.org/10.1044/1092-4388(2012/11-0257)

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*(3), 803–821. https://doi.org/10.2307/2532201

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

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer* (Version 6.0.50) [Computer program]. http://www.fon.hum.uva.nl/praat/

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

Brunner, J., Ghosh, S., Hoole, P., Matthies, M., Tiede, M., & Perkell, J. (2011). The influence of auditory acuity on acoustic variability and the use of motor equivalence during adaptation to a perturbation. *Journal of Speech, Language, and Hearing Research, 54*(3), 727–739. https://doi.org/10.1044/1092-4388(2010/09-0256)

Burris, C., Vorperian, H. K., Fourakis, M., Kent, R. D., & Bolt, D. M. (2014). Quantitative and descriptive comparison of four acoustic analysis systems: Vowel measurements. *Journal of Speech, Language, and Hearing Research, 57*(1), 26–45. https://doi.org/10.1044/1092-4388(2013/12-0103)

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

Carter, E. T., & Buck, M. (1958). Prognostic testing for functional articulation disorders among children in the first grade. *Journal of Speech and Hearing Disorders, 23*(2), 124–133. https://doi.org/10.1044/jshd.2302.124

Cialdella, L., Kabakoff, H., Preston, J. L., Dugan, S., Spencer, C., Boyce, S., Tiede, M., Whalen, D. H., & McAllister, T. (2021). 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

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

Culton, G. L. (1986). Speech disorders among college freshmen: A 13-year survey. *Journal of Speech and Hearing Disorders, 51*(1), 3–7. https://doi.org/10.1044/jshd.5101.03

Davis, S. M., & Drichta, C. E. (1980). Biofeedback: Theory and application to speech pathology. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 3,

pp. 283–308). Academic Press. https://doi.org/10.1016/B978-0-12-608603-4.50015-9

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

Dugan, S. H., Silbert, N., McAllister, T., Preston, J. L., Sotto, C., & Boyce, S. E. (2019). Modelling category goodness judgments in children with residual sound errors. *Clinical Linguistics & Phonetics, 33*(4), 295–315. https://doi.org/10.1080/02699206.2018.1477834

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

Farquharson, K., & Tambyraja, S. R. (2019). Describing how school-based SLPs determine eligibility for children with speech sound disorders. *Seminars in Speech and Language, 40*(02), 105–112. https://doi.org/10.1055/s-0039-1677761

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

Flipsen, P., Shriberg, L. D., Weismer, G., Karlsson, H. B., & McSweeny, J. L. (2001). Acoustic phenotypes for speech-genetics studies: Reference data for residual /ɝ/ distortions. *Clinical Linguistics & Phonetics, 15*(8), 603–630. https://doi.org/10.1080/02699200110069410

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578–588. https://doi.org/10.1093/comjnl/41.8.578

Fucci, D. J. (1972). Oral vibrotactile sensation: An evaluation of normal and defective speakers. *Journal of Speech and Hearing Research, 15*(1), 179–184. https://doi.org/10.1044/jshr.1501.179

Fucci, D. J., & Robertson, J. H. (1971). "Functional" defective articulation: An oral sensory disturbance. *Perceptual and Motor Skills, 33*(3), 711–714. https://doi.org/10.2466/pms.1971.33.3.711

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. (2000). *Goldman-Fristoe Test of Articulation–Second Edition*. Pearson/The Psychological Corporation.

Gritsyk, O., Kabakoff, H., Li, J. J., Ayala, S., Shiller, D. M., & McAllister, T. (2021). Toward an index of oral somatosensory acuity: Comparison of three measures in adults. *Perspectives of the ASHA Special Interest Groups, 6*(2), 500–512. https://doi.org/10.1044/2021_PERSP-20-00218

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

Guenther, F. H. (2016). *Neural control of speech*. Massachusetts Institute of Technology Press. https://mitpress.mit.edu/books/neural-control-speech, https://doi.org/10.7551/mitpress/10471.001.0001

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics, 28*(4), 377–396. https://doi.org/10.1006/jpho.2000.0121

Hearnshaw, S., Baker, E., & Munro, N. (2019). Speech perception skills of children with speech sound disorders: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research, 62*(10), 3771–3789. https://doi.org/10.1044/2019_JSLHR-S-18-0519

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews: Neuroscience, 13*(2), 135–145. https://doi.org/10.1038/nrn3158

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

Hitchcock, E., 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

Hodges, N. J., & Franks, I. M. (2001). Learning a coordination skill: Interactive effects of instruction and feedback. *Research Quarterly for Exercise and Sport, 72*(2), 132–142. https://doi.org/10.1080/02701367.2001.10608943

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *Test, 18*(1), 1–43. https://doi.org/10.1007/s11749-009-0138-x

Irwin, R. B., West, J. F., & Trombetta, M. A. (1966). Effectiveness of speech therapy for second grade children with misarticulations: Predictive factors. *Exceptional Children, 32*(7), 471–479. https://doi.org/10.1177/001440296603200705

Jacobs, R., Serhal, C. B., & van Steenberghe, D. (1998). Oral stereognosis: A review of the literature. *Clinical Oral Investigations, 2*(1), 3–10. https://doi.org/10.1007/s007840050035

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

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lee, J., Russell, C. G., Mohebbi, M., & Keast, R. (2022). Grating orientation task: A screening tool for determination of oral tactile acuity in children. *Food Quality and Preference, 95,* 104365. https://doi.org/10.1016/j.foodqual.2021.104365

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 script]. https://lennes.github.io/spect/scripts/collect_formant_data_from_files.praat

Lonegan, D. S. (1974). Vibrotactile thresholds and oral stereognosis in children. *Perceptual and Motor Skills, 38*(1), 11–14. https://doi.org/10.2466/pms.1974.38.1.11

Lukasewycz, L. D., & Mennella, J. A. (2012). Lingual tactile acuity and food texture preferences among children and their mothers. *Food Quality and Preference, 26*(1), 58–66. https://doi.org/10.1016/j.foodqual.2012.03.007

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

Madison, C. L., & Fucci, D. J. (1971). Speech-sound discrimination and tactile-kinesthetic discrimination in reference to speech production. *Perceptual and Motor Skills, 33*(3), 831–838. https://doi.org/10.2466/pms.1971.33.3.831

Matthews, T., Barbeau-Morrison, A., & Rvachew, S. (2021). Application of the challenge point framework during

treatment of speech sound disorders. *Journal of Speech, Language, and Hearing Research, 64*(10), 3769–3785. https://doi.org/10.1044/2021_JSLHR-20-00437

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

McAllister, T., Preston, J. L., Hitchcock, E., & Hill, J. (2020). Protocol for Correcting Residual Errors with Spectral, Ultrasound, Traditional Speech Therapy Randomized Controlled Trial (C-RESULTS RCT). *BMC Pediatrics, 20*(1), 66–14. https://doi.org/10.1186/s12887-020-1941-5

McAllister Byun, T., & Campbell, H. (2016). Differential effects of visual-acoustic biofeedback intervention for residual speech errors. *Frontiers in Human Neuroscience, 10*(567), 1–17. https://doi.org/10.3389/fnhum.2016.00567

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

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

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

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

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

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

Meyer, M. K., & Munson, B. (2021). Clinical experience and categorical perception of children's speech. *International Journal of Language & Communication Disorders, 56*(2), 374–388. https://doi.org/10.1111/1460-6984.12610

Miccio, A. W. (1995). Metaphon: Factors contributing to treatment outcomes. *Clinical Linguistics & Phonetics, 9*(1), 28–36. https://doi.org/10.3109/02699209508985321

Miccio, A. W., Elbert, M., & Forrest, K. (1999). The relationship between stimulability and phonological acquisition in children with normally developing and disordered phonologies. *American Journal of Speech-Language Pathology, 8*(4), 347–363. https://doi.org/10.1044/1058-0360.0804.347

Milisen, R. L. (1954). The disorder of articulation: A systematic clinical and experimental approach. *Journal of Speech and Hearing Disorders* (Monograph Supplement 4).

Moreau, V. K., & Lass, N. J. (1974). A correlational study of stimulability, oral form discrimination and auditory discrimination skills in children. *Journal of Communication Disorders, 7*(3), 269–277. https://doi.org/10.1016/0021-9924(74)90038-0

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

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *The Journal of the Acoustical Society of America, 113*(5), 2850–2860. https://doi.org/10.1121/1.1567280

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

Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2019). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLOS Computational Biology, 15*(9), Article e1007321. https://doi.org/10.1371/journal.pcbi.1007321

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., & Guenther, F. H. (2004). The distinctness of speakers' /s/–/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language, and Hearing Research, 47*(6), 1259–1269. https://doi.org/10.1044/1092-4388(2004/095)

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America, 130*(1), 461–472. https://doi.org/10.1121/1.3593366

Peterson, L., Savarese, C., Campbell, T., Ma, Z., Simpson, K. O., & McAllister, T. (2022). Telepractice treatment of residual rhotic errors using app-based biofeedback: A pilot study. *Language, Speech, and Hearing Services in Schools, 53*(2), 256–274. https://doi.org/10.1044/2021_LSHSS-21-00084

Powell, T. W., Elbert, M., & Dinnsen, D. A. (1991). Stimulability as a factor in the phonological generalization of misarticulating preschool children. *Journal of Speech and Hearing Research, 34*(6), 1318–1328. https://doi.org/10.1044/jshr.3406.1318

Powell, T. W., & Miccio, A. W. (1996). Stimulability: A useful clinical tool. *Journal of Communication Disorders, 29*(4), 237–253. https://doi.org/10.1016/0021-9924(96)00012-3

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., Hitchcock, E. R., & Leece, M. C. (2020). Auditory perception and ultrasound biofeedback treatment outcomes for children with residual /ɹ/ distortions: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research, 63*(2), 444–455. https://doi.org/10.1044/2019_JSLHR-19-00060

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 Research, 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. (2019). *R: A language and environment for statistical computing. In R Foundation for Statistical Computing* [Programming language]. R Foundation for Statistical Computing. https://www.R-project.org/

Ringel, R. L., Burk, K. W., & Scott, C. M. (1968). Tactile perception: Form discrimination in the mouth. *International Journal of Language & Communication Disorders, 3*(2), 150–155. https://doi.org/10.3109/13682826809011454

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

Ruscello, D. M., & Shelton, R. L. (1979). Planning and self-assessment in articulatory training. *Journal of Speech and Hearing Disorders, 44*(4), 504–512. https://doi.org/10.1044/jshd.4404.504

Rvachew, S., & 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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Scrucca, L., Fop, M., Murphy, T., & Raftery, A. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*(1), 289–317. https://doi.org/10.32614/RJ-2016-021

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition*. The Psychological Corporation.

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

Shriberg, L. D. (1975). A response evocation program for /ɝ/. *Journal of Speech and Hearing Disorders, 40*(1), 92–105. https://doi.org/10.1044/jshd.4001.92

Shriberg, L. D., Tomblin, J. B., & McSweeny, J. L. (1999). Prevalence of speech delay in 6-year-old children and comorbidity with language impairment. *Journal of Speech, Language, and Hearing Research, 42*(6), 1461–1481. https://doi.org/10.1044/jslhr.4206.1461

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

Sommers, R. K., Leiss, R. H., Delp, M., Gerber, A., Fundrella, D., Smith, R., Revucky, M., Ellis, D., & Haley, V. (1967). Factors related to the effectiveness of articulation therapy for kindergarten, first, and second grade children. *Journal of Speech and Hearing Research, 10*(3), 428–437. https://doi.org/10.1044/jshr.1003.428

Steele, C. M., Hill, L., Stokely, S., & Peladeau-Pigeon, M. (2014). Age and strength influences on lingual tactile acuity. *Journal of Texture Studies, 45*(4), 317–323. https://doi.org/10.1111/jtxs.12076

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

Tambyraja, S. R., Farquharson, K., & Justice, L. (2020). Reading risk in children with speech sound disorder: Prevalence, persistence, and predictors. *Journal of Speech, Language, and Hearing Research, 63*(11), 3714–3726. https://doi.org/10.1044/2020_JSLHR-20-00108

Thoonen, G., Maassen, B., Gabreels, F., & Schreuder, R. (1999). Validity of maximum performance tasks to diagnose motor speech disorders in children. *Clinical Linguistics & Phonetics, 13*(1), 1–23. https://doi.org/10.1080/026992099299211

Tiede, M., Boyce, S. E., Holland, C. K., & Choe, K. A. (2004). A new taxonomy of American English /ɹ/ using MRI and ultrasound. *The Journal of the Acoustical Society of America, 115*(5), 2633–2634. https://doi.org/10.1121/1.4784878

To, C. K. S., McLeod, S., Sam, K. L., & Law, T. (2022). Predicting which children will normalize without intervention for speech sound disorders. *Journal of Speech, Language, and Hearing Research, 65*(5), 1724–1741. https://doi.org/10.1044/2022_JSLHR-21-00444

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

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America, 122*(4), 2306–2319. https://doi.org/10.1121/1.2773966

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

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *Comprehensive Test of Phonological Processing–Second Edition (CTOPP-2)*. Pro-Ed.

Westbury, J. R., Hashi, M., & Lindstrom, M. J. (1998). Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication, 26*(3), 203–226. https://doi.org/10.1016/S0167-6393(98)00058-2

Wong, P. C., Vuong, L. C., & Liu, K. (2017). Personalized learning: From neurogenetics of behaviors to designing optimal language training. *Neuropsychologia, 98,* 192–200. https://doi.org/10.1016/j.neuropsychologia.2016.10.002