Phonomotor Training for Accent Modification

Lauril Michele Sachet

A thesis submitted in partial fulfillment for the requirements for the degree of

Master of Science

University of Washington

2013

Committee:

Diane Kendall, Chair

Kristie Spencer

Tanya Eadie

Program Authorized to Offer Degrees:

Speech and Hearing Sciences

Abstract

Phonomotor Training for Accent Modification

Lauril M. Sachet

Chair of the Supervisory Committee:
Dr. Diane Kendall, Ph.D., CCC-SLP
Speech and Hearing Sciences

There is a lack of evidence regarding efficacious training programs for modifying accent. The program developed for this study is based on a parallel distributed processing model of phonology and is expected to improve production of sounds in non-native speakers of English, and generalize to measures of intelligibility. The data gathered from this treatment study will inform models of language processing as well as inform the development of subsequent trials of the same program. The following research questions asked if intelligibility and listener effort improved following training and whether there was a difference in production of trained sounds and untrained contexts following the training program. Post-training results revealed a large effect size for phonemes in isolation thus supporting the hypothesis that non-native speakers can improve phoneme level production with targeted phonomotor training. However, generalization results beyond phonemes in isolation were variable suggesting that future participants would benefit from an increased training period and further program development to promote transfer of skills.

INTRODUCTION

I. Accent definition

A foreign accent is the result of a linguistic difference between a nonnative speaker's native language (NL) and that of the second or target language (TL) (Chakraborty, 2011). These discrepancies are often categorized as either segmental (i.e., relating to production of consonants and vowels) or suprasegmental (i.e., prosodic) (Flege, 1995). Accents often reflect one's cultural and linguistic background and can be a source of pride and identity. It is important to note that an accent is considered a linguistic difference and NOT a disorder (ASHA, 2012). However, accents can also contribute to decreased comprehension of a speaker's message, increased effort on the part of the listener, and may be associated with negative stereotyping (Gluszek, 2010). As a result, many nonnative speakers may seek out accent modification in order to improve communication for both personal and professional reasons.

II. Foundational theory

While several existing theories explain the phenomenon of foreign accents, those predominating the literature have focused on the perception-production paradigm. This paradigm emphasizes the various factors known to influence perception of the phonological structures of spoken language. In 1995 Winifred Strange reviewed research trends of second language pronunciation in order to synthesize what had been learned and what areas of research still needed to be investigated. Drawing from Liberman's early works (1967), Strange described the two fundamental research questions of the time 1) how does an individual's NL linguistic history affect TL perception, and 2) how much is perception influenced by the constraints of maturation and age of acquisition (Strange, 1995). The first question lead to what Liberman coined, "the consistency problem" referred to the allophonic quality of speech (Strange, 1995, p.

10). In other words, there is no 1:1 correspondence between an exact articulatory posture and consistent production of a sound. Rather, due to coarticulation and the dynamic nature of speech, each time we produce a sound it is slightly different yet still identifiable. This gave way to the concept of "categorical perception" which states that we have within category (i.e., allophone) and across category (i.e. phoneme) perception allowing for perceptual consistency.

Numerous classic studies from the 1960s and 70s supported Liberman's theory and demonstrated that the boundaries between sounds are based off contrastive features (i.e., voicing, manner, and place) and are indeed linguistically sensitive (Strange, 1995). Unfortunately, these studies were limited to the manipulation of synthesized sounds in a laboratory, decreasing both their ecological validity and clinical applicability. In addition, the majority of the studies focused solely on altering the voice onset time of stop-consonants and did not address additional features. Furthermore, the distinct categorical boundaries identified in these instances did not equally apply to vowels, which were found to have "blurred" categorical margins.

In 1995 James Flege developed a theory called the Speech Learning Model (SLM) based off the notion of categorical perception (Flege, 1995). By comparing the interaction between NL and TL phonetic categories, he generated a number of hypotheses to predict TL error types and degrees of learning difficulty. Like Liberman, he was most interested in the segmental features of accent modification and suggested that perception drives production, "without accurate perceptual 'targets' to guide sensorimotor learning of TL sounds, production of the TL sounds will be inaccurate" (Flege, 1995, p. 538). He asserted that we perceive TL through the "grid" of our NL meaning that unfamiliar TL phonemes will be assimilated to fit into familiar NL phonetic categories. Therefore, in Flege's view, accent modification consists of redefining preexisting phonetic categories and/or establishing new ones. He declared that although the ability to develop perceptual awareness decreases with age of learning (AOL), it is nevertheless malleable throughout the lifespan.

In addition to AOL, Flege suggests that the distance in "phonetic space" also contributes to ease of acquisition. He identified three distinct types of TL sounds: identical to the NL, similar, and novel. Identical sounds are thought to directly transfer to the TL with no interference. Similar sounds, on the other hand, present the greatest degree of perceptual difficulty. Like novel sounds, they too are assimilated into the closest NL phonetic category; however, since there is less distance or perceptual characteristics upon which to create a new category, these will be more challenging for English as a second language (ESL) speakers. Therefore, the greater the perceived difference between sounds, the greater the likelihood a new phonetic category will be created. For example, Japanese has one glide /r/, while English has two /ɹ/ and /l/; as a result, many Japanese inaccurately produce /ɹ/ and /l/ by treating them as allophones of /r/. However, since /r/ is acoustically closer to /l/, Flege argues that it will be easier for Japanese speakers to create a new category /ɹ/ than /l/ since it is perceptually farther away. In addition, because NL and TL represent a shared phonological system, there is potential for bidirectional affects as phonetic categories are being redefined.

The bidirectional hypothesis helps to illustrate how even once a new category has been established, an accent may still exist if the category of the ESL speaker is founded upon on a different feature (e.g., voice, manner, and/or place) than that of native speaker's. While Flege's SLM helps predict what types of sounds learners may struggle with, it does not clarify how much phonetic space is required for a sound to be considered similar versus novel. Nor does it take into account the articulatory motor learning associated with pronunciation.

An alternative perspective to target language phonetic category formation was presented by Eckman in 1977 (Eckman, 2008). Based off of Lado's contrastive analysis hypothesis (CAH) (Lado, 1957), which identified two TL sound types (same and different), Eckman, like Flege, felt the binary quality of the contrastive analysis hypothesis was insufficient to explain the varying degrees of difficulty observed in TL phoneme acquisition. He designed a weighted system that took into account universal frequency, NL phonological effects, and articulatory complexity calling it the markedness differential hypothesis (MDH). The markedness differential hypothesis assumes that the degree of difficulty an ESL speaker will have in acquiring a new sound depends on the quality of the sound and its level of 'markedness'. When describing how the markedness hypothesis can be applied to traditional dual phoneme comparisons with one sound being produced more accurately, Eckman states (Eckman, 2008, p. 96):

> One member of the opposition was assumed to be privileged in that it has a wider distribution, both within a given language and across languages… the member of the opposition that was more widely distributed than the other was designated as unmarked, indicating that it was, in some definable way, simpler, more basic and more natural than the other member of the opposition, which was in turn defined as the marked member.

The example commonly used to by Eckman to illustrate the pitfalls of the contrastive analysis hypothesis and merits of the markedness differential hypothesis examines the error patterns observed between two different TL groups. When looking at German speakers learning English, word final consonant stops are consistently devoiced because German phonotactic constraints prohibit voicing in the word final position; however, German voicing contrasts do occur in syllable onsets. On the other hand, English speakers learning French have no difficulty producing /ʒ/ word initially in French, as in 'je', although /ʒ/ only occurs word medially (e.g., 'measure') or finally (e.g., 'beige'). Eckman explains this discrepancy by applying the markedness differential hypothesis to show that voiced codas are more marked for German speakers than word initial /ʒ/ is for English speakers. He postulates that, "If a language has a voice contrast in the syllable coda position, it necessarily has this contrast in the syllable onset position, but not vice versa" (Eckman, 2008, p. 97). In the case of German, an onset contrast did not imply a coda contrast causing voiced codas to become more marked in English. Alternatively, the presence of /ʒ/ in the coda position did permit onset contrast in French making it less marked for English speakers. While the markedness differential hypothesis, like Flege's speech learning model, provides insight into *why* certain types of TL errors are observed and how NL influences TL production, it fails to inform how to effectively *teach* TL phoneme perception and production.

To summarize, much of the accent modification literature stemmed from questions linguists were asking about universal differences between first and second language acquisition. Because a correlation was observed between age of learning and degree of accentedness, most researchers steered their inquiry towards speech perception. Findings indicated two important points, 1) Perception and NL linguistic experience strongly influence TL production, and 2) novel phoneme perception is greatly affected, but not exclusively, by age of learning (Flege, 1995). Outcome measures used in early studies were restricted to perception at the segmental level and, as previously mentioned, the sounds themselves were often synthetically manipulated and scarcely reflected authentic speech. Although much energy was expended on developing means to predict TL error types, there was little discussion regarding the application of such analysis. In addition, researchers were not as concerned with the motor learning required to produce new sounds, nor the suprasegmental contributions to intelligibility and listener effort in connected speech.

III. Challenges in translating theory to practice

In addition to linguists, the field of accent modification is also comprised of applied phonetics, speech language pathology (SLP) and teaching English to speakers of other languages (TESOL). The resulting benefit is that many questions surrounding the topic have been asked from different perspectives. However, the challenge remains effectively translating theory and research findings into clinical practice. Couper (2006) summarized this common sentiment among ESL instructors stating,

> One of the difficulties is that research to date has given teachers little guidance as to how they should teach pronunciation, or indeed if they should teach it at all. As Derwing and Munro (2005: 382) point out, pronunciation 'remains a very marginalized topic in Applied Linguistics'. This has in turn meant that there have been very few studies that focus on the effectiveness of pronunciation teaching (Derwing and Munro 2005: 387). This lack of research and consequent lack of training has meant that some teachers have serious misgivings about the effectiveness of teaching pronunciation at all, while others lack the skills and confidence to tackle it in their classes (Macdonald 2002; Breitkreutz, Derwing and Rossiter 2002). (Couper, 2006, p. 47)

Along with the challenges of extrapolating clinical relevance, the mere existence of accent training in ESL curriculum has undergone vast pedagogical shifts.

Original teaching sentiments defined mastery as the ability to achieve native-like production (Burgess, 2000). However, with works coming out by linguists showing that both perception and AOL greatly impacted accentedness, instructors were faced with the fact that, for most, native targets were unattainable. Therefore, pronunciation followed the general shift in TESOL away from standards of perfection, and instead favoring functional communication (Burgess, 2000). As the pendulum swung in the opposite direction, pronunciation dropped out of the curriculum entirely. With the emphasis placed solely on communication, teaching 'speech' was left up to the discretion of individual programs and/or instructors. More recently, recognition that pronunciation can affect communication through intelligibility has resulted in the redefining and resurgence of pronunciation teaching (Sikorski, 2005). Error prediction theories, such as Flege's SLM and Eckman's MDH, combined with the functional communication approach, led to use of the term "functional load" (Munro, 2006; Saito, 2012). Instead of the target being native-like pronunciation, the role of the instructor became selecting which sounds would have the greatest overall affect on an individual's intelligibility. Two approaches are commonly used to make this selection (Burgess, 2000). One involves targeting sounds that are unique, Eckman would say more 'marked' in the TL. For instance, errors in the voiced and voiceless 'th' are highly salient and noticeable to native English listeners. On the other hand, these errors rarely lead to poor comprehension of the speaker's message. Instead, the second approach targets sounds that occur frequently in minimal pairs therefore increasing the probability that mispronunciation will hinder communication. Examples of this method include the /ɹ/ and /l/ distinction for Japanese and /v/ versus /b/ in Spanish. Due to the greater impact on overall communication, the latter tactic has become more widely adopted, yet there is little pedagogical continuity and consensus remains scant.

The concept of functional load has also been extended to examine the impact of consonant versus vowel training on accentedness. Arguably, vowels carry a greater functional load than consonants since they too exist in minimal pairs, are highly salient to the native listener, and carry the prosody and intonation of an utterance (Franklin, 2009). A doctoral dissertation study conducted by Franklin 2009 examined three groups of ESL speakers (Japanese, Korean and Spanish) before and after receiving eight weeks of vowel-focused treatment. Native listener rated for accuracy (i.e., correct vowel identification) and "goodness" a qualitative measure aimed as capturing the relative distance to the target. Results supported Flege's Speech Learning Model in that shared vowels were identified as accurate from baseline, while similar and novel vowels both showed significant improvement following training. Although further investigation is needed to parse the functional load of consonant versus vowel focused training on overall communication, both approaches have distinct merits to be considered when designing an accent modification program.

An additional product of placing communication and functional load at the forefront was that the language used to describe the accent phenomenon also changed. Munro and Derwing deconstructed the ambiguous term "accent" into three distinct measurements, 'accentedness', 'intelligibility' and 'comprehensibility'. Accentedness can be considered and umbrella term ascribing the degree to which the non-native speech differs from that of the native listener's. Intelligibility is defined as the quantity of message understood by the listener and is frequently calculated in percent units correct. Comprehensibility on the other hand, refers to the amount of effort required by the listener and is typically measured using an analogue scale. Because the term 'comprehensibility' has a similar yet slightly different connotation in the dysarthria literature, our study has chosen to use 'listener effort' as an equivalent alternative. While these factors are related, each captures a slightly different parameter contributing to the speaker/listener experience (Derwing, 1997). For example, it is possible for a non-native speaker to be highly intelligible yet have poor comprehensibility; meaning the listener was able to glean the entirety of the message but it took immense effort to do so. Development of explicit outcome measures that are both ecologically valid and discrete have allowed researchers to become more effective in designing clinically appropriate studies.

IV. Current research

As previously noted, rigorous studies focused on the effects of pronunciation instruction on accentedness are minimal; however, recent years have yielded a small but growing body of works aimed at approaching the perception/production paradigm from a functional-instructional perspective. Preliminary works such as those by Perlmutter and Derwing were primarily concerned with performing phase I, quasi-experimental studies to show that target populations of ESL speakers could indeed benefit from pronunciation training (Derwing, 1997; Perlmutter, 1989). Although lacking control, these studies documented significant improvement in accentedness after receiving instruction. Others focused on manipulating specific variables along the perception/production continuum to maximize functional load and improve intelligibility, comprehensibility, and/or accentedness showing that both input and feedback are crucial in producing change.

An insightful study done by Champagne-Muzar examined the relationship between segmental versus suprasegmental instruction pertaining to both speech perception and production. This study looked at adults in beginner level French as a second language, (FSL) students in the university setting. Throughout the 12 weeks of instruction, half were devoted to receptive skills only via identification and discrimination tasks of discrete phonemes, utterance intonation and utterance rhythm. The other half incorporated both perception and production training. Immediate post testing revealed a significant improvement in treatment group phoneme and intonation accuracy but not rhythm. Results indicated that, regardless of the underlying mechanism, adults can learn to improve their pronunciation when provided with adequate perceptual input, opportunity to produce, and explicit segmental and suprasegmental instruction. (Champagne-Muzar, 1993).

Saito and Lyster 2012 also targeted perception by examining the effects of "form-focused instruction" (FFI) versus form-focused instruction plus corrective feedback (FFI+CF) on native Japanese speakers' /ɹ/ productions. Form-focused refers to direct explanation of the features involved in creating the sound and is considered a "proactive" method of instruction. FFI+CF added another element by providing corrective feedback to inaccurate productions in the form of recasting. This is considered "reactive" instruction. While each group received equal hours of instruction, only the FFI+CF group showed significant improvements. This study illustrates how instruction alone is insufficient to drive perception/production and that reactive feedback plays a critical role in this process (Saito, 2012).

Similarly, a study by De Bot looked at the outcome of visual feedback alone and in conjunction with auditory feedback on intonation learning. While both groups significantly improved perception and production of targets, the audio-visual feedback group did so faster. A secondary outcome measure also inspected whether forty-five or ninety-minute treatment sessions had a greater influence on rate and accuracy of learning; however in this regard both groups performed equally suggesting that treatment duration is not as crucial as multimodal input. The authors also noted that audio-visual feedback correlated with increased production attempts thus implying that more input renders to more trials. (De Bot, 1983)

While some researchers were concerned with shaping perception through modifying input and feedback, others were occupied with determining which elements (i.e., segmental or suprasegmental) carried a greater functional load. A quasi-experimental study done by Derwing et al in 1998 looked at the effects of segmental versus suprasegmental (i.e., 'global') instruction at sentence and spontaneous speech levels following eleven weeks of training. Interestingly, at the sentence level the segmental group showed significant improvement; however, at the spontaneous speech level, only global instruction appeared to have an effect. The authors posited two explanations for their findings. First, that both narrow (segmental) and broad (global) instruction is necessary for significantly improving ESL accentedness and comprehensibility since both phonological skills and prosodic ones are used in different ways to repair communication breakdowns. The second explanation focused on the listeners suggesting that the findings may be a product of the different strategies listeners use for comprehending sentence level versus conversational discourse (Derwing, 1998).

One of the most sophisticated studies to date done by Couper (2006) approached functional load not strictly from a segmental versus suprasegmental stand point but rather by analyzing predominant error patterns displayed by a group of disparate non-native speakers. In doing so, he identified two errors in particular that had consequences at both the segmental and suprasegmental levels. By targeting both perception and production of epenthesis and phoneme omission, he was able to ask a number of explicit research questions: does target specific training improve perception/production, are these gains maintained long-term, and are they generalizable to new linguistic contexts? Interestingly, results provided unexpected insight into the question of perception/production. While production of the target items significantly improved immediately and one year following treatment, the same was not true for perception, which revealed no long-term maintenance. However, the improvements in production did in fact generalize to untrained contexts. Couper provided two explanations for the disparity in perception suggesting that it is harder to change perception than production and that perception is not the sole factor determining pronunciation errors. Other features may be native language phonological constraints and formulation time required of different speaking tasks, “In other words, learners may learn to apply some patterns or knowledge about pronunciation to their speaking, if they have some time to plan or rehearse, but still not be able to apply it to their listening” (Couper, 2006, p. 57). Therefore, these findings challenge the perspective that perception is exclusively responsible for driving production.

In summary, important findings to date indicate that pronunciation can indeed improve with instruction. More explicitly, proactive and reactive teaching methods are necessary to change behavior in addition to multimodal input. Both segmental and suprasegmental features are also important in enhancing comprehension and repairing communication breakdowns. Perception is harder to change than production and is not solely responsible for pronunciation errors. Finally, new behaviors can be maintained long-term and generalized to novel contexts. The task of future research is to duplicate these findings and further expand upon the type and quantity of input required to incite sustainable and transferable modifications. It is also necessary to scrutinize the interchange between segmental and suprasegmental features to maximize instruction as well as re-examine the degree to which perception shapes production.

V. Our study and the parallel distributed processing (PDP) model

While there remain many theoretical and pedagogical quandaries relating to accent modification, integrating what we know about impaired systems can provide insight into the underpinnings of speech production. Just as Flege and Eckman made inferences about phonological processing based on error patterns of non-native speakers, numerous models have also been presented to account for errors observed in individuals with compromised systems. Although it is likely that the phonological system of ESL learners functions differently compared to those with impaired systems (i.e., inaccurate TL productions are based off of intact transfer of NL representations as opposed to inaccurate productions due to degraded monolingual phonological representations), the processes for acquiring phonemes, phonological awareness and motor planning may be similar. Barrowing from the well-established models stemming from aphasia research can help explain the underlying mechanisms involved in phoneme perception and production, and thus inform phonological modification and habilitation of novel sounds in non-native speakers.

One such model generated by Nadeau (2001) lends itself particularly well to understanding the covert processes implicated in phoneme acquisition. The parallel distributed processing (PDP) model of phonology was developed to reconcile typical "slip of the tongue" errors, diverse speech and language impairments resulting form aphasia, and what is known about neurological architecture and function. Nadeau suggested that rather than serial processing of isolated units, information is comprised of representations stored in various domains relying on simultaneous or parallel activation. Therefore, a discrete phoneme representation is multisensory and made up of visual, auditory, and oral tactile-kinesthetic information contributing to both perception and production. Nadeau describes this as, "a given element is represented as a pattern of activity involving hundreds if not thousands or millions of neurons (or units in a model)" (Nadeau, 2001, p. 512). As such, individual phonemes do not exist as localized units, but are rather a product of simultaneous multimodal activation. To further illustrate phoneme representation in terms of parallel distributed processing Nadeau states,

> In a pure PDP model of phonology, there is no need to build in specific structures to account for specific phonological phenomena. The structure of the model is defined

> entirely in terms of the domains of information accessible to it and the necessary relationship of these domains to each other (Nadeau, 2001, p. 514).

The implication being that a phoneme representation is dynamic rather than static and is a manifestation of the domains being acted upon.

Three distinct aspects of the PDP model are useful in both supporting and enhancing our understanding of accent modification. The first aids our understanding of the accent phenomenon and modification by incorporating the principles of neural processing and neuroplasticity. According to the PDP, instantiation of representations (e.g., a thought, word or sound) depends on the strength of the neural connections. The probability that a representation will receive sufficient activation depends on the frequency of access and amount/diversity of neural input received (e.g., activation from multiple cognitive domains). Kendall (2008) describes this as,

> During learning of a language, the strengths of the connections between the units are gradually adjusted so that a pattern of activity involving the units in one domain elicits the correct pattern of activity in the units of another domain. The entire set of connections between the two domains forms a pattern associator network. (Kendall, 2008, p. 3)

As such, an accent results when preexisting native language phonological representations are activated in the context of the target language, resulting in perceived distortions, substitutions, omissions and additions in the TL.

While an accent might be a product of typical neural processing, accent modification can be thought of as the product of neuroplasticity, “the mechanism by which the brain encodes experience and learns new behaviors” (Kleim, 2008, p. 225). More specifically, Nadeau describes experience dependent neuroplasticity as involving the, “genesis of dendritic spines, synaptogenesis, and modification of synapses” (Nadeau, In Press, p. 3). Therefore, accent modification could be considered a process of altering existing patterns of activation or creating new ones by providing rich multimodal input and ample opportunities for strengthening those patterns.

In addition, the PDP model also bolster Eckman and Flege's notions of categorical perception that shape their theories of accent modification. Translating Eckman's concept of markedness to the vernacular of neural processing, a sound is deemed more "marked" because the pattern of activation is stronger either due to frequency of access or wealth of input. Flege's speech learning model is also accounted for, NL and TL shared sounds will have the strongest connections while similar phonemes with shared NL input will be difficult to modify because the likelihood of activating the NL representation is greater. In addition, novel sounds will present a challenge, as they require new multisensory input to develop the discrete patterns of activation. While Eckman and Flege viewed accent modification as manipulation of categorical boundaries, the PDP model approaches it from the perspective of altering and promoting neural connectivity.

Thus far the PDP model has helped to explain accent and accent modification in terms of neural processing and plasticity. It has also supported and supplemented the previous theories put forth in accent modification. Perhaps the most crucial piece the PDP offers is this multimodal framework for how to elicit change. While the theories previously posited by Lado (1957), Lieberman (1967), Eckman (1977) and Flege (1995) viewed phoneme perception as an auditory phenomenon responsible for driving production, they did not account for motor learning, nor did they explain improvements in production in the absence of perceptual discrimination as evidenced by Couper (2006). By broadening the definition of perception beyond auditory to also include visual, orthographic and tactile-kinesthetic information, the PDP accounts for changes in production in the absence of auditory perception and identifies additional avenues of instructional.

V. Research Questions

Considering the primary objective of accent modification is to alter and create novel phonological representations, which we aim to do through multimodal instruction, our research questions are concerned with the degree, quality and generalization of that change. Although the treatment takes place largely at the segmental level, it is important that change have a meaningful impact on the individual's overall communication. The accent modification training employed in this study incorporates both the learning principles and multimodal instruction provided by the PDP model. Adapting elements from the Lindamood Phoneme Sequencing (LiPS) program, Dr. Diane Kendall, Ph.D., CCC-SLP originally developed this therapy program for adults with acquired phonological impairments due to aphasia, calling it *phonomotor* treatment. Although this is the first time applying phonomotor training to accent modification, considering the multimodal instructional approach, and emphasis on intensity, repetition, and use of Socratic questioning, we predict this will be sufficient to promote change in TL pronunciation.

Therefore, our primary research question is, *is there a significant difference in intelligibility and listener effort following treatment?* Our secondary research questions target the distinct parameters of that change. The first being one of acquisitional accuracy, *is there a significant difference in production of trained sounds following training (i.e., phonemes in isolation)?* In addition, *is there a significant difference in production of trained sounds to untrained contexts following training (i.e., real words, nonwords, and sentences)?*

METHODS

Participant: The participant in this study was a 20 y/o, right hand dominant male from main land China. He spoke both a local dialect and Mandarin. He began learning English in grade school when he was 12 years old from a native Chinese speaker. He came to the United States to study at the university level in 2011 and reported typically spending 50% of his day speaking English. He expressed a desire to improve his pronunciation for general communication purposes and specifically wanted to gain more awareness regarding how to produce sounds accurately.

The participant demonstrated an accent as evidenced by frequent sound errors (e.g., substitutions, omissions, additions and/or distortions) that compromised intelligibility and listener effort in conversation as determined by experienced listeners (i.e., members of the research team accustomed to assessing speech sound errors). Exclusionary criteria included: 1) impairments in cognition, 2) high baseline intelligibility in conversation as judged by a speech language pathologist (SLP) and SLP master's student.

Characterization of included participant: In order to ascertain information pertaining to cognition, underlying phonological processing, hearing and self perception, the following measures were taken prior to the start of treatment: The Standardized Assessment of Phonology in Aphasia (SAPA) (Kendall et al 2010), Ravens Progressive Matrices (Raven, 1972), a hearing screening, and The Communication Participation Item Bank (CPIB) (Baylor, In Press). A description of each measure including stimuli, administration schedule, procedures, scoring and reliability is included below. Examples of measures can also be found in the appendix of this document.

Phonological awareness:

- *Stimuli:* Developed by Kendall in 2010, the SAPA is designed to examine the phonological awareness of individuals with aphasia. It consists of three subtests that examine: reading (real words, nonwords, pseudohomophones and irregular words); auditory phonological processing (real word rhyme, nonword rhyme, lexical decisions and minimal pairs); and repetition, parsing and blending (real word repetition, nonword repetition, real word blending, nonword blending, real word parsing and nonword parsing).
- *Administration schedule:* the entire SAPA was administered once during the pre test period and once again during post testing.
- *Procedures:* Prior to each subtest the participant was given practice items and the opportunity to ask questions and receive feedback. Once the test had begun no feedback was provided.
- *Scoring:* Responses are scored online. Each of the 151 items is valued at one point and a percent accuracy score is derived for each subtest. Incorrect responses include: phonological errors, semantic errors and no response. Sound distortions are considered correct. The SAPA has been validated on 37 individuals who are neurotypical. Normal control performance for Subtest 1 (Reading) was 44.7 (SD 4.32), Subtest 2 (Auditory Phonological Process) was 35.6 (SD 2.81) and for Subtest 3 (Repetition) was 23.7 (SD 3.34).
- *Reliability:* Test-retest reliability was performed on 10 of patients with aphasia. Paired t-tests results above 0.05 indicate differences between test sessions were not statistically significant for all three subtests (reading $p=0.2781$, perception $p=0.264$, repetition $p=0.8928$).

Quality of life measure:

- *Stimuli:* the CPIB is an equal appearing interval scale developed as a quality of life measure. It was originally intended to capture perceived effects of disordered communication on daily life and was adapted for this study to address effects of accent on non-native speakers. A short form version was used consisting of ten items. In each, the word "condition" was substituted for "accent". See appendix for original and adapted versions.
- *Administration schedule:* the CPIB was given once during the pre test period and once again at post testing.
- *Procedures:* the participant was instructed to complete the questionnaire based on how he felt on average by selecting from the following options: (3) not at all, (2) a little, (1) quite a bit, (0) very much.
- *Scoring:* a total was calculated by adding the numerical value of each response. A high score was considered desirable indicating no negative impact on communication, while a low score denoted greater difficulty. Although not a precise measurement tool, the CPIB offers some quantifiable insight into how the speaker felt communication was impacted.
- *Reliability:* the CPIB was not normed on non-native English speakers and therefore results are not directly interpretable; however, the CPIB does provide a general quantitative pre-and-post training sense regarding quality of life.

Hearing screen:

- *Stimuli:* an audiometer was used to introduce pure tones into each ear at 25 dB for each of the following frequencies: 500 Hz, 1k Hz, 2k Hz, 3k Hz
- *Administration schedule:* once at pre testing to rule out hearing loss
- *Procedures:* the participant was instructed to raise his hand every time he heard a tone delivered through headphones.
- *Scoring:* was based on whether or not a response was elicited.

Nonverbal reasoning:

- *Stimuli:* Ravens Progressive Matrices consisting of thirty-six colored items.
- *Administration schedule:* pre testing to rule out non-linguistic cognitive impairments.
- *Procedures:* the participant was shown a series of examples prior to test initiation. The examiner indicated that the participant should choose the appropriate design to continue the pattern. Each colored item shows three related designs with the fourth missing. The participant must deduce from the six options provided which is the appropriate design to complete the sequence. The test concludes when all thirty-six items have been administered.
- *Scoring:* a point value of one was allotted for each correct answer and zero for each incorrect response. A percent correct was then derived from the total number possible.

Study design: In the context of a single subject repeated probe design, the participant received probes prior to training, during training and immediately following completion of the program. Measures including the SAPA, CPIB, and both intelligibility and listener effort (sentence and discourse levels) were administered pre and post training over a span of 3 days each. Repeated probes were also performed five times at baseline over the 3-day testing period (2-hours per day). During training, probes were given following every two hours of instruction (nine times). Post training repeated probes were administered four times throughout the 3-day post-testing period (2-hours per day). All repeated probes included: trained sounds in isolation, trained sounds in real words, trained sounds in nonwords and the control measure. See *Table 1* for details.

Primary outcome measures:

The primary outcome measures of this study were intelligibility (percent of key words correctly understood by an unfamiliar listener) and listener effort (visual rating scale indicating no effort to maximum effort as determined by unfamiliar listeners), both of which were examined at the sentence and discourse levels. All measures were digitally recorded using Marantz Professional Solid State Recorder PMD671 recording system and AKG C535 EB Austria microphone. The following is a description of each measure including: stimuli, administration schedule, procedures, scoring and reliability.

Sentence level intelligibility:

- *Stimuli:* were derived from the Harvard Sentences, which include 72 lists of 10 sentences, each phonetically balanced to represent frequency of occurrence in English. Content is considered decontextualized thus requiring the listener to rely on word comprehension rather than logical deduction for understanding. Novel lists of 10 sentences were spoken by the participant and digitally recorded twice pre and twice post training for a total of 40 sentences. See appendix section for exact stimuli.

- *Administration schedule:* A certified Speech Language Pathologist (SLP), and member of the research team, administered the entire pre-test battery including the Harvard Sentences. A different certified SLP and member of the research team administered the entire post-test battery following the same protocol. Both examiners were different from the clinicians providing the training.
- *Participant procedures:* Since these sentences were to be scored for intelligibility, the objective was to sound as natural as possible and to not be hindered by reading or memory errors. Therefore, the following instructions were given to the participant, "I am going to read you some sentences. Please repeat each sentence as naturally as possible." The participant was provided with a verbal model as well as the corresponding written text. Reinstruction was permitted if necessary.
- *Unfamiliar listener procedures:* Unfamiliar listeners were set up with a computer program and given the following instructions, "When you start this program, it will immediately play a sample for you. Please transcribe the sample on the Excel spreadsheet on the left. Please do your best to write each word you hear - you may guess or leave a blank space if you are not sure. You may replay the sentence as many times as needed."
- *Scoring*: To comply with the Harvard Sentences scoring protocol, lexical accuracy was measured in percent keywords correctly understood by the unfamiliar listeners. Scoring was performed by three unfamiliar listeners (i.e., undergraduate students from the department of Speech and Hearing Sciences). The forty pre and post sentences were randomized and auditorily presented to each listener separately. The listeners then used orthographic transcription to capture what they had heard on a desktop computer. A different certified SLP and member of the research team then compared each

transcription to the original sentence to calculate number of key words correctly understood by each listener; sentences contained between 3 to 6 key words each. Key words were defined as adjectives, nouns, adverbs, verbs and negation (e.g., 'not'). Pre and post training averages were then derived by calculating percent key words correctly transcribed.

- *Reliability*: Inter-rater reliability for sentence stimuli was calculated by comparing the intelligibility score for each sentence from each listener to that of each other listener (using single-measures intraclass correlation coefficients; ICCs), and by comparing the score from each listener to the group mean for each sentence using average-measures ICCs (Shrout, 1979).

<u>Discourse level Intelligibility:</u> was elicited via a picture description task.

- *Stimulus*: the picnic scene from the Western Aphasia Battery (WAB).
- *Administration schedule:* Three-minute picture description samples were digitally recorded at both pre and post testing. Each sample was then spliced into four, thirty-second segments totaling 8 individual files.
- *Procedures:* The examiners provided the participant with the following instructions, "I'm going to show you a picture. Take a few minutes to look it over. When you're ready, please tell me about what's happening in the picture. I'm going to record you talking for 3 minutes. Do you have any questions before you begin?" Since the discourse sample was not meant to be a language exercise (e.g., how much vocabulary he knew), the participant was encouraged to look at the picture first and ask for any unknown words.

- *Scoring*: was performed by the same unfamiliar listeners and orthographic transcription was used for this task as well. The four pre and four post samples were randomized and auditorily presented to each listener individually. The unfamiliar listeners were instructed to exclude fillers such as 'um' in their transcriptions. The same researcher who scored the sentence level intelligibility also calculated discourse accuracy. The unfamiliar listeners' transcriptions were compared to that of a familiar listener. Percent keywords correctly understood was calculated based on total possible keywords identified by the familiar listener.
- *Agreement:* Because of the small number of picture description stimuli, a measure of inter-rater agreement was more appropriate than ICCs for these samples. Inter-rater variability, a unitless measure of the spread of scores around the mean value, was calculated for the 8 description stimuli. This measure shows the variation of ratings made by each listener without using an arbitrary cutoff point, such as "within 1 scale point" (Portney, 2000). The variability score is calculated by squaring the difference between each listener's rating and the mean of all ratings for a sample. Lower variability scores indicate that mean ratings by a particular listener were closer to the group average; higher variability scores indicate larger differences from group means (Chan & Yiu, 2002).

Listener effort (sentence and discourse):

- *Stimuli*: the same Harvard sentences and picture description samples were used for determining listener effort.
- *Administration schedule:* same as previously noted for intelligibility.

- *Participant procedure:* same as previously noted for intelligibility.
- *Unfamiliar listener procedure:* Unfamiliar listeners were given the following instructions, "After transcribing the sentence, please rate in this program the amount of effort it took to understand it. A 9-point scale is provided with 1 meaning no effort and 9 meaning extreme effort. Place the cursor carefully. When you submit the rating, the next sample will immediately begin. If you have questions, please ask them now."
- *Rating:* Immediately after transcribing each sentence or passage, listeners rated the amount of effort required to understand the sample using a 9 point equal appearing interval scale (1 = no effort, 9 = extreme effort), presented on a desktop computer.
- *Reliability & Agreement:* As above, inter-rater reliability for sentence stimuli was calculated using ICCs.

Secondary outcome measures:

Secondary outcome measures focused on acquisition of trained sounds in isolation and generalization to trained sounds in real words, nonwords and sentences. In order to ascertain if any positive effects of training were attributed to the program, pre-, repeated and post-probe control measures were administered. All measures were digitally recorded using the Marantz Professional Solid State Recorder PMD671 recording system and AKG C535 EB Austria microphone. The following is a description of each measure including: stimuli, administration schedule, procedures, scoring and reliability.

Sounds in isolation:

- *Stimuli*: Because multimodal input is integral to this training program, assessment also included both visual and auditory stimuli. A pre-recorded visual and verbal model of each phoneme was used to elicit sounds in isolation. This was presented on a computer monitor at a sound level that was comfortable to the participant. The stimuli was developed by the research team and included a video recording of one of the research members producing each of the 37 phonemes in isolation. To enhance visual input and reduce distraction, only the lower portion of the face was recorded. Three other members of the research team reviewed each sound to verify production quality.
- *Administration schedule*: All 37 phonemes were examined five times at baseline over the three-day pre-testing period. After every two hours of training for a total of nine 'during training' data points, and again four times following program completion during the three day post-testing period.
- *Procedures:* A microphone was set up at six inches from the participant to digitally capture all responses. Files were then saved to a secure database and all prompts were edited out.
- *Scoring*: was completed online by the three separate administrators, the pre-test examiner, the clinician delivering the training program, and the post-test examiner. A value of one point was given for any accurately produced phoneme. Distortions, substitutions, and additions (e.g., epenthesis) were considered inaccurate productions and given a score of zero. A percent accuracy score out of one hundred was then derived for each data point.
- *Reliability:* both inter and intra-rater reliability was performed on 20% of the entire corpus across all three-time points.

Real word production:

- *Stimuli*: sets of fifty words were selected from the Assessment of Intelligibility of Dysarthric Speech (AIDS) in order to determine phoneme accuracy in real words. The AIDS was chosen because it is phonetically balanced to represent phoneme occurrence in English.
- *Administration schedule*: novel lists of 50 words were presented five times at baseline over the three-day pre-testing period. After every two hours of training for a total of nine 'during training' data points, and again four times following program completion during the three day post-testing period.
- *Procedures:* to minimize reading errors, stimuli were presented with both verbal and written models. The participant was given the following instructions, "I'm going to say a word once, repeat it after me." The microphone was set up at approximately six inches from the participant to digitally capture all responses. Files were then saved to a secure database and all prompts were edited out.
- *Scoring*: occurred offline by a single member of the research team. A count of phonemes per word was conducted and summed to tabulate a grand total of number of phonemes possible for each set of fifty words. Rhotic vowels (e.g., 'er', 'ar, 'air') were considered a single phoneme. Each word was scored individually and a point was subtracted for any phoneme in error (i.e., sound substitutions, distortions and deletions). Additions were also considered an error and resulted in a point deduction. A percent accuracy score out of one hundred was then derived for each data point.

Non-word production:

- *Stimuli*: each probe was comprised of a full list containing one hundred forty-four nonwords. Due to the extensive list length, each testing point consisted of a half list of seventy-two nonwords. The nonwords consisted of phonological sequences of low phonotactic probability and high neighborhood density. Phonotactic probability was calculated using methods similar to Vitevitch and Luce (1999). All nonwords were phonotactially legal in English. A web-based interface was used to calculate phonotactic probabilities for the real and nonwords (Vitevitch, 2004). Neighborhood density was computed by counting the number of words in the dictionary that differed from the target by a one phoneme addition, deletion, or substitution. Phonotactic probability and neighborhood density were computed for stimuli and were categorized as high or low based on a median split (Strokel, 2006). Prerecorded video by a native English speaker provided an auditory and verbal model for each nonword. Again, the video contained just the lower portion of the speaker's face to support an accurate repetition by the participant.
- *Administration schedule*: For convenience, full lists were broken down into half lists of seventy-two nonwords (section A and B). Half lists were probed five times at baseline over the three-day testing period. Again after every two hours of training for a total of nine "during training" points, and four times immediately following completion of the program during the three-day post testing period. Since each list is phonetically balanced across the entire set (one hundred and forty-four nonwords), performance on half lists was collapsed to create one data point.

- *Procedures:* The total list of one hundred forty-four nonwords was split into half lists of seventy-two which were alternated for each probe. The participant was instructed to repeat each nonword as accurately as possible. All responses were digitally recorded.
- *Scoring*: each nonword was scored online by the three separate administrators, the pre-test examiner, the clinician delivering the training program, and the post-test examiner. Whole word scoring was applied to nonwords; therefore, if any phoneme was produced inaccurately, the whole word was considered wrong and received a score of zero. Correct responses required accurate production of all phonemes and received a score of one. A final score of percent whole word correct out of seventy-two was then calculated for each data point. Nonword half lists were collapsed to create a single data point, scores therefore reflect performance across the balanced sample.

<u>Sentence level phoneme accuracy:</u>

- *Stimuli*: the same Harvard sentences used for intelligibility and listener effort were examined for sentence level phoneme accuracy.
- *Administration schedule*: since the same stimuli were used, the same administration schedule applies: two novel lists of ten sentences each were presented prior to training and two novel lists of ten were presented at post training for a total of forty sentences.
- *Procedures:* see procedures for intelligibility and listener effort previously stated.
- *Scoring*: all forty pre and post sentences were entered into an Excel spreadsheet and assigned a random number. Pre and post identification was then hidden and a randomization function was performed to mix the data points and avoid biasing the rater's judgments. Each sentence was written in broad transcription using the

International Phonetic Alphabet (IPA). A total number of phonemes possible was calculated for each sentence. Rhotic /r/ was considered a single phoneme as were diphthongs. Additionally because the word 'and' is typically pronounced "an" in connected speech by native English speakers, the word was calculated as two phonemes instead of three. The primary scorer was a master's student and member of the research team (also the author of this paper). Discrete phonemes were examined at the sentence level using Praat software to facilitate the scorer's ability to control the listening experience. One point was subtracted for each inaccurate phoneme per sentence. Errors included sound distortions, deletions and substitutions. Occasional addition errors were made (e.g., epenthesis) resulting in a point being subtracted from the total number correct. For each list of ten sentences, a grand percentage of number of phonemes correct out of total number of phonemes possible was calculated. The two pre-list scores were then averaged, as were the two post-list scores to create two distinct data points, one pre and one post.

- *Reliability:* inter-rater reliability was performed on twenty percent of the stimuli (i.e., eight sentences) by the primary clinician who delivered the training program. Both scorers followed the same procedures.

<u>Control Measure:</u>

- *Stimuli:* the control probe was an auditory grammatical judgment subtest from the Test of Adult and Adolescent Language (TOAL-3) (Hammill, 1994). A total number of thirty-five items were randomly organized into eighteen separate lists of ten items each. This control measure was chosen because although syntactic knowledge is still a linguistic skill, it was not expected to change in relation to phonological training.

- *Administration schedule:* each list of ten items was presented five times at baseline over the three-day pre-testing period. After every two hours of training for a total of nine 'during training' data points, and again four times following program completion during the three day post-testing period.
- *Procedures:* the examining clinician read each of the ten items aloud. The participant was instructed to listen to three separate statements per item and select which of the three (option A, B, or C) was grammatically correct.
- *Scoring:* occurred online by the examining clinician. A percentage of number correct out of ten was derived for each time point.
- *Normative data:* This test was normed on 3,056 individuals from the United States. The adult ages ranged from 18 years to 24-11 years. For more details see TOAL-3 manual.

Phonomotor program protocol:

- Stimuli: Stimuli were comprised of phonemes in isolation, 40 non-words, and 40 real words characterized by low phonotactic probability and high neighborhood density. Vowel (V), consonant-vowel (CV) was constructed in 1- and 2-syllable real and non-word combinations. The choice to use low phonotactic probability stimuli was based on the concept that training atypical exemplars of a domain increased knowledge relevant to *both* atypical *and* typical exemplars, whereas training only typical exemplars benefits *only* typical exemplars (Kiran, 2003; Plaut, 1996) an observation that is entirely consistent with principles of neural network function captured in PDP models (Nadeau, 2012; Thompson, 2007). Storkel (2006) has provided empirical support for this concept in the domain of phonology. The choice to use stimuli with high neighborhood density has been made to maximize the number of word concepts that might engage trained phonemes and phonological sequences, thereby aiding the development of semantic-phonologic connectivity. Phonotactic probability was calculated using methods similar to Vitevitch and Luce (1999). Criteria for moving from one stage to the next were based on reaching 80% accuracy for both perception and production tasks for each phoneme.

- Stage 1 – Consonants in Isolation: The purpose of Stage One is to engage individual sounds by teaching a) motor descriptions (e.g., the tip of your tongue is behind your front teeth and taps to make the sound /t/); b) perceptual discrimination (e.g., do /t/ and /d/ sound the same or different?); and c) production (e.g., repeat after me…say /t/). Orthography is occasionally employed in this program solely as an additional way to access the domain of phonological sequence knowledge, not with the goal of improving reading. A mirror was placed on the table for the participant to use for visual feedback for recognition and correction of errors. Each sound was represented by a picture of a mouth in the corresponding posture.
- Stage 1-Task 1: Exploration of sounds: The participant was shown a mouth picture of a sound and asked to look in the mirror and repeat after the therapist to make the sound. Knowledge of results (KR) were initially given at 100% frequency following each production then faded to 30% across trials. Following production, the therapist asked the participant what they saw and felt when the sound was made. Socratic questioning was used to enable the participant to “discover” the auditory, visual, articulatory and tactile/kinesthetic attributes of the sounds (e.g., “What do you feel when you make that sound?”). Through practice and

repetition the participant will become adept at recognizing what they actually need to feel, see, hear and do to make the sound.

- Stage 1-Task 2: Motor description: A description of each sound was provided. The therapist described what articulators were moving and how they moved (e.g., for /p/ the lips come together and blow apart, the voice box is turned off, the tongue is not moving). The subject was asked to repeat the sound and then asked to describe how the sound was made. For example, "Do your lips or tongue move to make that sound?"
- Stage 1-Task 3: Perception Task: The therapist made a sound (e.g., /p/) and asked the participant to choose that sound from an array of pictures (e.g., /f/, /g/, /p/). Socratic questioning was used for correct and incorrect responses.
- Stage 1-Task 4: Production Tasks: Production of sounds was elicited auditorily (repetition), visually (mouth picture), and via motor description (e.g., "make the sound where your lips come together and blow apart"). Socratic questioning was used for correct and incorrect responses. For example, "you said /b/ is that the sound where your tongue taps the roof of your mouth?"
- Stage 2 – Syllables: The purpose of this stage was to extend skills acquired in Stage 1 to various phoneme sequences. Production, perception and graphemic tasks remained the same with the one difference that sounds were produced in combinations rather than isolation. Training progresses hierarchically (e.g., VC, CV, CVC, CCV, VCC, CCVC, CVCC, CCVCC). Upon mastery of 1-syllable stimuli, 2-syllable stimuli were composed using various combinations of 1-syllable stimuli. Sound combinations (both real- and non-words) consisted of phonemes and phonological sequences with low phonotactic probabilities. Both real and nonwords were trained using the same procedures detailed below.

- Stage 2-Task 1: Perception Task: The therapist produced a real word or nonword sound combination (e.g., VC or VCC-VC). The therapist asked the participant to arrange pictures or graphemes to depict the target. For example, if the subject heard the VC "ip", they would select the graphemes /i/ and /p/.

RESULTS

To orient the reader to the results, Figure 1 of the appendix displays results prior to the phonomotor program, probe performance during training, and immediately following. At this time no three-month follow up data was available for analysis of maintenance. Figure 1 (a) represents sounds in isolation, Figure 1 (b) represents nonword repetition and corresponds, Figure 1 (c) represents real word accuracy and corresponds, Figure 1 (d) is the control measure.

In order to measure the strength of the change in behavior to the treatment being provided, effect size was measured using Cohen's d. A small effect size was interpreted as ≤ 2.60, medium ≤ 3.90, large ≤ 5.80 (Beeson, 2006). Calculation of d was achieved by taking the mean of A2 (post training value) minus the mean of A1 (pre training value) divided by the standard deviation of A1. Additionally, for the pre and post measures, change was calculated using paired students t-tests with significance set at $p = < .05$. Results regarding each research question are detailed below.

Research Question 1: Intelligibility and listener effort

Research question 1 addressed the functional outcomes of phonomotor accent modification training as measured by intelligibility and listener effort. In order to determine the extent to which generalization occurred, intelligibility and listener effort data were collected at both the sentence and discourse levels.

Intelligibility (sentences): The Harvard Sentences pre-training data revealed the participant was initially 82% intelligible (SD=3%) with post training increasing to 87% (SD=6%) (p= .123).

Intelligibility: Sentences	
Pre-training	82% (SD=3%)
Post-training	87% (SD=6%)
P	.123 (not significant)

Reliability: The single-measures ICC for intelligibility scores was .665, indicating moderate agreement among the listeners; the average-measures ICC was .856, indicating very strong agreement with the mean (Portney & Watkins, 2000).

Intelligibility (discourse): was measured using percent key words correctly identified by unfamiliar listeners during a picture description, and revealed pre-training intelligibility at 95% (SD=2.52%) and post-training intelligibility at 95% (SD=1%) (p = .678).

Intelligibility: Discourse	
Pre-training	95% (SD=2.52%)
Post-training	95% (SD=1%)
P	.678 (not significant)

Reliability: Mean inter-rater variability for the 8 descriptive stimuli was 0.0013, with a range of 0.0006 to 0.0022 for the three listeners. These very small variability scores indicate considerable adherence to the mean.

Listener effort (sentences): was scored using a 9-point rating scale performed by unfamiliar listeners. Pre-training data at the sentence level revealed an average score of 3.87 (SD = .36) and post 2.33 (SD = .58) resulting in a significant difference (p =. 002).

Listener Effort: Sentence	
Pre-training	3.87 (SD = .36)
Post-training	2.33 (SD = .58)
P	. 002 (significant)

Reliability: The single measures ICC for listener effort was .438, indicating fair agreement among all judges; the average measures ICC was .700, indicating strong agreement with the mean.

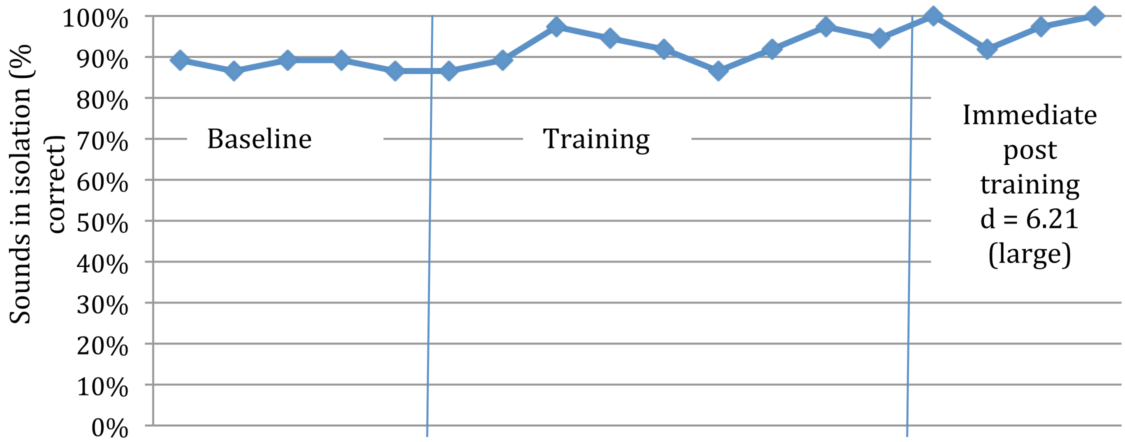
Listener effort (discourse): the same rating tool used for the discourse level picture description revealed pre 4.17 (SD=0.52) and post 4.00 (SD=1.15) with (p = .76) indicating no significance.

Listener Effort: Discourse	
Pre-training	4.17 (SD=0.52)
Post-training	4.00 (SD=1.15)
P	.76 (not significant)

Reliability: Inter-rater variability was calculated as above for the picture description task. Mean inter-rater variability for the 8 descriptive stimuli was 0.7778, with a range of 0.1806 to 1.2222 for the three listeners. These relatively low variability scores suggest adherence to the mean.

<u>Research Question 2: Acquisition of trained sounds</u>

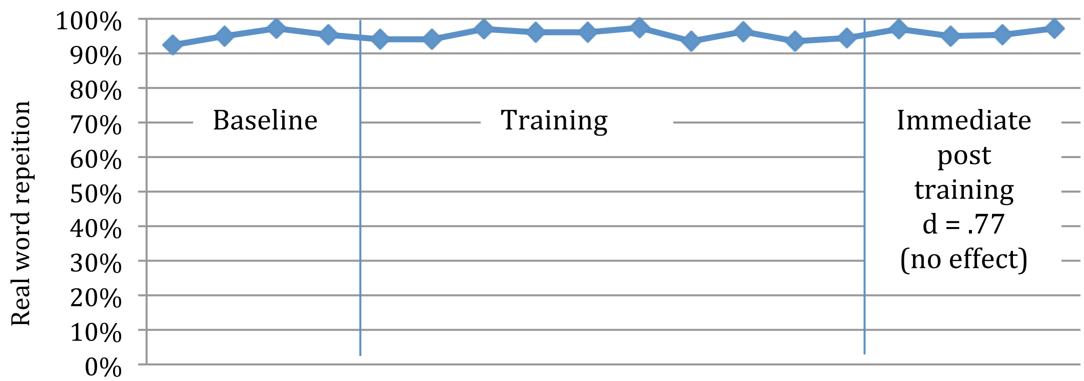
Accuracy of trained sounds prior to training was 88% (SD 1%) and post training was 97% resulting in a large effect size (ES = 6.21).



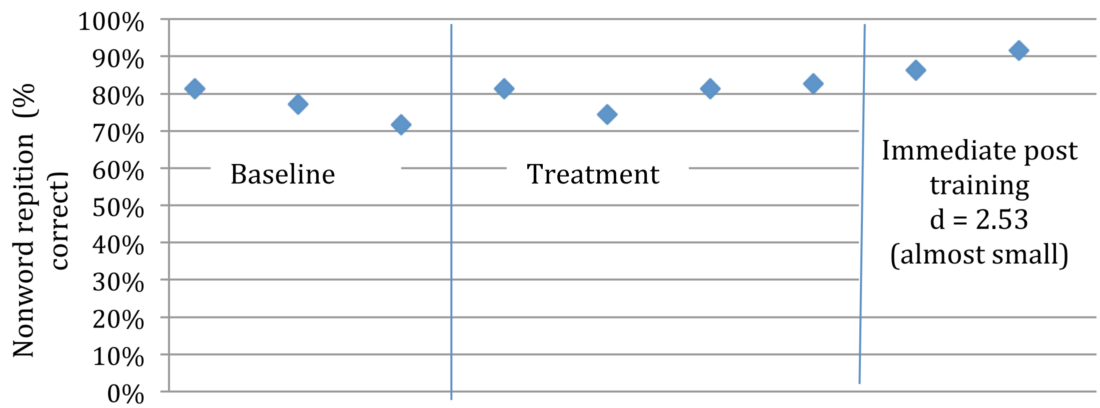
Reliability: Intrarater reliability was found to be 83% while interrater reliability was 60%.

Research Question 3: Generalization of trained sounds to untrained contexts

Trained phonemes embedded in real words: were 95% (SD = 2%) accurate pre training and 96% post training yielding a no effect size of (d = .77).



Percent correct of trained phonemes in nonwords: pre training was 78% (SD = 6%) and 89% post training. At (d = 2.53) the effect size fell slightly short of meeting the criteria for a small effect.



Phonemes embedded in sentences: elicited via repetition and supplemented with written stimuli were 76% (SD = .35) accurate pre training and 82% (SD = 1.45) post training (p = .092).

Percent correct phonemes embedded in sentences	
Pre-training	76% (SD = .35)
Post-training	82% (SD = 1.45)
P	.092 (not significant)
Inter-judge reliability	83.33%
Intra-judge reliability	92.40%

Reliability: Inter-judge reliability was 83.33% for pre-test measures and 92.40% post-testing.

<u>Pre and Post Training Standardization Tests</u>

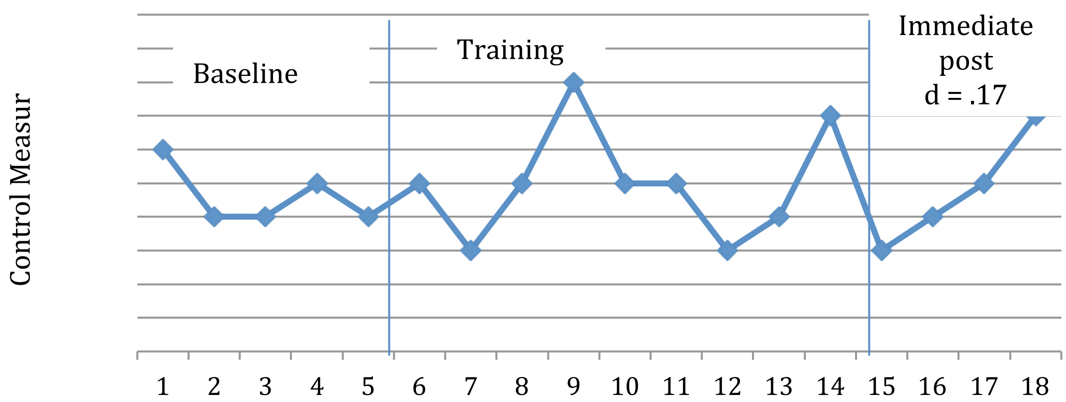
The SAPA was given as both a pre and post training formal measure to evaluate phonological awareness. In pre-testing the participant received a score of 91/151 (60%) and in post testing the score increased to 105/151 (70%).

SAPA	Pre-Testing	Post-Testing
Subtest 3: Repetition, Parsing, Blending	56%	76%
Subtest 1: Reading	44%	46%
Subtest 2: Auditory Phonological Processing	75%	85%

An addition pre/post measure the CPIB was used to capture the participant's own perception regarding the effects of his accent on communicating in different contexts. At 66% pre- to 66% there was no change in scores post-training.

Control Measure

Our intent was to find a linguistic control measure that would not be expected to change in relation to the phonomotor training. Pre-training scores on the sentence grammaticality judgment test revealed 46% (SD = 9) accuracy. With post-treatment accuracy at 48% (d = .17), no statistical change was noted. Considering the mean score on the TOAL-3 was approximately 47% correct with relatively consistent variability across probes, it appears that this was an effective and appropriate control measure.



DISCUSSION

The purpose of this Phase I, proof of concept, study was to investigate whether a non-native speaker could improve production of TL phonemes trained in isolation and to explore how far those changes would extend along the speech hierarchy (i.e., sounds in isolation, real words, nonwords, sentences, and discourse). Although a large effect size was found for trained sounds in isolation, results beyond that were variable. The discussion that follows examines possible explanations for these findings and identifies further areas of investigation needed to more thoroughly answer the research questions presented in this study. In order to follow the progression of the speech hierarchy, the discussion will begin with research question 2 “acquisition of trained sounds”, move to research question 3 “generalization of trained sounds to untrained contexts”, address research question 1 “intelligibility and listener effort”, and finally consider some of the additional measures. It is important to note that although training progressed through all stages presented in the program protocol, only consonant-vowel-consonant syllable structures were mastered in perception/production tasks. In the last four hours of treatment, the participant reached 60% accuracy with production of consonant-vowel-consonant-consonant (CVCC) structures. In addition, he also showed increased awareness of errors with lax vowels.

Acquisition of trained sounds

Based on the combined theories stemming from the accent modification research (Strange, 1995), principles of neuroplasticity (Kleim, 2008) and the PDP model (Nadeau, 2001) we hypothesized that the multisensory phonomotor approach to teaching phoneme perception and production would be sufficient to significantly improve TL phoneme production in ESL speakers following twenty hours of intensive instruction. Positive results revealing a large effect size strongly support this hypothesis indicating that ESL speakers can improve TL phoneme level production with phonomotor training. These results are consistent with prior accent modification studies (Chakraborty, 2011; Couper, 2006; De Bot, 1983; Derwing, 1998; Flege, 1995; Saito, 2012) and principles of neuroplasticity (Kleim, 2008; Nadeau, In Press). Indeed, given sufficient stimulation, feedback and targeted perception/production opportunities, especially in a neurotypical brain, change does occur.

While all sounds responded positively to training, select phonemes were deemed less 'stable' following training. In other words, select phonemes were more variable compared to others across linguistic contexts following training. For example, /r/, /ɪ/ and diphthongs greatly improved in isolation but regressed to resemble pre-training production in words, sentences and discourse. Two primary reasons are offered to explain this phenomenon. First, due to their inherent degree of complexity (i.e., markedness), these sounds are introduced later in the program thus receiving less treatment time. It is possible that with more training time these sounds would show greater consistency as well. The second explanation harkens back to Flege's notion of categorical perception and the degree of difficulty associated with learning novel versus similar TL sounds as opposed to those with NL to TL direct transfer. Because phoneme analysis of the participant's NL was not performed, it is still unclear how NL representations influence the modification and development of TL phonemes. In spite of this, according to the theories previously discussed, we would anticipate that with additional training, input and repetition, these sounds too would show more stable improvements. Additional research is needed to verify these claims.

Generalization of trained sounds to untrained contexts

Three separate linguistic contexts were assessed to examine the degree to which phoneme training in isolation generalized. These measures were designed to capture phoneme stability with increased contextual complexity. In order to compare the integrity of the TL phonological network being manipulated, we examined phoneme production in real, nonwords, and sentences.

In regard to trained sounds embedded in real word results, phoneme accuracy was assessed to be 95% accurate at baseline and increased to 96% post testing. No change was detected in part because the baseline presentation was too high to show a statistically significant level of change. In addition, improved phoneme accuracy of sounds embedded in real words, compared to isolation suggests that performance was indeed boosted by lexical/semantic knowledge.

Trained phonemes in nonwords, which have no lexical/semantic input and are purely a phonomotor task, approached near significance ($d = 2.53$). These findings suggest that intensive phonomotor practice improved phonomotor representations and are consistent with study done by Kendall et al. 2008. It is anticipated that with more practice and training, nonword production would also reach significant levels of improvement. As mentioned above, the discrepancy between nonword and real word performance is similar to that of sounds in isolation and real words, indicating that without lexical/semantic involvement the phonological system remains compromised.

Continuing along the speech hierarchy, phoneme accuracy was also examined at the sentence level with pre- (76%) and post-training (82%). Again, while this is not enough to reach statistical significance ($p = .092$), it does indicate a positive trend, providing further support for improvements in TL phonological awareness. It is interesting to note that phoneme accuracy in sentences was lower than phonemes embedded in real words. This finding coincides with the increase in errors typically seen with an increase in linguistic level due to greater complexity and demand on cognitive resources. That said more phoneme errors in sentences (compared to real words in isolation) might also be the product of NL phonological constraints becoming more prevalent at the sentence level. Recall that in the MDH example of German versus French (Eckman, 2008, p. 97), the German speakers consistently devoiced word final consonant stops because this followed the phonological rules of that language. It is plausible that our participant was less susceptible to these errors in isolation and in words, but more prone in sentences. This may be due to the pervasive effects of NL on TL and the decreased ability to control those effects as more resources are allocated to producing a lengthy utterance rather than the discrete units of which it is comprised. Further detailed error analysis and comparison to NL phonological constraints are needed to provide insight.

Intelligibility and listener effort

Since spontaneous communication is the combination of sounds, words, and sentences, we measured intelligibility and listener effort at both the sentence and discourse levels to obtain a more ecologically valid picture of how training affected overall communicative performance.

Intelligibility and listener effort are discussed independently within each linguistic setting since results suggested a context dependent relationship.

While sentence level intelligibility increased marginally, it was insufficient to denote any significant change. However, listener effort did decrease significantly (p = .002). In other words, while the unfamiliar listeners' comprehension of the speaker's message did not change, the amount of effort expended to comprehend that message was deemed less effortful. It is likely that this signifies a general shift in phoneme accuracy towards improved approximations of the targets. It is possible that rather than an entire corpus shift in phoneme accuracy, it is the same pervasive sounds identified in isolation that negatively affected sentence level intelligibility. Further error analysis of the unfamiliar listeners' broad transcription would be required to determine whether this is the case. In addition, investigation into the comparative saliency of these sounds on listeners' comprehension would be beneficial.

Finally, there were no significant changes seen in either intelligibility nor listener effort in discourse. The most likely conclusion is that treatment of sounds in isolation did not generalize to spontaneous speech with the given amount of training. Considering the positive trends at all other linguistic levels, it is likely that with more training and advancement through the syllable structure hierarchy, improvement would generalize to spontaneous speech. It is also possible that suprasegmental factors such as prosody are more prevalent in discourse and have a greater effect on intelligibility and listener effort. A more profound understanding of the impacts of suprasegmental versus segmental errors on intelligibility and listener effort is needed to better understand generalization at this level.

Pre- and post-training performance on standardized tests

This individual showed improvement in phonological awareness as evidenced by a 10-percentage points gain on the SAPA. Not only does this overall improvement on the SAPA corroborate a general positive trend, but also gains were noted across all subtests.

Lack of improvement in CPIB scores can be interpreted one of two ways. Either the participant truly felt the training made no difference in his everyday communication, or the measure was not sensitive enough to detect the quality of change., The fact that the participant expressed feeling more confident and better understood by unfamiliar listeners when ordering coffee points to the sensitivity of the items capturing overall improvement. Additionally, he felt that gains were made during the week but noticed a regression of skills following weekends when only the NL was spoken. Since the CPIB was intended for individuals with acquired communication disorders, it is possible that further adaptation is required to accurately capture the communication difficulties experienced by ESL speakers.

<u>Control Measure</u>

Limitations and Future Directions

The intention of this Phase I study was to verify the efficacy of treatment. Robey defines 'efficacy' as, "The researcher's task is to assure that only the effects of the independent variable (i.e., the treatment protocol) on the dependent variable (i.e., the clinical-outcome) plausibly accounts for observed change in the outcome measure" (Robey, 2004, p. 402). While our study was successful in this regard, further Phase II research is necessary to determine its treatment effectiveness or, "potency of a treatment protocol for brining about change in the real world" (Robey, 2004, p. 402). It is possible that extrinsic variables such as the participant's native language and level of English proficiency influence the effectiveness of this program; therefore, further replication is needed across individuals with differing linguistic backgrounds and degrees of accentedness will help to identify optimal dosage and participant characteristics. Further areas of refinement for the treatment protocol and measures should include:

- Early identification of difficult TL sounds to maximize training program and devote more time to difficult sounds.
- Formal examination of error type behavior as linguistic level increases (e.g., are errors at more complex contexts consistent with NL phonotactic constraints, challenging TL sounds, or both?).
- Perform an exit interview with the participant to qualitatively capture how he/she felt the training changed his/her behaviors. From that a more sensitive short form measure could be developed to replace the CPIB.
- Develop a data collection tool to measure perceptual changes of familiar conversation partners (e.g., family, classmates, friends…).
- Investigation of suprasegmental effects, such as prosody, on intelligibility and listener effort: is it prudent to incorporate direct suprasegmental training into the phonomotor program?
- Although 3-month follow-up testing was scheduled for after the date of submission for this document, maintenance data is needed to determine the long-term effects of training.

Acknowledgements

First and foremost, I would like to thank the participant for volunteering his time and eagerness to partake in this study. Additionally, I owe a debt of gratitude to the entire research team. To Dr. Diane Kendall Ph.D., CCC-SLP, my advisor and mentor for her willingness to take on this project and extend her research in a new direction. Also, for showing me the excitement and rewards of research and for guiding me through the investigative process. A huge thanks to Kathy Nagle Ph.C., CCC-SLP for developing the primary outcome measures and to her undergraduate students for fulfilling the role of unfamiliar listeners. Both Kathy and Dr. Amber Franklin Ph.D., CCC-SLP, were instrumental in directing my review of the accent modification literature and collaborating to bring a novel approach to accent modification. To the additional members of the UW Aphasia lab Liz Brookshire M.S., CCC-SLP, Lauren Bislick Ph.C., CCC-SLP, and Megan Oelke M.S., CCC-SLP I extend my gratitude for their involvement throughout the research process and constant emotional support. A special thanks to Megan for teaching me the ropes of the phonomotor program and answering countless questions. A final thinks to the members of my thesis committee Dr. Kristie Spencer, Ph.D., CCC-SLP and Dr. Tanya Eadie, Ph.D., CCC-SLP for their revisions and insight.

REFERENCES

ASHA. (2012, 09/25/12). Accent Modification Retrieved 10/11/12, 2012, from http://www.asha.org/public/speech/development/Accent-Modification/

Baylor, C., Yorkston, K., Eadie, T., Chung, H., & Amtmann, D. . (In Press). The Communicative Participation Item Bank (CPIB): Item bank calibration and development of a disorder-generic short form. *Journal of Speech Language and Hearing Research, In Press*.

Beeson, P., & Robey, R. (2006). Evaluating Single-Subject Treatment Research: Lessons Learned from the Aphasia Literature. *Neuropsychology Rev, 16*, 161-169. doi: 10.1007/s11065-006-9013-7

Burgess, J., & Spencer, S. . (2000). Phonology and Pronunciation in Integrated Language Teaching and Teacher Education *Elsevier Science Ltd, System 28*, 191-215.

Chakraborty, R., Somsch, C., & Gonzales, M.D. . (2011). Articulatory Behaviors of Nonnative Speakers: Role of L2 Proficiency and Accent Modification. *Perceptual and Motor Skills, 113*(1), 311-330. doi: 10.2466/11.17.21.28.PMS.113.4.311-330

Champagne-Muzar, C., E. I. Scheiderman, & J. S. Bourdages. (1993). Second Language Accent: The role of the pedagogical Environment *International Review of Applied Linguistics, 31*(2), 143-160.

Couper, G. (2006). The Short and Long-Term Effects of Pronunciation Instruction *Prospect, 21*(1), 46-66.

De Bot, K. (1983). Visual Feedback of Intonation I: Effectiveness and Induced Practice Behavior. *Language and Speech, 26*(4), 331-350.

Derwing, T., Munro, M., & Wiebe, G. (1997). Pronunciation Instruction for "Fossilized" Learners Can it Help? *Applied Language Learning, 8*(2), 217-235.

Derwing, T., Munro, M., & Wiebe, G. (1998). Evidence in Favor of a Broad Framework for Pronunciation Instruction *Language Learning, 48*(3), 393-410.

Eckman, F. (2008). Typological markedness and second language phonology. In J. G. H. E. a. M. L. Zampini (Ed.), *Phonology and Second Language Acquisition* (pp. 95-116). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Flege, J. E. (1995). Second Language Speech Learning *Theory, Findings, and Problems*. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233-277). Maryland: York Press.

Franklin, A. (2009). English Vowel Production in Japanese, Korean, and Spanish ESL Speakers After Vowel-Focused Pronunciation Training.

Gluszek, A., & Dovidio, J.F. . (2010). The Way *They* Speak: A Social Psychological Perspective on the Stigma of Nonnative Accents in Communication *Personality and Social Psychology Review, 14*(2), 214-237. doi: 10.1177/1088868309359288

Hammill, D., Brown, V., Larsen, S., & Wiederholt, L. . (1994). Test of Adolescent and Adult Language *Assessing Linguistic Aspects of Listenting, Speaking, Reading, and Writing* Austin, TX: Pro.ed.

Kendall, D., Rosenbek, J., Heilman, K., Conway, T., Klenberg, K., Gonzalez, L., Nadeau, S. . (2008). Phoneme-based Rehabilitation of Anomia in Aphasia. *Brain and Language, 105*, 1-17.

Kiran, S. T., C.K. . (2003). Effect of typicality on online category verification of animate category exemplars in aphasia. *Brain and Language, 85*, 3441-3450. doi: 10.1016/S0093-934X(03)00064-6

Kleim, J. (2008). Principles of Experience-Dependent Neural Plasticity: Implications for Rehabilitation After Brain Damage. *Journal of Speech Language and Hearing Research, 51*, S225-S239.

Munro, M. J., & Derwing, T.M. (2006). The Functional Load Principle in ESL Pronunciation instruction: An exploratory Study. *System, 34*, 520-531. doi: 10.1016/j.system.2006.09.004

Nadeau, S. E. (2001). Phonology: a Review and Proposals for a Connectionist Perspective. *Brain and Language, 79*, 511-579. doi: 10.1006/brln.2001.2566

Nadeau, S. E. (2012). *The neural architecture of grammar*. Cambridge, Mass.: MIT Press.

Nadeau, S. E. (In Press). *Neuroplastic Mechanisms of Language Recovery after Stroke*.

Perlmutter, M. (1989). Intelligibility Rating of L2 Speech Pre-and Postintervention *Perceptual and Motor Skills, 68*, 515-521.

Plaut, D. C. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language, 52*, 25-82. doi: 10.1006/brln.1996.0004

Portney, L. W., M. (2000). *Foundations of clinical research: applications to practice* (2nd ed.): Upper Saddle River: Prentice Hall Health

Robey, R. (2004). A Five-Phase Model for Clinical-Outcome Research. *Journal of Communication Disorders, 37*(5), 401-411.

Saito, K., & Lyster, R. (2012). Effects of Form-Focused Instruction and Corrective Feedback on L2 Pronunciation Development of /®/ by Japanese Learners of English. *Language Learning 62*(2), 595-633.

Shrout, P. F., JL. (1979). Intraclass correlation: uses in assessing rater reliability. . *Psychological Bulletin, 86*(2), 420-428.

Sikorski, L. D. (2005). Foreign Accents: Suggested Competencies for Improving Communicative Pronunication *Seminars in Speech and Language, 26*(2), 126-130.

Strange, W. (1995). Cross-Language Studies of Speech Perception a Historical Review. In W. Strange (Ed.), *Speech Perception and Linguistic Experience* (pp. 3-45). Timonium, Maryland: York Press.

Strokel, H. L., Armbruster, J., & Hogan, T.P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech Language and Hearing Research 47*, 1175-1192. doi: 10.1044/1092-4388(2006/085)

Thompson, C. K. S., L.P. . (2007). Complexity of treatment of syntactic deficits. *American Journal of Speech-Language Pathology, 16*, 30-42. doi: 10.1044/1058-0360(2007/005)

Vitevitch, M. S. L., P.A. . (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods Instruments & Computers, 36*, 481-487. doi: 10.3758/BF03195594

APPENDIX

Table 1- primary and secondary outcome measures:

Aims	Research Question	Outcome Measure	Administration schedule
Primary Aim:	1. Is there a significant improvement in intelligibility and listener effort following training?	1. Intelligibility: percent of key words correctly understood by an unfamiliar listener in: sentences (Harvard Sentences) and conversational speech (3 min discourse sample based off of a picture description from the Western Aphasia Battery) 2. Listener effort: visual rating scale indicating no effort to maximum effort as determined by unfamiliar listeners (sentences and picture description)	Pre- and post-training
Secondary Aims	1. Acquisition: Is there a significant difference in production of trained sounds following treatment?	Accuracy of repetition of trained sounds in isolation	Repeated probe
	2. Generalization: Is there a significant difference in production in untrained stimuli?	1. Percent correct of trained phonemes embedded in real words and elicited during single word repetition 2. Percent correct of trained phonemes embedded in nonwords and elicited via repetition 3. Percent correct of trained phonemes embedded in sentences and elicited via repetition and supplemented with written stimuli	Repeated probe Repeated probe Pre/post
	3. Control: Is there a significant difference in grammatical judgments at the sentence level?	Percent correct grammatical judgments of sentences presented auditorily from the TOAL	Repeated probe

Standardized Assessment of Phonological Awareness (SAPA):
Pre/Post measure describing the participant's phonological awareness

	RAW SCORE	% ACCURACY
Subtest 1 – Reading		
Real words (total 21)		
Nonwords (total 12)		
Pseudohomophones (total 8)		
Irregular words (total 11)		
TOTAL 52 points		
Subtest 2 – Auditory Phonological Processing		
Real word rhyme (total 15)		
Nonword rhyme (total 23)		
Lexical decision (total 12)		
Minimal pairs (total 15)		
TOTAL 65 points		
Subtest 3 – Repetition, Parsing, Blending		
Real word repetition (total 9)		
Nonword repetition (total 6)		
Real word blending (total 5)		
Nonword blending (total 4)		
Real word parsing (total 5)		
Nonword parsing (total 5)		
TOTAL 34 points		
TOTAL SAPA SCORE (151 points)		

<u>Communication Participation Item Bank (CPIB) ORIGINAL:</u>
Quality of life Pre/Post measure

The Communicative Participation Item Bank – General Short Form

Instructions:
The following questions describe a variety of situations in which you might need to speak to others. For each question, please mark how much your condition interferes with your participation in that situation. If your speech varies, think about an AVERAGE day for your speech – not your best or your worst days.

	Not at all (3)	A little (2)	Quite a bit (1)	Very much (0)
1. Does your condition interfere with... **…talking with people you know?**	○	○	○	○
2. Does your condition interfere with... **…communicating when you need to say something quickly?**	○	○	○	○
3. Does your condition interfere with... **…talking with people you do NOT know?**	○	○	○	○
4. Does your condition interfere with... **…communicating when you are out in your community (e.g. errands; appointments)?**	○	○	○	○
5. Does your condition interfere with... **…asking questions in a conversation?**	○	○	○	○
6. Does your condition interfere with... **…communicating in a small group of people?**	○	○	○	○
7. Does your condition interfere with... **…having a long conversation with someone you know about a book, movie, show or sports event?**	○	○	○	○
8. Does your condition interfere with... **… giving someone DETAILED information?**	○	○	○	○
9. Does your condition interfere with... **…getting your turn in a fast-moving conversation?**	○	○	○	○

10. Does your condition interfere with... **…trying to persuade a friend or family member to see a different point of view?**	○	○	○	○

Scoring guide for the CPIB General Short Form

To score the short form, add the scores for the ten items to obtain a summary score (Not at all = 3; A little = 2; Quite a bit = 1; Very much = 0). The summary score will range from 0 – 30. High scores are more favorable, meaning that high scores indicate less interference in participation. Using the table below, the summary scores can be converted to IRT theta values (logit scale). On the logit scale, scores generally range from -3.0 to +3.0 with 0 logits representing the mean for the calibration sample. Again, high scores are preferable. The table also includes a conversion to standard T scores (mean = 50; standard deviation = 10). **VERY IMPORTANT: This score translation table is ONLY valid for the 10 item short form presented in this manuscript.** Remember that in IRT, the person score is based on the parameters of the individual items and on how the person answers the items. This scoring table has been generated using the item parameters for the ten items in this short form, and these parameters would differ for different items. A new score translation table must be created for any other combination of items.

CPIB 10-Item General Short Form Scoring Table

Summary	Theta	T score	Summary	Theta	T score
0	-2.58	24.20	16	-0.22	47.80
1	-2.18	28.20	17	-0.10	49.00
2	-1.94	30.60	18	0.03	50.30
3	-1.76	32.40	19	0.15	51.50

CPIB 10-Item General Short Form Scoring Table

Summary	Theta	T score	Summary	Theta	T score
4	-1.60	34.00	20	0.27	52.70
5	-1.46	35.40	21	0.40	54.00
6	-1.34	36.60	22	0.53	55.30
7	-1.22	37.80	23	0.65	56.50
8	-1.10	39.00	24	0.78	57.80
9	-0.99	40.10	25	0.92	59.20
10	-0.89	41.10	26	1.06	60.60
11	-0.78	42.20	27	1.22	62.20
12	-0.67	43.30	28	1.42	64.20
13	-0.56	44.40	29	1.67	66.70
14	-0.45	45.50	30	2.10	71.00
15	-0.33	46.70			

<u>Communication Participation Item Bank (CPIB) ADAPTED VERSION:</u>
Quality of life Pre/Post measure

Accent Questionnaire

Instructions:
The following questions describe a variety of situations in which you might need to speak to others. For each question, please mark how much your accent affects your participation in that situation. If your accent varies, think about an AVERAGE day for your speech – not your best or your worst days. We are asking about your accent or the way your speech sounds, NOT the words you use or grammar.

	Not at all (3)	**A little (2)**	**Quite a bit (1)**	**Very much (0)**
1. Does your accent interfere with... **…talking with people you know?**	○	○	○	○
2. Does your accent interfere with... **…communicating when you need to say something quickly?**	○	○	○	○
3. Does your accent interfere with... **…talking with people you do NOT know?**	○	○	○	○
4. Does your accent interfere with... **…communicating when you are out in your community (e.g. errands; appointments)?**	○	○	○	○
5. Does your accent interfere with... **…asking questions in a conversation?**	○	○	○	○
6. Does your accent interfere with... **…communicating in a small group of people?**	○	○	○	○
7. Does your accent interfere with... **…having a long conversation with someone you know about a book, movie, show or sports event?**	○	○	○	○
8. Does your accent interfere with... **… giving someone DETAILED information?**	○	○	○	○
9. Does your accent interfere with... **…getting your turn in a fast-moving conversation?**	○	○	○	○
10. Does your accent interfere with... **…trying to persuade a friend or family member to see a different point of view?**	○	○	○	○

Nonword probes (half list): Secondary outcome measure pre/during/post measure. One data point equals a balanced full list containing 2 half lists.

NW PROBES		Pronunciation
1	ain	AIN
2	chewtee	**CHEW**-TEE
3	doi	D-OY
4	af	AE-F (rhyme with at)
5	tuse	T-OOS
6	pomb	P-OOM
7	zheree	**ZURE**-EE
8	foko	**FOE**-KOE
9	lehber	**L-EH**-BER (rhyme webber)
10	wurkee	**WUR**-KEE
11	kotoe	**KO**-TOE
12	gee	G-EE (like /g/ in goat)
13	sheiv	SH-EEV
14	hage	HAY-J
15	doem	**DOUGH**-EM (rhyme poem)
16	mefow	**MEH-**FO
17	shever	**SHEV**-ER (rhyme ever)
18	feether	**FEE**-**TH**UR (rhyme breather)
19	toyler	**TOY**-LER
20	wayzer	**WAY**-ZER
21	ek	EH-K
22	datch	D-ATCH
23	peans	P-EEN-Z
24	izul	**EYE**-ZUL
25	shaybee	**SHAY**-BEE
26	veeder	**VEE**-DER
27	poa	**PO**-AH
28	rooit	**ROO**-IT
29	loy	L-OY
30	meeth	M-EETH
31	heag	H-EEG
32	zower	**ZOW**-ER (rhyme tower)
33	tawthee	**TAW**-THEE
34	jihver	**JIH**-VER
35	wooter	**WOO**-TER
36	ri	R-IH
37	ish	IH-SH
38	whup	WH-UP
39	breek	BR-EE-K
40	dungee	**DUNG**-EE
41	turmee	**TUR**-MEE
42	leczure	**L-EHK**-ZURE

43	lehkey	**<u>LEH</u>**-KEY
44	voo	V-OO
45	eepe	EE-P
46	reesh	R-EESH
47	jong	J-ONG
48	juro	**<u>JUR</u>**-O
49	shaso	SHASSO (rhyme lasso)
50	hoiter	HOY-TER
51	sayvay	SAY-VAY
52	nuy	N-EYE
53	aidg	EYE-J
54	zine	Z-EYE-N
55	broise	B-R-OY-Z
56	fooer	**<u>FOO</u>**-ER
57	neenee	NEE-NEE
58	rayzel	**<u>RAY</u>**-ZUL
59	laybee	**<u>LAY</u>**-BEE
60	thag	TH-AG
61	oit	OY-T
62	kurr	K-ER
63	phroose	FROOSE
64	high-ger	**<u>HIGH</u>**-GER (rhyme with tiger)
65	wowah	WO-WUH
66	unger	UNG-ER
67	grazy	GRAY-ZEE
68	poi	P-OY
69	aub	AW-B
70	jeef	J-EEF
71	grake	G-R-AKE
72	ehkee	**<u>EH</u>**-KEY

<u>Assessment of Intelligibility of Dysarthric Speech (AIDS):</u>
Secondary outcome measure pre/during/post measurement
List 1

1. Swore
2. Trace
3. Fear
4. Root
5. Pitting
6. Darn
7. Dead
8. Glitter
9. Port
10. Dead
11. Pete
12. Renter
13. Burn
14. Weave
15. Gable
16. Damp
17. Beat
18. Fall
19. Sprain
20. Mall
21. Soul
22. Seat
23. Grade
24. Less
25. Shut
26. Serve
27. Intact
28. So
29. Bit
30. Grape
31. Wake
32. Wit
33. Wire
34. Shop
35. Par
36. Ought
37. Crawl
38. Soul
39. Lake
40. Teacher
41. Lie
42. Defy
43. Live
44. Tanner
45. Grab
46. Spark
47. Glass
48. Jade
49. Pace
50. Near

<u>Harvard Sentences:</u> used for both primary and secondary outcome measures. Administered at pre/post testing.

List 1

1. The birch canoe slid on the smooth planks.
2. Glue the sheet to the dark blue background.
3. It's easy to tell the depth of a well.
4. These days a chicken leg is a rare dish.
5. Rice is often served in round bowls.
6. The juice of lemons makes fine punch.
7. The box was thrown beside the parked truck.
8. The hogs were fed chopped corn and garbage.
9. Four hours of steady work faced us.
10. Large size in stockings is hard to sell.

List 2

1. The boy was there when the sun rose.
2. A rod is used to catch pink salmon.
3. The source of the huge river is the clear spring.
4. Kick the ball straight and follow through.
5. Help the woman get back to her feet.
6. A pot of tea helps to pass the evening.
7. Smoky fires lack flame and heat.
8. The soft cushion broke the man's fall.
9. The salt breeze came across from the sea.
10. The girl at the booth sold fifty bonds.

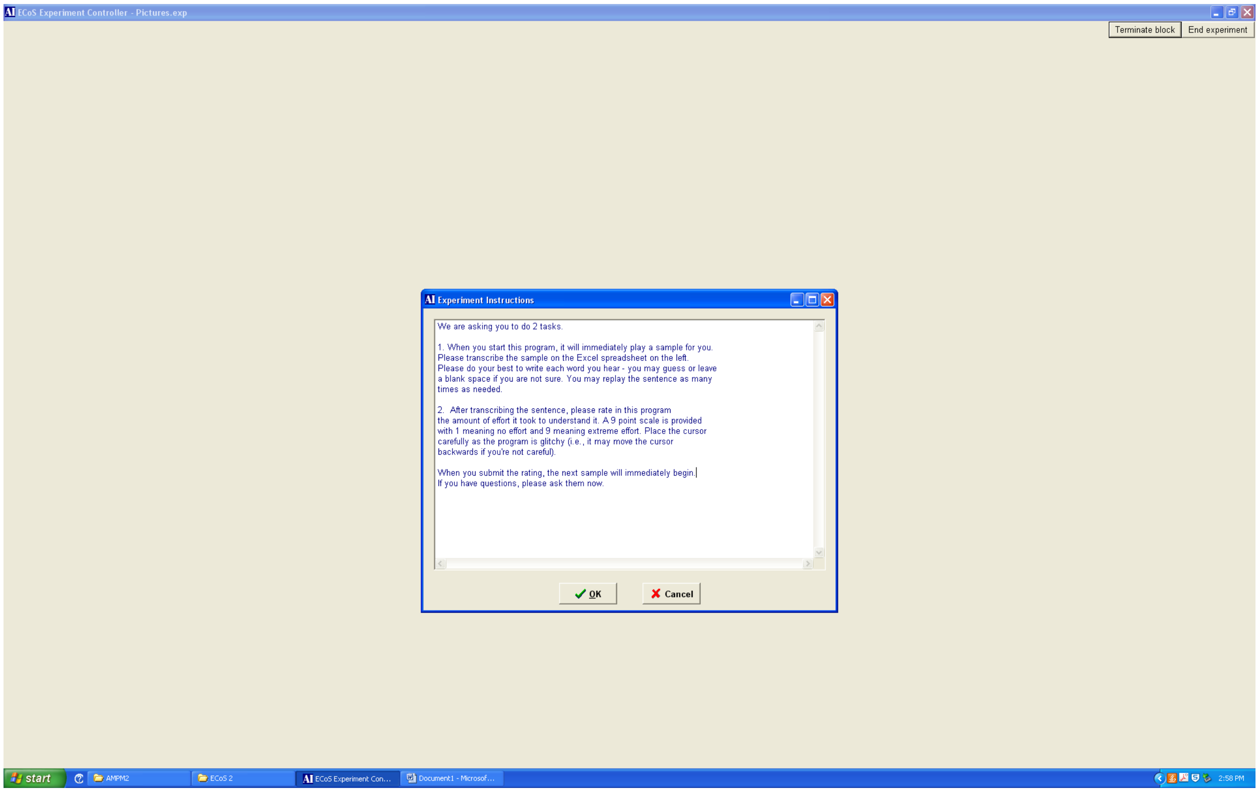
List 3

1. The small pup gnawed a hole in the sock.
2. The fish twisted and turned on the bent hook.
3. Press the pants and sew a button on the vest.
4. The swan dive was far short of perfect.
5. The beauty of the view stunned the young boy.
6. Two blue fish swam in the tank.
7. Her purse was full of useless trash.
8. The colt reared and threw the tall rider.
9. It snowed, rained, and hailed the same morning.
10. Read verse out loud for pleasure.

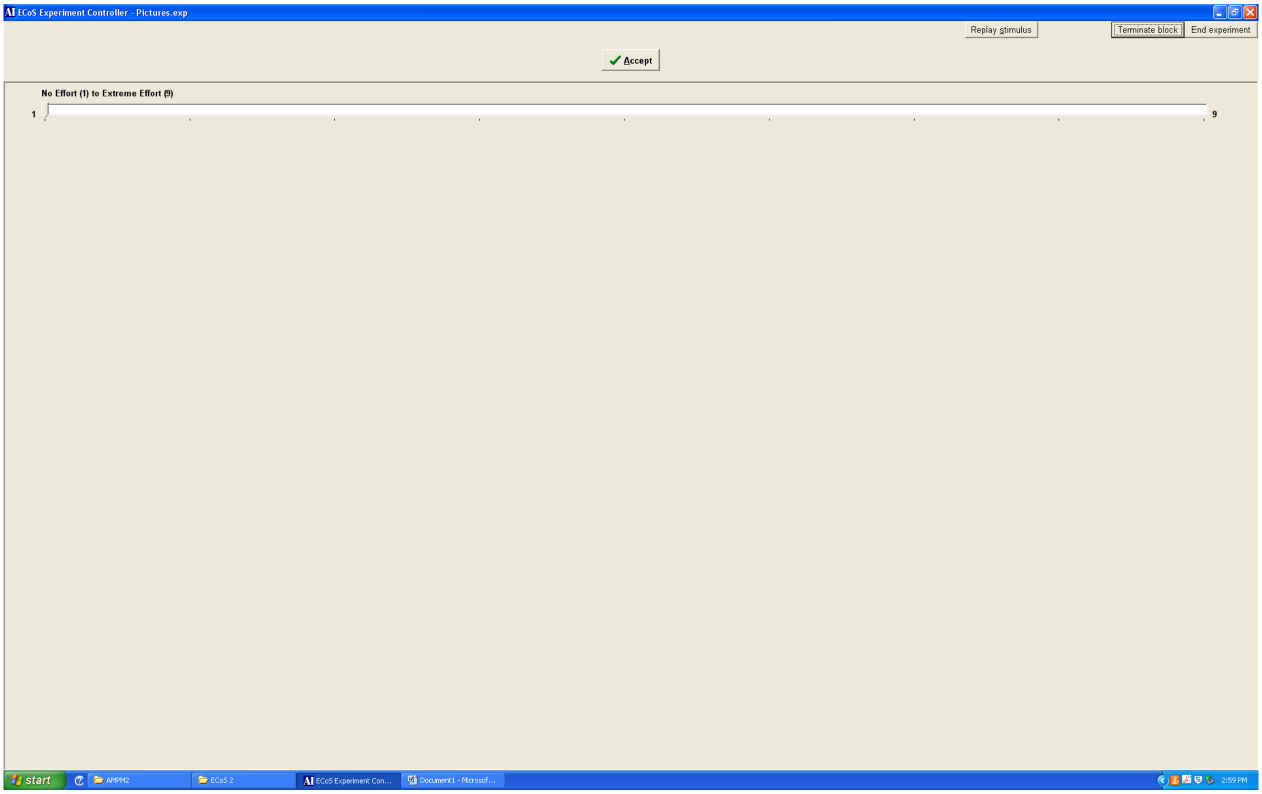
List 4

1. Hoist the load to your left shoulder.
2. Take the winding path to reach the lake.
3. Note closely the size of the gas tank.
4. Wipe the grease off his dirty face.
5. Mend the coat before you go out.
6. The wrist was badly strained and hung limp.
7. The stray cat gave birth to kittens.
8. The young girl gave no clear response.
9. The meal was cooked before the bell rang.
10. What joy there is in living.

Unfamiliar Listener Instructions for Harvard Sentences & Picture Descriptions: Primary outcome measure, scored by unfamiliar listeners.



<u>Unfamiliar Listener Scale used for Harvard Sentence & Picture Descriptions:</u> Primary outcome measure, scored by unfamiliar listeners.



<u>Western Aphasia Battery Picture:</u> used for picture description stimulus for primary outcome measure

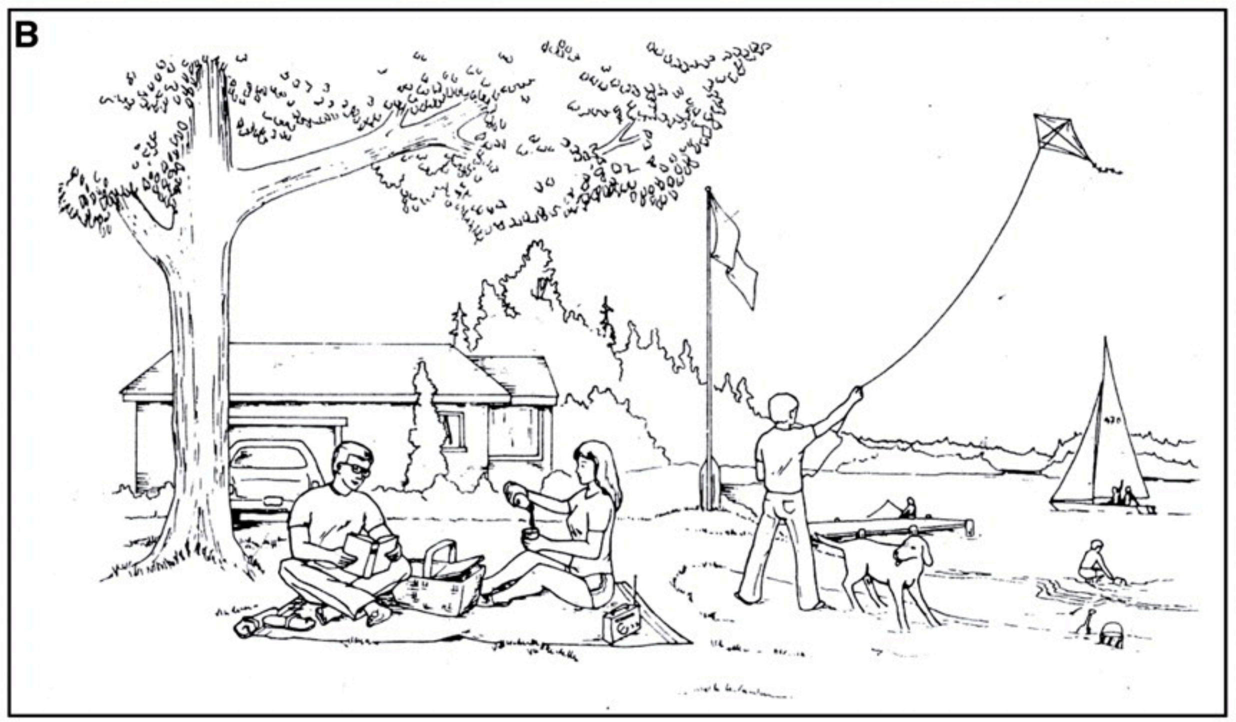


Figure 1: Repeated probe data