A Comparison of Equal-Appearing Interval Scaling and Direct Magnitude Estimation of Nasal Voice Quality

Richard I. Zraick
University of Arkansas for Medical Sciences
Little Rock, AR

Julie M. Liss
Arizona State University
Tempe, AZ

Listeners rated the nasality of synthesized vowels using two psychophysical scaling methods (equal-appearing interval scaling and direct magnitude estimation). A curvilinear relationship between equal-appearing interval ratings and direct magnitude estimations of nasality indicated that nasality is a prothetic rather than metathetic dimension. It also was shown that the use of direct magnitude estimation results in nasality ratings that are more consistent and reliable. The results of this experiment are discussed in relation to other studies that have examined the validity and reliability of equal-appearing interval scaling of voice quality. Additionally, there is a discussion of methodological issues for future research and the implications of the findings for clinical and research purposes.

KEY WORDS: voice, perception, nasality, scaling

Most protocols for evaluating pathological voices include perceptual assessment of quality (American Speech-Language-Hearing Association, 1993). Various instruments have been proposed for perceptual evaluation (e.g., Hirano, 1981; Wilson, 1977; Wirz & Beck, 1995). However, none of these protocols has been widely accepted because of concerns about the validity of the scales used and about the reliability with which listeners can rate voices (Kent, 1996; Kreiman & Gerratt, 1996).

Concerns about scale validity and rater reliability are well founded. Investigators have shown that perception of many variants of abnormal voice/speech quality (e.g., roughness, breathiness, naturalness) is multidimensional; that is, a listener's perception of quality involves integrative judgments of more than one dimension (Kempster, Kistler, & Hillenbrand, 1991; Kreiman, Gerratt, & Berke, 1994). It also appears that the salience of these dimensions to the listener may be dependent upon a number of factors, including the acoustic properties of the stimulus, context effects, and listener experience and bias, all of which may affect agreement and reliability of ratings by individuals and groups (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). As Kreiman et al. (1993) state, "…it is unclear which of the many scales, procedures, and statistics that have appeared in the literature are best suited to measuring voice quality and evaluating the reliability of such measurements" (p. 21).

In their tutorial on the perceptual evaluation of voice quality, Kreiman et al. (1993) reviewed the study design of 57 relevant studies published between the years 1951 and1990 and reported that all but 10 (nearly 83%) used equal-appearing interval (EAI) scaling; additionally, they reported that 18 of the subset of 47 studies employed a 7-point scale and that 16 of the studies employed a 5-point scale. One common measure of rater reliability when using EAI scales is the number of ratings that fall within plus-or-minus one scale point, so the greater the range of points along the continuum, the less likely it will be that agreement will occur due to chance probability (Dunn-Rankin, 1983).

The continued prominence of EAI scaling in the measurement of perception is puzzling considering the robust challenges to the validity of its use. There is a convincing body of experimental psychology literature from the past 25 years which has demonstrated that the perception of some sensory dimensions is poorly represented by use of an EAI scale (see Stevens, 1975, for an overview). In particular, Stevens (1974) has demonstrated that respondents tend to exhibit a systematic bias toward subdividing the lower end of the continuum into smaller intervals when attempting to partition certain dimensions into equal intervals; that is, they do not perceive intervals as equal at different locations on the scale. Furthermore, Stevens (1974) has shown that the prescribed nature of an EAI scale may not capture a respondent's full range of perception; that is, the scale may limit the response given. It also has been shown that, in an EAI task, respondents tend to assign stimuli to categories so that all categories are used equally as often (Gescheider, 1976).

Given the limitations of EAI scaling described above, an alternative scaling method known as *direct magnitude estimation* (DME) has been proposed (Stevens & Galanter, 1957). First used in the study of human communication disorders by Cullinan, Prather, and Williams (1963) and Martin (1965), DME has steadily gained more widespread acceptance in our field (Coleman, 1971; Emanuel & Smith, 1974; Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Heiberger & Horii, 1982; Kreiman et al., 1993; Metz, Schiavetti, & Sacco, 1990; Schiavetti, Metz, & Sitler, 1981; Schiavetti, Sacco, Metz, & Sitler, 1983; Southwood, 1996; Toner & Emanuel, 1989). In DME, listeners scale individual speech samples relative to each other or to a standard stimulus (a.k.a. modulus), which is usually obtained from the middle of the range of stimuli. This modulus is typically assigned a numerical value (for example, 100). Following presentation of the modulus, listeners assign a numerical value to each subsequent token that is relative to the modulus. If they believed that a subsequent token exhibited twice as much of the relative dimension to be rated (for example, nasality), then a value of 200 would be assigned to that token. Likewise, if they believed that the token was one half as nasal, then a value of 50 would be assigned. Advantages of DME over EAI scaling are that DME does not assume a linear partition of the continuum (Schiavetti et al., 1981), and DME is not bound by fixed minimum/maximum values (Stevens, 1975). Thus, there are no constraints on the scale (Southwood, 1996).

Central to the preceding review of EAI and DME scaling is an understanding of the nature of the dimensions to be scaled. Stevens (1975) has described two classes of dimensions to be scaled, metathetic and prothetic. A metathetic dimension is one that varies in terms of a change in quality and is sometimes described as substitutive. Pitch is often cited as an example of a metathetically scaled dimension. For example, as pitch increases, one perceives a change in quality, rather than quantity, of the stimulus (Stevens, 1975). A prothetic dimension, on the other hand, is one that does vary in terms of a change in degrees of quantity or magnitude, and is therefore sometimes described as additive. Loudness is an example of a prothetically scaled dimension. Stevens (1974) has shown that a prothetic continuum is not amenable to linear partitioning, that is, EAI scaling. For example, when listeners try to partition loudness judgments into equal intervals, there is a systematic bias to partition the lower end of the continuum into smaller intervals, resulting in a continuum that is unequal by nature. With a metathetic dimension, however, Stevens (1974) has shown that listeners are able to divide the continuum into equal intervals. That is, the listeners' naturally occurring perceptual intervals are equal.

Stevens (1975) has outlined a procedure that compares EAI and DME scale judgments to determine whether a dimension falls along a metathetic or prothetic continuum. First, listeners judge a set of stimuli along the dimension of interest (for example, nasality), using EAI and DME scaling. Second, the mean EAI scale scores are plotted against the mean DME scores, and the nature of the relationship of these two sets of scores is examined. A linear relationship indicates that the listener assigned equal perceptual space to the intervals on the EAI scale, suggesting a metathetic continuum. A nonlinear relationship suggests a prothetic continuum, for which DME would be the appropriate scaling method.

Using the methodology outlined by Stevens (1975) and others (see, e.g., Barry & Kidd, 1981), investigators have shown that many of the perceptual dimensions commonly scaled in the speech and voice arena are prothetic rather than metathetic. Schiavetti et al. (1981) had 20 listeners scale the intelligibility of 20 speaker with hearing impairment whose speech ranged in severity from mildly to severely unintelligible across a

hierarchy of speech tasks. Their examination of the plot of the means obtained via a 7-point EAI scale and DME with a modulus of 10 revealed a nonlinear relationship, suggesting that their dimension of interest was prothetic, and thus most validly measured by DME. In another similar study, Schiavetti et al. (1983) had 15 listeners scale the stuttering severity of 20 speakers varying in degree of fluency. In addition to comparing a 7-point EAI scale to a DME with a modulus of 10, these investigators also plotted the means obtained via an additional response condition of DME without a given modulus (i.e., the modulus was self-generated by each listener). Each of the DME conditions yielded scale values that were related to the EAI scale values in the nonlinear fashion indicative of prothetic continua, supporting the use of DME to measure listeners' perceptions of this dimension. Speech naturalness has been investigated in the speech of persons with amyotrophic lateral sclerosis (Southwood, 1996; Southwood & Weismer, 1993). In Southwood and Weismer (1993), the relationship between speech intelligibility and the dimensions of bizarreness, acceptability, naturalness, and normalcy was investigated using DME without a given modulus. Five listeners scaled connected speech produced by two different speaker groups (normal and dysarthric). It was reported that the four dimensions were highly correlated with each other and with speech intelligibility; that is, there was a great deal of shared variance between all the dimensions, particularly as intelligibility increased, suggesting perceptual fusion of the presumably separate dimensions by some listeners. One explanation for this offered by the authors was the possibility that the dimensions may not have been prothetic, thus leading to Southwood (1996), which used Stevens' (1975) methodology to examine the dimensions of speech naturalness and bizarreness. Twelve listeners scaled these dimensions using DME without a given modulus and a 7-point EAI scale, and both dimensions were conservatively determined to be prothetic, though some response bias was thought to have occurred because of the use of self-generated moduli. Lastly, in an investigation that is more in line with the current study, Toner and Emanuel (1989) examined the dimension of roughness in 10 speakers who produced sustained vowels varying in degree of roughness. Twenty listeners performed both DME with a modulus of 100 and 5-point EAI scaling, and the relationship among the plotted means was found to be nonlinear, suggesting that listeners' perceptions of this particular vowel quality is most validly measured using DME. Taken together, results from these studies strongly suggest that other perceptual phenomena related to speech and voice may be prothetic.

To date, there have been no published studies that have examined the construct validity of DME versus EAI scaling in measuring listeners' perception of nasality. Such an investigation may help to determine which scaling method is most appropriate for this particular perceptual dimension. Inappropriate use of a particular scaling method has serious ramifications, most notably, the potential for misclassification of persons for research and/or clinical purposes. Severity of impairment is a typical descriptor for an individual or population of persons, and it is commonly used as an inclusion criterion. If the measure of severity is based, in whole or part, on subjective ratings (expressed via scaling), validity of the method used must be established. Also, if perceived change is an outcome measure, then the representation of that perception must be accurate.

Nasality is a quality judgment of interest for a number of reasons. First is its place in the clinical evaluation of the speech of a variety of patients. Nasal quality may result from either congenital or acquired disorders and may be noted in both children and adults. As such, clinicians are routinely called upon to make perceptual judgments of nasality in those patients who present for evaluation of velopharyngeal sufficiency for speech. A second point of interest is its place in clinical treatment across disorders. The influence of abnormal nasality on the overall intelligibility of speech is well documented (Griffiths & Bough, 1989; McWilliams, Morris, & Shelton, 1984; Seikel, Wilcox, & Davis, 1990; Yorkston & Beukelman, 1981), as is the deleterious effect decreased speech intelligibility has on speakers' attitudes and communication strategies (Berry, Evans, & Lane, 1990; Yorkston, Bombardier, & Hammen, 1994). In routine clinical practice, the primary treatment goal of improved verbal communication is sometimes achieved via direct interventions aimed at decreasing nasal quality and, by extension, increasing speech intelligibility.

Voice synthesis has been employed to study the acoustic correlates of voice qualities such as breathiness (Klatt & Klatt, 1990) and steadiness (Rozyspal & Millar, 1979), and it has been employed to study the acoustic correlates of non-vowel articulatory features as well (Brend, 1975; Shadle, 1987). Martin, Fitch, and Wolfe (1995) used synthetic stimuli created to represent three different voice qualities (breathy, rough, hoarse) to train student clinicians to recognize these qualities in natural voices. Listeners were asked to classify the voice quality and rate its severity using a 7-point EAI scale, with high agreement and reliability reported. Gerratt et al. (1993) used synthesized rough vowels as the stimuli in their study comparing internal and external standards on voice quality ratings. They used the Klatt formant synthesizer (Klatt, 1980; Klatt & Klatt, 1990) to synthesize 22 variations of /a/ along a continuum of severity. Five of these tokens were chosen from the continuum to serve as anchors on their anchored rating task. Criteria for choosing these five tokens were that (a) they were discriminated with 100% accuracy in pilot tests;

(b) they spanned the entire range of roughness represented by the synthetic stimuli; and (c) they were approximately perceptually equidistant, as judged by the authors.

The traditional rationale for using vowel segments in studies of voice quality is that a listener's impression of voice quality in connected speech may be affected by non-voice factors. These include a speaker's dialect, speaking rate, intonation, and idiosyncratic articulatory behavior (Fritzell, Hammarberg, Gauffin, Karlsson, & Sundberg, 1986; Hollien, Michael, & Dougherty, 1973; Murry & Dougherty, 1980). On the other hand, it has been argued that connected speech segments are better suited for perceptual evaluation because connected speech is the more typical voice behavior and would also allow for more detailed description of deviant voice quality characteristics (Hammarberg, Fritzell, & Schiratzki, 1984). De Krom (1994) conducted a perception experiment in which listeners were asked to rate voice segments obtained from a variety of speakers on breathiness and roughness. Four different types of stimuli were presented to each listener: (1) a connected speech fragment, (2) a 200-ms vowel onset, (3) a 1,000-ms steady vowel segment, and (4) a 200-ms vowel postonset. It was reported that stimulus type had no significant effect on rater agreement, though reliability was reportedly higher for the whole vowel segment, leading to the author's conclusion that the more cumbersome and difficult acoustic analyses of connected speech is not warranted. In a related study, deKrom (1995) investigated the perceptually relevant acoustic correlates of breathiness and roughness and whether these correlates were different across speech segment. For the same stimuli in his previous study, the author reported that there was no segment effect for breathiness, but a significant segment effect for roughness was reported. That is, the percentage of variance in breathiness ratings that could be accounted for by the correlated acoustic parameters was essentially the same for all segments, but for roughness ratings, the most variance was reported for the vowel onset and whole vowel segments. This was interpreted as evidence that breathiness is a salient perceptual cue regardless of segment type (connected speech versus any of the vowel subsegments) and that roughness is more salient for the vowel segments than the connected speech segments.

The use of synthesized stimuli in this investigation allows for controlled modification of those parameters corresponding to the known acoustic correlates of the voice quality of interest. It is important that there be as much control over the characteristics of the stimuli as possible. Therefore, any changes in the dependent variables of interest can be best explained by the systematic manipulation of the operationally defined independent variables. Although the generalizability of results to other voice populations will be limited, this is a first study that is exploratory in nature. At this point in time, there is no empirical basis for examining natural human voices, and it is highly unlikely that such an examination would yield meaningful and readily interpretable results—thus, the historical trend (see Gerratt et al., 1993) of using synthesized vowels to establish a foundation and direction for the exploration of perception of human voice.

The purpose of this study, then, is to investigate the validity of EAI scaling of nasality, in an effort to determine whether this quality falls along a metathetic or prothetic continuum. The null hypothesis is that nasality is a metathetic dimension. Rejection of the null hypothesis and acceptance of an alternative hypothesis would suggest that EAI scaling of nasality is inappropriate and that the nature of the multiple percepts underlying this judgment warrants further investigation.

Method

Synthetic Vowel Stimuli

All vowel stimuli were synthesized using the Klatt formant synthesizer, KLSYN88 (Klatt & Klatt, 1990), in cascade mode. Initially, a base stimulus was created (the vowel /i/, sustained for 1.5 seconds, with a fixed amplitude throughout). A 1.5-second stimulus duration was chosen to give the listener an opportunity to attune auditorily to the stimulus. This base stimulus was then synthesized at five fundamental frequencies (80 Hz, 120 Hz, 180 Hz, 220 Hz, and 300 Hz), chosen because of the approximation of the range of speaking fundamental frequencies observed in male and female adults and children (Case, 1996). The resulting five synthesized vowels were designated collectively as *oral vowels*. For each oral vowel, four different nasalized cohorts were then synthesized. These nasalized cohorts simulated a continuum of perceived nasality, from mildly to severely nasal. These four additional sets of five stimuli were designated collectively as Nasal Vowels A, B, C, and D, respectively. Synthesis parameters for the nasal vowels are reported in Table 1.

There is evidence from the literature on speech synthesis that the spectral relationship between the frequency and bandwidth of the nasal poles and zeros, and the frequency and bandwidth of F1–F3, form the acoustic basis for the perception of nasalization (Beddor, 1993; Beddor & Hawkins, 1990; Chen, 1995; Hawkins & Stevens, 1985; Huffman, 1990; Maeda, 1982). Collectively, these studies provide the basis for the synthesis of the vowel stimuli used in the current investigation.

Hawkins and Stevens (1985) synthesized five vowels /i, e, a, o, u/, each along an oral-nasal continuum,

Table 1. Major KLSYN88 (Klatt & Klatt, 1990) synthesis control parameters and values (in hertz) for the vowel (/i/).

Parameter	Oral	Nasal A	Nasal B	Nasal C	Nasal D
F1	270	300	275	250	225
B1	60	300	250	200	150
F2	1500	2100	2250	2400	2550
FNP	500	700	800	900	1000
FNZ	500	1200	1300	1400	1500
BNP	90	150	150	150	150
BNZ	90	250	250	250	250

Note. Duration of all vowels = 1,500 ms. Non-listed control parameters at default values. F1 = frequency of the first formant. B1 = bandwidth of the first formant. F2 = frequency of the second formant. FNP = frequency of the nasal pole. FNZ = frequency of the nasal zero. FNZ = frequency of the nasal zero. BNP = bandwidth of the nasal pole. BNZ = bandwidth of the nasal zero.

and presented these stimuli to a group of naive listeners in a series of identification and discrimination experiments, which ultimately revealed that an oral-nasal distinction for each of these vowels could be synthesized. Each synthetic nasal vowel differed from its oral counterpart in either one or two ways: all nasal vowels contained an additional pole-zero pair, and in some cases the frequency of F1 differed from that of the oral counterpart. General characteristics of their nasal vowels were that the first formant was shifted to a higher frequency (relative to its oral starting point), and the nasal zero was about midway between this shifted F1 and the nasal pole. The resultant vowel spectra were marked by a broad low-frequency prominence (i.e., nasal formant). Intermediate stimuli on the continuum were synthesized by interpolating in equal steps between values of F1, FNZ, and FNP for the oral and nasal extremes.

Chen (1995) analyzed the nasal vowels produced by normal speakers and hearing-impaired speakers and, by spectral matching of the fundamental frequency contour and the first five formant frequencies and bandwidths of these speakers, was able to synthesize nasal stimuli for all the major vowels. As is the case in Hawkins and Stevens (1985), the nasal vowel was marked by a broad low-frequency nasal formant and also a widening of the first formant bandwidth, which is consistent with predictions from acoustic theory (Fant, 1960).

Huffman (1990) used an articulatory synthesizer to generate vowels of different heights and coupling sizes and elicited nasality judgments from listeners. He found that, although high vowels required more coupling than high vowels to be labeled as nasal, such differences could be explained in terms of the effects of coupling on F1 intensity. Listeners' judgments correlated with the magnitude of F1 amplitude reduction and bandwidth increase. Similar results were reported by Maeda (1982), who concluded that the vowel-independent acoustic correlate of vowel nasalization was low-frequency spectral flattening and or spreading.

Pilot Studies

Two pilot studies were undertaken to establish the validity of the vowel stimuli. The purpose of the first study was to confirm that the descriptive labels attributed to the stimuli ("oral" vs. "nasal") were appropriate. The purpose of the second study was to confirm that the stimuli fell into five discrete categories along a continuum of perceived nasality (from oral/non-nasal to severely nasal).

Subjects in both pilot studies were 4 graduate-student clinicians majoring in speech-language pathology. Subjects were either enrolled in a graduate-level course in voice disorders, or had recently successfully completed such a course. These listeners had no reported history of any hearing, speech, voice, or language difficulties and were screened for the ability to detect pure tones bilaterally at 20 dB HL at octave frequencies from 500 Hz to 8 kHz. The same 4 subjects participated in both pilot studies.

For the first pilot study, a listening tape was constructed consisting of a subset of the original vowels described above, as well as additional vowels synthesized to represent the quality of roughness. These latter stimuli were included in the stimulus set because it was necessary to demonstrate that the non-oral stimuli could be identified as nasal, and not some other abnormal quality. Rough vowels, in particular, were included because (a) rough quality was a judgment that listeners were familiar with and (b) they were readily synthesized (see Gerratt et al., 1993). All stimuli were equated for fundamental frequency (120 Hz), duration (1.5 seconds) and intensity (70 dB), so as to control for extraneous factors that potentially may have distracted listeners from their task of focusing on vowel quality. Stimuli were presented binaurally via headphones. Presentation order was randomized, and there was a 5.0-second interval of silence between each vowel, which was presented two times not consecutively to allow for assessment of reliability of ratings. After being presented with a written description of each voice quality, and hearing multiple exemplars of each prior to the task, subjects were instructed to identify the general voice quality of each vowel, circling one of three choices (oral, nasal, rough) on a response sheet. Mean accuracy for this task was 96% (range = 90–100%); interrater reliability was .95; and intrarater reliability was .98, indicating that there was nearly perfect identification and high rater agreement and reliability.

For the second pilot study, a listening tape was constructed, which consisted of pairs of vowels covering all

possible combinations of nasal severity at one given fundamental frequency (120 Hz) and intensity (70 dB). This resulted in 10 pairs of forward-ordered vowels; to control for possible order effects, a reverse ordering of these combinations was also undertaken, yielding 10 additional pairs. The presentation order of the resultant 20 vowel pairs was randomized, and four pairs were repeated to allow for assessment of reliability. The interval of silence within each pair of voices was 1.0 seconds, and each pair of voices was separated by 5.0 seconds. After training to the task, subjects were instructed to judge whether the voices heard in each pair were "same or different" and to record their judgments on a response sheet. Mean accuracy for this task was 91.25% (range = 80–100%); interrater reliability was .89; and intrarater reliability was .85. Qualitative examination of individual errors for this task revealed that no specific pair was missed by more than one listener and that the few errors involved pairs of voices that were within one step on the presumed continuum of nasality. Because of this minimal ambiguity, it was assumed that the continuum of nasality had perceptual reality to listeners.

Stimulus Tapes

Upon completion of the pilot studies, the 25 vowel stimuli were dubbed onto a digital audiotape. Presentation order was randomized, and there was a 5.0 second interval between stimuli. Each vowel stimulus was presented two times to allow for assessment of intrarater reliability of ratings.

Subjects

Twelve graduate student-clinicians participated in the listening tasks. All students were majoring in speech-language pathology and were enrolled in a graduate-level course in voice disorders, or had recently successfully completed such a course. Listeners had no reported history of any hearing, speech, voice, or language difficulties and were screened for the ability to detect pure tones bilaterally at 20 dB HL at octave frequencies from 500 Hz to 8 kHz. None of the 12 listeners had participated in the pilot studies just described.

Procedure

Listeners were informed that the experimenters were interested in the perception of nasality, and that two different rating methods would be compared over two sessions, to be held 24 hours apart. Order of task presentation was randomized across listeners. Since the listeners had prior exposure to synthesized speech and the vowel stimuli used in this study from recent participation in another study, familiarization with the nature of the stimuli was not necessary.

To obtain severity ratings using the EAI scaling method, listeners were instructed to rate individually the perceived nasality of the vowel stimuli on a 5-point EAI scale, with a rating of 1 indicating least nasal and a rating of 5 indicating most nasal. Additionally, they were instructed to not rate the nasality between scale points.

A 5-point EAI scale was used because of its correspondence with the 5 discreet points along the continuum of nasality. Anchor stimuli determined by the investigators to be representative of the 5 scale points were played prior to the listening task, and after presentation of every 10 stimuli. This was done in an effort to control two major sources of rating error (1) differences among listeners in their internal standards for different quality judgments and (2) context-related variability. Gerratt et al. (1993) have demonstrated that the use of fixed, external referents increases intra- and interrater reliability of ratings and decreases variability in responses due to context effects.

To obtain severity ratings using the DME method, listeners were instructed to rate individually the perceived nasality of the stimuli relative to a standard stimulus. This standard stimulus was chosen from the middle of the continuum of stimuli (corresponding to Point 3 on the 5-point EAI scale of nasality). The standard stimulus was played for listeners, and they were informed that it was arbitrarily assigned a scale value of 100 by the experimenters, and they were informed that this standard stimulus would henceforth be referred to as a "modulus." Listeners were then instructed to assign a value to the subsequent stimuli relative to the modulus. Per Gerratt et al. (1993), the modulus was reintroduced after every 10 stimuli to prevent difficulty recalling the modulus and causing a shift in the listeners' internal standard for the judgment under question. Prior to rating the voices, listeners were trained to the task with nasal stimuli not from the experimental set. These were vowels that differed from the experimental stimuli only in fundamental frequency.

Data Analysis

The mean of the EAI nasality ratings (over listeners) for each vowel stimuli was compared to the mean of the DME nasality ratings (over listeners) for each vowel stimuli. The linearity of the relationship between the means of the EAI ratings and the DME ratings was estimated by simple linear regression analysis. To determine if the relationship between the two sets of ratings could better be described by a curve or a straight line, higher order polynomials were fit to the rating data until two consecutive nonsignificant improvements in fit were

obtained. The degree of polynomial that resulted in the least significant improvement in fit was considered to be the most appropriate model.

Cronbach's coefficient alpha (Cronbach, 1971) is a measure of the internal consistency of the stimulus items in a set, and analyses of internal consistency seek to determine the degree to which the items are interrelated (Brown, 1983). If the scores (in this case, ratings) on the various items comprising a set intercorrelate positively high, the set is considered homogenous (DuBois, 1970). Coefficient alpha was calculated for each scaling method (EAI vs. DME) and used to determine which method most consistently represented listeners' judgments of the nasality of the vowel stimuli.

Interrater reliability for each scaling method was assessed using the intraclass correlation coefficient (ICC; Bartko, 1966; Ebel, 1951). The ICC reflects the overall coherence of an entire group of listeners, and is an appropriate statistic for assessing reliability between raters (Shrout & Fleiss, 1979).

Results

Table 2 presents the results of fitting a second-degree polynomial to the DME and EAI ratings for the synthesized /i/ vowel. Specifically, this table shows the analysis of variance table for EAI nasality ratings regressed on DME nasality ratings. From this table, note the statistically significant F ratio ($p < .01$), which indicates that a curvilinear model accounted for a statistically significantly amount of the variance present, above and beyond that accounted for by a simple linear model. Also note the line of best fit, which allows one to predict DME ratings from EAI ratings.

Table 3 presents Cronbach's coefficient alpha (Cronbach, 1971) and the interclass correlation (ICC) coefficients for both scaling methods, for the entire set of vowel stimuli and by subsets of stimuli based on nasality level.

Examination of this table reveals that the internal consistency of the vowel stimuli as determined by the DME scaling method was extremely good (>.9), whereas that of the EAI scaling method was extremely poor (<.2). Additionally, interjudge reliability for the DME method was considerably better than that for the EAI method.

Discussion

The significant curvilinear relationship between EAI and DME nasality ratings (see Table 2) indicates that a prothetic rather than a metathetic continuum best represents the perceived nasality of the vowel stimuli. This finding is not unexpected, given that many

Table 2. Analysis of variance table for comparison of curvilinear and linear models for EAI ratings regressed on DME ratings.

Source	df	Mean square	Sum of squares	*F*
Regression	2	36.522	73.045	1611.46*
Residual	47	00.022	00.430	

Line of best fit: EAI = .328 + (.01929)DME + (–.0000715)DME^2

*$p < .01$

Table 3. Internal consistency (coefficient alpha) and reliability (intraclass correlation coefficient) of DME and EAI ratings for all vowels and by nasality level.

	Coefficient alpha		Intraclass correlation coefficient	
Nasality level	DME	EAI	DME	EAI
Oral vowels	.7758	.1785	.75	.52
Nasal Vowels A	.9332	.1773	.82	.51
Nasal Vowels B	.9256	.1875	.80	.53
Nasal Vowels C	.9858	.1653	.86	.58
Nasal Vowels D	.9537	.1564	.84	.56
All vowels	.9765	.1623	.82	.54

of the perceptual dimensions commonly scaled in speech and voice have been shown to be prothetic rather than metathetic (Berry & Silverman, 1972; Schiavetti et al., 1981; Schiavetti et al., 1983; Toner & Emanuel, 1989). It appears that nasality is one more dimension that cannot be validly rated using EAI scaling. Instead, DME appears to permit more valid rating of perceived nasality. Additionally, listeners' ratings of nasality were more consistent and reliable using DME than EAI scaling (see Table 3), casting further doubt upon the traditional use of EAI scaling to rate some perceptual aspects of speech and voice.

The approach and design of this study sets a framework for the continued systematic investigation of the validity of EAI scaling of nasality. By design, a fairly limited set of synthetic stimuli was used. One could expand the stimuli to include vowels other than /i/, which is a high-front vowel and, as such, has its own unique acoustic properties. There is a growing body of literature which suggests that the specific characteristics of the vowel spectra influences listeners' perception of vowel height (Beddor & Hawkins, 1990; Wright, 1986), vowel backness (Ladefoged, 1982; Lindau, 1978; Stevens, 1989), and vowel distinctiveness (Mohr & Wang, 1968; Wright, 1986), as well as vowel duration (Whalen & Beddor, 1989) and vowel context (Kawasaki, 1986). For example, studies with synthetic stimuli have shown that low vowels require more nasal coupling than high vowels to elicit nasal percepts (Abramson, Nye, Henderson,

& Marshall, 1981; House & Stevens, 1956; Maeda, 1982). Duration of the vowel stimulus could be investigated, to include stimuli longer than 1.5 seconds. Using synthetic speech, Whalen and Beddor (1989) found that longer stimuli were consistently perceived as more nasal, a finding that was vowel independent, and which has been shown to also be language independent (Krakow, Beddor, Goldstein, & Fowler, 1988). Phonetic context could also be investigated, to include vowels not in isolation. Krakow and Beddor (1991) have found that nasal vowels were more often correctly judged as nasal when spliced out of nasal contexts and presented in isolation or in an oral context. Future studies may also wish to examine stimulus variables such as fundamental frequency and intensity, as well as listener variables such as familiarity with nasal voice. In the current study, five fundamental frequency levels were chosen to approximate the range of speaking fundamental frequencies heard in human voices. Likewise, only two intensity levels were examined. Those fundamental frequencies and intensities were not allowed to vary within a specific vowel utterance; certainly, this does not approximate the natural variations in human voice. Subtle variations in fundamental frequency and intensity may well influence the perception of nasality, particularly with synthetic stimuli. Last, but certainly not least, the validity of EAI scaling of an even wider range of nasality levels could be investigated. In the current study, only five levels of nasality were investigated, thus limiting the generalizability of the results to other utterances of even lesser, greater, or intermediate degrees of nasality.

In general, the use of EAI scaling to rate nasality has a number of limitations for both research purposes and routine clinical use. First is the bias exhibited by listeners who attempt to partition their perception of nasality into equal intervals. As Stevens (1974) has demonstrated with a variety of prothetic dimensions, listeners do not perceive intervals as equal at different locations on the scale. In the case of nasality, the difference in magnitude between a nasality rating of 1 and 2 (for example) may not be the same magnitude of difference as a nasality rating of 2 versus 3 (for example) or 6 versus 7 (for example). Thus, it is difficult to interpret relative comparisons between scale values assigned either within individuals (e.g., pre/post-therapy) or across individuals (e.g., Intervention A vs. Intervention B). A second limitation of EAI scaling of nasality stems from Stevens's (1974) finding that the prescribed nature of an EAI scale may not capture a respondent's full range of perception. That is, the scale (regardless of its size) may limit the response given. Therefore, in terms of construct validity, it doesn't matter if (for example) one uses a 7-point EAI scale or a 15-point EAI scale—neither may be adequate. The listener is still constrained by the nature of the scale in use. Thus, subtle changes in perceived nasality may not be adequately represented by the specific values on an EAI scale. A third limitation of EAI scaling of nasality stems from Gescheider's (1976) finding that listeners tend to assign stimuli to categories so that all categories are used equally as often. If a large number of individuals are rated, there may be an artificial dispersion of ratings. The use of DME avoids the aforementioned limitations of EAI scaling, thus providing a potentially more valid and reliable method for representing listeners perceptions of nasality.

The present results provide preliminary information about one possible objective criterion for selecting an appropriate psychophysical method for scaling vowel nasality. Although these results suggest that the nasality continuum is prothetic, they do not provide sufficient basis for a firm conclusion about the validity of EAI scaling of nasality in human speech. One might find EAI scaling to be an inappropriate method, particularly as synthetic nasal stimuli become more complex and approximate the human voice. Ultimately, the current study may lead to a series of studies using human voices, in an attempt to better understand the acoustic-perceptual correlates of nasality.

Acknowledgements

This paper is based in part on a doctoral dissertation completed by the first author while at Arizona State University, under the direction of the second author. Thanks to dissertation committee members Leonard LaPointe, James Case, Michael Dorman, and Stephen Beals. Thanks also for statistical consulting to Levent Dumenci of the University of Arkansas for Medical Sciences. Lastly, we would like to thank the student clinicians at Arizona State University who participated in the listening tasks.

References

Abramson, A. S., Nye, P. W., Henderson, J., & Marshall, C. W. (1981). Vowel height and the perception of consonantal nasality. *Journal of the Acoustical Society of America, 70,* 329–339.

American Speech-Language-Hearing Association. (1993). Preferred practice patterns for the professions of speech-language pathology and audiology: Voice assessment. *Asha, 35,* (Suppl. 1), 69–70.

Barry, S. J., & Kidd, G. (1981). Psychophysical scaling of distorted speech. *Journal of Speech and Hearing Research, 46,* 44–47.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19,* 3–11.

Beddor, P. S. (1993). The perception of nasal vowels. In M. Huffman & R. Krakow (Eds.), *Phonetics and phonology: Nasals, nasalization and the velum* (pp. 171–196). New York: Academic Press.

Beddor, P. S., & Hawkins, S. (1990). Influence of spectral prominence on perceived vowel quality. *Journal of the*

Acoustical Society of America, 87, 2684–2704.

Berry, W., Evans, Y., & Lane, A. (1990). *The importance of patient attitude variables in dysarthria rehabilitation*. Paper presented at the Clinical Dysarthria Conference, San Antonio, TX.

Berry, R. C., & Silverman, F. H. (1972). Equality of intervals on the Lewis-Sherman Scale of Stuttering Severity. *Journal of Speech and Hearing Research, 15*, 185–188.

Brend, R. M. (1975). Male-female intonation patterns in American English. In B. Thorn & N. Henly (Eds.), *Language and sex: Difference and dominance* (pp. 84–87). Boston, MA: Newbury House.

Brown, F. (1983). *Principles of educational and psychological testing*. New York: Holt, Rhinehart and Winston.

Case, J. L. (1996). *Clinical management of voice disorders* (3rd ed.). Austin, TX: Pro-Ed.

Chen, M. (1995). Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers. *Journal of the Acoustical Society of America, 98*, 2443–2453.

Coleman, R. F. (1971). Effect of waveform changes upon roughness perception. *Folia Phoniatrica, 23*, 314–322.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 48–51). Washington, DC: American Council on Education.

Cullinan, W. L., Prather, E. M., & Williams, D. E. (1963). Comparison of procedures for scaling severity of stuttering. *Journal of Speech and Hearing Research, 6*, 187–194.

de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research, 37*, 985–1000.

de Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research, 38*, 821–827.

Dubois, P. H. (1970). Varieties of psychological test homogeneity. *American Psychologist, 25*, 532–536.

Dunn-Rankin, P. (1983). *Scaling methods*. Hillsdale, NJ: Lawrence Erlbaum.

Ebel, R. (1951). Estimation of the reliability of ratings. *Psychometrica, 16*, 407–425.

Eisler, H. (1963). How prothetic is the continuum of smell? *Scandinavian Journal of Psychology, 4*, 29–32.

Emanuel, F., & Smith, W. (1974). Pitch effects on vowel roughness and spectral noise. *Journal of Phonetics, 2*, 247–253.

Fant, G. (1960). Descriptive analysis of the acoustic aspects of speech. *Logos, 5*, 3–17.

Fritzell, B., Hammarberg, B., Gauffin, J., Karlsson, I., & Sundberg, J. (1986). Breathiness and insufficient vocal fold closure. *Journal of Phonetics, 14*, 549–553.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research, 36*, 14–20.

Gescheider, G. A. (1976). *Psychophysics, method and theory*. Hillsdale, NJ: Lawrence Erlbaum.

Griffiths, C., & Bough, I. D. (1989). Neurologic diseases and their effect on voice. *Journal of Voice, 3*, 148–156.

Hammarberg, B., Fritzell, B., & Schiratzki, H. (1984). Teflon injection in 16 patients with paralytic dysphonia: Perceptual and acoustic evaluations. *Journal of Speech and Hearing Disorders, 49*, 72–82.

Hawkins, S., & Stevens, K. N. (1985). Acoustic and perceptual correlates of nonnasal-nasal distinction for vowels. *Journal of the Acoustical Society of America, 77*, 1560–1575.

Heiberger, V. L., & Horii, Y. (1982). Jitter and shimmer in sustained phonation. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 7, pp. 299–332). New York: Academic Press.

Hirano, M. (1981). *Clinical examination of voice*. New York: Springer-Verlag.

Hollien, H., Michael, J., & Dougherty, E. T. (1973). A method for analyzing vocal jitter in sustained phonation. *Journal of Phonetics, 1*, 85–91.

House, A. S., & Stevens, K. N. (1956). Analog studies of the nasalization of vowels. *Journal of Speech and Hearing Disorders, 21*, 218–232.

Huffman, M. (1990). The role of F1 amplitude in producing nasal percepts. *Journal of the Acoustical Society of America, 88* (Suppl. 1), S54.

Kawasaki, H. (1986). Phonetic explanation for phonological universals: The case of distinctive vowel nasalization. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology* (pp. 81–103). Orlando, FL: Academic Press.

Kempster, G. B., Kistler, D. J., & Hillenbrand, J. (1991). Multidimensional scaling analysis of dysphonia in two speaker groups. *Journal of Speech and Hearing Research, 34*, 534–543.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice. *American Journal of Speech-Language Pathology, 5*(3), 7–23.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*, 820–857.

Krakow, R. A., & Beddor, P. S. (1991). Coarticulation and the perception of nasality. *Proceedings of the 12th International Congress of Phonetic Sciences, 5*, 38–41.

Krakow, R. A., Beddor, P. S., Goldstein, L. M., & Fowler, C. A. (1988). Coarticulatory influences on the perceived height of nasal vowels. *Journal of the Acoustical Society of America, 83*, 1146–1158.

Kreiman, J., & Gerratt, B. R. (1996). The perceptual structure of pathologic voice quality. *Journal of the Acoustical Society of America, 100*, 1787–1795.

Kreiman, J., Gerratt, B. R., & Berke, G. (1994). The multidimensional nature of pathologic voice quality. *Journal of the Acoustical Society of America, 96*, 1291–1302.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36*, 21–40.

Ladefoged, P. (1982). *A course in phonetics* (2nd ed.). New York: Harcourt Brace Jovanovich.

Lindau, M. (1978). Vowel features. *Language, 54*, 541–563.

Maeda, S. (1982). Acoustic correlates of vowel nasalization: A simulation study. *Journal of the Acoustical Society of America, 72*, S102.

Martin, D. (1965). Direct magnitude estimation judgments of stuttering severity using audible and audible-visible speech samples. *Speech Monographs, 32*, 169–177.

Martin, D., Fitch, J., & Wolfe, V. (1995). Pathologic voice type and the acoustic prediction of severity. *Journal of Speech and Hearing Research, 38*, 765–771.

McWilliams, B., Morris, H., & Shelton, R. (1984). *Cleft palate speech*. Philadelphia: B. C. Decker.

Metz, D. E., Schiavetti, N., & Sacco, P. R. (1990). Acoustic and psychophysical dimensions of the perceived naturalness of no stutterers and posttreatment stutterers. *Journal of Speech and Hearing Disorders, 55*, 516–525.

Mohr, B., & Wang, W. S. (1968). Perceptual distance and the specification of phonological features. *Phonetica, 18*, 31–45.

Murry, T., & Dougherty, E. T. (1980). Selected acoustic characteristics of pathological and normal speakers. *Journal of Speech and Hearing Research, 23*, 361–369.

Rozsypal, A., & Millar, B. (1979). Perception of jitter and shimmer in synthetic vowels. *Journal of Phonetics, 7*, 343–355.

Schiavetti, N., Metz, D. E., & Sitler, R. W. (1981). Construct validity of direct magnitude estimation and interval scaling: Evidence from a study of the hearing-impaired. *Journal of Speech and Hearing Research, 24*, 441–445.

Schiavetti, N., Sacco, P. R., Metz, D. E., & Sitler, R. W. (1983). Direct magnitude estimation and interval scaling of stuttering severity. *Journal of Speech and Hearing Research, 26*, 568–573.

Seikel, J. A., Wilcox, K. A., & Davis, P. J. (1990). Dysarthria of motor neuron disease: Clinicians judgments of severity. *Journal of Communication Disorders, 23*, 417–431.

Shadle, C. (1987). *The acoustics of fricative consonants*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Southwood, H. P. (1996). Direct magnitude estimation and interval scaling of naturalness and bizarreness of the dysarthria associated with amyotrophic lateral sclerosis. *Journal of Medical Speech-Language Pathology, 4*, 13–25.

Southwood, H. P., & Weismer, G. (1993). Listener judgments of the bizarreness, acceptability, naturalness, and normalcy of the dysarthria associated with amyotrophic lateral sclerosis. *Journal of Medical Speech-Language Pathology, 1*, 151–161.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3–45.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153–181.

Stevens, S. S. (1974). Perceptual magnitude and its measurement. In E. C. Caterette & M. P. Friedman (Eds.), *Handbook of Perception* (Vol. 2, pp. 22–40). New York: Academic Press.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects*. New York: Wiley.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology, 54*, 377–411.

Toner, M. A., & Emanuel, F. W. (1989). Direct magnitude estimation and equal appearing interval scaling of vowel roughness. *Journal of Speech and Hearing Research, 32*, 78–82.

Whalen, D. H., & Beddor, P. S. (1989). Connections between nasality and vowel duration and height. *Language, 65*, 457–486.

Wilson, F. (1977). *Voice disorders*. Austin: Learning Concepts.

Wirz, S., & Beck, J. M. (1995). Assessment of voice quality: The vocal profile analysis scheme. In S. Wirz (Ed.), *Perceptual approaches to communication disorders* (pp. 39–55). London: Whurr.

Wright, J. T. (1986). The behavior of nasalized vowels in the perceptual vowel space. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental psychology* (pp. 45–67). Orlando, FL: Academic Press.

Yorkston, K., & Beukelman, D. R. (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders, 46*, 296–301.

Yorkston, K., Bombardier, C., & Hammen, V. L. (1994). Dysarthria from the viewpoint of individuals with dysarthria. In J. A. Till, K. M. Yorkston, & D. R. Beukelman (Eds.), *Motor speech disorders: Advances in assessment and treatment* (pp. 19–35). Baltimore: Brookes.

Received April 30, 1999

Accepted December 14, 1999

Contact author: Richard I. Zraick, PhD, Univerisity of Arkansas, Department of Audiology & Speech Pathology, 2801 South University, Little Rock, AR 72204. Email: rizraick@exchange.uams.edu

A Comparison of Equal-Appearing Interval Scaling and Direct Magnitude Estimation of Nasal Voice Quality

Richard I. Zraick, and Julie M. Liss
J Speech Lang Hear Res 2000;43;979-988