Perceptual judgment of hypernasality and audible nasal emission in cleft palate speakers

Downing, Kerri

http://hdl.handle.net/2144/13658
Boston University
PERCEPTUAL JUDGMENT OF HYPERNASALITY AND AUDIBLE NASAL EMISSION IN CLEFT PALATE SPEAKERS

by

KERRI DOWNING

B.A., University of Connecticut, 2010

Submitted in partial fulfillment of the requirements for the degree of Master of Science

2015
Approved by

First Reader
Cara E. Stepp, Ph.D.
Assistant Professor of Speech, Language & Hearing Sciences
Assistant Professor of Biomedical Engineering

Second Reader
Barbara B. Oppenheimer, M.A., CCC-SLP
Clinical Associate Professor

Third Reader
Melanie L. Matthies, Ph.D.
Senior Associate Dean
Associate Professor of Speech, Language & Hearing Sciences
PERCEPTUAL JUDGMENT OF HYPERNASALITY AND AUDIBLE NASAL EMISSION IN CLEFT PALATE SPEAKERS

KERRI DOWNING

ABSTRACT

Objective: The purpose of this study is to determine whether a novel, user-friendly rating system, visual sort and rate (VSR) provides comparable ratings to the currently used direct magnitude estimation (DME) rating system for rating perceptions of audible nasal emission (ANE) and hypernasality in cleft palate speakers.

Methods: Twelve naïve listeners rated 152 speech samples of speakers with cleft palate across four conditions: rating hypernasality and ANE using either a VSR or DME rating scale. Raters were provided with a short training session, prior to rating each day. Inter- and intra-rater reliabilities, as well the line of best fit between scores using VSR and scores using DME was calculated to determine usability of VSR as a novel rating system.

Results: Direct magnitude estimation resulted in the highest levels of inter-rater reliability, when rating hypernasality (DME r=.48; VSR r=.14), as well as ANE (DME r=.27; VSR r=.15). Most raters demonstrated high intra-rater reliabilities across conditions. A curvilinear line of best fit most accurately captured the relationship between DME and VSR scores when rating hypernasality (r=.64) and ANE (r=.66).

Conclusions: A curvilinear relationship between ratings suggests that both variables are prothetic, and therefore, best captured using a DME rating scale (Eadie & Doyle, 2002). The use of DME is supported for continued use rating hypernasality, even
amongst naïve listeners given a training session. Rating ANE was difficult, as ratings yielded low inter-rater reliabilities, regardless of the scale used. Further research regarding perceptions of audible nasal emission is warranted.
TABLE OF CONTENTS

ABSTRACT................................................................. iv

TABLE OF CONTENTS....................................................... vi

LIST OF TABLES............................................................ ix

LIST OF FIGURES.......................................................... x

LIST OF ABBREVIATIONS.................................................. xi

BACKGROUND ............................................................... 1

Prevalence and Characteristics of Cleft Palate ................................ 1
Resonance Patterns in Individuals with Palatal Cleft Repair................... 2
Management of Cleft Palate.................................................. 2
Nasal Air Emission ................................................................ 4
Nasometry as an Objective Measure of Nasal Resonance....................... 5
Perceptual Judgments of Nasal Resonance......................................... 7
Inter- and Intra-Rater Reliability ............................................... 9
How Experience Affects Perceptual Judgments................................. 10
Hypernasality and Nasal Emission as a Coninuum?............................. 11
Methodologies Used to Collect Perceptual Ratings............................ 12
Aligning Perceptual Judgments with Objective Measurements............... 15

STIMULI ........................................................................... 16

Sample Collection................................................................ 16
LIST OF TABLES

Table 1. Speaker Characteristics.......................................................... 17
Table 2. Inter-rater Reliability. ............................................................ 28
Table 3. Mean Differences in VSR Ratings............................................ 29
LIST OF FIGURES

Figure 1. Training Condition .............................................................................................................. 21
Figure 2. DME Condition .................................................................................................................... 22
Figure 3. VSR Condition ...................................................................................................................... 23
Figure 4. Assigning Numerical Values in VSR Condition ................................................................. 24
Figure 5. Pearson Correlations........................................................................................................... 27
Figure 6. Linear and Polynomial Models .......................................................................................... 31
LIST OF ABBREVIATIONS

ANE ............................................................ Audible Nasal Emission
DME .................................................................. Direct Magnitude Estimation
VSR ................................................................. Visual Sort and Rate
Background:

Prevalence and Characteristics of Cleft Palate:

Craniofacial defects including cleft lip and palate, are one of the most common birth defects (Centers for Disease Control and Prevention [CDC], 2013). Estimates from 30 countries around the world, indicate that cleft palate is a prevalent, worldwide occurrence (“Prevalence at Birth of Cleft Lip With or Without Cleft Palate,” 2011). Therefore, classifying speech abnormalities in this population is of importance. According to Kummer (2008), a cleft is described as, “an abnormal opening of a fissure in an anatomical structure that is normally closed” (p.37). These abnormal openings can affect the complete formation of the lip or palate resulting in cleft lip and/or palate, resulting in velopharyngeal insufficiency.

Velopharyngeal insufficiency is defined as a structural or functional abnormality that prevents velopharyngeal closure (Kummer, 2008). Individuals usually opt for corrective surgery; however, even after surgery, individuals often continue to experience velopharyngeal insufficiency (Timmons, Wyatt, & Murphy, 2001), incompetence (neuromotor or physiological disorder that prevents the proper velopharyngeal movement), or mislearning (incorrect velopharyngeal movement caused by articulation problems). Velopharyngeal dysfunction is an umbrella term which includes insufficiency, incompetency, and mislearning. Velopharyngeal dysfunction (VPD) is defined as, “a condition where the velopharyngeal valve does not close consistently and completely during the production of oral sounds” (Kummer, 2008, p. 178). Velopharyngeal dysfunction can cause: hypernasal speech, hyponasal speech, audible nasal emission
(ANE), cul-de-sac resonance, and/or compensatory articulations (Kummer, 2008).

**Resonance Patterns in Individuals with Palatal Cleft Repair:**

Resonance, as described by Kuehn and Moller (2000), is described as, “…the direct effect of coupling of the nasal space with the oral-pharyngeal space during vowel and vocalic productions” (p.348–4). There are several common resonance patterns amongst individuals with velopharyngeal dysfunction. According to Henningsson et al. (2008), the five most important dimensions of cleft speech ratings are hypernasality, hyponasality, nasal emission, compensatory articulations, and voice disorder. *Hypernasality*, or the presence of increased nasal resonance during non-nasal sounds, is a common result of velopharyngeal insufficiency. After cleft palate repair surgery, children’s voices are still often perceived as hypernasal (Persson, Lohmander, & Elander, 2006). This becomes problematic as resonance problems can affect speech intelligibility, thereby warranting further therapy (Kuehn & Moller, 2000; Persson et al., 2006). Even five years after surgery, hypernasality remained significant when compared to controls (Paliobei, Psifidis, & Anagnostopoulos, 2005).

**Management of Cleft palate:**

Velopharyngeal Insufficiency (VPI) can be corrected through prosthetic use, surgery, and/or speech therapy (Kummer, 2008). Kummer states that while the use of prosthetics previously was the predominant treatment for cleft-palate, surgery has become more prominent (2008). Surgeries that correct deformities in the palate are called palatoplasty, while any surgical procedure of the pharynx aimed to treat VPI is called
pharyngoplasty. Kummer suggests that repair of the palate typically requires two surgeries to account for the growth of the palate through adulthood. The first of the two surgeries typically occurs between 6 and 15 months of age.

Early repair is conducted to reduce velopharyngeal mislearning and compensatory articulations (Kummer, 2008). Several types of surgeries exist for cleft palate repair including: Von Langenbeck surgery, Wardill-Kilner surgery, intraveloplasty, and furlow palatoplasty (Hopper, Tse, Smartt, Swanson, & Kinter, 2014; Woo, 2012). Secondary surgeries, or pharyngeal wall augmentation, has been known to improve VPI (Kummer, 2008; Seagle, Mazaheri, Dixon-Wood, & Williams, 2002). Two types of pharyngoplasty, pharyngeal flap and sphincter pharyngoplasty are used. Currently, the pharyngeal flap palatoplasty is the most common surgical technique known to improve VPI (Armour, Fischbach, Klaiman, & Fisher, 2005; Cable, Canady, Karnell, Karnell, & Malick, 2004). Prosthetics have become less necessary as surgeries are typically completed at younger ages (Gardner & Parr, 1996; Kummer, 2008). However, for those individuals who do not qualify for surgery, prosthetics can be important for treating velopharyngeal disorders. Three types of prosthetics can be used: palatal lift, palatal obturator, and a speech bulb obturator. The palatal lift is useful for treating velopharyngeal incompetence, as it aids the palate to lift during the production of oral sounds (Kummer, 2008). A palatal obturator covers any openings in the palate to stop airflow through the open palate, and through the nasal cavity. Lastly, a speech bulb obturator aids velopharyngeal insufficiency, as it can fill the empty space between the velum and posterior pharyngeal wall during velar closure.
As noted by Kummer (2008), speech therapy is useful once prosthetic or surgical techniques have been completed to aid patients in managing the newly acquired oral pressure. Similarly, compensatory articulations (e.g. pharyngeal fricatives, glottal stops, and nasal shunting) may develop in children with VPI; speech therapists can resolve compensatory articulations. Hardin-Jones and Jones (2005) found that sixty-seven percent of preschoolers with repaired cleft-palate are currently enrolled or were previously enrolled in speech therapy. Their findings indicated aberrant articulations persisted after palatal repair including glottal stops, pharyngeal fricatives, middorsum palatal stops, and nasal substitutions.

**Nasal Air Emission:**

Velopharyngeal insufficiency often results in *nasal emission*, or air leaking through the nose during high pressure oral consonants such as /b/, /p/, and /d/ (Wyatt et al., 1996). The most severe type of audible nasal emission (ANE) is typically called nasal turbulence or rustle (Kummer, Curtis, Wiggs, Lee, & Strife, 1992). While resonance is amongst one of the most studied voice variables in perceptual assessments of individuals with cleft palate, nasal air emission is less studied, as fewer than half of recent studies reviewed, assessed nasal emission (Baylis, Munson, & Moller, 2011; Lohmander & Olsson, 2004). Nasal emission is more subtle than other measures of nasality and occurs from a different velopharyngeal mechanism than hypernasality; therefore, further research on nasal emission is warranted to better understand its effect on speech intelligibility.
Hypernasality and nasal emission occur through different velopharyngeal mechanisms. As stated by Sell, Harding, and Grunwell (1999), “hypernasality modifies vocal tone, whereas nasal emission reflects nasal airflow which is unrelated to voice…” nasal emission is most readily perceived on the production of voiceless consonants” (p. 19). Nasal emission and nasality are also associated with different sizes of the opening of the nasal port. Using videofluoroscopy, Kummer et al. (1992) found that a large opening of the velopharyngeal gap is often associated with hypernasality whereas a small opening is typically associated with nasal rustle. It is believed that the small opening causes air to whistle through, creating a high frequency nasal emission. Meanwhile, the large opening of the port can result in an overabundance of nasalization during typically non-nasal sounds.

**Nasometry as an Objective Measure of Nasal Resonance:**

Currently, perceptual judgments are the gold standard for classifying speech differences amongst individuals with cleft palate (Kuehn & Moller, 2000; Lee, Whitehill, & Ciocca, 2009). However, there is conflicting evidence regarding the effects of experience on perception, and a lack of consistent inter- and intra-rater reliabilities. Therefore, the nasometer was designed as an objective measure to identify speech differences. A nasometer is a non-invasive headset with a plate separating the nasal and oral areas. Two microphones, one placed on each side of the plate make simultaneous recordings and create a ratio, comparing the acoustic energy coming from the oral and nasal areas, which is provided as a measure of nasalance (KayPentax, 2011). According to Dalston, Warren, and Dalston (1991b), a high nasalance score results when there is
increased nasal resonance.

However, the nasometer cannot be used in isolation, without perceptual judgments because it is not reliably able to distinguish between normal nasalance during typical nasal consonants and hypernasality (Watterson, Hinton, & McFarlane, 1996). As found by Sweeney and Sell (2008), the presence of high nasalance readings from a nasometer did not always reflect high perceptual ratings of hypernasality. In addition, the presence of nasal emission changes the accuracy of the nasometer. When detecting hypernasality including samples of individuals with nasal emission, sensitivity was 0.3; excluding samples with nasal emission increased the sensitivity of the nasometer to 0.67 (Dalston, Warren, & Dalston, 1991a). In addition, total nasalance scores increase when audio samples include nasal emission (Karnell, 1995). The nasometer cannot differentiate between audible nasal emission and hypernasality, as both are perceived as an excess of nasalance. Discrimination between hypernasality and ANEs is therefore impossible without the use of an additional measure, such as perceptual judgment.

Considering that hypernasality is a result of a largely open velopharyngeal port, and ANEs are due to a slightly open funneling of air, the inability for the nasometer to detect these differences can lead to problems in creating a treatment plan for these children post-surgery. Although an individual with hypernasality and an individual with ANE may both receive the same nasometer scores, they would each require different treatment plans, as an individual with a large opening may require additional surgeries, while a small opening may be considered a successful surgery and necessitate speech therapy for resolution (Kummer et al., 1992). Therefore, perceptual judgments that
differentiate between hypernasality and the presence of ANEs would provide crucial treatment information about children with cleft palate.

_Perceptual Judgments of Nasal Resonance:_

Several measurement systems have been employed to assess perceptual judgment. Four main rating scales are typically used in rating voice, Equal-appearing interval (EAI), Visual Analog scale (VAS), Direct Magnitude Estimation (DME), and Paired comparison. In EAI, individuals are asked to rate a voice sample from 1 to n (depending on the scale) with equal intervals between each rating. Typically these scales are seven point scales (with some nine or five point scales) and end points of the scales which represent the range of dysfunction of the samples. So for instance, a one may quantify the most typical rating, with the other endpoint, a seven, representing the most severe.

Direct Magnitude Estimation asks individuals to provide a rating for the speech sample that is typically unrestricted, and is usually accompanied by a modulus to provide a marker of the average of the perceptual continua for the quality being assessed (Eadie & Doyle, 2002). The modulus provided is given a rating of 100, and raters can assign any rating they desire to capture the stimulus by comparison. This rating system provides magnitudes of how stimuli differ from one another. Another rating system, paired comparison ratings, asks raters to compare the two samples on an aspect of voice (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). This creates a forced choice between which sample sounds more dysfunctional. Although this rating provides information regarding which sample is more impaired along a certain dimension, it does not provide information about how much more deviant the sample is.
Visual Analog Scaling provides raters with a line in which ratings are made by placing a mark along the continuum. The line acts as a visual cue for raters to place their score along the continuum; numerical ratings are typically gathered by measuring the location of the tick mark relative to the length of the line. A variant of VAS scaling, visual sort and rate (VSR), asks individuals to sort samples based on severity, then assign values to them using a VAS interface. The VSR provides raters with the opportunity to place stimuli in order from most disordered to least disordered, as well as to place each sample along a visual axis to capture severity (as in a visual analog scale), providing a numerical rating. Sound samples placed closer to one another represent more similar samples, while samples further from one another represent samples different from one another.

Qualities of voice and resonance can be considered either a prothetic or metathetic quality. A prothetic quality is one that varies in magnitude, and is representative of a quantitative change, such as loudness. A metathetic quality varies in quality, such as pitch (Eadie & Doyle, 2002). Certain qualities, depending on whether they are prothetic or metathetic, can be best captured using different rating systems. As stated by Eadie and Doyle (2002), “a prothetic continuum… is best scaled with DME because observers cannot subdivide a prothetic continuum into equal intervals (p. 3015).” Therefore, prothetic variables are typically best investigated using DME, as EAI scaling are not a valid way of ascertaining ratings of hypernasality (Whitehill, Lee, & Chun, 2002; Zraick & Liss, 2000). Therefore, careful consideration of the type of scale used to collect ratings is necessary for valid perceptual ratings.
**Inter- and Intra- Rater Reliability**

Inter- and Intra- rater reliabilities are used to determine agreement of perceptual ratings. Intra-rater reliability measures how stable a given judges’ score is. Additionally, inter-rater reliability describes how well a group of listeners agrees on scoring. Past research has shown poor inter- and intra- rater reliability on ratings of nasality and ANE (Brunnegård, Lohmander, & van Doorn, 2012). Amongst experienced speech language pathologists, there was low inter-rater reliability on ratings of hypernasality during oral sentences, ranging from 42% to 62%. Ratings of hyponasality had even larger ranges, with percentages of inter-rater reliability ranging from 38% to 96%. Ratings of ANE revealed inter-rater reliability ranging from 7% to 99%.

Ratings of ANEs are often not reported even when they are collected (Baylis et al., 2011). Several studies have asked listeners to provide opinions regarding ANEs but few have reported inter and intra-rater reliabilities (Kummer, Briggs, & Lee, 2003; Kummer et al., 1992; McWilliams et al., 1996; Randall et al., 2000; Witt, Berry, Marsh, Grames, & Pilgram, 1996). Amongst the studies that have reported inter- and intra- rater reliabilities, inter-rater reliabilities are moderate at best (ranging from .58–.70) and intra-rater reliabilities ranging from .42–.94 (Baylis et al., 2011; John, Sell, Sweeney, Harding-Bell, & Williams, 2006; Keuning, Wieneke, van Wijngaarden, & Dejonckere, 2002; Persson et al., 2006; Pinborough-Zimmerman, Canady, Yamashiro, & Morales, 1998). Given the large range of rater reliabilities, studies have focused on qualitative differences between raters, such as experience amongst listeners, to explain variance in ratings.
How Experience Affects Perceptual Judgments:

An individual’s level of expertise may affect perceptual judgments. Previous studies have found inconsistent results regarding the effects of experience. Some studies state that experienced listeners are more likely to use an internalized scale built from client’s voices to produce judgments (Kummer, Clark, Redle, Thomsen, & Billmire, 2012). These subjective measurements are at risk of being colored by professional experiences over many years of clinical practice. Professionals may use previous patients as an example of the most severe type of voice dysfunction when making perceptual judgments. Studies have shown that experience leads to greater inter-rater reliabilities compared to naïve listeners (Brunnegård et al., 2012). For example, when rating nasality, inter-rater reliability for trained speech language pathologists was between 48–60% and was 30–62% for non-expert SLPs, illustrating significant differences in ratings between experienced SLPs and naïve raters. However, contrasting these results, one study found lower inter-rater reliabilities amongst expert listeners (r=.27–.73) compared to naïve listeners (r=.42–.66) when rating pathological voices (Kreiman, Gerratt, Precoda, & Berke, 1992). The expert listeners showed a larger range in reliabilities, with some speech pathologists agreeing more than others.

Often, inexperienced listeners are more likely to rate severity as more severe than experienced listeners (Helou et al., 2010; Laczi, Sussman, Stathopoulos, & Huber, 2005; Lewis, Watterson, & Houghton, 2003). However, Kreiman and Gerratt (2000) suggested that raters are not reliably able to isolate different aspects of speech. Therefore, perhaps they are using other voice qualities to create their ratings while experienced listeners are
better at isolating and scoring hypernasality independently. In order to account for these differences, it has been suggested that naïve listeners should receive training sessions to improve reliability (Eadie & Baylor, 2006; Lee et al., 2009).

**Hypernasality and Nasal Emission as a Continuum?:**

Studies have found conflicting evidence whether measures of nasality should be rated on a continuum or whether they represent discrete mechanisms. For example, are hyper and hyponasality two ends of the same scale of nasalance (Fletcher, 1976; Henningsson et al., 2008)? Other studies have suggested these to be discrete disorders and support separate scoring criteria (Keuning, Wieneke, & Dejonckere, 1999; Sell et al., 1999). Scoring nasalance can be difficult as raters often group hyper and hyponasality together, suggesting that perceived increases in hypernasality often increase reports of increased hyponasality (Brunnegård et al., 2012; Nellis, Neiman, & Lehman, 1992). Typically, as nasalance scores increase, perception of hypernasality increases and therefore, perception of hyponasality should decrease.¹

Similar problems have been cited when rating nasal emission. This subtle emission can be categorized as a nasal rustle, nasal turbulence, and auditory nasal emission. While some studies have suggested that varying amounts of nasal emission affect the label of nasal turbulence or rustle (Baylis et al., 2011; Kummer et al., 1992), other studies have suggested that nasal rustle and turbulence are completely distinct perceptions and should be measured on their own scales (Sell et al., 1999). However,

---

¹ For the proposed study, hypernasality will be assessed in isolation, without asking raters to assess hyponasality.
most studies do not differentiate the types of nasal emission (Keuning et al., 2002; Keuning et al., 1999; Persson et al., 2006; Randall et al., 2000; Witt et al., 1996).  

**Methodologies Used to Collect Perceptual Ratings:**

Research by Zraick and Liss (2000) concluded that nasality is a prothetic variable, and therefore, would be best measured using a Direct Magnitude Estimation. As stated by Stevens (1975), “Whenever given a prothetic continuum, the matching procedure employed is such that the observer is forced out of ratioing and into differencing, so to speak, there results a distorted scale” (p.135). Therefore, the use of a DME scale allows users to use a scale that befits the stimuli, while decreasing the possibility that the extreme ends of the spectrum will be poorly represented. Similarly, Whitehill et al. (2002), found that EAI ratings had a curvilinear relationship to DME ratings suggesting DME as the most appropriate rating system. A novel study by Baylis et al. (2011) examined nasal emission as prothetic or metathetic quality. Their results support for nasal emission as a prothetic quality. They concluded that nasal emission is a continuum of impairment, and therefore, best captured using DME measurements. Additionally, using DME can increase intra-rater reliability (Baylis et al., 2011). In a later study by Baylis, Chapman, Whitehill, and Group (2014), further support was provided for DME and ANE as prothetic variables. Findings revealed a curvilinear relationship when comparing ratings using DME compared to EAI. From this evidence, it seems as if DME is the best measurement for nasality.

2 Listeners will not be asked to differentiate among the different types of ANE and will provide one unique rating for ANEs as a group.
Overall, most perceptual continua are prothetic, and therefore, best measured using DME (Stevens, 1975). In addition, “magnitude estimation permits the free assignment of numbers so that magnitudes, as well as ratios among magnitudes, can be reflected in the numbers that the observer emits to depict his subjective impression” (Stevens, 1975, p. 135). In addition, Eadie and Doyle (2002) assessed perceptual ratings of a prothetic variable, voice severity, and a metathetic variable, voice pleasantness. They found that EAI or DME could be accurately used to provide ratings for the metathetic variable, while only DME was appropriate for the prothetic variable. Therefore, the use of DME may be superior to using an equal interval scale, as it can adequately measure both prothetic and metathetic values.

Visual sort and rate (VSR), has not been investigated in as many studies evaluating resonance. Higher correlation coefficients were found using VSR than when using VAS ratings to rate synthetic voices on the presence of high frequencies, and natural voices on breathiness, roughness, and hyperfunctionality. (Granqvist, 2003). Since VSR allows listeners to compare each sample to a previous sample, there is no need for a modulus. This VSR system can decrease the effects of professional experience, as naïve listeners have direct comparison of each of the stimuli to one another. Trained professionals will be less likely to compare these samples by using internal standards of a disordered voice, since they are asked to rank the samples in severity in relation to one another. Although EAI is the most common rating scale when assessing nasality (Kreiman et al., 1993), recent findings suggest nasality is better captured using DME (Whitehill et al., 2002; Zraick & Liss, 2000). However, other studies have found that
DME is inferior to more user friendly rating systems (Brancamp, Lewis, & Watterson, 2010). Therefore, VSR could provide a user friendly alternative to DME.

When using an EAI scale to rate roughness of pathological voices, raters tended to increase severity ratings as raters heard more samples. Kreiman et al. (1993) posited that raters were using previous samples as comparisons, resulting in higher overall severity for later trials than earlier ones. These methodological differences were not true when raters were using a VAS scale (Kreiman et al., 1993). Furthermore, a study by Keuning, et al. (1999), found that visual analog scales often resulted in raters differing in their use of the scale with ratings skewed to the extremes of the scale. Therefore, the use of VSR, in which voices are ordered by virtue of dysfunction, should reduce methodological concerns about the use of endpoints.

In addition, individuals often differ in how they weigh various aspects of speech. For instance, some raters find fundamental frequency as the most important indicator of a pathological voice, while others note perceptions of shimmer and roughness (Kreiman, Gerratt, & Precoda, 1990). Since individuals tend to use idiosyncratic anchor referents and compare it against their own mental representations of atypical voices, the authors suggested the use of an anchor stimulus after every few stimulus presentations to lessen the effects of internal mental representations (Kreiman et al., 1992). Therefore, it is advisable to provide a modulus for raters when using the DME rating system.³

³ A modulus will be provided for raters when rating samples using DME.
Aligning Perceptual Judgments with Objective Measurements:

Studies investigating how well perceptual judgments align with objective measurements of nasalance have found conflicting results. One study found that experience mattered, with expert speech-language pathologists having higher correlations with objective measures than naïve listeners. Expert speech-language pathologists had strong correlations (.67–.74) between their perceptual ratings of nasality and the nasalance scores, while untrained speech-language pathologists and untrained listeners had moderate correlations ranging from .55–.76 and .42–.48 respectively (Brunnegård et al., 2012). Other studies have found good agreement between perceptual judgments and nasometer scores across naive listeners. In one study, when asking naïve listeners to assign ratings for how much nasality is present, listener judgment ratings correlated with nasalance scores with a correlation of .91 (Fletcher, 1976).

Additionally, comparing ratings of the Temple Street scale, an EAI scale used to rate nasality, to nasalance scores from a nasometer, the nasometer had a sensitivity of .83–.88 and specificity ranging from .78–.95 when using perceptual ratings as diagnostic judgment (Sweeney & Sell, 2008). Other studies have also found similar sensitivity and specificity, ranging from .70 to .79 (Dalston, Neiman, & Gonzalezlanda, 1993; Watterson, McFarlane, & Wright, 1993). These moderately high sensitivities and specificities of nasometry suggest that it could be an effective non-invasive useful tool in objectively identifying speech differences. However the nasometer is not sufficient to diagnose voice perceptions on its own. As stated by Kuehn and Moller (2000), “suffice it to say that at the present time, there is no technique that demonstrates a sufficient
relationship to perceived nasality to eliminate the use of perceptual judgments in
satisfactory description of speech” as objective measurements require further research
regarding efficacy.

Overall, there has been a wide range of research on nasality and how these scores
should be measured. However, studies on nasal emission have been underrepresented in
the literature, especially in-depth studies on natural speech and how they are best
categorized. Direct magnitude estimation has been used to provide ratings for both
metathetic and prothetic variables. However, few studies have studied using visual sort
and rate for providing ratings of nasality. An initial study investigating voice quality
found high correlations and highlighted it as a user friendly interface (Granqvist, 2003).
Given the literature summarized above, the goals of the current study are:

1) To determine which methodology results in the highest inter and intra-rater reliability
   for hypernasality and ANE.

2) To investigate the validity of VSR as a novel rating system of hypernasality and ANE
   perceptions relative to DME.

**Stimuli:**

*Sample Collection*

Speech Samples were collected from children with cleft palate at the University of
Wisconsin Madison, at the American Family Children’s Hospital Pediatric Cleft and
Craniofacial Anomalies Clinic through a jointly-approved human studies protocol.

Samples were recorded from 28 native English-speaking children (aged 4–16, 8 females)
with one of the following disorders: repaired sub-mucous cleft palate (SMCP), repaired
unilateral cleft lip and palate, or repaired bilateral cleft lip and palate (CL+P). Some children had syndromic comorbidities commonly associated with cleft palate (e.g. hemifacial microsomia, Kabuki syndrome, Robin sequence). Some children had a secondary surgery, including pharyngeal flap, complete revisions, and fistula repair. See Table 1 for more details regarding speaker characteristics.

Table 1: Speaker characteristics

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Age</th>
<th>Gender</th>
<th>Diagnosis</th>
<th>Secondary repair</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>M</td>
<td>Unilateral CL+P</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>M</td>
<td>SMCP</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>M</td>
<td>SMCP</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>M</td>
<td>CL+P; Hemifacial microsomia</td>
<td>Pharyngeal Flap</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>F</td>
<td>Bilateral CL+P</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>M</td>
<td>Bilateral CL+P</td>
<td>Complete revision</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>M</td>
<td>CP</td>
<td>Sphincter pharyngoplasty</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>M</td>
<td>CL+P</td>
<td>Complete revision</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>M</td>
<td>Unilateral CL+P</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>F</td>
<td>Unilateral CL+P</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>F</td>
<td>CL+P</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>F</td>
<td>SMCP; 22Q deletion</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>16</td>
<td>F</td>
<td>CP; Kabuki Syndrome; DD</td>
<td>Pharyngeal Flap</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>M</td>
<td>Unilateral CL+P</td>
<td>Fistula Repair and Palatal Revision</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>M</td>
<td>SMCP; Kabuki Syndrome</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>M</td>
<td>CP; Robin Sequence</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>16</td>
<td>M</td>
<td>CL+P</td>
<td>Sphincter Pharyngoplasty</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>M</td>
<td>CL+P</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>11</td>
<td>M</td>
<td>CL+P</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>5</td>
<td>M</td>
<td>CL+P</td>
<td>Facial Artery Musculomucosal Flap</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
<td>F</td>
<td>CP</td>
<td>Palatal revision and Pharyngeal flap</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>F</td>
<td>CP</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>M</td>
<td>CP</td>
<td>Furlow Palatoplasty</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
<td>M</td>
<td>CL+P</td>
<td>Palatal repair</td>
</tr>
<tr>
<td>27</td>
<td>9</td>
<td>F</td>
<td>CL+P</td>
<td>Pharyngeal flap</td>
</tr>
<tr>
<td>28</td>
<td>5</td>
<td>M</td>
<td>SMCP</td>
<td></td>
</tr>
</tbody>
</table>
One of two different sets of instrumentation was used to obtain speech samples.
The first set-up used a Sennheiser PC131 headset and the second set-up utilized a WH20
XLR microphone. Samples were collected in quiet clinic rooms. Speakers completed the
MacKay-Kummer SNAP Test-R (SNAP test-R). The SNAP Test-R elicits consonant
vowel (CV) syllable repetitions with /a/ and /i/ vowels (e.g. /pa/ or /pi/), eliciting
sentences loaded with high pressure plosives, nasals, and sibilants, as well as a short
paragraph reading loaded with plosives and sibilants (Kummer, 2005).

**Sample Selection:**

Since the main goal of this study was to determine the most accurate paradigm to
rate hypernasality and ANE, isolated CV syllable repetitions were chosen as stimuli to
reduce the risk of elevated scores due to mis-articulations during connected speech.
Furthermore, two of the 28 children were excluded from stimuli selection, as their
samples had a markedly lower signal to noise ratio than other samples, which could
reduce the accuracy of rating the percepts in question. It was also found that amplifying
samples resulted in distortion of the samples; ultimately, a total of 26 speakers were
utilized. Their samples were peak-amplitude normalized. Fourteen phoneme sets were
available (e.g. /pi/, /ti/, /ʃi/, etc.). The five CV syllables with the highest number of usable
samples across speakers were selected (/pi/, /ki/, /ʃi/, /ti/, /si/). Phonemes with /i/ were
selected, as an independent speech-language pathologist noted instances of hypernasality
and audible nasal emission during the production of some of those phoneme productions.
Since raters were rating nasality, nasal phonemes were not rated.
The use of both sibilants and plosives provided opportunities to capture hypernasality and nasal emission. Any phoneme set in which the speaker did not complete the full six repetitions of the phoneme, or there was measurable background noise were excluded. Ultimately, 9/133 or 6.2% of phoneme sets were excluded from the ratings. Lastly, 20% of samples were repeated, in order to determine intra-rater reliability of ratings for both rating scales. A total of 152 phoneme sets per condition, or a total of 608 stimuli samples across four conditions were rated: DME-hypernasality, DME-ANE, VSR-hypernasality, and VSR-ANE. Since the DME paradigm requires a standard to compare the audio samples against, a modulus had to be selected. In order to determine the modulus, four pilot participants rated samples, via each of the four conditions. The four pilot subject’s scores on the VSR paradigms were averaged to create an average hypernasality and average ANE score for each sample. The mid-point of the range was selected and the audio sample with an average score in the middle of the range was selected as the modulus.

**Participants:**

Twelve participants were recruited (aged 20–28; 3 males). Since this study required careful review of audio samples, participants passed a pure-tone hearing screening at 25 dB hearing level (HL) for pure tones at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz bilaterally. As the purpose of this study was to assess hypernasality and perceptual judgments amongst naïve listeners, any graduate students studying speech-language pathology were excluded from participation to ensure limited exposure to the
percepts. Participants were compensated for their participation and attended two, two hour rating sessions, scheduled within a two week period.

**Procedures:**

**Training:**

Every participant completed a brief, ten-minute training session at the start of each day. To begin, participants read definitions of hypernasality, and audible nasal emission (see Appendix A) adapted from Kummer (2008). As hyponasality is a common resonance disorder accompanying cleft-palate, participants were provided with the definition of hyponasality and asked to disregard instances in their ratings.

Participants also participated in a brief listening training on the computer in which they listened to samples containing varying severity of hypernasality and ANE (see Figure 1). Unused samples from the cleft speakers (e.g. /kɑ/, /tɑ/) were chosen as training stimuli to prevent exposure to the rating stimuli. Nine samples were chosen with varying resonance and ANE severity to represent a range of typical resonance, mild, moderate, and severe instances of hypernasality and ANE. An experienced speech language pathologist confirmed the presence of each percept in the audio samples and determined its severity. Severity of hypernasality or ANE was indicated with each sample presentation for the listener to reference. Listeners were instructed to listen to each sample as many times as they desire, and select the next sample when they felt they understood the provided percept.
Figure 1: Training condition: Listeners were asked to listen to a range of samples including typical, mild, moderate, and severe hypernasality and ANE.

Perceptual Ratings:

Participants rated each audio sample using two different rating systems, on two different voice percepts: hypernasality and nasal emission. Therefore, each participant participated in four conditions: 1: Rating hypernasality using VSR, 2: Rating hypernasality using DME, 3: rating ANE using VSR, 4: rating ANE using DME. To avoid rater confusion between hypernasality and ANE, all participants completed hypernasality ratings on day 1 and ANE on day 2. The order of rating system (VSR and DME) was counterbalanced across participants. Listeners rated the same audio samples across all conditions to determine whether either rating system provided different ratings.

Participants listened to audio samples on a computer using Sennheiser HD 280
pro headphones and provided ratings. For the DME condition (see Figure 2), listeners were instructed to listen to the modulus stimulus first. They were told that the modulus represented 100, and that it represented the middle of the range of ANE or hypernasality. Next, participants listened to each auditory stimulus and assigned a rating based on its comparison with the modulus. This rating was unrestricted so that they could assign any number to a sample to capture their percept of the sample. For example, if they perceived a speech sample as twice that of the modulus, they should have assigned the sample a 200. If they believed it to be half as nasal or contain half as severe ANE, they should have assigned the stimulus a 50. Raters listened to the modulus and sample as many times as they wished.

Figure 2: DME condition: Listeners were asked to rate a sample audio sample in comparison to a modulus set at 100. The DME score was unrestricted.

For the VSR rating system, participants were provided with a set of eight samples
per trial. Sets were created based on pilot ratings using DME for each percept and were established to include typical, mild, moderate, and severe samples. Given a set of 8 sound samples, listeners were instructed to play the audio sample as many times as needed to provide their rating. They were asked to rank the samples from least to most severe (see Figure 3). They moved the audio samples around the screen so that they sorted samples from least severe to most severe cases imaginable for the given percept. Once the samples are placed in order of severity, they moved the audio stimulus along a numbered axis to assign a numeric value on a visual analog scale from 0–100 (see Figure 4). Raters progressed through 19 sets, providing ratings for all 152 samples. The set order (1–19), as well as the order of samples (1–8) within each set was randomized across participants.

Figure 3: The VSR condition: Participants moved each sample, represented by the black dot, around the screen. They were instructed to order the samples from most (top of axis) to least severe (bottom of axis). Participants assigned a numerical value to each sample by its location on the axis.
As listeners moved the samples around the screen, a numerical value populated on the y-axis. Raters were asked to ensure that the numerical value most accurately captured the stimulus.

**Statistics:**

**Inter- and Intra-rater Reliability of DME vs. VSR:**

Each rater randomly scored 20% of samples (30 samples) twice and participants were not alerted to when a sample was repeated. Pearson correlations were calculated to compare their rating on the first iteration of the sample, to the second iteration of the sample for each condition (Baylis et al., 2011; Eadie & Doyle, 2002). The intra-rater reliabilities for each condition were compared to determine which rating scale yielded the highest intra-rater reliabilities for each percept. In order to determine inter-rater reliability, type 2 intra-class correlations were calculated for each condition (Baylis et al., 2011; Whitehill et al., 2002). This resulted in a unique inter-rater reliability measure for
each of the four conditions (e.g. hypernasality using DME). The inter- and intra-rater reliabilities across all four conditions were compared to determine which rating system provided the highest levels of inter- and intra-rater reliability for the two percepts. In order to assess whether raters experienced a ceiling effect for interval rating, the mean difference was calculated for each participant for the VSR scale comparative to the group mean (see Table 3) for both hypernasality and ANE. It was not necessary to assess DME ratings, as there is no ceiling effect because ratings are completely unrestricted.

**Does VSR Supply Valid Estimates of ANE and Hypernasality?**

To compare ratings across rating systems, a single average score across listeners was created for each sample. This overall impression rating was calculated for the DME scale by taking the geometric mean of each listener’s rating for each sample. The geometric mean has historically been used as DME scores are completely unrestricted and occasionally result in large scores (Baylis et al., 2011; Schiavetti, Martin, Haroldson, & Metz, 1994; Stevens, 1975; Whitehill et al., 2002). Since scores from the visual sort and rate were out of 100, the arithmetic mean scores of each sample across raters were taken to provide one overall impression rating of hypernasality or ANE per phoneme set in the VSR rating system. This overall impression rating was calculated for each condition and was used to compare scores from each rating system.

To compare the ratings between the DME and VSR scales for each percept, the sample’s geometric means from DME was plotted against the arithmetic mean ratings in VSR. The line of best fit was calculated. If there was a linear relationship with a positive
slope between DME ratings and VSR ratings, then VSR was most likely acting in the same fashion as DME and could be used as a suitable replacement. This relationship indicates that the variables are metathetic. As stated by Stevens (1975), metathetic qualities are thought to be measurements that are representative of changes in quality. These changes can be partitioned into equal intervals. However, if the line of best fit is best represented by a non-linear relationship, then differences are thought to be considered additive effects, or variation in magnitude. These variables are said to be prothetic variables. When scores are best captured by curvilinear relationships, it is said that interval scaling (e.g. VSR) is not appropriate for use.

**Results:**

*Intra-rater reliability:*

Most raters demonstrated strong intra-rater reliabilities (grater than .40) across all four conditions (Figure 5). Some listeners had poor intra-rater reliabilities for one condition, but did not demonstrate this across all four conditions. For example, as noted in Figure 5, participant 10 (blue triangle) had weak inter-rater reliability for rating hypernasality with DME but a strong intra-rater reliability for detecting hypernasality with VSR. Thus, despite having strong intra-rater consistency for hypernasality with one scale, his/her consistency was not strong across both rating scales. Ratings of hypernasality using DME reflected intra-rater reliabilities ranging between r (120) = .10–.79. When rating hypernasality using VSR, intra-rater reliabilities were comparable ranging between r (120) = .14–.79. Similarly, when rating ANE using DME, Pearson correlations ranged between r (120) = .31–.89. When rating ANE using VSR, Pearson
correlations ranged between \( r(120) = .21 - .83 \).

![Figure 5: Comparison of intra-rater Pearson correlation values for each listener across conditions.](image)

**Inter-rater reliability:**

To compute inter-rater reliability, ratings for each stimulus across listeners were compared using an intraclass correlation ICC(2,k). Overall, inter-rater agreement was moderate to poor across conditions ranging from .14–.48. See Table 2 for inter-class correlations by condition. The greatest inter-rater agreement existed between raters when assessing hypernasality using DME (\( r = .48 \)). Overall, the DME conditions yielded higher intraclass correlations (.48 and .27) than VSR conditions (.14 and .15). The lowest ICC
resulted when rating hypernasality using VSR, indicating that it would not be a suitable rating system for hypernasality. Poor inter-rater reliability was found rating ANE using VSR with an ICC of .15. Rating ANE using DME resulted in a higher ICC of .27.

<table>
<thead>
<tr>
<th>Inter-rater reliability</th>
<th>DME Hyper</th>
<th>VSR Hyper</th>
<th>ANE DME</th>
<th>ANE VSR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.48</td>
<td>.14</td>
<td>.27</td>
<td>.15</td>
</tr>
</tbody>
</table>

Table 2: Inter-rater reliability across conditions using ICC (2)

To determine whether inter-rater reliability was impacted by ratings from those with weak intra-rater reliability on multiple paradigms, inter-rater reliability was re-run removing subjects. Subjects 3, 7, 8, and 12 were observed to have low intra-rater reliability from visual inspection of Figure 5. When removing their ratings from analyses, inter-rater reliability improved for the DME conditions (ICC values of .65 when rating hypernasality and .38 for ANE). However, ICC values remained the same or decreased for the VSR conditions (ICC values of .07 when rating hypernasality and .18 for ANE).

Mean differences, standard deviations, and minimum and maximum ratings were calculated for the VSR scale to investigate whether a ceiling effect restricted results. It was not necessary to assess these values for DME ratings, as ratings are completely unrestricted. As depicted from the mean differences, standard deviations, and minimum and maximum values, listeners used the VSR scale in a similar fashion. Most raters were within 10 points of average ratings (9 of 12 raters). Minimum and maximum ratings indicate that most users utilized the entire range to capture their ratings. Standard deviations listed below represent the variation in the ratings used by each rater. Higher
standard deviations represent increased use of the scale, which would be consistent with capturing scores from mild to severe ratings.

Table 3: The tables below represent the mean differences of ratings across samples for hypernasality (top table) and audible nasal emission (bottom table) for the VSR scale.

<table>
<thead>
<tr>
<th>Hypernasality</th>
<th>Mean Difference</th>
<th>Standard Deviation</th>
<th>Minimum Rating</th>
<th>Maximum Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>6.5</td>
<td>24.1</td>
<td>5.4</td>
<td>95.6</td>
</tr>
<tr>
<td>Participant 2</td>
<td>11.5</td>
<td>19.7</td>
<td>16.0</td>
<td>99.4</td>
</tr>
<tr>
<td>Participant 3</td>
<td>11.2</td>
<td>17.5</td>
<td>10.6</td>
<td>100</td>
</tr>
<tr>
<td>Participant 4</td>
<td>-2.0</td>
<td>28.5</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Participant 5</td>
<td>-1.2</td>
<td>18.8</td>
<td>8.4</td>
<td>96.6</td>
</tr>
<tr>
<td>Participant 6</td>
<td>-9.0</td>
<td>25.8</td>
<td>0</td>
<td>94.7</td>
</tr>
<tr>
<td>Participant 7</td>
<td>-2.5</td>
<td>31.3</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Participant 8</td>
<td>-22.7</td>
<td>30.1</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Participant 9</td>
<td>6.4</td>
<td>25.3</td>
<td>2.2</td>
<td>96.9</td>
</tr>
<tr>
<td>Participant 10</td>
<td>-6.4</td>
<td>23.2</td>
<td>3.1</td>
<td>97.0</td>
</tr>
<tr>
<td>Participant 11</td>
<td>2.8</td>
<td>16.0</td>
<td>17.5</td>
<td>92.8</td>
</tr>
<tr>
<td>Participant 12</td>
<td>5.5</td>
<td>19.6</td>
<td>4.2</td>
<td>95.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANE</th>
<th>Mean Difference</th>
<th>Standard Deviation</th>
<th>Minimum Rating</th>
<th>Maximum Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>-1.1</td>
<td>28.2</td>
<td>0</td>
<td>99.2</td>
</tr>
<tr>
<td>Participant 2</td>
<td>7.6</td>
<td>18.1</td>
<td>15.2</td>
<td>91.0</td>
</tr>
<tr>
<td>Participant 3</td>
<td>3.0</td>
<td>6.1</td>
<td>22.7</td>
<td>70.7</td>
</tr>
<tr>
<td>Participant 4</td>
<td>-3.5</td>
<td>25.2</td>
<td>3.4</td>
<td>100</td>
</tr>
<tr>
<td>Participant 5</td>
<td>-2.7</td>
<td>17.9</td>
<td>10.4</td>
<td>95.1</td>
</tr>
<tr>
<td>Participant 6</td>
<td>2.6</td>
<td>24.9</td>
<td>0</td>
<td>95.4</td>
</tr>
<tr>
<td>Participant 7</td>
<td>-2.8</td>
<td>23.8</td>
<td>3.7</td>
<td>94.9</td>
</tr>
<tr>
<td>Participant 8</td>
<td>-8.4</td>
<td>26.0</td>
<td>0</td>
<td>89.9</td>
</tr>
<tr>
<td>Participant 9</td>
<td>3.1</td>
<td>18.1</td>
<td>5.6</td>
<td>92.9</td>
</tr>
<tr>
<td>Participant 10</td>
<td>-25.0</td>
<td>24.2</td>
<td>0</td>
<td>85.9</td>
</tr>
<tr>
<td>Participant 11</td>
<td>10.0</td>
<td>17.6</td>
<td>19.8</td>
<td>90.4</td>
</tr>
<tr>
<td>Participant 12</td>
<td>17.3</td>
<td>12.8</td>
<td>31.1</td>
<td>90.5</td>
</tr>
</tbody>
</table>
Does VSR Supply Valid Estimates of ANE and Hypernasality?

Average ratings of each sound sample were calculated by averaging all listeners’ ratings of each sample to produce a single perceptual average of each sample. This average rating was computed for all four conditions. To assess whether there was a linear or non-linear relationship between the values, each sample’s average rating using the DME scale was plotted against its average rating in VSR. A line of best fit was calculated to assess the relationship. Figure 6A shows the linear line of best fit of DME rating compared to VSR ratings of hypernasality. The ratings were correlated with an $r^2=.62$. Figure 6B shows the same data using a curvilinear, polynomial model reflecting an $r^2=.64$. Figure 6C and 6D represents the ratings for ANE, plotting DME ratings against VSR ratings. Results show that the curvilinear line (polynomial) of best fit, with an $r^2=.68$, more efficiently captures the data than the linear best fit line, $r^2=.64$. Both percepts were more accurately captured using a polynomial model.
Figure 6: Figures A, B represent average hypernasality VSR scores plotted against average DME scores. Figure A is the linear relationship between these values. Figure B is the polynomial (curvilinear) line of best fit. Figure C and D represent average ANE scores for DME and VSR scales. Figure C is the linear relationship between those values, while figure D is the polynomial (curvilinear) line of best fit.

**Discussion:**

**Validity and Reliability of Scores:**

Intra-rater reliabilities ranged between $r=0.10–0.89$, indicating a large range of reliability within participants across conditions. As seen in Figure 5, Pearson correlations were variable across listener and condition, with no participant demonstrating poor correlations on all four conditions. Two participants showed negligible Pearson correlations for both scales when rating hypernasality. However, their intra-rater
reliabilities were moderate to strong when rating ANE. Thirty-eight of 48 intra-rater reliabilities were strong (correlations greater than .4). Strong intra-rater reliability suggests that individuals were attentive when rating speech samples, as they were likely to score repeated samples with high consistency. Overall, though intra-rater reliabilities varied, most participants demonstrated strong intra-rater reliabilities.

Regarding inter-rater reliability, listeners demonstrated moderate reliability in the DME hypernasality condition (r=.48), but negligible to weak reliability across other conditions (r=.14–.27), indicating little agreement between listeners. Variability in inter-rater reliability is common amongst perceptual listening studies when rating nasality and nasal emission (John et al., 2006; Lee et al., 2009). In a study by Lee et al. (2009), speech-pathology students had weak to moderate inter-judge reliability, even with training (ranging from .29–.42). In regards to audible nasal emission, inter-rater reliabilities (Table 2) were negligible to weak regardless of the rating system used. This percept was difficult for raters to discern, as evidenced by low inter-rater reliability. Poor inter-rater reliability suggests that VSR would not be a suitable replacement for DME ratings. Poorest inter-rater reliabilities were found when raters were using the VSR scale.

In order to further investigate the VSR scale, average ratings were compared to determine whether raters utilized the same portions of the scale. As depicted by Table 3, most of the raters utilized the VSR scale similarly, providing ratings within 10 points of the average. Ratings across samples and across listeners were around 50, which were expected considering representations of samples of various severities (53.0 capturing
hypernasality, and 50.2 capturing ANE). Standard deviations in this case, represent the distance of ratings from the mean. So in this case, higher standard deviations, represent that the listener used a large range surrounding the mean ratings to capture extreme ratings. In the two cases in which the mean difference was greater than 20, the two raters had negative mean differences, indicating that their average ratings were lower than the remainder of the cohort. These findings support that raters were not constricted in their rating by a ceiling effect.

Strong intra-rater reliabilities suggest strong internal consistency. However, weak inter-rater values suggest inconsistency between raters. Perhaps internal representations of these perceptions differed across raters. Raters may have utilized other aspects of the speech samples than hypernasality and ANE. Or, perhaps raters were using other aspects of recording (e.g. voice quality or misarticulations) to determine severity. This suggests that further training may raise inter-rater agreement.

**Benefits of a Training Session:**

Previous studies have supported the use of naïve listeners, lending support to using naïve listeners in the current study. Findings by Brunnegard, Lohmander, and van Doorn (2009) suggested that untrained listeners rated hypernasality as well as trained speech-language pathologists, using a five point rating scale. However, naïve listeners were less sensitive in detecting nasal emission with expert listeners yielding higher inter-rater reliability of ratings. Therefore, perhaps raters in the current study were not as adept when rating ANE. However, naïve listeners in the current study achieved mostly strong
intra-rater reliabilities, suggesting success of the training. As all listeners received a training session, exact benefits of training are unknown.

The current training was a passive listening task. Previous research by Lee, et al. (2009), did not find significant differences between the use of a training session with and without contingent feedback. Training resulted in improved reliability of hypernasality ratings regardless of the type of training. However, considering the low inter-rater reliabilities, perhaps participants were not rating percepts in the same manner. A training with contingent feedback of severity of training stimuli may help increase inter-rater performance.

**ANE and Hypernasality as Prothetic Variables:**

Plotting DME ratings against VSR ratings resulted in a curvilinear relationship, which may indicate that nasality and nasal emission are prothetic variables. This finding is consistent with previous findings of nasality as a prothetic variable (Baylis et al., 2011; Zraick & Liss, 2000). Caution is warranted, as the polynomial model only improved the model slightly (from an \(r^2\) of .61 to .64). Additional samples on the severe end should be added to provide further information regarding the classification of these percepts.

Nasal emission was best represented by a curvilinear relationship, consistent with findings by Baylis et al. (2011). This suggests these variables may be best captured using magnitude estimation, rather than interval scaling, as magnitude scaling is more appropriate when scoring prothetic variables (Stevens, 1975).
Use of VSR as a Novel Scale:

Direct magnitude estimation yielded higher inter-rater reliability than VSR in both conditions. Visual sort and rate conditions yielded the lowest inter-rater reliabilities, regardless of the percept being scored. Therefore, VSR may not be a suitable scale to measure these percepts. A curvilinear relationship was found between ratings on DME and VRS, which suggests they are best captured using ratio scaling (such as DME). Direct Magnitude Estimation has historically been used to measure prothetic variables. Therefore, DME should continue to be used when rating hypernasality.

In regards to ANE, low inter-rater reliability regardless of rating scale used suggests difficulty providing ANE ratings. Perhaps rating nasal emission on a numerical scale proved too challenging. Clinically, nasal emission is typically rated as present or absent, described by the frequency of nasal emission (consistent, intermittent), and whether they are audible, visible, or turbulent in nature (McWilliams & Philips, 1979). John et al. (2006) rated audible nasal emission and nasal turbulence as absent, occasional or frequent, rather then on a numerical scale. In the John et al. study, seven expert speech language pathologists had difficulty providing reliable ratings of ANE, as evidenced by low inter-rater reliability (inter-rater reliability range 16–64%) when rating nasal emission and turbulence using an EAI measure, the Cleft Audit Protocol for Speech-Augmented (CAPS-A). As expert speech-pathologists had difficulty providing reliable ratings with one another, it is not surprising that naïve listeners in the current study had such low reliabilities. Therefore, perhaps rating ANE perceptions in a continuous manner is not ideal. More research is needed regarding a more appropriate scale for rating ANE.
**Limitations:**

It was assumed that rating hypernasality in isolated syllables would be more effective, perhaps more naturalistic speech is needed to provide accurate nasality ratings. Findings by Brunnegård et al. (2012) indicated higher inter-rater reliabilities when rating hypernasality across naïve listeners (46–57%) than the current study (48%). Their stimuli consisted of connected speech rather than isolated syllable strings. Therefore, future studies should investigate the use of connected speech over syllable phoneme repetitions as stimuli samples.

In regards to stimuli, samples were collected from children as young as four. Therefore, the quality of samples was not consistent. Although samples with overt noise in the background were excluded, children may not have delivered the syllable sets reliably. Similarly, children with articulatory or phonological processes were not excluded. As shown in Paliobei et al. (2005), the presence of articulation errors in samples resulted in higher severity ratings of hypernasality. Therefore, further research is warranted to separate perceptions of disordered speech production from nasality and nasal air emission.

**Conclusions:**

Overall, intra-rater reliability was moderate to strong amongst most participants across conditions, with negligible/weak inter-rater reliabilities when using VSR. This indicates that raters were reliable at rating these percepts but may have different internal representations of them. Rating hypernasality using the DME scale resulted in the highest
inter-rater reliability, which supports current literature in the field. Therefore, the use of VSR as a novel rating scale for these percepts is not supported by current findings, as VSR yielded lower inter-rater reliability. A curvilinear relationship was observed between ratings of DME and VSR for both hypernasality and audible nasal emission. This suggests that these variables may be prothetic, and better captured by magnitude estimation (Stevens, 1975). Further research is needed with additional samples in the severe range. Overall, ANE was difficult to characterize and additional work is needed to investigate the best way to categorize nasal emission.
Appendix A: Training Materials

For the purposes of this study, you will be rating hypernasality and audible nasal emission. Both of these sound qualities can be very subtle and hard to hear. Some of the perceptions you will rate today have been described as “static” noise in the background. However, all of the samples are high quality, and you should assume any deviations may be potential instances of nasal emission, or hypernasality.

Hypernasality

Hypernasality is a speech disorder that occurs when sounds that are supposed to be produced only in the mouth/oral cavity are also transmitted through the nose/nasal cavity. In this listening task, you will hear speech sounds that are all typically produced orally, without any sound from the nose. Therefore, if you hear a sound that appears to be coming from the nasal cavity, please consider it hypernasal. Then, please rate how much of that quality you hear in the sample using the rating scale/task provided.

Audible Nasal Emission:

Nasal emission (ANE) can be described as the perception of “extra noise” in the speech signal. The noise associated with ANE is air escaping through the nasal passageway. Audible nasal emission are typically perceived in these speakers during the production of consonants that require high air pressure in the mouth (such as /pa/,/sha/) and often results in a high pitched whistle or large amount of air, due to air escaping out of the nose. Nasal emission can be very loud and distracting, or very soft and barely audible.

For the purposes of this study, please do not consider de-nasalized speech (also called hyponasality). Hyponasality can often be confused with hypernasality, as both often have been described as “muffled speech” and both often results in less intelligible speech.

For your reference the definition is included below.

PLEASE IGNORE ANY HYPONASALITY THAT YOU HEAR. DO YOUR BEST TO ONLY JUDGE THE ATTRIBUTE (HYPERNASALITY OR ANE) THAT YOU ARE ASKED TO JUDGE.

Hyponasality:

Hyponasality is a resonance disorder when there is a reduction in the amount of sound coming from the nose when typically needed. Sounds such as /m/ /n/ /ng/- as in “ring” always include sounds from the nose to produce those sounds. Place your finger on your nostril, and say the word “man.” Notice how you can feel your nose vibrating when you
say it. Those individuals who have hyponasality have a reduction in sounds from the nose. Hyponasality is often described as if they person sounds as if “they are stuffed up” or “that they have a cold.”

BIBLIOGRAPHY


Cleft Palate-Craniofacial Journal, 28(2), 184–188; discussion 188–189.

doi:10.1016/j.jvoice.2005.08.007


VITA

Kerri Downing was born in 1988, in Long Beach, NY. She graduated from the University of Connecticut (UCONN) in May of 2010 with a Bachelor of the Arts (B.A) in Psychology. She subsequently worked on a collaborative longitudinal research study at Boston University and Boston Children’s Hospital, investigating the development of infants with a sibling with an autism spectrum disorder (ASD) and specific Language Impairment (SLI). She is expected to graduate from Boston University with a Master of Science (M.S) in Speech Language Pathology in 2015.