EVALUATIONS OF /r/ ATTEMPTS OF CHILDREN IN SPEECH THERAPY BY SPEECH-LANGUAGE PATHOLOGISTS AND CHILD EDUCATORS

by

BOSKO RADANOV

B.A., University of British Columbia, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Audiology and Speech Sciences)

THE UNIVERSITY OF BRITISH COLUMBIA

October 2007

Abstract

Background:

Previous studies of treatment for English /r/ (designated with the North American symbol /r/) have mainly used Speech-Language Pathologists (SLPs) as expert listeners and scalar rating methods (e.g. Chaney, 1998). Tasks have involved rank order judgment of natural or synthesized speech stimuli, with a variety of trained and untrained adult and child listeners.

Aims:

The present study set out to compare expert and untrained listener evaluations of different /r/ attempts by children. The two comparison groups were SLPs and Educators (teachers or child care workers). A secondary objective was to compare an identification listening task with a paired comparison task.

Methods and Procedures:

Sixteen /r/ syllables ([ræ], [ar]) were extracted from pre- and post-treatment field recordings of four Canadian English-speaking children. The two tasks (identification of tokens as /r/ or not /r/, and a forced choice comparison of /r/ pairs) were presented through Microsoft Powerpoint under headphones. Twenty SLPs and eighteen Educators judged the quality of the /r/ attempts. Formant analyses were also made of the stimuli.

Outcomes and Results:

The expert listeners (SLPs) showed higher intra-rater reliability: 91% on the pair-wise comparison task and 81% on the identification task, compared with 84% and 78% for the untrained listeners respectively. Inter-rater reliability on single measures (ICC Educators=.51 in comparison, .21 in identification; SLPs=.42 in comparison; .31 in

identification) was lower than that of average measures (ICC Educators=.96 in comparison, .87 in identification; SLPs=.95 in comparison; .92 in identification)

Rank order of sample ratings as on- or off-target was similar between the two groups. The rankings matched the normative formant data for /r/ published in Guenther et al. (1999) and Flipsen et al. (2000, 2001) for the best tokens, with SLPs providing a ranking closer to the acoustic norms.

Conclusions and Implications:

Trained listeners appeared to be better able to identify nuances in /r/ quality, as confirmed by acoustic analysis of /r/ tokens. Intra-rater reliability was higher for SLPs despite greater disagreement among SLPs for single measures of inter-rater reliability. The paired comparison task had higher reliability scores than the identification task for both listener groups.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my sincere gratitude to Prof. Barbara May Bernhardt, for her guidance and support. I thank her also for providing me an opportunity to learn and grow as a student and SLP in the unique research and practice environment she creates.

I also thank Dr. Bryan Gick and the Interdisciplinary Speech Research Laboratory for supporting me through this thesis and giving me my start in research. Many thanks for technical and non-technical advice, support, and editing.

Special thanks and appreciation to Dr. Jeff Small for all his contributions, support and editing, all of which enabled me to finish this thesis in a timely manner.

I would also like to express sincere thanks to Dr. Penelope Bacsfalvi, Marcy Adler-Bock and Geeta Modha who have traveled around British Columbia in order to collect the data that was used in this study.

Great appreciation to Maja Grubisic for her hard work, advice and many late hours of help. I am also thankful to Benjamin Perry for his support, input and sharing of ideas.

I owe a big thank you to my parents, who provided the item of greatest worth – the possibility. Thank you for standing by me through the many trials and decisions of my educational career.

Finally, I would like to thank my teachers, Barbara Bächmann and Walter Liebel, who got me excited about learning, who believed in me and who continue to believe in me, regardless of the distances involved.

CHAPTER 1

INTRODUCTION

Of all the phonemes used in the English language, /r/[1] is arguably one of the most difficult to master (Janzen & Shriberg, 1977; Ruscello, 1995, 1995b; Shriberg, Flipsen, Karlsson & McSweeny, 2000, 2001; Shuster, Ruscello & Smith, 1992). The primary goal of this paper was to determine whether different groups of adults have different capabilities in deciding what is an on-target /r/ attempt, based on their training, experience as listeners or some other factor. The first group of listeners consisted of SLPs and the second consisted of Educators (teachers and child care workers). A secondary goal of this study was to investigate whether the way in which listeners are presented with data affects judgment reliability, i.e., specifically comparing an identification task with a pair-wise comparison task. The following introduction reviews the characteristics and perception of the North American /r/, past research and listener variables and ends with the research questions raised in the current study.

Background

Articulation and Acoustics of /r/

The North American /r/ varies among speakers. The variance is due to differing articulatory and motor components and the type of /r/ produced. The /r/ can be produced as

[1] Although /r/ is the IPA symbol for the alveolar trill, it is often used to describe the North American rhotic. It is used in this paper to represent the North American English /r/.

a retroflexed' or 'bunched' /r/, or some alternative posture between these two (Alwan, Narayanan, & Haker, 1997; Delattre & Freeman, 1968; Hagiwara, 1995; Westbury, Hashi & Lindstrom, 1998). The /r/ has three constrictions above the larynx coupled with lateral tongue bracing. The retroflex /r/ is made with the tongue tip curled back towards the palate, in comparison with a bunched /r/ which is produced with the tongue body approaching the palate (Westbury et al., 1998). In both cases the tongue is retracted towards the rear of the pharynx (Delattre & Freeman, 1968). The lateral tongue bracing helps the midline of the tongue lower by anchoring the tongue on the back upper molars (Alwan et al., 1997).

Although visual perception affects perception of speech (Vatikiotis-Bateson, Munhall, Kasahara, Garcia & Yehia, 1996), the major contributor to speech perception is the auditory signal (Liberman, Harris, Hoffman & Griffith, 1957; Nygaard & Pisoni, 1995). Typically, the acoustic signal for /r/ shows a characteristic lowering, or rising (in the case of initial /r/) of the third formant (F3). There also exists a relatively small gap between the second (F2) and third (F3) formants (Delattre & Freeman 1968; Guenther, Espy-Wilson, Boyce, Matthies, Zandipour, & Perkell, 1999; Westbury et al. 1998). In a study centering on adolescents, Flipsen, Shriberg, Weismer, Karlsson & McSweeny (2000) reported an average F3 value of 1934 Hz and an average F2 of 1337 Hz. Measurements in F2 and F3 have been used to determine if a phoneme is an /r/ or if it has rhotic qualities. In most recent studies, researchers have gone one step further by looking at what kind of /r/ (retroflexed or bunched) is associated with which formant values (Espy-Wilson, 1992; Espy-Wilson & Boyce, 1999; Hashi, Honda &Westbury, 2003). Using MRI images and acoustic analysis Zhou, Espy-Wilson, Tiede & Boyce (2007) found that F2, F3 and fundamental frequency

values were similar for both /r/ variants. However, a retroflexed /r/ had a larger difference between F5-F4 (1400Hz) than a bunched /r/ (700 Hz).

In other studies, factors such as the difference between F3 and F2, F2 transition rate and duration, F0, F1, F2, F3 and F4 frequency have all been suggested as the most important factor in /r/ identification (Flipsen et al., 2000, 2001). The difference in /r/ F3-F2 reported by Flipsen et al. (2001) ranged from 303Hz to 700Hz (ages 9-15) in normally developing female and male children. Generally, formant frequencies lower as children develop. In fMRI studies, F3 was related to front cavity resonance, and variations in F1 and F2 were related to changes in mid- and back oral cavity geometries (Espy-Wilson, Boyce, Jackson, Narayanan, & Alwan, 2000)

In order to recognize any /r/-like sound as an accurate representative of /r/, listeners have to be able to distinguish among many different speech sounds, including being able to tell 'good' exemplars from less proficient ones. When we actively listen to speech, we are involved in the processes of perceptual differentiation between speech sounds (perceiving /p/ differently from /b/). We perceive speech sounds categorically, that is to say, we notice more easily the differences between categories of phonemes (/p/ versus /b/) than within categories (/t^h/ versus /t/, versus /t^j/, versus /t^l/, etc.).

Theories of Speech Perception

It is very difficult to find a reliable, constant example of a phoneme to serve as an "exemplary model" in everyday speech. There are several explanations for the lack of a clearly observable 'prototype.' Because of the continous nature of speech, all phonemes spoken in a word, sentence, etc. are subject to context-induced variations. The back vowel /u/ can become fronted between coronal consonants (Hillenbrand, Clark, & Nearey; 2001).

Voice onset times also may vary, depending on the phoneme position within the syllable (Lisker & Abramson; 1967). Other important considerations are changing speech conditions such as rate and stress levels. Finally, just as unique as fingerprints, every person's version of a given phoneme is distinct due to the uniqueness of their physical and psychological properties. Together, the above considerations are commonly decribed as the 'lack of invariance' problem in speech perception.

When the lack of invariance is considered, one wonders how people manage to agree on any single phoneme category. It has been argued that in the light of lack of invariance, there must exist another mechanism that helps listeners sort phonemes into appropriate categories. One proposed solution involves the mechanisms of perceptual constancy and normalization. In perceptual normalization, variances in frequency (between male and female voices), vocal tract size, and other such factors are all considered to be forms of 'noise,' which are filtered out by attention to formant ratios or their means, rather than some set values. By normalizing for vocal tract variances, as well as speech rate variances, listeners are able to arrive at an underlying category for any sound (Johnson, 2005; Strange, 1999; Syrdal & Gopal, 1986). The existence and precise shape of normalization is still debated. Several speech theories try to explain how speech is perceived. These theories warrant consideration before investigating /r/ perception.

The Fuzzy-Logic Model (FLM) of speech perception argues for an internal prototype representation of phonemes in listeners (Massaro, 1989) This means that listeners decide wether X sounds like /r/ based on the relative goodness of the match between all the information taken in about X and the listeners' own internal values for a prototypical /r/ (Hayward, 2000). The advantage of the FLM is that it allows for the inclusion of all

information surrounding the speech signal, including non-auditory information. Computer models of FLM have been used to demonstrate behavior corresponding to that of human listeners (Oden & Massaro, 1978).

Another proposal for speech perception was forwarded by Kenneth Stevens. Stevens based his theory on the interaction between phonological features and articulatory gestures (Stevens, 2002). According to his theory of Acoustic Landmarks and Distinctive Features (ALDF), listeners concentrate on acoustic landmarks (such as formant values F1, F2, F3, etc.) within the speech signal that carry information about the gestures that produced them. Acoustic properties of the landmarks are the basis upon which distinctive features are established. Bundles of distinctive features make a phoneme, in turn building words, sentences, etc. Because the gestures which the listeners decode are made by the speakers by using their vocal apparatus with all its limitations, the Lack of Invariance problem does not exist in this model.

There are many other theories of speech perception (Prototype, Network, Motor, Direct-Realist, Exemplar theory, TRACE, etc.), too many to all summarize here. However, all consider aspects of Categorical Perception (CP) in their logic (Nygaard & Pisoni, 1995). In some of these models CP is probabilistic, in others deterministic in nature. Further discussion of CP is presented in the following section.

Perception of North American English /r/

"The world is full of things that vary in their similarity and interconfusability" (Pevtzow & Harnad, 1997). One strategy with which listeners resolve such uncertainties (such as between an accurate and mispronounced /r/) is by using CP. When we are listening

to multiple examples of one phoneme (such as /r/), we are always engaging in within-category discrimination. Therefore, Categorical Perception (CP) is to be considered in explanations of the listeners' varying performance.

According to current definitions, CP occurs whenever perceived within-category differences are compressed and/or between-category differences are separated, relative to some baseline of comparison (Harnad, 1987). In the case of /r/, FLM theory (Massaro, 1989) would call that baseline of comparison every person's prototype of /r/ with which they compare all other perceived productions of /r/. Originally, CP was introduced as part of the Motor Theory of speech perception, with the explanation that the abruptly switching perception between /pa/ and /ba/ categories is attributable to the anatomy of speech production (Liberman et al., 1957). Stops (such as /p/ and /b/) are produced abruptly and produce a clear CP effect. Sounds that are produced in a continuous manner (such as vowels or /r/) show a weaker CP effect. Since the phoneme /r/ belongs in the group of glides, laterals and rhotics, which is between continuous sounds (such as vowels) and plosive sounds (stops), we would expect to see more graded CP effects, rather than sudden jumps between categories as is the case with stops. CP effects have since been shown in animals in addition to humans, and are today understood to be not only relevant for speech or color perception, but to be far more general (Harnad, 1987). But, are CP effects equal across groups of listeners?

Participants in more recent research have demonstrated different boundaries for CP in different language groups. Flege and Efting (1988) compared Spanish and English voice-onset time perception and imitation and found evidence for different phonetic category formation, suggesting that stimuli were in fact categorized. These and similar studies have

been undertaken testing the belief that CP is innate (Eimas, Siqueland, Jusczyk & Vigorito, 1971). However, can different groups of people speaking the same language have different /r/ categories, depending on their education, training or some other non-language related factor? So far, CP was induced by learning alone on non-language related tasks (Goldstone, Lippa & Shiffrin, 2001; Lane, 1965; Lawrence, 1950). The Goldstone, Lane and Lawrence studies show that the stimuli to which participants make the same response (e.g. judging them as better) tend to become more similar (or as in CP, closer in category). Similarly, the stimuli to which a repeated different response is made, become more distinctive.

The above findings suggest that listeners who hear many accurate or inaccurate /r/s could have a better defined category or prototype of acceptable /r/ on which to base comparative decisions. This means that they should be better able to categorize /r/ stimuli that to untrained ears might sound ambiguous (on the accuracy dimension) and sort them into appropriate groups (accurate/inaccurate). Past research indicates differences in perceptual ability in listener judges. That suggests that group differences in /r/ specimens/category boundaries do exist. These studies are reviewed below.

Perceptual Tasks in Past Research

Previous perceptual research has utilized both synthesized and natural stimuli in a variety of /r/ listening and evaluation studies. Shelton, Johnson and Arndt (1974) established that some of the untrained adult listeners in their study (listening repeatedly to variants of /r/ produced naturally) were more reliable judges than others. Subsequently, Sharf, Ohde and Lehman (1988) used synthesized acoustic /r/ tokens of child-like speech, which either varied in second and third formant onset frequencies along the /r-w/

continuum, or were distorted. They found that not all adult listeners (eight SLP students) were able to make reliable judgments about the presented items, (i.e., finding similar results to Shelton et al., 1974). An additional point of interest for Sharf et al. (1988) is that in an older study, Sharf and Benson (1982) found that adult listeners did tend to stabilize in their ratings across sessions when they were given feedback to let them know when they have rated correctly, based on the experimenters' criteria. This finding suggests that the degree of training which listeners receive might play a role in the accuracy of their judgments.

Subsequent studies involving SLPs showed that these expert listeners were more strict in their rating of appropriate /r/ production than untrained listeners (parents, as in Chaney, 1988). A professional bias could have prevented the SLPs from rating positively weaker attempts at /r/. In contrast, the parent group may have thought that every attempt at /r/ should count, and rated each attempt that way. Using natural stimuli, Chaney had parents and SLPs evaluate different types of children's productions of /r/. The author found that parents of children who misarticulated /r/ were more likely to judge ambiguous /r/ productions as /r/. In contrast, the SLPs were more likely to judge those same productions as /w/. This might be explained by the fact that parents inherently possess a more lenient attitude towards their child, while SLPs may be more result-oriented, and therefore more critical. In a more recent study (Wolfe, Martin, Borton & Youngblood, 2003), SLP students without clinical experience demonstrated reduced sensitivity to the acoustic cues for /r/. These students showed weaker phonetic percepts for /r/ and /w/ than did the students with practicum experience. The authors suggested that a task based on intra- and inter- phonemic differences, such as cue trading, could be useful in assessing perceptual sensitivity of misarticulated /r/. Cue trading is a perceptual paradigm in which a change in the setting or

value of one phonetic cue, which leads to a change in the phonetic perception, can be offset by an opposed setting of a change in another phonetic cue so as to maintain the original phonetic perception (Repp, 1982; Moore, 1997).

The literature thus shows variability in judgment of accuracy of phoneme productions across listeners, regardless of their specialty (e.g., Sharf et al., 1988; Shelton et al., 1974), but also points to experience/learning, as having an effect in how good of a judge the rater can become (Wolfe et al., 2003).

The observed inconsistencies in perceptual judgments in adults show that more reliable techniques for qualitative judgment of phoneme productions are needed. Wolfe et al. (2003) suggest that SLPs be assessed and/or trained for this kind of work. While this may be useful, the SLPs are not the professionals who spend the most time with children.

The current study set out to compare judgments of SLPs with those of other adult professionals (teachers and child care workers) who spend more time with individual children. Further discussion of listener groups is presented below.

Speech-Language Pathologists and Educators as Listeners

In their undergraduate and graduate training, SLPs have education in phonetics, phonology, acoustics, oral-motor mechanics, speech and language acquisition and other relevant areas. In addition, many SLP students in Canada start graduate school with a specialization, or a concentration in linguistics. Lastly, before becoming a practicing SLP, students in many countries are required to take a formal federal certification exam. A graduate of an SLP program is assumed by the professional regulating bodies and

universities to have the clinical competence to make accurate field judgments on speech production. It is further assumed in the clinical world that experience and continuing education may improve that ability.

In the treatment process, when a person has a problem with /r/ production, the SLP will typically attempt a training method s/he is most comfortable with to teach the child how to produce the target phone. Once the child has undergone therapy, the length of which is usually determined by the SLP, the child typically improves at producing the target phone. Sometimes, the /r/ proves to be so difficult for an individual to acquire, that pronunciation difficulties persist into adulthood (Janzen & Shriberg, 1977; Ruscello, 1995a; Shuster et al. 1992; Shriberg et al., 2001).

One issue of interest for many SLPs is just how to determine when the target phoneme (in this case /r/) has been reached. Because the SLP decides when the therapy is over, he or she is also the one who makes the ultimate decision on accuracy of production. Although acoustic analysis programs are now more readily available free from the internet, SLPs seldom use acoustic analysis to determine whether the phone produced by the child after therapy has reached documented norms (in formant values or otherwise). This is often due to time restrictions, availability of appropriate technology (i.e. a computer), and lack of specific training. The SLP instead relies on his or her own perceptual ability, internalized 'prototypes' and possibly other factors (based on personal and professional experience and training) to make the determination that therapy should cease, either because of perceived success or lack thereof over time. In other words, SLPs are by definition and by job description expert listeners.

During school years, however, most children spend more time with Educators than with SLPs. Teachers often make the initial referral for a child to an SLP and are the ones who ultimately judge the child's performance in school. They help determine the degree of the childrens' academic success, with serious implications for their future (Bennett & Runyan, 1982; Fujiki & Brinton, 1984). Child care workers play a big part in children's lives, similar to teachers. They observe children in informal settings during before- or after-school care. In these situations children often produce most of their creative, self-initiated language because they are outside of the structured class environment. Their job provides teachers and child care workers with a great variety of children's speech, upon which they may develop standards on what is acceptable and what is not in child language. Both types of Educators therefore have very important roles in child language development (aside from the instructional aspects), and can make very important observations about the children in their care.

Past research has documented that SLPs could be better listeners for speech than untrained listeners (Chaney, 1988). Teachers and child care workers are included in this study because they also could arguably be called 'expert' listeners. They do not receive the same training and education in phonetics or language development as the SLPs, but they do listen to a greater number of children on a daily basis. To relate back to theoretical discussion of CP (Goldstone, 1994; Lane, 1965; Lawrence, 1950), the Educators have even more exemplars of any given phoneme available to them (because they listen to more children repeatedly, whereas the SLP listens to the children with impairment in a similar, if not greater ratio than to the normally developing children). The time spent perceiving repeated items could imply that the continuum between 'accurate' and 'inaccurate' versions

of /r/ might be less differentiated in Educators than in SLPs, assuming that there exists a 'sum effect' in prototype building. If more accurate productions of /r/ are heard than inaccurate ones, the prototype representation of /r/ should be closer to the accurate version. Thus, for the current study, the question was whether the perceptual expertise of the Educators would match those of the SLP group. In addition to listener type, the current study is also concerned with the type of task as discussed below.

Listening Task Type

Typically, listener studies are designed as rating tasks on some kind of rank-order scale. A listener hears a number of different samples of words, speech sounds or other material and makes a decision on the 'goodness' of the sample heard. The rating is usually scalar, and either numerical (1-10) or qualitative (poor, good, best) or based on some other rating scheme. By gathering many ratings, researchers are able to determine the preferred and less preferred items. With such scales it is difficult to draw any conclusions about why some tokens are preferred over others, especially the tokens which are not clearly 'best' or 'worst.' An additional complication of using scales is that there are biases inherent to any scale. Judges shy away from rating on one or the other extreme, and show a tendency to rate towards the middle (Andrich, 1978; Edwards & Cronbach, 1952). Both of these trends can severely limit the amount of information gained from a listener judgment task. For the current study, another type of task was used in addition to a scale: a pair-wise comparison task (Bradley & Terry, 1952; Luce, 1959; Thurstone, 1927, 1929), in which listeners attended to competing tokens and decided which they preferred.

The aim of including two types of tasks was to compare listener reliability. The identification task has been useful in previous studies (Adler-Bock, Bernhardt, Gick & Bacsfalvi, 2007; Bernhardt, Bacsfalvi, Gick, Radanov, & Williams, R., 2005; Bernhardt, Gick, Bacsfalvi & Ashdown, 2003;). If the current data follow the same trends in both tasks (identification and comparison), then it would be logical to assume that the comparison task may also be useful. The tasks are discussed further in the Methods section below.

Summary of Research Questions and Predictions

The literature on speech perception, particularly as it concerns /r/ as pronounced by children with speech production impairments, led to the following questions and predictions for the current study:

(1) Do SLPs and Educators differ in the way they rate the same /r/ attempts?

The FLM (Massaro, 1989) put forward the establishment of a prototype which serves as a litmus test for whether some sound passes as phoneme X or not. If that is true, then both groups of listeners should show a preference toward the same stimuli which is closest to that prototype. If the establishment of the prototype functions in a sum fashion (i.e. adding all the phonemes X ever heard together) then the Educators should have an /r/ prototype closer to the norm, since they listen to more normal /r/s than the SLPs. This logic would predict that SLPs should have a more diffuse version of an /r/ prototype; however, their training and knowledge could counteract such tendencies. The Acoustic Landmarks and Distinctive Features theory (Stevens, 2002) conceives listeners inspecting the incoming speech signal for acoustic landmarks, and on that basis establishing distinctive features,

which in turn specify phonetic segments. If this is indeed the case, then the way in which the SLPs and Educators rank tokens should neatly match up with a ranking based on acoustic measurements of formants in all stimuli based on previously investigated and established norms (Flipsen et al., 2001; Shriberg et al., 2001).

The reviewed literature (Chaney, 1988; Goldstone, 1994; Lane, 1965; Lawrence, 1950; Sharf & Benson, 1982; Wolfe et al., 2003) suggests that there should be an expert effect in favor of the SLPs, since the 'accurate' and 'inaccurate' categories should be more clearly defined in SLPs, due to their training and the nature of their profession. A related sub-question was:

(1a) Does some other factor, such as age, first language, being a parent, etc., affect perceptual judgements?

Every speech perception theory considers some aspect of experience in playing a role in perceptive ability. Usually the two are positively related, meaning that the more experience one has in listening, the better perceptive skills one should have, provided learning continues in a natural manner. Still, with experience and learning time becomes a factor. In a long period of time there is a higher likelihood of establishing personal factors, inherent to the individual that may also play a role in speech perception (Massaro, 1989). Previously published research by Chaney (1988) found that SLPs differed as judges from another group of parents, even though both had access to the same group of children, the parents more so than the SLPs. As far as any additional demographic factors, no previous research has provided sufficient evidence as a basis for prediction.

(2) Do rank-order and paired comparison tasks result in different levels of reliability?

Identification tasks and rank order scales have shown success in past research (Adler-Bock et al. 2007; Bernhardt et al., 2003, 2005). Paired comparison has not yet been used in speech research in /r/ production. The question of interest was whether a paired comparison task would prove to be of equal, lesser or better value to future researchers than the identification task. In other disciplines, differences among procedures in survey methods have produced notable differences in results (Ebel, 1951; Thurstone, 1959).

CHAPTER 2

METHOD

Participants

The listeners for this study were recruited through the BC Association of Speech Language Pathologists newsletter, the Vancouver YMCA Child Care division and by word of mouth (snowballing effect). The numbers of participants in each group, although not balanced for gender, age or language are representative for the Canadian populations of SLPs and Educators, where males are rare, years of work experience vary and backgrounds are fairly diverse. All Educators work in elementary schools; 11 SLPs work in elementary schools and 9 work in health units, where they work primarily with children.

Table 1. Information About Listeners

Participant Variables	**Group 1 – SLPs N=20**	**Group 2 – Educators N=18**
Number of participants	19 females, 1 male	14 females, 4 males
Age range	20-65 years	20-60 years
Mean age	45 years	38 years
English as first language	20	16; 2 are bilingual
Speak another language	5	9
Are parents	10	9
Range of years of work experience with children under age 12 years	1.5-35 years	1-34 years
Average years of experience with children under age 12 years	16 years	14 years

Stimuli

The speech tokens used in this study consisted of 16 /r/ syllables from previously collected data. 16 syllables were used because the stimuli needed to be diverse, yet the task needed to be short. The phoneme /r/ was selected in word-initial and word-final position in order to minimize possible lexical effects. The syllable from the word rabbit was cut at the beginning of the vocal signal and at the middle of the vowel formants for /æ/, leaving a [ræ] syllable. The word star was cut at the beginning of the vowel formants' steady state to the end of the vocal signal, leaving a [ɑr] syllable. The original recordings were of field quality with variable amplitude of speech signal and noise. Attempts at altering the stimuli by reducing noise or boosting signal quality resulted in poorer overall quality samples, increasing idiosyncracy of the stimuli. Therefore stimuli were left unaltered.

The syllables were extracted from speech samples of four children recorded during research for Bernhardt et al. (2007). The study was undertaken to evaluate the effects of consultative treatment with ultrasound in rural communities. SLPs trained in ultrasound use provided ultrasound consultation in rural communities to school-aged children (including the ones who provided the stimuli for the present study), all of whom had residual speech impairments. Words from both pre- and post-treatment conditions were selected in order to create an acoustically diverse sample of /r/s. The children and speech samples that were selected for the current study were the most homogenous in terms of ages and treatment targets.

Table 2. Description of Children Who Provided the Stimuli

Child	Gender	Age	Oral Mechanism Evaluation:	Treatment targets
NA1	M	11[a]	Unremarkable	/r/
NA2	M	8;1	Braces, Jaw-assisted tongue movement?	/r/
NA4	F	12[a]	High arched palate, Lip-assisted tongue movement?	/r/
NA5	M	7[a]	Mouth breather, Dentalized alveolars, Slight lateralization on up and down movement of tongue	/s, z, r/

It is important to note that the children who provided the stimuli for this study were selected because they presented many different versions of /r/. It was not the aim of the current study to investigate treatment effects.

Study Design

The study was composed of two tasks: an identification task and a paired comparison task. Both tasks were constructed in Microsoft PowerPoint 2003 as an interactive presentation with some SimpleBasic controls added. The two tasks were combined into one presentation, and given to listeners in four versions, varying the identification and comparison order, and the order of tokens. Before each listener started the listening task, he or she was briefed on examples of accurate and poor productions of /r/ in both syllable initial /ræ/ and syllable final /ɑr/ contexts. In total, the presentation consisted of 164 slides.

Identification Task

In the identification task the judges listened to a randomized set of slides presenting all 16 tokens, with 8 tokens (first and last four of the original 16) repeated in order to evaluate later intra-rater reliability. Interspersed throughout the order of the slides with sounds clips were slides with amusing cartoon animations which had no sound. The purpose of these slides was to give the listeners a break, so that they could maintain attention to task more easily. Every slide would present the listener with the same token three times, upon which the listener was to answer the question "Is this an /r/?" Listeners would then check their selection and the slideshow would move to the next slide. Slide transition was timed at 7 seconds, or could be initiated by the listener by clicking the > right arrow key.

Presented below is an example of an identification task slide:

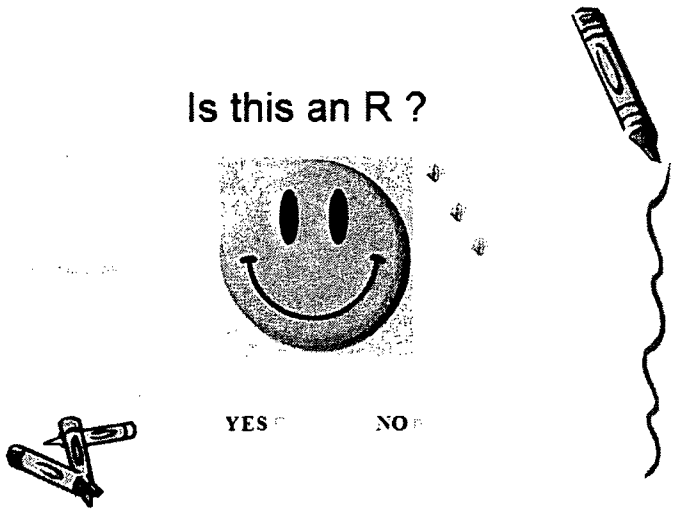


Figure 1: Identification Task Screenshot

Comparison Task

In the paired comparison task the judges listened to a permutatively arranged set of slides. There were two utterances present on each slide. Playback would move from Smiley1 to Smiley2, for a total play of three times for each token. Upon hearing the audio cues, the listeners were to decide which of the tokens presented to them was a more accurate presentation of /r/. All listeners were informed beforehand that there could be no ties between tokens, although some might sound very much alike. A random 10% of the slides were repeated in task to ensure later intra-rater reliability measures. Interspersed throughout the order of the slides with sounds clips were slides with amusing cartoon animations which had no sound. The purpose of these slides was to give the listeners and their ears a break, enabling them to maintain attention to task more easily. Slide transition was timed at 7 seconds, or could be initiated by the listener by clicking the right arrow key. Presented below in Figure 2 is an example of a comparison task slide:

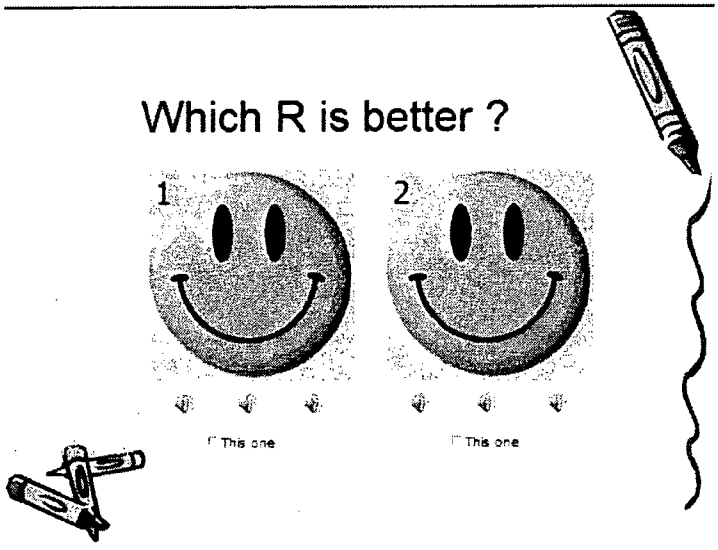


Figure 2: Paired Comparison Task Screenshot

Data Collection

The data were presented through Microsoft PowerPoint (2003), either on a laptop or a desktop computer capable of running Windows XP with Microsoft Office. The participants were equipped with high quality headphones (JVC HA-G55 or similar), and listened to the presentations at a level that was comfortable to them. All data collection took place between 8am and 4pm, the usual working time for both SLPs and Educators. The collections took place in the quietest location available at the collecting site, usually an office or similar space. All data were recorded into Excel (2003) spreadsheets for later processing. In addition to the responses to the auditory stimuli, a basic demographic questionnaire was filled out by each participant (see APPENDIX 5).

Data Analysis

All data were entered into Excel (2003) spreadsheets. Basic percentages of intra-rater reliabilities were calculated in Excel. More complicated statistical procedures (such as Pearson's r) were calculated using the SPSS graduate student service software (SPSS 12.0, Graduate Student Version, Chicago III). The most complex level of analysis was the calculation of Multivariate Analysis of Variance (MANOVA), which was carried out in the R statistical software (R 2.3.1, 2006). In order to measure reliability between groups, an Intraclass Coefficient (ICC) was calculated for single and multiple measures.

Acoustic Analysis

Acoustic analysis of stimuli was performed using Praat version 4.4.13 (http://www.fon.hum.uva.nl/praat/, Boersma & Weenink 2006). Stimuli acoustics were determined by two trained judges using spectrograms obtained from Praat. Initially, formant height was determined by measuring an instantaneous slice at the point of greatest constriction between F2 and F3. This approach resulted in considerable differences between raters. In order to improve inter-rater reliability, an average formant value for the whole /r/ was obtained (mean F value in Praat query), resulting in greater inter-rater agreement (average difference of 22Hz).

Table 3. Acoustic Information About Stimuli

Stimuli	F1 (Hz)	F2	F3	F3-F2	/r/ duration	Vowel duration	Total duration
1NA2ar	524	1294	2672	1378	0.2 secs.	0.15 secs.	0.35 secs.
1NA2ra	474	1275	2745	1470	0.32 secs.	0.09 secs.	0.41 secs.
1NA4ar	665	1583	2245	662	0.11 secs.	0.2 secs.	0.31 secs.
1NA4ra	545	1932	2446	514	0.2 secs.	0.12 secs.	0.32 secs.
1NA5ar	782	1817	2793	976	0.14 secs.	0.21 secs.	0.35 secs.
1NA5ra	450	1296	3007	1711	0.38 secs.	0.11 secs.	0.49 secs.
1NA1ar	437	1328	2530	1202	0.37 secs.	0.13 secs.	0.5 secs.
1NA1ra	430	1249	2685	1436	0.2 secs.	0.09 secs.	0.29 secs.
2NA2ar	639	1190	2196	1006	0.15 secs.	0.19 secs.	0.34 secs.
2NA2ra	584	1620	2824	1204	0.12 secs.	0.08 secs.	0.2 secs.
2NA4ar	669	1710	2243	534	0.3 secs.	0.36 secs.	0.66 secs.
2NA4ra	577	2012	2463	451	0.32 secs.	0.22 secs.	0.54 secs.
2NA5ar	807	1952	3310	1359	0.26 secs.	0.22 secs.	0.48 secs.
2NA5ra	557	1872	3121	1249	0.1 secs.	0.1 secs.	0.2 secs.
2NA1ar	598	1530	2639	1110	0.26 secs.	0.18 secs.	0.44 secs.
2NA1ra	488	1519	2608	1090	0.14 secs.	0.11 secs.	0.25 secs.

Note. The stimulus name format can be interpreted as follows:
NA: from a large northern town in British Columbia
NA1xx, NA2xx, etc. = speaker number and token, i.e. ra or /ræ/ and ar or /ɑr/
1N, etc.: 1 = pre-treatment, 2 = post-treatment

CHAPTER 3

RESULTS

This study aimed to measure the extent to which two or more listeners agree when rating the same stimulus. Before that question can be answered, it was necessary to determine whether the judges were performing the task assigned to them by checking their responses in a test-retest measure.

Intra-Rater Reliability

One of the first questions to answer is whether the individual judges included in this study were rating the task reliably. To evaluate this query, intra-rater reliability was examined. Table 4 on the following page shows the intra-rater reliability based on what percentage of the time the same decision was made in the test-retest aspect of the identification task:

Table 4. Intra-Rater Reliability Percentages on Both Tasks

Identification Task				Paired Comparison Task			
SLP1	1	Teacher 1	1	SLP1	0.9	Teacher 1	1
SLP2	0.88	Teacher 2	1	SLP2	0.9	Teacher 2	1
SLP3	0.88	Teacher 3	0.88	SLP3	0.9	Teacher 3	0.95
SLP4	0.88	Teacher 4	0.88	SLP4	0.85	Teacher 4	0.9
SLP5	0.88	Teacher 5	0.88	SLP5	0.9	Teacher 5	0.89
SLP6	0.88	Teacher 6	0.86	SLP6	0.9	Teacher 6	0.89
SLP7	0.88	Teacher 7	0.86	SLP7	0.95	Teacher 7	0.88
SLP8	0.88	Teacher 8	0.86	SLP8	0.95	Teacher 8	0.85
SLP9	0.88	Teacher 9	0.83	SLP9	0.95	Teacher 9	0.85
SLP10	0.88	Teacher 10	0.75	SLP10	0.9	Teacher 10	0.84
SLP11	0.88	Teacher 11	0.75	SLP11	0.95	Teacher 11	0.84
SLP12	0.88	Teacher 12	0.75	SLP12	0.95	Teacher 12	0.83
SLP13	0.75	Teacher 13	0.75	SLP13	0.9	Teacher 13	0.83
SLP14	0.75	Teacher 14	0.75	SLP14	0.9	Teacher 14	0.75
SLP15	0.75	Teacher 15	0.63	SLP15	0.8	Teacher 15	0.75
SLP16	0.75	Teacher 16	0.63	SLP16	0.95	Teacher 16	0.74
SLP17	0.75	Teacher 17	0.57	SLP17	0.9	Teacher 17	0.74
SLP18	0.75	Teacher 18	0.43	SLP18	0.9	Teacher 18	0.6
SLP19	0.63			SLP19	0.85		
SLP20	0.38			SLP20	0.95		
Average	0.81	Average	0.78	Average	0.91	Average	0.84

SLPs on average had higher intra-rater reliability than Educators on both tasks, although there were individuals in both groups with low intra-rater reliability and high intra-rater reliability. Both groups showed improved reliability on the comparison task. The only statistically significant finding discovered using a t-test was that the SLP reliability score significantly improved on the comparison task ($p=0.002$)

Considering that the listeners were overall consistent and similar in that regard (making the same decision 84% of the time across groups and tasks), the analysis thus proceeded.

Inter-Rater Reliability

If the SLPs/educators agreed with each other, then the correlation between the ratings given by one SLP/educator and those given by the other SLP/educator should have been high. In order to assess inter-rater agreement of the two listeners ratings the correlations of the two SLPs/educators ratings were measured. However, it is important to remember that two SLPs/educators ratings could be highly correlated but have little or no agreement. The remedy to this problem is to calculate the Intraclass Correlation Coefficient (ICC).

Intraclass Correlation Coefficient

The ICC was calculated using an Intraclass Coefficient calculator (http://www.med-ed-online.org/rating/reliability.html; accessed May 28th 2007). ICC is used when data are collected using only one listener at a time, but there are two or (preferably) more listeners on a subset of the data for purposes of estimating inter-rater reliability. In SPSS this statistic is called the single measure intraclass correlation (Figure 3 and Table 5.). If the reliability for all the listeners averaged together is required, the Spearman-Brown correction needs to be applied. SPSS calls this statistic the average measure intraclass correlation, but it is also called the inter-rater reliability coefficient by other researchers (MacLennon, 1993). The following figure and table show ICC calculated for all of the listeners and data used in this study. Predictably, as the number of listeners increases, so does their agreement with eachother. The same results are presented in Table 5 on the next page.

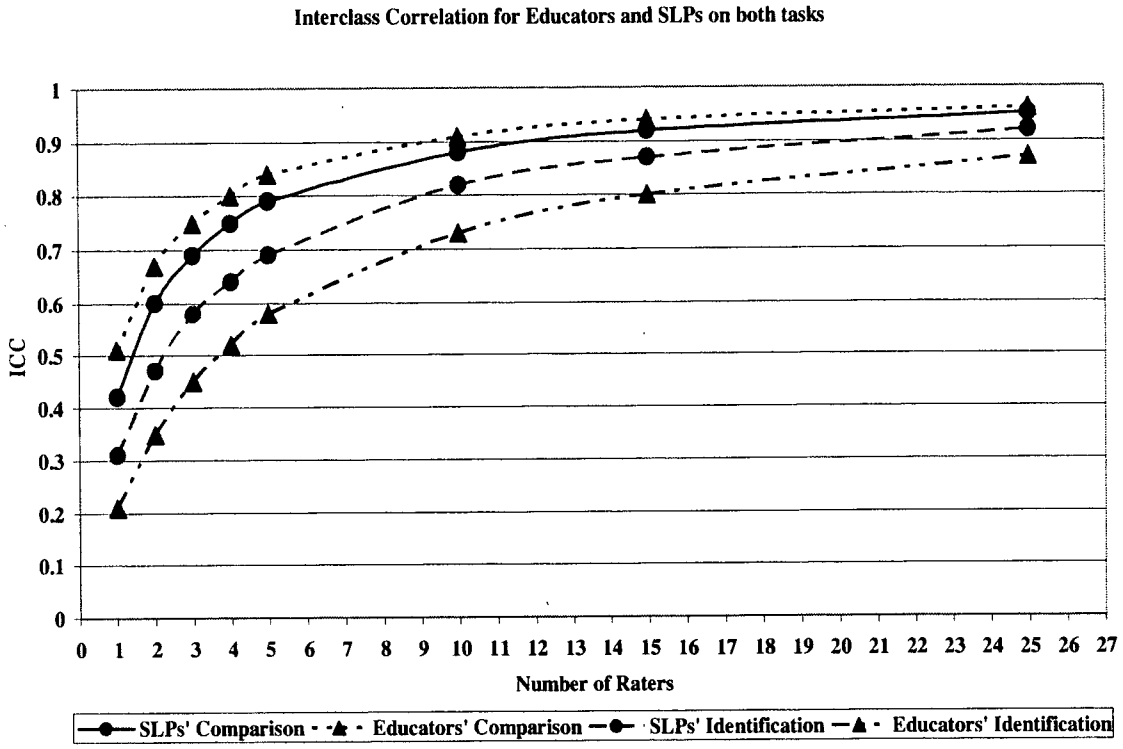


Figure 3. Interclass Correlation Calculations Represented Graphically

Table 5. ICC for Both Tasks by Number of Listeners

Task type	Numbers of listeners	1	2	3	4	5	10	15	25
Comp.	SLPs	0.42	0.6	0.69	0.75	0.79	0.88	0.92	0.95
Comp.	Educators	0.51	0.67	0.75	0.8	0.84	0.91	0.94	0.96
ID	SLPs	0.31	0.47	0.58	0.64	0.69	0.82	0.87	0.92
ID	Educators	0.21	0.35	0.45	0.52	0.58	0.73	0.8	0.87

Note. Comp= comparison; ID = identification.

ICC Interpretation

SLPs had a better score on the identification task than the educators (ICC 0.31>0.21). This means that if an SLP rated the quality of some /r/, his or her judgments would agree with another random SLP listener slightly better than the educators would agree with each other.

The educators have a better ICC than SLPs (.51>.42) on the comparison task. This implies that if only one teacher and one SLP rated the quality of some /r/ when being asked "Which one of these two is better?", the teacher would agree with other random teacher judges slightly better than the SLP with other random SLPs.

An important consideration to keep in mind is that the ICC Coefficient shows the uniformity, but not the correctness of judgments. Implications of these data will be discussed in sections ahead. The important observation here is that a group of 20 listeners can rate very reliably (high .90), after which their performance improves only marginally.

Comparison of Stimuli Rankings Using Paired t- tests

As the next step in analysis, Educator ratings were compared to SLP ratings by token across tasks through t-tests with two independent samples, assuming different population variances. The purpose of this analysis was to find out if the two groups rated some stimuli significantly different from each other. See Figures 4 and 5 on the next page.

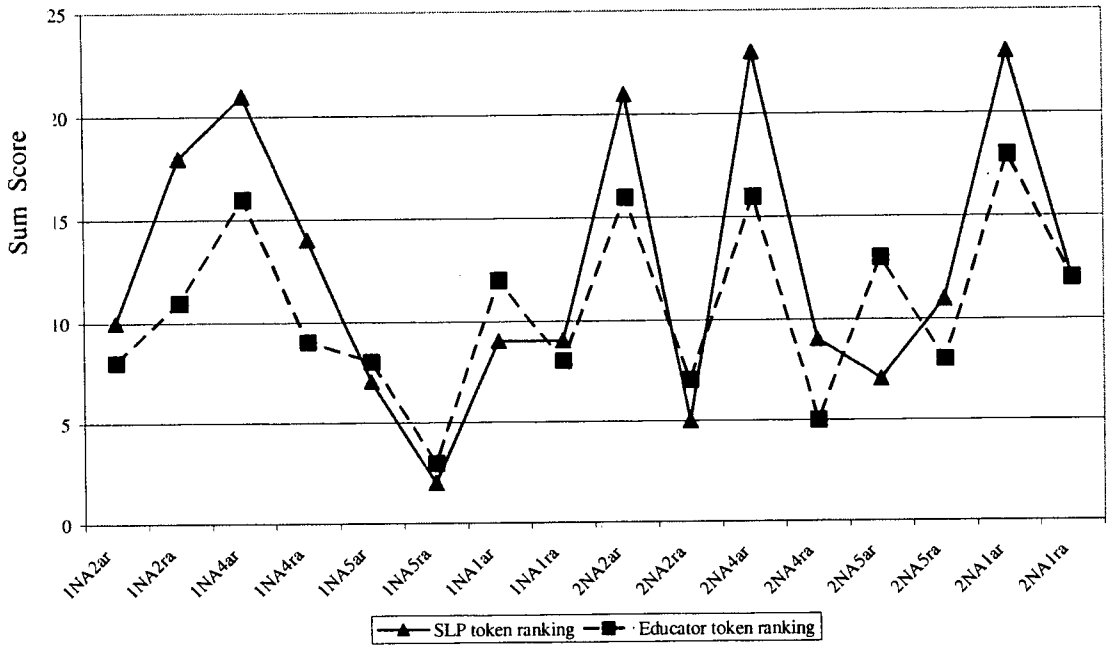


Figure 4. Sum of Scores Graph, Identification Task, Grouped by Stimuli

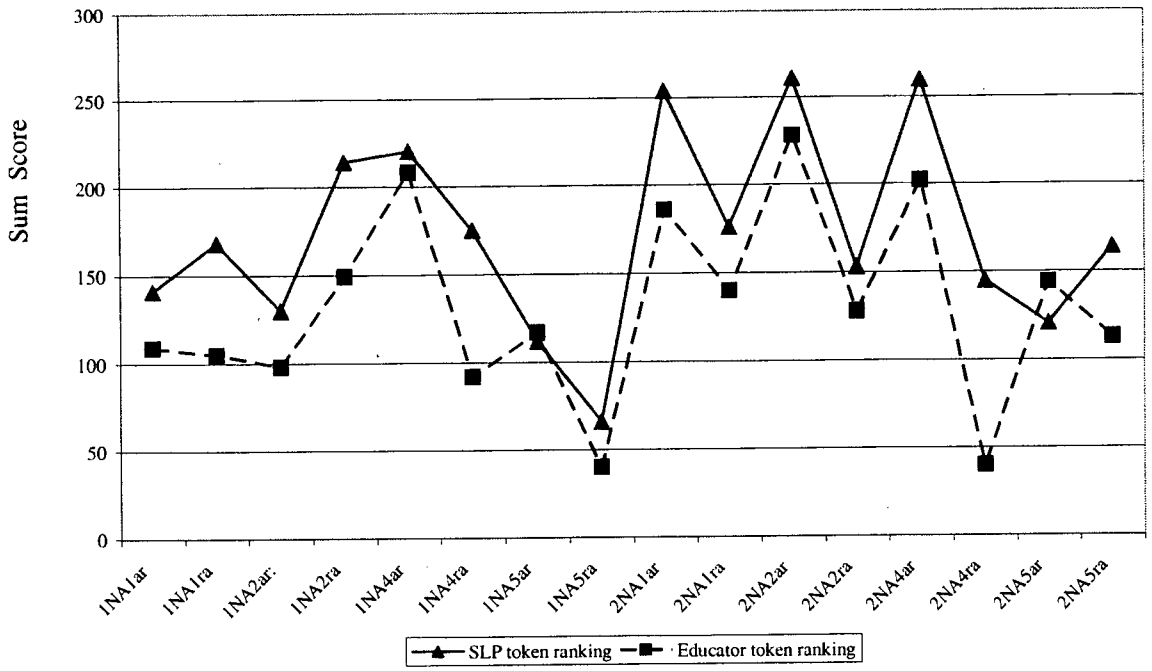


Figure 5. Sum of Scores Graph, Comparison Task, Grouped by Stimuli

Only two stimuli stood out as being rated significantly different between teacher and SLPs as shown in Table 6.

Table 6. Differently Rated Stimuli as Identified By t-tests

Token #	Stimuli	p-value	95% CI
4	1NA4ra	0.002121	(-4.819391, -1.158386)
12	2NA4ra	0.00000977	(-5.755546, -2.500010)

The p-value representing the probability of the difference between the two groups needs to be less than 0.003125 to be significant. Since 16 simultaneous t-tests were administered, a Bonferroni's correction was applied. If no correction is applied there would have been a 55.99% chance of finding one or more significant differences in 16 tests. To obtain an alpha level overall of 0.05, alpha had to be lowered for each test to 0.003125 (obtained by dividing alpha of 0.05 by 16 tests).

Relating the findings above to the results of acoustic analysis, it can be seen that the two tokens were rated differently from others across the two groups of listeners. One interesting observation is that these two tokens had the lowest F3-F2 value out of the total set of 16 tokens. (See Table 3.)

Multivariate Analysis of Variance

The Multivariate Analysis of Variance (MANOVA) was calculated for all listeners (SLPs + educators) in order to determine whether some other variable that was not yet

considered played a role in the listener ratings. The MANOVA matrix considered six additional variables: age of listener, being a native speaker of English, speaking other languages, being a parent, years of experience with children under 12 years of age and the order of the test given (to eliminate the possibility that one of the four versions used in the study had training effects).

Table 7. MANOVA Results for SLPs and Educators, Comparison Task

Stimuli and Factors (N=26)	**Sum Sq**	**Mean Sq**	**F value**	**Pr (>F)**
1NA4PRra				
Age	90.785	22.696	3.1790	0.02978*
Parent	57.250	57.250	8.0187	0.00882**
1NA5PRar				
Order of Test	67.421	22.474	3.3474	0.03436*
2NA2PRar				
Age	124.020	31.005	5.0304	0.003874**
2NA5PRra				
Born in Canada	51.523	51.523	5.4910	0.02704*

There were no significant systematic results found for any of the considered variables. One of each of the considered additional factors shows up for four stimuli, which are all different.

Ranking Stimuli in Order of Preference

Finally, the ranking of the speaker stimuli based on the total score they achieved is presented in Table 7, i.e., summing all the times they were identified as an /r/ on the identification task, or all the times they won in competition with another token on the paired comparison task. This demonstrates that the tokens were sufficiently different in terms of the accuracy continuum.

Table 8. Stimuli Ranking, Sum Scores for SLPs and Educators: Identification Task

SLP ranking – Identification			**Teacher ranking– Identification**		
1	2NA4ar	23	1	2NA1ar	18
2	2NA1ar	23	2	1NA4ar	16
3	1NA4ar	21	3	2NA2ar	16
4	2NA2ar	21	4	2NA4ar	16
5	1NA2ra	18	5	2NA5ar	13
6	1NA4ra	14	6	2NA1ra	12
7	2NA1ra	12	7	1NA1ar	12
8	2NA5ra	11	8	1NA2ra	11
9	1NA2ar	10	9	1NA4ra	9
10	1NA1ar	9	10	2NA5ra	8
11	1NA1ra	9	11	1NA5ar	8
12	2NA4ra	9	12	1NA1ra	8
13	1NA5ar	7	13	1NA2ar	8
14	2NA5ar	7	14	2NA2ra	7
15	2NA2ra	5	15	2NA4ra	5
16	1NA5ra	2	16	1NA5ra	3

Table 9. Stimuli Ranking, Sum Scores for SLPs and Educators: Comparison Task

Comparison task – SLPs			Comparison task – Educators		
1	2NA2ar	261	1	2NA2ar	228
2	2NA4ar	260	2	1NA4ar	208
3	2NA1ar	254	3	2NA4ar	202
4	1NA4ar	220	4	2NA1ar	186
5	1NA2ra	214	5	1NA2ra	149
6	2NA1ra	176	6	2NA5ar	144
7	1NA4ra	175	7	2NA1ra	140
8	1NA1ra	168	8	2NA2ra	128
9	2NA5ra	164	9	1NA5ar	117
10	2NA2ra	153	10	2NA5ra	113
11	2NA4ra	145	11	1NA1ar	109
12	1NA1ar	141	12	1NA1ra	105
13	1NA2ar	130	13	1NA2ar	98
14	2NA5ar	121	14	1NA4ra	92
15	1NA5ar	112	15	1NA5ra	40
16	1NA5ra	66	16	2NA4ra	40

From the tables above we can observe that both the SLPs and the Educators agreed about which tokens should be ranked best, and which should be ranked worst, except for a few. Two disagreements that both stand out have to do with the same tokens identified earlier by the t-tests: 1NA4ra and 2NA4ra.

There is one more important comparison, that of the groups rankings considering acoustics, especially in the light of past research information. Flipsen et al. (2000, 2001) describe an average F3 of 1934Hz for the adolescents in their study and an average F2 of 1337 Hz (a slightly higher formant value would be expected in this study because the children are younger). Flipsen et al. also report a F3-F2 acceptable range of 303Hz to 700Hz (for children aged 9-15). The two stimuli in disagreement are highlighted in Table 10 on the next page.

Table 10. Stimuli Ranking, Including Acoustics for SLPs and Educators: Comparison Task

SLP Rank		F1	F2	F3	F3-F2	/r/ Length	Vowel Length	Total Length
1	2NA2ar	626	1199	2190	991	0.15	0.19	0.34
2	2NA4ar	667	1718	2234	516	0.3	0.36	0.66
3	2NA1ar	603	1581	2708	1127	0.26	0.18	0.44
4	1NA4ar	664	1597	2239	642	0.11	0.2	0.31
5	1NA2ra	473	1258	2707	1449	0.32	0.09	0.41
6	2NA1ra	485	1518	2608	1090	0.14	0.11	0.25
7	**1NA4ra**	**537**	**1934**	**2434**	**500**	**0.2**	**0.12**	**0.32**
8	1NA1ra	429	1313	2553	1240	0.37	0.13	0.5
9	2NA5ra	542	1876	3144	1268	0.1	0.1	0.2
10	2NA2ra	573	1606	2827	1221	0.12	0.08	0.2
11	**2NA4ra**	**583**	**2024**	**2460**	**436**	**0.32**	**0.22**	**0.54**
12	1NA1ar	429	1313	2553	1240	0.37	0.13	0.5
13	1NA2ar:	527	1293	2675	1382	0.2	0.15	0.35
14	2NA5ar	817	1945	3301	1356	0.26	0.22	0.48
15	1NA5ar	777	1814	2656	842	0.14	0.21	0.35
16	1NA5ra	449	1293	3003	1710	0.38	0.11	0.49

Educator Rank		F1	F2	F3	F3-F2	/r/ Length	Vowel Length	Total Length
1	2NA2ar	626	1199	2190	991	0.15	0.19	0.34
2	1NA4ar	664	1597	2239	642	0.11	0.2	0.31
3	2NA4ar	667	1718	2234	516	0.3	0.36	0.66
4	2NA1ar	603	1581	2708	1127	0.26	0.18	0.44
5	1NA2ra	473	1258	2707	1449	0.32	0.09	0.41
6	2NA5ar	817	1945	3301	1356	0.26	0.22	0.48
7	2NA1ra	485	1518	2608	1090	0.14	0.11	0.25
8	2NA2ra	573	1606	2827	1221	0.12	0.08	0.2
9	1NA5ar	777	1814	2656	842	0.14	0.21	0.35
10	2NA5ra	542	1876	3144	1268	0.1	0.1	0.2
11	1NA1ar	429	1313	2553	1240	0.37	0.13	0.5
12	1NA1ra	429	1313	2553	1240	0.37	0.13	0.5
13	1NA2ar	527	1293	2675	1382	0.2	0.15	0.35
14	**1NA4ra**	**537**	**1934**	**2434**	**500**	**0.2**	**0.12**	**0.32**
15	1NA5ra	449	1293	3003	1710	0.38	0.11	0.49
16	**2NA4ra**	**583**	**2024**	**2460**	**436**	**0.32**	**0.22**	**0.54**

CHAPTER 4

DISCUSSION

The objectives of the present study were to determine whether SLPs and Educators differed in the way they rate the same /r/ attempts, if there were any other significant factors that influence listeners, and finally, if the way that the task was constructed plays a role in listener ratings. Each of those questions is discussed below.

Listener Differences and Similarities

The first evidence of difference between the two groups seems to be shown in Table 2, with the Intra-Rater percentages. SLPs seem to be more consistent in a test-retest situation. This could be a by-product of experience accumulated administering standardized tests to a higher degree than teachers or child care staff, or a result of professional education factors. SLPs could be more comfortable making a decisions on what is and what is not an /r/. Educators and child care workers most likely have minimal experience rating a child's /r/ in terms of its quality.

According to the FLM (Massaro, 1989) and the establishment of a prototype idea, we speculated that both groups of listeners would have an established prototype in mind while performing the experimental task. If this was the case, all intra-rater reliability scores would in theory be 100% (listeners making same decisions all the time). Clearly this is not so. It is very interesting to see a wide range of variability in both groups. In each group

there are individuals who are very consistent judges and others who are not so reliable. These findings could arguably mean that some people have a varying prototype, or alternatively that they did not perform well on the task due to some other factor (concentration, tiredness, etc.)

Next we have the Intraclass Correlation Coefficient graph (Figure 3, Table 5), which at first glance seems to show more reliable performance among the Educators compared with the SLPs. It is important, however, to interpret the ICC correctly. The ICC is generally regarded as an improvement over Pearson's $\underline{r}$ and Spearman's $\underline{\rho}$. This is because the ICC takes into account the differences in ratings for individual segments, along with the correlation between raters. The range of ICC is between -1.0 and 1.0, meaning that ICC will be high (closer to 1.0) when there is little variation between the scores given to each item by the raters (i.e. if all raters give the same, or similar scores to each of the items rated). In the context of the presented data, this means that, although the teacher group was a little lower on Intra-Rater reliability, the scores distributed by the group across items were more similar to each other than those of the SLPs. On any kind of scale, ICC would presumably be very high when a large central tendency bias is present (raters avoid using extreme responses, hence all answers clustering towards the middle). From the sum of scores graphs (Figure 4 and 5) it can be seen that in both tasks the SLPs gave more extreme ratings to stimuli. If stimuli were rated high, then they were usually rated higher by the SLPs. If stimuli were rated low, then they were usually rated lower by the SLPs. In addition, the SLP group had a few outliers in the intra-listener reliability. These SLPs also contributed to less homogeneity in overall ratings, perhaps more so than the less reliable Educators.

Finally, both the ranking by score (Tables 7 and 8) and the t-test results show slight discrepancies in preference about what makes a good /r/ among SLPs and Educators. As predicted by FLM theory and the idea of a prototype, both groups in general agree on which of the tokens can be described as the 'best' /r/s (2NA2ar, 2NA4ar, 2NA1ar, 1NA4ar and 1NA2ra). The theory of categorical perception implies that between-group differences (in this case, between an accurate /r/ and an obviously inaccurate one) are easier to perceive than within-category differences, such as deciding which of two inaccurate /r/s is better, when they are both close to the boundary of the /r/ category. The tokens 1NA4ra and 2NA4ra are just two such examples. Their values for F2 are close to 2000 (1934, 2024Hz) which is to high for a good F2 for /r/. Yet, at the same time their F3-F2 values fall perfectly in the 300-700Hz range (500, 436Hz . This makes the two stimuli (1NA4ra & 2NA4ra) ambiguous with respect to the Flipsen et al. (2000, 2001) and Shriberg et al. (2001) data.

How tighly concentrated or scattered are the groups prototypes of /r/? We argued before that if the establishment of the prototype proceeds in simple sum fashion then the Educators should have an /r/ prototype closer to the norm, since they listen to more normal /r/s than the SLPs. This logic would predict that SLPs should have a more scattered version of an /r/ prototype. We also discussed that their training and knowledge could counteract such tendencies. In light of the Acoustic Landmarks and Distinctive Features theory (Stevens, 2002), if listeners inspect the incoming speech signal for acoustic landmarks, and on that basis establish distinctive features (that in turn specify phonetic segments), then the way in which the SLPs and Educators rank tokens should neatly match up with a ranking based on acoustic measurements of formants in all stimuli based on previously investigated and established norms (Flipsen et al., 2000, 2001; Shriberg et al., 2001).

In reviewing the token ranking by groups in light of acoustic data, we can observe that time or duration of the stimuli components does not matter to either group of the listeners. F1 can similarly be discounted as a factor. F2 does not seem to matter by itself. However, an appropriate F3 (around 2200Hz – slightly higher than Flipsen et al's 1934 Hz, due to the younger average age of children) is present in the tokens that both groups agree should be ranked high. The second feature published in past research is the difference between F3-F2. Flipsen et al. (2001) gives a 300-700Hz range for a similar age group to the one in this study. 1NA4ra has an F3-F2 of 500Hz and 2NA4ra has an F3-F2 of 436Hz. Both values fit nicely in that range, yet these tokens were ranked lowest by the Educators. In contrast the SLPs appeared to pick up on this feature since they rated the two tokens (1NA4ra & 2NA4ra) higher. The other NA4 tokens are ranked even higher by the SLPs (and the Educators) but in these tokens the F3 fits the norms in addition to F3-F2 being acceptable. This finding suggests that the F3 value might be a more robust point of orientation than F3-F2, but that listeners consider both (with the F3-F2 possibly being a helping factor in decision-making for at least the SLP group). As Acoustic Landmarks and Distinctive Features theory (Stevens, 2002) predicted, the SLPs seemed to orient their ranking based on acoustic landmarks (i.e. F3), whereas the Educators showed no such consistent trends. This could partially be due to the specific education that the SLP receive in contrast to the Educators.

It would be interesting to investigate the trends in F4 and F5 considering the rankings in both groups with respect to Zhou et al.'s (2007) data. However, due to the field recording quality of the signal in all stimuli, no reliable F4 and F5 measurements could be taken.

In addition to the differences between the Educator and SLP groups, this study also looked at additional variables (age of listener, native English skill, being multilingual, being a parent, years of experience with children younger than 12 and test version given) as possibly being a factor in perceptive skills of the listeners. In theory, the more experience one has in listening, the better perceptive skills one should have, provided learning continues in a natural manner. There is also the factor of time and established personal factors, to consider (Massaro, 1989). MANOVA results show no factors of significance in addition to F3. It should be noted however, that the small number of subjects within factors that were considered (age groups, ESL speakers, etc.) limit the generalizibility of statements about the factors considered (see Table 1).

The greatest similarity between the two groups of listeners are the trends of scores in their ranking of the stimuli (see Figures 4 and 5), regardless of the task. This phenomenon supports the idea of an internal prototypical category for phonemes (in this case /r/) or at least similar overall experience in a similar dialect area, with listening to similar children. The individual rankings, and ranking in groups can be and are slightly different, but overall we can observe similar trends for all listeners. The second greatest similarity (perhaps surprisingly so) is the similarity in the variety observed in intra-listener reliability. The Educators, performing this type of task were almost as reliable as SLPs who perform the same type of task on a daily basis.

Listening Task Type

The last question that the present study aimed to answer was whether the task design mattered for judgment of /r/. Identification tasks and rank order scales have been successfully employed in past research (Adler-Bock et al. 2007, Bernhardt et al., 2003, 2005), and they have also been shown to work in this study. However, the pair-wise comparison task had several advantages. It enabled even untrained listeners to make more internally reliable judgments, it provided a higher ICC on inter-reliability, and it enabled the researchers to score the gray area of preference on the stimuli investigated (between the very clearly best and worst stimuli, close but not quite clearly the prototype that the listeners internalized, or on the edge of the category boundary in CP terms).

In closing, it should be mentioned that the listener feedback on the study design was very positive and relevant. All participants found each listening task easy to do, and in general preferred the paired comparison component. They explained this preference by saying that it was easier to tell which /r/ was better. This might have to do with the fact that the comparison task did not need an 'authoritative' judgement. In the identification task, listeners felt that they had to refer to some exemplary version of /r/ (since the question was: is it an /r/ or is it not?), which some were not sure that they had internalized. The portability of the study was also praised, as was the relatively short amount of time in which the test could be completed (30 minutes.). Most of the feedback from the expert listeners (SLPs) had to do with the fact that the tokens they listened to were not of equal length (which is a normal occurrence in everyday life). No listener commented negatively about the fact that only short, extracted syllables were presented.

Another unintended significant finding of this study was that, although as a group, the SLPs reached a solid and arguably valid ranking of the 'goodness' of /r/, there still was considerable variability among raters. This can be explained by the fact that there exists no unified training which all SLPs receive on what makes a good /r/. This leaves the SLPs to rely on their own experiences, which are highly variable.

Study Limitations and Future Research Implications

The Pearson and Spearman correlations measures (APPENDIX 3) confirm what previous researchers found: SLPs orient their ranking process around F3 perception (p=.624) for paired comparison and (p=.605) for the identification task. There was no significance for Educators with respect to acoustics and their stimuli ranking. One of the drawbacks of the paired comparison listening task is its permutative nature, which makes studies grow exponentially with additions of tokens to be investigated. This means that the total number of stimuli investigated in one study is always relatively small (up to 20 stimuli if the task is to remain under one hour); for larger numbers of stimuli other options need to be investigated.

One way to tease apart learned CP effects in the SLPs would be to compare younger SLPs with the more experienced ones. The number of young SLPs is, however, too small to yield significant findings (N=2). In future research the comparison of these two groups could answer the question if it is education alone, or education and experience that creates SLP expertise as listeners.

In terms of further variables that may influence listeners, this study considered listener age, native English skills, knowing other languages, being a parent, years of experience with children, and the version of the test given. Future studies might consider other possible variables over larger samples.

Clinical Implications

The results of this study show that SLPs are subject to the same differences in perception as other listeners. Particularly for tasks such as judging the correctness of a phoneme such as /r/, SLPs responses ideally would be more homogeneous. Such standardization could be accomplished by workshops that center on difficult to correct phonemes and phonetic retraining that would serve to standardize the way that phones are judged.

In theory, research has already provided us with ranges of values in acoustics for almost every single speech sound used in North American English. Based on these measurements, fairly simple acoustic software screening programs could be created that calculate the deviation of a questionable phoneme sample from the norm in age-equivalent terms. In an expansion, such a program could also relate the formant targets for the produced /r/ (or any other phoneme) to concrete oro-mechanical gestures and concrete physical phenomena. This type of feedback for the SLP would further demystify the question of how to deal with /r/ correction in future generations of speech-language pathologists.

REFERENCES

Alwan, A., Narayanan, S. & Haker, K. (1997). Towards articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. Journal of the Acoustical Society of America, 1012, 1078-1089.

Adler-Bock, M., Bernhardt, B., Gick, B., & Bacsfalvi, P.(2007). The use of ultrasound in remediation of /r/ in adolescents. Amer. Journal of Speech-Language Pathology, 16(2), 128-139.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 357-74.

Bernhardt, B., Gick, B., Bacsfalvi, P., & Ashdown, J. (2003). Speech habituation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners. Clinical Linguistics and Phonetics, 17(3), 199-216.

Bernhardt, B., Bacsfalvi, P., Gick, B., Radanov, B., & Williams, R. (2005a). Exploring electropalatography and ultrasound in speech habilitation. Journal of Speech-Language Pathology and Audiology, 29, 169-182.

Bernhardt, B.M., Bacsfalvi, P., Adler-Bock, M., Shimizu, R., Cheney, A., Giesbrecht, N., O'Connell, M., Sirianni, J. & Radanov, B. (2007). Ultrasound as visual feedback in speech therapy: Exploring consultative use in rural British Columbia. Unpublished paper.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, Vol. 39, No. 3/4, pp. 324-345.

Chaney C., (1998). Acoustical analysis of correct and misarticulated semivowels. Journal of Speech and Hearing Research, 31, 275-287.

Delattre, P. & Freeman, D. (1968). A dialect study of American r's by x-ray motion picture. Linguistics, An International Review, 44, 29-68.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psychometrika, 16, 407-424.

Edwards, A. & Cronbach, L. J. (1952) Experimental design for research in psychotherapy. Journal of Clinical Psychology, 8 (1), 51-9

Eimas, P.D., Siqueland, E.R., Jusczyk, P.W., & Vigorito, J. (1971). Speech perception in infants. Science, 171, 303-306.

Espy-Wilson, C. Y. (1992). Acoutic measures for linguistic features distinguishing the semivowels /w j r l/ in American English. Journal of the Acoustical Society of America, 92, 736-757

Espy-Wilson, C.Y., Boyce, S., Jackson, M., Narayanan, S. & Alwan, A. (1999). Acoustic modeling of American English /r/. Journal of the Acoustical Society of America, 108 (1), 343-356.

Flege, J.E., Efting, W., (1988). Imitation of a VOT continuum by native speakers of English and Spanish: evidence for phonetic category formation. Journal of the Acoustical Society of America, 83, 729-740.

Flipsen, P., Shriberg, L. D., Weismer, G., Karlsson, H. B., & McSweeny, J. L.(2000). Acoustic data for American English /r/ and /v/ in typically speaking adolescents (Tech. Rep. No. 10). Phonology Project, Waisman Center, University of Wisconsin-Madison. http://www.waisman.wisc.edu/phonology/bib/bib.htm.

Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H. B., & McSweeny, J. L. (2001). Acoustic phenotypes for speech-genetics: Reference data for residual /r/ distortions. Clinical Linguistics and Phonetics, 15(8), 603-630.

Fujiki, M. & Brinton, B. (1984). Supplementing language therapy - Working with the Classroom Teacher. Language, Speech, and Hearing Services in Schools, 15, 98-109.

Goldstone, R. L, Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. Cognition, 78, 27-43.

Goldstone, R. L.. (1999). Similarity. In R.A. Wilson & F. C. Keil. (Eds.), MIT encyclopedia of the cognitive sciences. (pp. 763-765). Cambridge, MA: MIT Press.

Guenther, F., Espy-Wilson, C., Boyce, S., Matthies, M., Zandipour, M., & Perkell, J. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. Journal of the Acoustical Society of America, 105(5), 2854-2865.

Hagiwara, R., (1995). Acoustic realizations of American /r/ as produced by women and men. UCLA Working Papers in Phonetics, 90, 1-187.

Hagiwara, R., Meyers Fosnot, S. & Alessi, D. M. (2002). Acoustic phonetics in a clinical setting: a case study of /r/-distortion therapy with surgical intervention. Clinical Linguistics & Phonetics, 6(6), 425 - 441

Harnad, S. (ed.) (1987). Categorical Perception: The groundwork of cognition. New York: Cambridge University Press.

Hashi, M., Honda, K., and Westbury, J. (2003). Time varying acoustic and articulatory characteristics of American English [r]: a cross-speaker study. Journal of Phonetics, 31, 3-22

Hayward, Katrina (2000). Experimental phonetics: An introduction. Harlow: Longman.

Hillenbrand, J.M., Clark, M.J., Nearey, T.M. (2001). Effects of consonant environment on vowel formant patterns. Journal of the Acoustical Society of America, 109(2), 748–763

Inter-Rater Reliability Calculator (2007) [computer software]. Accessed May 28th, 2007, from: http://www.med-ed-online.org/rating/reliability.html.

Janzen, K., & Shriberg, L. (1977). How to evoke and generalize "R": A compendium of 36 evocation and phonetic context cues. Madison, Wisconsin: The University Bookstore.

Johnson, K. (2005). Speaker normalization in speech perception. In D.B. Pisoni & R. Remez, (Eds.) The handbook of speech perception. (pp.363-389). Oxford: Blackwell Publishers.

Lane, H. (1965). The motor theory of speech perception: A critical review. Psychological Review, 72, 275-309.

Lawrence, D. H. (1950). Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. Journal of Experimental Psychology, 40, 175-188.

Liberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 54, 358-368.

Lisker, L., Abramson, A.S. (1967). Some effects of context on voice onset time in English stops. Language and Speech, 10, 1-28.

Livingston, K. Andrews & Harnad, S. (1998). Categorical perception effects induced by category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition 24(3), 732-753.

Luce, R.D. (1959). Individual Choice Behaviours: A Theoretical Analysis. New York: J. Wiley.

MacLennon, R. N., (1993). Interrater reliability with SPSS for Windows 5.0. The American Statistician, 47, 292-296

Massaro, D.W. (1989). Testing between the TRACE Model and the Fuzzy Logical Model of Speech perception. Cognitive Psychology 21, 398-421.

Moore, B.C.J. (1997). An Introduction to the Psychology of Hearing. (4th ed.) San Diego, CA: Academic Press.

Nygaard, L.C., Pisoni, D.B. (1995). Speech perception: new directions in research and theory.
In J.L. Miller & P.D. Eimas (Eds.), Handbook of Perception and Cognition. Volume II: Speech, Language and Communication. (pp. 63-96). New York: Academic Press.

Oden, G. C., Massaro, D. W. (1978). Integration of featural information in speech perception. Psychological Review, 85, 172-191

Ohde, R. N. & Sharf, D.J.(1988). Perceptual categorization and consistency of synthesized (r-w) continua by adults, normal children and (r)-misarticulating children. Journal of Speech and Hearing Research, 31, 556 – 568.

Pevtzow, R. & Harnad, S.. (1997). Warping similarity space in category learning by human subjects: The role of task difficulty. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain, (Eds.), Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization (pp. 189-195). Department of Artificial Intelligence, Edinburgh University.

Boersma, P. and Weenink, D. 2003, Praat. Version 4.0.49 [computer software]. Amsterdam,
The Netherlands: University of Amsterdam.

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.

Ruscello, D. (1984). Motor learning as a model for articulation instruction. In J. Costello (Eds.), Speech disorders in children, recent advances. (pp. 129-156). San Diego, CA: College Hill Press.

Ruscello, D. (1993). A motor skill learning treatment program for sound system disorders. Seminars in Speech and Language, 14(2),106-118.

Ruscello, D. (1995a). Speech appliances in the treatment of phonological disorders. Journal of Communication Disorders, 28, 331-353.

Ruscello, D. (1995b). Visual feedback in treatment of residual phonological disorders. Journal of Communication Disorders, 28, 279-302.

Ruscello, D. & Shelton, R. (1979). Planning and self-assessment in articulatory training. Journal of Speech and Hearing Disorders, 44, 504-512.

Ruscello, D. M., Louis, K. O. and Mason, N. (1991). School-aged children with phonologic disorders: Coexistence with other speech/language disorders. Journal of Speech and Hearing Research, 34, 236-242.

Sharf, D.J. & Benson, P.J. (1982). Identification of synthesized /r-w/ continua for adult and child speakers. Journal of the Acoustical Society of America, 71, (4) 1008 – 1015.

Sharf, D. J. & Benson, P.J. (1983). Comparison of speech-language pathologists' and naive listeners' identification of synthesized /r-w/ continua. Journal of Speech and Hearing Research, 26 pp 525-530.

Sharf, D.J., Ohde, R.N. & Lehman, M.E. (1988). Relationship between the discrimination of /w-r/ and /t-d/ continua and the identification of distorted /r/. Journal of Speech and Hearing Research, 31, 193 – 206.

Shelton, R.L., Johnson, A., & Arndt, W.B.(1974). Variability in judgments of articulation when observer listens repeatedly to the same phone. Perceptual and Motor Skills, 39, 327-332.

Shriberg, L. (1975). A response evocation program for /r/. Journal of Speech and Hearing Disorders, 40(1), 92-105.

Shriberg, L. (1980). An intervention procedure for children with persistent /r/ errors. Language, Speech, and Hearing Services in Schools, 11, 102-110.

Shriberg, L., Flipsen, P., Karlsson, H. & McSweeny, J. (2001). Acoustic phenotypes for speech-genetics studies: An acoustic marker for residual /r/ distortions. Clinical Linguistics and Phonetics, 158, 631-650.

Shuster, L., Ruscello, D. & Smith, K. (1992). Evoking /r/ using visual feedback. American Journal of Speech-Language Pathology, 1, 29-34.

Shuster L., Ruscello, D. & Toth, A. (1995). The use of visual feedback to elicit correct /r/. American Journal of Speech-Language Pathology, 4, 37-44.

Stevens, K.N. (2002). Toward a model of lexical access based on acoustic landmarks and distinctive features (PDF). Journal of the Acoustical Society of America, 111(4), 1872-1891.

Strange, W. (1999). Perception of vowels: Dynamic constancy. In J.M. Pickett (Ed.). The acoustics of speech communication: Fundamentals, speech perception theory, and technology. (pp. 153-165). Needham Heights (MA): Allyn & Bacon.

Syrdal, A.K., Gopal, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. Journal of the Acoustical Society of America, 79, 1086-1100.

Thurstone, L.L. (1927). A law of comparative judgement. Psychological Review, 34, 278-286.

Thurstone, L.L. (1929). The measurement of psychological value. In T.V. Smith & W.K. Wright (Eds.), Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago. Chicago: Open Court.

Thurstone, L.L. (1959). The measurement of values. Chicago: The University of Chicago Press.

Vatikiotis-Bateson, E., Munhall, K. G., Kasahara, Y., Garcia, F.& Yehia, H. (1996). Characterizing audiovisual information during speech, Proceedings of the International Conference on Spoken Language Processing-1996, 1485-1488.

Westbury, J., Hashi, M. & Lindstrom, M., (1998). Differences among speakers in lingual articulation for American English /r/. Speech Communication, 26, 203-226.

Wolfe, V. Martin, D., Borton, M. & Youngblood, H.C. (2003) The effects of clinical training on cue trading for the /r-w/ contrast. American Journal of Speech-Language Pathology, 12, 221 – 228.

Zhou, X., Espy-Wilson, C., Tiede, M. and Boyce, S. (2007). Acoustic cues of "retroflex" and "bunched" American English rhotic sound. Journal of the Acoustic Society of America, 121. (5), Pt.2: 3168.

APPENDIX 1

University of British Columbia Ethics Approval

The University of British Columbia
Office of Research Services
Behavioural Research Ethics Board
Suite 102, 6190 Agronomy Road, Vancouver, B.C. V6T 1Z3

CERTIFICATE OF APPROVAL - FULL BOARD

PRINCIPAL INVESTIGATOR:	INSTITUTION / DEPARTMENT:	UBC BREB NUMBER:
Barbara M. Bernhardt	UBC/Medicine, Faculty of/Audiology & Speech Sciences	H06-03610

INSTITUTION(S) WHERE RESEARCH WILL BE CARRIED OUT:

Institution	Site
UBC	Point Grey Site

Other locations where the research will be conducted:
James Mather Building, University of British Columbia, Rooms 222, 221, 223 Schools in the Lower Mainland (after receiving permission from the school boards, local schools), YMCA Child Care after-school program which takes place in the local schools (i.e., approval from both the school and the after-school program is being sought once conditional approval has been obtained from BREB)

CO-INVESTIGATOR(S):

Bosko Radanov
Benjamin Perry

SPONSORING AGENCIES:
N/A

PROJECT TITLE:
Perception of children's /r/ production by other children, teachers and speech-language pathologists

REB MEETING DATE:	CERTIFICATE EXPIRY DATE:
November 23, 2006	**November 23, 2007**

DOCUMENTS INCLUDED IN THIS APPROVAL: **DATE APPROVED: January 8, 2007**

Document Name	Version	Date
Consent Forms:		
Teacherconsent	Version 2	December 11, 2006
SLP Consent Form	Version 2	December 11, 2006
Parent consent	Version 2	December 11, 2006
Assent Forms:		
Child assent	Version 2	December 11, 2006
Advertisements:		
AdSLP	Version 2	December 11, 2006
Adchildparent	Versioin 2	December 11, 2006
Adteacher	Version 2	December 11, 2006
Letter of Initial Contact:		
Initial Contact to Schools	Version 1	November 9, 2006
Initial Contact to YMCA Child Care Program, Vancouver	Version 1	December 11, 2006

The application for ethical review and the document(s) listed above have been reviewed and the procedures were found to be acceptable on ethical grounds for research involving human subjects.

Approval is issued on behalf of the Behavioural Research Ethics Board and signed electronically by one of the following:

Dr. Peter Suedfeld, Chair
Dr. Jim Rupert, Associate Chair
Dr. Arminee Kazanjian, Associate Chair
Dr. M. Judith Lynam, Associate Chair

APPENDIX 2

2. Ranking, Acoustics, Pearsons and Spearmans Correlations

(Educators tables left out because of no significance)

SLP comparison ranking	**F1**	**F2**	**F3**	**F3-F2**	**Rdur**	**Vdur**	**TOTALdur**
Pearson Correlation	.018	.160	**.624****	.402	.203	-.244	.024
SLP comparison ranking	**F1**	**F2**	**F3**	**F3-F2**	**Rdur**	**Vdur**	**TOTALdur**
Spearman's rho	-.162	.211	**.562***	.415	.210	-.066	.134
SLP identification task	**F1**	**F2**	**F3**	**F3-F2**	**Rdur**	**Vdur**	**TOTALdur**
Pearson Correlation	-.035	.109	**.605***	.424	.179	-.352	-.054
SLP identification task	**F1**	**F2**	**F3**	**F3-F2**	**Rdur**	**Vdur**	**TOTALdur**
Spearman's rho	-.190	.177	**.543***	.375	.193	-.197	.062

Note. dur = duration, V = Vowel

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

In both cases the only factor that varied significantly with the way the stimuli were ranked was F3. This significant result only showed up for educators.

APPENDIX 3

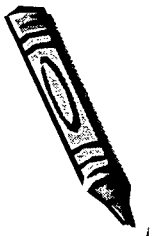
3. **Order of Stimuli Presentation** During Experiment

Test Version 1	Test Version 2	Test Version 3	Test Version 4
4 slides test-retest ID	20 slides test-retest paired comparison	50 slides paired comparison only	4 slides test-retest ID
16 slides ID	100 slides paired comparison only	20 slides test-retest paired comparison	16 slides ID
4 slides test-retest ID	20 slides test-retest paired comparison	20 slides test-retest paired comparison	4 slides test-retest ID
20 slides test-retest paired comparison	4 slides test-retest ID	50 slides paired comparison only	50 slides paired comparison only
100 slides paired comparison only	16 slides ID	4 slides test-retest ID	20 slides test-retest paired comparison
20 slides test-retest paired comparison	4 slides test-retest ID	16 slides ID	20 slides test-retest paired comparison
		4 slides test-retest ID	50 slides paired comparison only

APPENDIX 4

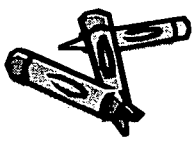
4. Demographics Questionnaire

Slide 1



Please take a moment to give us a little information about yourself:

1. I am female ☑ I am male ☐
2. I am 20-30 ☐ 31-40 ☐ 41-50 ☑ 51-60 ☐ 60+ ☐ years old.
3. I was born in Canada. YES ☑ NO ☐
4. I came to Canada ____ years ago.
5. I speak: English



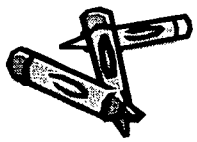
At home we speak:

English

Slide 2

1. I live in Richmond (city, country)
2. Job title SLP Work setting School District
3. I am a parent YES ☑ NO ☐
4. I have 1 children, age/gender 11/M
5. I have 23 years experience with kids under 12.

☑ Currently - or - ☐ This many years ago ____



Slide 3

1. Are you a parent of a child who has, or has had difficulties pronouncing sounds after kindergarten? YES ☐ NO ☐

2. Have you ever had any difficulties with speech sounds?
YES ☐ NO ☐

3. How is your hearing?
- ☐ No problems
- ☐ Undiagnosed problem
- ☐ Diagnosed problem

4. Have you had you hearing checked, and how long ago since the check-up? No

I would like to have my hearing checked please ☐

Distracter slide example – no sound played