Journal of Speech and Hearing Research, Volume 36, 21–40, February 1993

Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research

Jody Kreiman
Bruce R. Gerratt
VA Medical Center, West Los Angeles and UCLA School of Medicine Los Angeles, CA

Gail B. Kempster
Governors State University University Park, IL

Andrew Erman
Gerald S. Berke
VA Medical Center, West Los Angeles and UCLA School of Medicine Los Angeles, CA

The reliability of listeners' ratings of voice quality is a central issue in voice research because of the clinical primacy of such ratings and because they are the standard against which other measures are evaluated. However, an extensive literature review indicates that both intrarater and interrater reliability fluctuate greatly from study to study. Further, our own data indicate that ratings of vocal roughness vary widely across individual clinicians, with a single voice often receiving nearly the full range of possible ratings. No model or theoretical framework currently exists to explain these variations, although such a model might guide development of efficient, valid, and standardized clinical protocols for voice evaluation. We propose a theoretical framework that attributes variability in ratings to several sources (including listeners' backgrounds and biases, the task used to gather ratings, interactions between listeners and tasks, and random error). This framework may guide development of new clinical voice and speech evaluation protocols, ultimately leading to more reliable perceptual ratings and a better understanding of the perceptual qualities of pathological voices.

KEY WORDS: voice quality, perception (voice), agreement, reliability

Voices can be objectively measured in many ways (see, e.g., Baken, 1987; Hirano, 1981). However, voice *quality* is fundamentally perceptual in nature. Patients seek treatment for voice disorders because they do not sound normal, and they often decide on whether treatment has been successful based on whether they sound better or not. For this and other reasons, speech clinicians use and value perceptual measures of voice and speech far more than instrumental measures (Gerratt, Till, Rosenbek, Wertz, & Boysen, 1991). Further, listeners' judgments are usually the standard against which other measures of voice (acoustic, aerodynamic, and so on) are evaluated (e.g., Coleman, 1969, 1971; Deal & Emanuel, 1978; Emanuel & Sansone, 1969; Fukazawa & El-Assuooty, 1988; Hillenbrand, 1988; Klatt & Klatt, 1990; Kojima, Gould, Lambiase, & Isshiki, 1980; Ladefoged, Maddieson, & Jackson, 1988; Sansone & Emanuel, 1970; Takahashi & Koike, 1975; Wendahl, 1966b; Wendler, Doherty & Hollien, 1980). For perceptual ratings to be meaningful, listeners must use scales consistently: A given rater must rate a voice sample the same way every time he or she hears it. Additionally, for ratings to be clinically useful, interrater agreement must be high: Each rater who hears a voice sample must rate it similarly. Thus, the reliability of such judgments is a central issue in the study of voice quality and voice disorders.

Unfortunately, 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. Research in this area has proceeded without benefit of a consistent theoretical approach to the voice perception process. Consequently, authors have had no particular basis for selecting one or another of the many possible protocols and analyses when designing studies. Both methodology

0022-4685/93/3601-0021$01.00/0

and estimates of listener reliability have varied widely from study to study. However, because research questions have not been motivated by a model or unified approach to the problems at hand, basic questions such as "How reliably *can* listeners judge voices in clinical settings?" and "What is the *best* method for gathering voice ratings and measuring rater agreement?" remain unanswered. A better understanding of the factors that affect intrarater and interrater agreement and reliability would lead to improved protocols for voice quality assessment, a cornerstone of the diagnosis and treatment of voice disorders.

This paper examines the issues of how voice quality can most appropriately be assessed, what reasonable standards for intra- and interrater reliability of perceptual judgments might be, and how reliability of ratings and levels of agreement within and among listeners might be maximized. A review of the literature on voice quality judgments describes the range of reported levels of agreement and variability, examines how methodological variables might affect listener agreement and reliability, and attempts to determine what de facto model of voice ratings underlies previous research. We describe the various statistics and scales that have appeared in the literature and discuss how appropriate these different measures of reliability might be for generalizing results to clinical settings. Experiments designed to evaluate the adequacy of the models implied by previous research are reported, and the suitability of the various methods for gathering voice quality ratings is evaluated. Finally, we propose an alternate theoretical framework that accounts for the present findings and for others in the literature.

Methodological Preliminaries

Because we are concerned with the most effective measurement of perceived vocal quality, we begin by reviewing a number of statistical concepts that are important for deciding how to gather and evaluate listeners' ratings.

Types of Rating Scales

Perceived voice quality has been measured using a variety of tasks. *Categorical ratings* involve assigning speech or voice samples to discrete, unordered categories (e.g., breathy, rough). *Equal-appearing interval* (EAI) scales require listeners to assign a number between 1 and n to a voice sample, where n in the number of points on the scale. Points on EAI scales are *assumed* to be equidistant, so measurements are generally treated as interval level and parametric statistics applied. *Visual analog* (VA) scales are undifferentiated lines, often 100 mm long. Listeners rate voices on these scales by making a mark on the line to indicate the extent to which a voice possesses a given characteristic. In *direct magnitude estimation* (DME), listeners assign a number to a voice sample to indicate the extent to which it possesses the quality being rated. The range of possible numbers generally is not restricted. In anchored DME, listeners are provided with a reference voice sample assigned a specific magnitude (usually 100) of the quality being rated. In unanchored paradigms, listeners make their judgments with reference to their own criteria. Finally, in *paired comparison* tasks listeners compare two stimuli. They may judge the extent of difference on some dimension, similarity/difference, relative roughness, and so on.

Other tasks may be used to measure perceived vocal quality (see, e.g., Dunn-Rankin, 1983, for discussion of many psychometric procedures). For example, Kempster, Kistler, and Hillenbrand (1991) recently used triadic comparisons to evaluate the similarity of dysphonic female voices. However, these other methods are not common in voice quality research and did not appear in any of the 57 papers reviewed below.

Agreement and Reliability of Judgments

The terms *reliability* and *agreement* are frequently treated as synonymous in common usage and in the literature, and are often referred to jointly as "reliability." However, in statistical usage they are technical terms with different meanings (e.g., Berk, 1979). Listeners are in *agreement* to the extent that they make exactly the same judgments about the voices rated. Ratings are *reliable* when the relationship of one rated voice to another is constant (i.e., when voice ratings are parallel or correlated), although the absolute rating may differ from listener to listener. The number of ratings that agree exactly or within ± one scale value is a common measure of agreement; the intraclass correlation is one measure of rating reliability (Ebel, 1951; Tinsley & Weiss, 1975).

The distinction between these two concepts is clinically as well as statistically important. *Agreement* implies that two listeners assign identical meanings to each scale point: their idea of what constitutes extreme roughness or breathiness (for example), their definitions of normal (nonrough or nonbreathy), and the distance between intervening points on the scale are all the same. *Reliability* implies only that listeners rate the voices in a parallel fashion, without implying that scale values have the same meaning. Thus, knowing that a set of ratings is reliable does not necessarily allow prediction of one listener's ratings from another's, unless their relative levels of rating are also known (e.g., A always rates voices one point lower than B).

There is no necessary relationship between levels of agreement and reliability. A rater whose standards drift within a listening session may have high test-retest reliability but low test-retest agreement. Similarly, two raters who use a scale in parallel ways but whose ratings consistently differ by some number of scale values will not be in agreement, but will be reliable relative to one another. Further, good agreement does not necessarily imply good reliability. If the range of ratings is restricted (for example, because listeners consistently avoid the endpoints of an EAI scale, or if stimuli vary very little with respect to the quality rated), reliability coefficients may be low, even if raters agree well with one another (Tinsley & Weiss, 1975).

In general, speech and voice researchers have required high levels of intrarater agreement from their listeners. Reliability, rather than agreement, has often been the standard for between-rater evaluations. In this case, agreement has

been viewed by some as too strict a requirement (e.g., Sheard, Adams, & Davis, 1991). However, in clinical situations where voices are evaluated one at a time, information about a clinician's general level of rating is not usually available. Because the absolute level of the ratings must be meaningful to another rater for rating information to be clinically useful, reporting reliability without including a measure of agreement may be inappropriate for studies with clinical applications.

Focus of the Evaluation

Depending on the experimenter's purpose, measures of reliability and agreement may compare ratings for pairs of raters or may reflect the overall coherence of an entire group of listeners.[1] For example, Pearson's *r* is frequently calculated across all possible pairs of raters and then reported as a single, average value (e.g., Moran & Gilbert, 1984; Nieboer, De Graaf, & Schutte, 1988; Samar & Metz, 1988). Calculation of agreement/reliability across all possible pairs of raters can reveal the number of raters who disagree with the majority and the extent of their disagreement, both of which may be masked by use of a single estimator of interrater agreement/reliability. Use of pairwise measures can also help identify which raters differ substantially from the group. However, averaging many pairwise measures to produce a single measure of reliability tends to amplify the error inherent in the individual statistics. These measures are also subject to inflation or deflation due to several factors, including the probability of chance agreement and levels of occurrence of the quality being rated, as discussed in the following section (see also Berk, 1979).

In contrast, the intraclass correlation (ICC) always reflects the overall coherence of an entire group of listeners. Two formulas are commonly used (although several others exist; see Shrout & Fleiss, 1979). ICC(2,1) reflects the average agreement between one observer and another (also called the reliability of a single listener).[2] ICC(2,k) reflects the reliability of the average of k observations (or the average agreement between the single random sample of observers used in a study and a theoretical set of other random samples drawn from the same universe of raters). This second measure is sometimes referred to as the reliability of the average rating. The value of ICC(2,k) will always be larger than that of ICC(2,1), provided ICC(2,1) is positive (Shrout & Fleiss, 1979). In addition, increasing the number of raters will increase the value of ICC(2,k), because the larger sample size improves the estimate of the mean rating. Because of this, it is misleading to report ICC(2,k) without identifying it as a measure of the reliability of the mean and without justifying the use of mean ratings as the unit of reliability.

Because individual raters, rather than average ratings, are usually of interest in clinical voice evaluations, as a rule ICC(2,k) is not appropriate for measuring reliability in this area. In research applications, however, experimenters may sometimes wish to generate a single average rating of a voice (for example, to study the correlation between average voice ratings and objective measures of voice). If use of mean ratings is appropriate, ICC(2,k) should also be used to assess the reliability of the estimate of those means. Because considerable interrater variability can underlie a stable mean, especially if *n* is large, we feel this statistic is not a substitute for ICC(2,1), even in research applications.

Chance Levels of Agreement and Reliability

Statistically significant coefficients do not necessarily demonstrate a strong relationship between two sets of measurements, just as large correlations or percent agreement values are not necessarily statistically significant. When evaluating a measure of agreement or reliability, it is important to consider the chance level associated with the statistic and the precision of the estimate of that measure (usually with reference to the 95% confidence interval). Particularly when samples are small, reliability coefficients are only guesses (e.g., Ebel, 1951). Confidence intervals and chance values permit assessment of the goodness of the guess.

Unfortunately, most measures of listener agreement do not correct for chance agreement and are subject to inflation or deflation if the set of voices includes very many or very few voices at the extremes of the scale (Berk, 1979). It has been shown (Hopkins & Hermann, 1977; Kearns & Simmons, 1988) that measures of agreement and reliability may also be inflated or deflated due to the rate of occurrence of the quality being rated. That is, listeners frequently agree about what constitutes normal phonation or severe pathology, but disagree more about the extent of mild-to-moderate behaviors. If the stimulus set includes a large number of normal voices, agreement levels may be misleadingly high.

Because listeners seldom agree perfectly, agreement on equal-appearing interval (EAI) scales is often measured as the percentage of ratings that agree within plus or minus one scale value. Assuming all scale values are equally probable (used equally often by each listener), the chance probability of listeners responding within ± one scale value equals $[n + 2(n - 1)]/n^2$, where n = the number of points in the scale (N. Antonanzas-Barroso, personal communication, November 1990). Thus for a 7-point EAI scale, 38.8% of ratings would be expected to fall within ± one scale value by chance alone. For a 5-point scale, chance equals 52%. Thus, "70% agreement within ± one scale value" is at most 18% above chance for a 5-point EAI scale. In practice, many listeners avoid scale extremes; chance levels of agreement are even higher in these cases. When the 95% confidence intervals about "70% agreement" are calculated, it is possible that agreement levels will not differ significantly from chance (see Cullinan, Prather, & Williams, 1963).

[1]Measures may also reflect the reliability with which an individual voice is rated, averaged across raters, rather than the usual focus on the reliability of a rater, averaged across voices. This is not commonly done (but see Kearns & Simmons, 1988).

[2]In this notation, the first number refers to the ANOVA model used to calculate the ICC. A value of 2 indicates that judges are included as a factor in a 2-way ANOVA (vs. model 1, which uses a 1-way ANOVA) and that judges are treated as a random effect (vs. model 3, which treats judges as fixed). The second value represents the unit of reliability: a single judge (1), or the mean of k judges (k). See Shrout & Fleiss (1979) for discussion of other models.

Methods for calculating chance probability levels and confidence intervals are not available for every statistic used to evaluate reliability and agreement (Tinsley & Weiss, 1975). Because the intraclass correlation is based on analysis of variance, confidence intervals can be calculated for this statistic (Ebel, 1951; Shrout & Fleiss, 1979). Additionally, the confidence intervals around the mean rating for a single voice can be calculated, and can indicate the extent to which average ratings reflect the underlying data (e.g., Cullinan et al., 1963).

Fixed Versus Random Samples of Judges and Voices

Whether measures of reliability and agreement can be generalized beyond a given study depends in part on whether the voices and raters are treated statistically as fixed or random factors (Ebel, 1951; Shrout & Fleiss, 1979). When samples of raters or voices are treated as fixed, we assume that we have sampled the entire population of interest, that is, every voice or every rater. Thus no conclusions can be drawn about other groups of raters or voices. When samples are treated as random, we attempt to draw conclusions about differences between raters in general (for example) based on a limited sample selected at random. (See, e.g., Hays, 1973, or Winer, 1971, for more discussion of these issues.)

Most statistics used to assess reliability and agreement do not distinguish between fixed and random effect models (Berk, 1979). However, the intraclass correlation can be calculated to reflect fully fixed, fully random, or mixed effects models, as appropriate (Ebel, 1951), and thus is often the preferred measure of reliability (Berk, 1979). Many authors recommend treating judges as a random effect and stimuli (e.g., voices) as a fixed effect (e.g., Shrout & Fleiss, 1979; Tinsley & Weiss, 1975). However, application of experimental results to clinical settings may in fact require a fully random ANOVA model, so that results will be generalizable to other raters *and* other voices. In any case, authors should specify and justify the ANOVA model used. Unfortunately, this information often is not supplied (see below).

Review of the Literature

This review included 57 papers randomly selected from those published between 1951 and 1990. All papers were concerned with assessment of speech and voice quality, although the studies had many different purposes and utilized a wide variety of qualities, speakers, listeners, and rating tasks. Speech, phonetics, and otolaryngology journals were included in the survey, which was meant to be thorough but not exhaustive. Details of the studies reviewed are given in the Appendix. The discussion below summarizes research practices during this period.

Study Designs

Listeners. The number of raters included in the studies reviewed ranged from 1 to 461, with a median of 11, a mean of 30, and a standard deviation of 63. The distribution of group sizes was roughly trimodal, with modes at 5, 8, and 11 raters. Judges have included experts (experienced clinicians, teachers, and scientists: 28 papers), graduate students in speech pathology (19 papers), undergraduate students (14 papers), naive adults (3 papers), and children (1 paper).[3] One paper did not identify the listeners used.

Rating instruments. "Categorical" ratings,[4] equal-appearing interval scales (with 4, 5, 6, 7, 8, or 9 points), visual analog scales, direct magnitude estimations (with and without anchor stimuli), and paired comparison tasks were all used in the studies reviewed here to gather perceptual judgments of speech and voice quality. EAI scales were by far the most prevalent, appearing in 82.5% of the studies (47/57). Seven-point scales were most common (18 studies), followed by 5-point scales (16 studies). Four studies varied the number of scale points with the quality being rated. Anchored DME was used in 7 studies; one of these also included an unanchored DME task. Six studies used paired comparison tasks; two used "categorical" ratings, and one used a visual analog scale.

Listener training. Task-specific listener training varied from none to years of formal practice and discussion. To simplify our analyses, training was divided into three levels: none, orientation, and extensive training. Orientation was defined as providing listeners with definitions of scale terms, sample or anchor stimuli, and/or a limited number of practice trials (fewer than 20, or less than 25% of experimental trials). Extensive training was any training in excess of orientation. Six studies (10.5%) provided listeners with no training, 29 (50.9%) provided orientation, and 8 (14.0%) included extensive training. Training procedures were not reported in nearly 30% of the papers sampled (15 studies).

Intrarater Agreement and Reliability

The number of repeated trials underlying the different estimates of intrarater agreement/reliability ranged from 3 to 224 (3.3 to 100% of trials). The mean across studies was 49.4% of trials repeated, with a standard deviation of 38.5%. Two studies did not specify the number of repeated trials, and 17 (29.8%) did not report test-retest reliability.

Seven different statistics were used to evaluate intrarater agreement and reliability. The most common was Pearson's *r*, the correlation between the first and second ratings (19 papers). Other statistics included the number of trials whose ratings agreed exactly or within ± one scale value (14 studies), Pearson's *r* for the means or medians of the two sets of ratings (7 studies), an "ANOVA technique" or "test-retest procedure" (4 studies), a t-test or one-way ANOVA for a (directional) difference between the means of the two sets of ratings (3 studies), Kendall's rank-order coefficient (1

[3]Throughout this discussion, totals may exceed 57 because some papers included more than one listener group or protocol.

[4]Although the authors of these studies (Anders, Hollien, Hurme, Sonninen, & Wendler, 1988; Sapir, Aronson, & Thomas, 1986) describe their rating instrument as categorical, both use parametric statistics to evaluate listener reliability. This suggests they actually used equal-appearing interval scales. However, we have maintained the original terminology here.

study), and the Phi correlation, a nonparametric measure of association related to Chi-squared analysis (e.g., Hays, 1973; 1 study). Seven studies reported both Pearson's *r* and the percentage of ratings within ± one scale value. The statistic(s) used varied within as well as across scales. All seven statistics were used with EAI scales, and four of the seven were used with other rating instruments.

Intrarater reliability was not obviously related to the scale type or statistic used, although comparisons must be made with caution because many scales and statistics appeared infrequently. For EAI scales, Pearson's *r* ranged from .66 to .74 for a 4-point scale (one study), from .37 to .98 across seven studies using 5-point scales, and from .73 to .97 for 7-point scales (5 studies). The percentage of ratings within ± one scale value ranged from 55 to 100% for 5-point EAI scales (7 studies), and from 80 to 95% for 7-point scales (2 studies). A single study reported 85% agreement within ± one scale value for a 9-point scale. Pearson's *r* for the means or medians of the first and second ratings ranged from .93 to .98 across 5 studies using 5-point EAI scales. The possible increase in reliability with the number of points in an EAI scale is probably artifactual. Values for Pearson's *r* for the 4-point scale may be depressed by the limited range of the scale; and results for the 5-point scale include one study using very young children as subjects (Deal & Belcher, 1990).

No obvious effect of listener training on intrarater reliability is apparent from the literature, although conclusions are again limited by the small number of observations in some cells. In particular, studies in which listeners were extensively trained did not report higher levels of intrarater reliability than did studies that employed only orientation to the task and/or stimuli. Across scales, Pearson's *r* varied from .86 to .98 for extensively trained listeners (3 studies), from .66 to .98 for listeners who received orientation (9 studies, not including one using small children as subjects (Deal & Belcher, 1990)). Similarly, 86 to 100% of ratings agreed within ± one scale value for extensively trained subjects (2 studies), and 80 to 96% agreed within ± one scale value for subjects who received orientation training (4 studies). A single study reported 85% agreement within ± one scale value for subjects who received no training.

Finally, no clear relationship emerged between intrarater reliability and a rater's experience. Correlations between the first and second rating (Pearson's *r*) ranged from .66 to .98 for expert listeners (10 papers), and from .59 to .95 for graduate students (5 papers). A single paper reported a correlation of .97 for undergraduate students. Similarly, 80 to 95% of ratings were within ± one scale value for experts (3 papers), and 55 to 100% for graduate students (6 papers). A single paper reported 85% agreement within ± one scale value for undergraduates.

Interrater Reliability

Nine different statistics were used to measure interrater reliability and agreement in the papers reviewed. The most common were intraclass correlations (12 papers), the percentage of ratings within ± one scale value (10 papers), and Pearson's *r* for pairs of listeners (7 papers). Other statistics included Cronbach's alpha and coefficient Q (2 papers each), Spearman's rho, Kendall's Coefficient of Concordance, Friedman 2-way ANOVAs, and the Phi correlation (one paper each). All nine statistics were used with EAI scales. Only intraclass correlations were used with VA and DME scales. Interrater reliability was not reported in 40% of papers reviewed (23/57).

The relationship between scale type and interrater reliability is difficult to determine, due to the variety of methods used and the limited number of observations in each cell. For EAI scales, values of Pearson's *r* for pairs of raters varied widely both within and across studies. For 5-point scales, values ranged from .56 to .98 across three studies. A single study using a 7-point scale reported values ranging from .18 to .66, and two studies using 8-point scales reported values from .17 to .87. Across six studies using 5-point scales, listener agreement ranged from 72 to 100% of ratings within ± one scale value. For 7-point scales, values ranged from 58 to 84% (3 studies); a single study using a 9-point scale reported 75% interrater agreement. Values of ICC(2,1) varied from .57 to .83 across five studies that distinguished between ICC models. Listener reliability for direct magnitude estimation tasks was comparable to that for EAI scales: ICC(2,1) values ranged from .58 to .65 (2 studies).

Task-specific training does not obviously improve agreement among raters. Values of Pearson's *r* ranged from .62 to .89 for extensively trained listeners, from .18 to .98 for listeners who received orientation training, and from .17 to .78 for listeners who received no training. Values were derived from at most two studies in each of these cases, and differences probably represent sampling error. Interrater agreement levels (± one scale value) support this suggestion. For extensively trained listeners, values ranged from 58 to 82% (2 studies), and from 75 to 100% for listeners receiving orientation (6 studies). A single study reported 75% agreement within ± one scale value for listeners who received no training.

Finally, no consistent relationship between listeners' background and levels of interrater reliability is apparent from the studies reviewed. For experts, a single study reported ICC(2,1) = .79. Values for graduate students ranged from .58 to .83 across 3 studies, and a single study using undergraduate students reported ICC values from .57 to .74. Values for other statistics followed similar patterns (see Appendix for details).

Discussion

An answer to the question, "Can listeners make reliable judgments about voice quality?," is not readily forthcoming from the literature. Across scales and statistics, reliability and agreement levels ranged from very low (a correlation of .18 indicates less than 4% of variance common to the raters) to extremely high (100% of ratings within ± one scale value). However, methods also varied widely across studies, and neither intrarater nor interrater reliability varied consistently with any of the methodological factors examined here. Un-

fortunately, it is not possible to conclude that any one factor does or does not contribute to rating reliability.

Thus the literature does not further our understanding of the sources of variability in ratings of voice quality in any substantial way. Several factors contribute to the present situation. First, methodological problems are not uncommon in this research area. For example, many studies failed to report reliability at all; others used the authors as the only raters. Second, inadequate or inappropriate statistical techniques for estimation of rater reliability and agreement limit the conclusions that may be drawn from the literature. In particular, the lack of confidence interval data and/or consideration of chance probability levels seriously limits the conclusions that may be drawn from previous studies.

Finally, the literature as a whole lacks a clear theoretical approach to perceptual and rating processes. Although in some cases design features are motivated by study concerns (e.g., Bassich & Ludlow, 1986), in the majority of cases the choice between scale types, rater population, sample size, and statistical approaches is not based on clearly stated goals or theoretical considerations. Further, with the exception of studies using ICC(2,1), previous research has treated all intra- and interrater variability as unpartitioned error. Because possible sources of consistent variation in reliability and/or agreement are not considered, authors have no basis for deciding what procedures are likely to result in maximally stable and reliable ratings. A better understanding of this variability is a necessary step in the development of methods that yield judgments that are reliable and valid across clinicians, voices, and occasions.

In order to begin systematically examining the sources and extent of rater variability, we gathered ratings of vocal roughness from experienced listeners. This study examines the effects of differences among voices, listeners, and tasks on rating reliability. By applying a variety of statistical approaches to the data, we hoped to determine (a) which sources of intra- and interrater variability may be consistent and thus potentially controllable, and (b) which statistical approaches best reflect both the central trend and the extent of variation in listeners' judgments. Finally, these data will serve as an empirical basis for a more adequate approach to voice quality perception.

EXPERIMENT 1

Method

Listeners

Twenty clinicians participated in the listening tests. Four were otolaryngologists, and 16 were speech pathologists. Each listener had a minimum of 2 years' postgraduate experience evaluating and treating voice quality disorders, with a mean of 9.0 years (*SD* = 6.4 years; range = 2–20 years).

Stimuli

The voices of 22 males with voice disorders were selected at random from a library of recordings. Speakers ranged in age from 24 to 77 years (mean age = 54.6 years; *SD* = 14.4 years). In the judgment of the authors, these voices sampled the entire range of vocal roughness: Five were mildly rough, 6 were severely rough, and 11 were moderately rough. (This impression was later confirmed by examination of listeners' mean roughness ratings, which spanned the entire scale and were approximately normally distributed.) An additional 8 voices were selected at random from a similar library of normal voices, except that the normal speakers were roughly matched in age to the pathological speakers (range 30–70 years; mean age = 56.6 years; *SD* = 14.2 years). All speakers were recorded using the same procedures and equipment. They were asked (among other tasks) to sustain the vowel /ɑ/ at comfortable levels of pitch and loudness for as long as possible.

Voice samples were low-pass filtered at 8 kHz and then sampled at 20 kHz using a 12-bit A/D converter. A 3-sec segment was extracted from the middle of each speaker's vowel. These digitized segments were normalized for peak voltage. The onsets and offsets were multiplied by a 50 msec ramp to reduce click artifacts. Stimuli were then output at 20 kHz through a 12-bit D/A converter, again using an 8 kHz low-pass filter.

Tapes

Two stimulus tapes were constructed. Each included two repetitions of each stimulus voice (so test-retest reliability could be evaluated) for a total of 60 trials per tape. For each tape, the 30 stimulus voices were randomized, played out, rerandomized, and played out again. Trials were separated by 6 sec. Listeners were not informed that any voices were repeated.

Tasks

Each listener participated in two listening sessions. At the first, he or she judged the roughness of the voices using a 7-point equal-appearing interval (EAI) scale. This scale was selected because it is commonly used in voice perception research. At the second listening session, listeners judged the voices using a 10 cm visual analog (VA) scale. This scale was selected because pilot studies (e.g., Kempster, 1987) suggested it is more reliable than EAI scales.

Procedure

Because knowing how (and how much) listeners differ in their voice ratings is a first step toward understanding why they differ, our primary interest in this study was to compare the voice ratings of individual listeners. Given this perspective, it was important that differences among listeners within a task not be confounded with possible learning effects associated with different task presentation orders. For this reason, tape and task presentation orders were not counter-

balanced across listeners in this first experiment. Instead, all listeners made judgments using the EAI scale first, followed by VA scale judgments; and all listeners heard tape 1 before tape 2. Thus the effects of task order and tape order, if any, were constant across listeners, and all listeners within a condition could be compared without any confounding effects. (These effects are examined separately in Experiment 2 below.) Test sessions were separated by at least 1 week to minimize learning effects across tasks.

Listeners received no formal training or instruction in the use of either scale. For the EAI scale task, they were asked to circle a single number to indicate the degree of roughness for each voice sample. They were asked not to circle between numbers and to try to use the whole scale. For the VA scale they were asked to make a single clear, unambiguous mark on an undifferentiated 10 cm line to indicate the roughness of each sample. The left and right endpoints of both the EAI and VA scales were labeled "not rough at all" and "extremely rough," respectively. However, listeners were offered no definition of roughness, but were asked to apply whatever standards or criteria they normally used in their practices when making their judgments.

For both tasks, separate answer sheets were provided for every voice, to minimize interactions among ratings. Listeners were asked not to refer to any previous ratings when judging the voices, and they were not permitted to change answers after hearing subsequent voices.

Listeners were tested individually in a sound-treated booth. Tapes were played at a comfortable listening level (approximately 75 dB SPL) in free field over two speakers equidistant from the listener. Each experimental session lasted approximately 15 minutes.

Results

Intrarater Reliability

Table 1 lists values of the most common measures of test-retest agreement and reliability, calculated from our data. On the average, ratings of these voices did not vary much from first to second rating. Squared values of Pearson's *r* (which measure the amount of variance common to two sets of ratings) suggest that overall about 74% of variance is consistent across rating occasions. Most statistics give approximately the same estimates of intrarater agreement and reliability, with the exception of Pearson's *r* for the medians of the first and second ratings. This statistic inflated reliability levels relative to the others examined here, because calculating median values prior to correlations removes most of the variance from the data before measuring reliability. Thus this statistic is not a good measure of intrarater reliability.

Although average statistics indicate reasonable intrarater agreement and reliability, individual listeners did vary in their performance. For both scales, squared correlations between the first and second ratings were below .75 for 8 of the 20 raters. Squared correlations were above .9 for only one rater. For the worst rater, r^2 was .55 for the EAI scale and only .41 for the VA scale.

TABLE 1. Measures of intrarater (test-retest) reliability.

Statistic	EAI ratings	VA ratings
Exact agreement	47.5%	
SD	11.1%	
Range	20–63.3%	
Mean ratings ± one scale value	84.7%	
SD	8.39%	
Range	68–93%	
Mean Pearson's *r*	.86	.86
SD	.06	.07
Range	.74–.95	.64–.97
Pearson's *r* for medians of 1st and 2nd rating	.95	.98
Mean Spearman's rho	.86	.86
SD	.06	.06
Range	.74–.96	.69–.96
Mean gamma coefficient	.86	.71
SD	.07	.08
Range	.72–.95	.50–.88
Significant *t*-test: 1st vs. 2nd rating	9/20	0/20

T-tests reveal a consistent trend in the EAI ratings for many listeners: Voices were rated as rougher when presented the second time during that listening session than they were the first time they were rated. Differences were significant when all listeners were combined in a single analysis (paired samples *t*-test, $t(359) = -6.65$, $p < .01$) and for 9 out of 20 individual listeners. Eight of these 9 clinicians had less than 3 years' experience evaluating voice quality. Ratings for the other 11 listeners showed the same trend, but the differences did not reach statistical significance. VA scale ratings showed no such effects, either for the group (paired samples $t(359) = -0.57$, $p > .01$) or for any individual listeners.

No other significant relationship was found between measured intrarater agreement and reliability and the amount of experience listeners had evaluating voice quality, for either scale. For the EAI ratings, the correlation between agreement (± one scale value) and years experience was .38 ($p > .01$). For exact agreement, the correlation was .31 ($p > .01$). For reliability measured with Pearson's *r*, the correlation was .38 ($p > .01$). For the VA scale, the correlation between Pearson's *r* values and years of experience was .29 ($p > .01$). Scatterplots confirmed that listeners varied in intrarater reliability at every level of experience. It was not the case that listeners with relatively little experience varied in reliability and more experienced listeners were consistently reliable.

Interrater Agreement and Reliability

Measures of interrater agreement and reliability are given in Table 2. The intraclass correlation was calculated using the formula for ICC(2,1) and a fully random effects ANOVA

TABLE 2. Measures of interrater reliability.

Statistic	EAI ratings	VA ratings
Mean exact agreement	33.7%	
SD	8.3%	
Range	6.7–56.7%	
Mean ratings ± one scale value	74.5%	
SD	8.52%	
Range	50–91.7%	
Mean Pearson's *r*	.78	.78
SD	.07	.07
Range	.55–.92	.56–.92
Mean Spearman's rho	.78	.78
SD	.07	.06
Range	.55–.92	.57–.90
Intraclass correlation		
Reliability of single rating	.79	.78
95% confidence interval	$.70 < \rho < .87$	$.69 < \rho < .87$
Reliability of mean rating	.99	.99
Kendall's Coefficient of Concordance	0.76	0.79

model. Squaring values of ICC(2,1) and average Pearson's *r* indicates that roughly 60% of variance in roughness ratings is explainable in terms of shared variance. Note that the 95% confidence intervals for the intraclass correlation suggest that our estimate of this parameter is not particularly precise: The true ICC value for the EAI ratings is probably between .7 and .87, and for the VA ratings it is probably between .69 and .87. Levels of exact agreement between pairs of listeners are consistently lower than are correlation values, indicating that listeners agreed about the relative roughness of the voices, but differed in the precise meaning assigned to each scale value. The value of ICC(2,20), which measures the reliability of the mean ratings of the voices, is nearly unity for both scales, suggesting that stable average ratings may be obtained with fewer than 20 listeners. Note that these values are considerably higher than those of other measures of the reliability of individual raters, as discussed above.

Despite the levels of average agreement, the range of values for these statistics indicates that many pairs of listeners are not in good agreement with one another in their judgments of vocal roughness. Only 12 out of 190 (6.3%) pairs of raters agreed with r^2 greater than .75 for the EAI scale. For the VA scale, only 20 of 190 pairs of raters (10.5%) agreed at r^2 levels greater than .75. In contrast, for the EAI scale 29 of 190 pairs of raters (15.3%) agreed at r^2 levels below .5 (less than half the variance in ratings shared by the two raters); 26 of 190 (13.7%) pairs of raters shared less than half the variance in ratings for the VA scale. Measures that represent an average of many pairwise comparisons (e.g., average Pearson's *r* or the average number of ratings within ± one scale value) mask the low probability of two individuals agreeing closely and the substantial probability of poor agreement between individuals.

TABLE 3. Intrarater agreement for EAI ratings of pathological and normal voices.*

	Pathological voices	Normal voices
% Exact Agreement	42.7	60.6
SD	11.1	20.8
Range	18.2–63.6	25.0–100
% ± One Scale Value	81.6	93.1
SD	9.1	11.1
Range	63.6–95.5	62.5–100

*Based on 20 individual raters. For the pathological voices, $n = 22$; for the normal voices, $n = 8$.

Pathological Versus Normal Voices

Table 3 shows intrarater agreement levels for EAI ratings of pathological and normal voices. Levels of exact test-retest agreement are given first, followed by agreement within ± one scale value. Due to the highly restricted range of ratings for the normal voices, other measures of agreement/reliability, such as Pearson's *r*, could not be meaningfully calculated. These analyses therefore include only ratings on the EAI scale. Although ranges of agreement levels overlap considerably, test-retest agreement was significantly greater for normal than for pathological voices, for both measures of agreement (exact agreement: paired samples $t = -3.90$, $df = 19$, $p < .01$; agreement within ± one scale value: paired samples $t = -4.71$, $df = 19$, $p < .01$).

Table 4 shows levels of interrater agreement for EAI ratings of pathological and normal voices. As above, agreement levels were significantly greater for normal than for pathological voices (exact agreement: paired samples $t = -12.46$, $df = 189$, $p < .01$; agreement within ± one scale value: paired samples $t = -27.18$, $df = 189$, $p < .01$).

Agreement Levels for Individual Voices

When examined voice by voice, listeners' ratings appear extremely variable. Ratings for 4 of the 22 pathological voices spanned the entire 7-point scale. Six additional voices received nearly the full possible range of ratings; ratings ranged from 2 to 7 for one voice, and from 1 to 6 for

TABLE 4. Interrater agreement for pathological and normal voices.*

	Pathological voices	Normal voices
% Exact Agreement	29.3	46.1
SD	9.7	16.3
Range	6.8–52.3	6.3–81.3
% ± One Scale Value	68.0	90.5
SD	10.1	8.5
Range	40.9–88.6	62.5–100

*Based on 190 pairs of raters. For the pathological voices, $n = 44$ (22 voices, each rated twice); for the normal voices, $n = 16$ (8 voices, each rated twice).

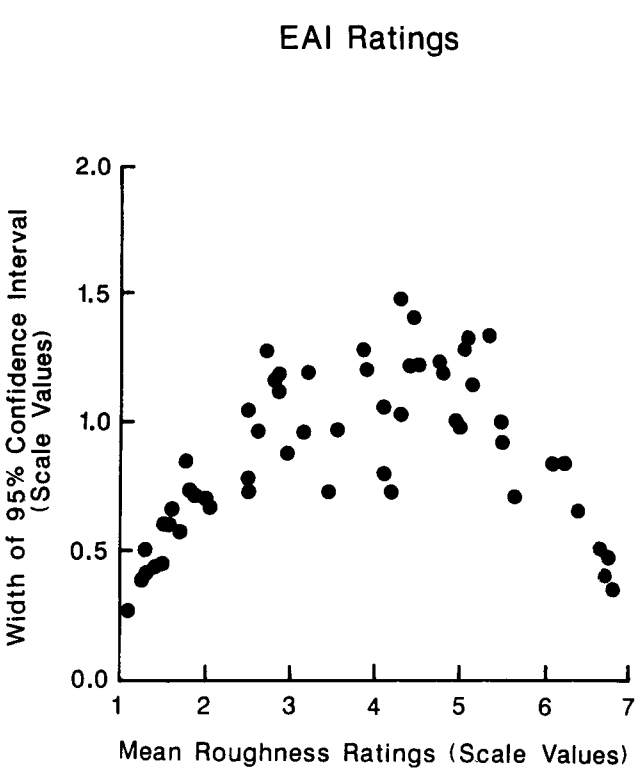

FIGURE 1. The width of the 95% confidence interval for a voice's mean rating vs. the mean rating for that voice on the EAI scale. Wider confidence intervals indicate more variance in ratings across listeners. Listeners agreed best (i.e., ratings were least variable) for voices rated normal or extremely rough; ratings varied most (confidence intervals were widest) for voices whose average rating was near the middle of the scale.

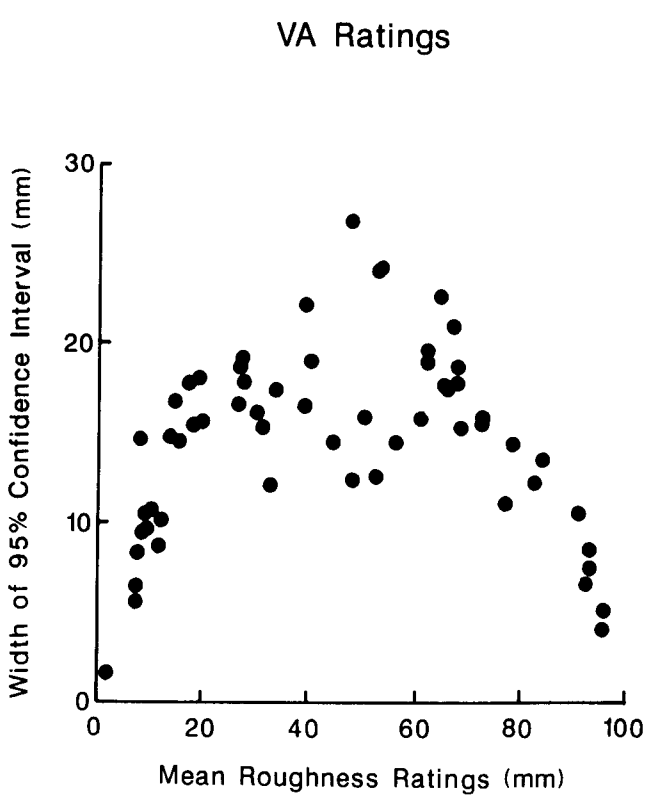

FIGURE 2. The width of the 95% confidence interval for a voice's mean rating vs. the mean rating for that voice on the VA scale. As for the EAI scale, ratings are most variable in the middle range of average roughness and least variable for voices that are near-normal or extremely rough.

five others. Across listeners, the range of ratings was three scale values or less for only three pathological voices (one whose ratings ranged from 1 to 3, one from 5 to 7, and one from 6 to 7). Agreement was not appreciably better for the normal voices. Ratings spanned three scale values or less for only 4 of the 8 speakers (from 1 to 2 or from 1 to 3).

The average range for the VA ratings of pathological voices was 69.5 mm (out of a possible 100). Ratings for three voices spanned more than 90 mm. The maximum range was 95.5 mm, and the minimum range was 32.5 mm. Ratings for the normal voices were not less variable than those for the pathological voices (an average range of 59.6 mm; $F(1, 28) = 2.05$, $p > .01$). The minimum range for the normal voices was 34 mm; the maximum was 75 mm.

Confidence intervals were calculated as a measure of the variability in ratings for each individual voice, using the large-sample formula: 95% confidence interval = mean rating ± (1.96* standard error of the mean) (e.g., Hays, 1973). For EAI ratings, confidence intervals ranged in width from .27 scale values to 1.48 scale values. For the VA scale, they ranged from 1.71 mm to 26.77 mm. The average confidence interval widths (.85 scale values for the EAI ratings and 14.4 mm for the VA ratings) again reflect the overall variability in ratings of individual voices across raters.

Figures 1 and 2 show the 95% confidence interval plotted against the mean rating for each voice for the EAI and VA scales, respectively. These figures show that ratings are much more variable in the midrange of vocal roughness (cf. Arends, Povel, van Os, & Speth, 1990; Kearns & Simmons, 1988). Confidence intervals for near-normal and extremely rough voices are markedly narrower than are those for mildly to moderately rough voices.

Discussion

As suggested by the literature review above, listeners varied widely in their levels of reliability and agreement, and some individual voices were much more likely to receive consistent ratings across listeners than others. Because all listeners in this experiment performed the tasks in the same order, reliability and agreement scores within a task condition are not confounded with presentation orders. Thus, comparisons among subjects and voices in reliability/agreement levels within a task reflect the range of expert skills on that scale under identical conditions. Had task order been randomized across subjects, within-task variability in reliability would include effects of different task orders, making it difficult to draw conclusions about the range of listener performance within a task.

Listener performance appeared quite similar overall on the EAI and VA tasks. However, between-task comparisons in this study are confounded with the fixed presentation order. For example, reliability for VA ratings may be inflated by learning effects, making findings of similar reliability levels for

TABLE 5. Measures of intrarater (test-retest) reliability: Experiment 2.

Statistic	EAI ratings	VA ratings
% Exact agreement	39.7	
SD	13.57	
Range	26.7–66.7	
Mean % ratings ± one scale value	81.0	
SD	13.70	
Range	56.7–100	
Mean Pearson's *r*	.80	.81
SD	.09	.05
Range	.66–.90	.73–.88
Pearson's *r* for medians of 1st and 2nd rating	.95	.96
Mean Spearman's rho	.80	.79
SD	.10	.06
Range	.59–.91	.68–.88
Mean gamma coefficient	.80	.62
SD	.12	.07
Range	.56–.96	.5–.67
Significant *t*-test: 1st vs. 2nd rating	4/10	1/10

the two scales incorrect. Further, drift may have occurred in the EAI task because it was presented first, whereas learning effects prevented such drift in the VA task. Thus it is not possible to determine from the present results if drift is a characteristic of EAI scales or an artifact of our design. Experiment 2 was undertaken to examine directly the role of task order in our results.

EXPERIMENT 2

Method

Ten additional subjects participated in this experiment. Eight were otolaryngologists and two were speech-language pathologists/voice researchers. None had participated in Experiment 1. They averaged 6 years' postgraduate experience evaluating and treating voice disorders (*SD* = 5.37 years; range = 3 to 20 years).

These listeners heard the same experimental tapes as in Experiment 1, but performed the tasks in the reverse order (i.e., VA task first, followed by the EAI task). Test sessions were again separated by at least 1 week. All other methods were as described in Experiment 1 above.

Results and Discussion

Intrarater reliability and agreement levels for Experiment 2 are reported in Table 5. Ranges of scores are comparable to those observed in Experiment 1. One-way ANOVAs showed no significant effect of task presentation order on levels of within-subject agreement on the EAI task (exact agreement: $F(1, 28) = 0.83$, $p > .01$; agreement within ± one scale value: $F(1, 28) = 2.87$, $p > .01$). A two-way ANOVA (experimental order × task) compared average values of Pearson's *r* for intrarater reliability on the VA and EAI tasks for the two presentation orders. The subjects in Experiment 2 were significantly less reliable overall on this measure than those in Experiment 1 [$F(1, 56) = 13.39$, $p < .01$]. However, the two tasks did not differ in intrarater reliability [$F(1, 56 = 0.06$, $p > .01$], and no experiment-by-task interaction was observed. Note that 95% confidence intervals for ICC(2,1) in Table 6 blanket those in Table 2, indicating that on this measure the two subject groups did not differ significantly in overall interrater reliability for either task.

Table 6 shows interrater agreement and reliability for Experiment 2. Again, ranges of scores are comparable to those in Experiment 1. For the EAI scale, listeners in this second experiment agreed less well on the average than those in Experiment 1, for both exact agreement [$F(1, 233) = 14.63$, $p < .01$] and agreement within ± one scale value [$F(1, 233) = 13.06$, $p < .01$]. A two-way ANOVA (experiment × task) showed that subjects in Experiment 2 were less reliable overall than those in Experiment 1 [$F(1, 466) = 80.45$, $p < .01$]. The two tasks did not differ significantly [$F(1, 466) = 0.24$, $p > .01$], and there was no task × experiment interaction [$F(1, 466) = 0.90$, $p > .01$]. In contrast, 95% confidence intervals for the ICC show no significant differences between tasks in levels of interrater reliability.

Finally, ratings showed the same patterns of "drift" when tasks were presented in the reverse order. Across the entire set of raters, voices were rated rougher the second time they were heard than the first for the EAI task (matched pairs *t* =

TABLE 6. Measures of interrater reliability: Experiment 2.

Statistic	EAI ratings	VA ratings
Mean % exact agreement	28.4	
SD	8.05	
Range	10.0–46.7	
Mean % ratings ± one scale value	68.8	
SD	13.01	
Range	33.3–93.3	
Mean Pearson's *r*	.71	.69
SD	.11	.08
Range	.39–.89	.44–.83
Mean Spearman's rho	.71	.70
SD	.12	.09
Range	.36–.90	.45–.84
Intraclass correlation		
Reliability of single rating	.71	.64
95% confidence interval	$.57 < \rho < .83$	$.48 < \rho < .79$
Reliability of mean rating	.96	.95
Kendall's Coefficient of Concordance	.71	.73

-4.23, $df = 179$, $p < .01$), but not for the VA task (matched pairs $t = -0.75$, $df = 179$, $p > .01$). Results were significant for 4 of the 10 individual raters for the EAI task (compared to 9 of 20 in Experiment 1), and for only 1 rater in the VA task (versus 0 in Experiment 1).

These results confirm our finding that experienced listeners vary widely in their ratings of the same voices. No evidence of a significant effect of presentation order was observed. We therefore conclude that differences between the two tasks in patterns of drift were due to differences between the rating scales themselves and were not artifacts of a fixed presentation order.

GENERAL DISCUSSION

Our findings may be summarized as follows. Previous studies using listener ratings to evaluate speech and voice quality have varied widely in the methods applied. Many studies suffer from methodological weaknesses, including failure to evaluate listener reliability, inadequate samples of raters, too few repeated trials for intrarater reliability estimates, failure to report confidence intervals, and use of inappropriate statistics. Across studies, no clear relationship between methods and reliability/agreement has emerged, and no model of voice perception or the rating process has been developed.

Although the present research examined ratings of a single perceptual quality (roughness), many of our findings are consistent with the literature on voice and speech ratings. Thus, we feel our findings may be generalized to ratings of other voice and speech qualities. Our results suggest that average levels of intra- and interrater reliability are relatively high, but that considerable variability underlies average values (see, e.g., Cullinan et al., 1963; Kempster, 1984). EAI ratings—but not VA ratings—drifted significantly during a single listening session, with voices sounding rougher the second time they were evaluated. Ratings for individual voices varied widely across raters. Ratings varied more for pathological than for normal voices and more for mild-to-moderately rough voices than for voices at scale extremes. Order of task presentation had no significant effect on listener performance.

To account for these findings, we propose the following descriptive theoretical framework for perceptual evaluations of voice quality. It includes factors drawn from the literature and from the studies described above. Although this framework is clearly preliminary and includes elements that are speculative, we feel it provides a useful format for conceptualizing the voice perception process and may allow a coherent experimental approach to be developed.

A Conceptual Framework for Voice Quality Perception

When listeners rate a voice on some dimension (e.g., breathiness or roughness), they compare the stimulus presented to an internal standard or scale. We suggest that listeners develop individual internal standards for vocal qualities through their experience with voices (Kreiman, Gerratt, Precoda, & Berke, 1992). These standards represent "average" or "typical" exemplars for some level of the quality being rated, based on that listener's experience.

Listeners may differ in where along the severity continuum they place their internal standards. Individuals may also differ in the amount of detail present in their internal representation of vocal qualities. Accordingly, the precise number of severity levels represented in memory probably differs from listener to listener. Scant evidence regarding the nature of internal representations for voices has been published. To date, arguments have been based on patterns of confusions in short-term memory (Kreiman & Papcun, 1991) and long-term memory (Papcun, Kreiman, & Davis, 1989) and spatial representations from multidimensional scaling (Kreiman, Gerratt, & Precoda, 1990; Kreiman et al., 1992). Clearly many more such studies are needed.

Existing evidence from multidimensional scaling provides support for the existence of individual internal standards for voice quality. Kreiman et al. (1992) have argued that all listeners have similar, relatively stable internal standards for "normal" voice quality, because all listeners have extensive and approximately equal experience with normal voices (through their everyday contact with normal speakers). Therefore, similarity ratings (and ratings of specific qualities in the present study) consistently show relatively little variation within and across listeners for normal and near-normal voices.

Expert listeners (particularly less-experienced ones) differ in their experience with pathological voices, so internal standards for pathological qualities differ from listener to listener (Kreiman et al., 1992). Naive listeners, on the other hand, have no formal exposure to pathological voices and thus appear to lack specific internal standards for judging pathological qualities. Naive listeners apparently judge pathological speakers according to standards better suited to normal voices. Their perceptual strategies are therefore quite similar. In contrast, expert listeners vary widely in their perceptions of pathological voices (Kreiman et al., 1990; Kreiman et al., 1992).

The results presented in this paper further suggest that internal standards for vocal qualities are inherently unstable and that a listener's notion of the extent of roughness or breathiness may be influenced by factors other than the acoustic characteristics of the voice being judged. As reported above, EAI ratings drifted significantly in a consistent direction within a listening session. The stimulus set used here contained a number of normal (8/30) and near-normal (5/30) voices. We suggest that this unbalanced context systematically influenced listeners' internal standards for roughness. That is, over time the rougher voices began to sound worse by comparison to the large number of normal and near-normal samples. Although all listeners in this study had at least 2 years' experience rating voices, eight of the nine clinicians whose ratings drifted significantly in Experiment 1 (and two of the four in Experiment 2) were relatively inexperienced. This suggests listeners need many years to develop a stable set of criteria for this kind of rating. This is consistent with the failure of extensive training to improve the reliability of voice quality ratings in the literature reviewed above.

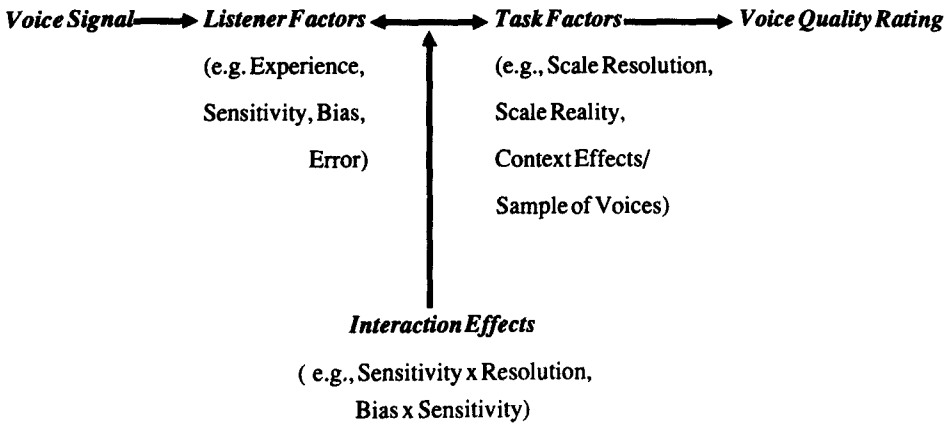

FIGURE 3. Factors involved in mapping an acoustic signal onto a voice quality rating.

Similar drift in ratings was not observed for the VA scale. This asymmetry, which has also been reported in studies evaluating the quality of synthetic speech (Rossi, Pavlovic, & Espesser, 1990), may be related to the relative coarseness/fineness of the two scale types. The EAI scale required listeners to distinguish seven levels of roughness; however, the VA scale (an undifferentiated line) in principle assumes listeners can distinguish a very large number of levels of roughness. When a scale is fine relative to a listener's ability to make judgments, measurements will include some random error related to the mismatch between scale resolution and the listener's perceptual acuity. For example, if a listener can reliably distinguish only five levels of breathiness, but breathiness is measured on a 100-point scale, the difference between ratings of 40, 45, and 52 may not be meaningful, and these numbers may be assigned to the same stimulus. This sort of error may have made drift difficult to detect on a VA scale.

The number of levels of roughness listeners can reliably distinguish and identify is not known. The 7-point EAI scale is probably slightly coarse relative to listeners' acuity, and the continuous 100 mm VA scale (with responses measured to the nearest 0.5 mm) is probably too fine. Further research may clarify these issues.

Sources of Variability in Voice Quality Ratings

The above discussion suggests that several factors are involved in mapping an acoustic signal onto a voice quality rating (Figure 3). The first is the acoustic voice signal being rated. Most studies of voice quality in the literature apparently assume that mapping between physical signals and psychological qualities is a constant, linear process and thus treat any variation in ratings as random error by raters. However, our findings indicate that several consistent—and thus potentially controllable—factors also contribute to observed voice ratings.

The rated quality a listener derives from a signal may be systematically affected by several factors related to the listeners. These factors include listeners' experience with voices (which will shape their particular internal standard[s] for the quality being judged), their individual perceptual habits and biases (Kreiman et al., 1990; Kreiman et al., 1992), and presumably overall sensitivity to the quality being judged. These factors change relatively slowly over time and thus hypothetically affect interrater reliability more than intrarater reliability. Additional factors related to listeners include fatigue, attention lapses, and mistakes. These "error" terms should affect both intra- and interrater reliability.

Factors related to the task of rating also systematically affect measurements of voice quality. If the quality to be rated is poorly defined or lacks perceptual reality, listeners will not be able to rate it consistently. Our results suggest perceptual context may cause systematic drift in ratings, presumably because of its effect of altering a listener's internal standards. We test this hypothesis directly in a companion paper (Kreiman, Gerratt, & Berke, 1993). These (and possibly other) factors can affect ratings within a given session and thus affect both intra- and interrater reliability.

Several systematic interactions among listener and task factors may also occur. Listener sensitivity may interact with scale resolution, and mismatches may add noise to the data or result in information loss. Also, listener biases may interact with "scale specificity." If the quality being rated is multidimensional in nature but is rated on a unidimensional scale, listeners may selectively focus on one dimension or another, reducing apparent agreement levels. For example, recent findings (Kreiman, Gerratt, & Berke, 1992) suggest that breathiness and roughness are related, multidimensional constructs. These investigators demonstrated that listeners' differential attention to various aspects of each quality is a significant source of interrater unreliability in voice quality ratings.

Conclusions and Implications for Research Design

Our findings paint a rather bleak picture of the current state of voice quality ratings. A review of the literature suggests that existing protocols do not consistently result in reliable ratings. Although intrarater and interrater reliability varied considerably across studies, they did not vary consistently with any of the methodological factors reviewed, including levels of task-specific training or the actual rating task used.

The present experimental results suggest that even highly experienced listeners frequently disagree completely about what they hear. Finally, the presence of an error term in our theoretical framework suggests that perfect reliability and agreement are not achievable, even in theory. It appears that nearly 30 years of research have returned us to the point articulated by Jensen in 1965: ". . . the assumption that the perceptual characteristics of a person's voice provide sufficient information for a reliable description of vocal deviation and its severity appears to be hazardous" (p. 82).

Nevertheless, the theoretical framework presented above does offer a potential remedy for this situation. It suggests that variability in voice quality ratings might be reduced by replacing listeners' idiosyncratic, unstable, internal standards with fixed external standards or "reference voices" for different vocal qualities. A voice rating protocol using fixed reference voices would reduce listener-related rating variability by providing all raters with a constant set of perceptual referents. Similar protocols using explicitly presented standards have improved listener performance in other kinds of auditory judgment tasks (for example, ratings of the intensity of tones: Berliner, Durlach, & Braida, 1978). Such a protocol for rating voice quality would also control context-related variability, because external standards remain constant from trial to trial.[5] (See the companion paper, Gerratt et al., 1993, for further discussion of anchored protocols for voice quality ratings.)

Pending development of new protocols, several other improvements may be made to existing research techniques. First, both intra- and interrater reliability must be reported in all studies using voice quality ratings. Our findings and the discussion above suggest that there is no one best measure of "listener concordance." Rather, statistics should be chosen on the basis of the characteristics of the application and of the data. In cases where the exact value of the ratings is not important, reliability should be assessed with the intraclass correlation (using the appropriate ANOVA model). However, if it is important that the meaning of each scale value be constant across raters (for example, in studies with strong clinical implications), agreement should also be assessed. If the stimulus voices vary little on a given quality (for example, if patients with a single diagnosis are studied), then statistics like Pearson's *r* that are sensitive to the range of values should be avoided. In all cases, authors should report ranges and confidence intervals for any measures used and should explicitly justify their selection of reliability and/or agreement measures.

Our results suggest that traditional voice rating methods may never generate ratings that consistently meet strict standards for reliability. However, new rating protocols may be developed to control some of the sources of variability in listeners' perceptions of vocal quality. In particular, main and interaction effects involving task factors can be eliminated by designing scales and protocols (possibly using fixed reference voices) that eliminate context effects and unnecessary rating noise from measures of voice quality. Such protocols might also benefit research aimed at developing instrumental measures of voice. For example, more valid and reliable perceptual measures of voice quality may facilitate the search for acoustic correlates of perceived vocal qualities. As our methods of voice assessment improve with further research, voice ratings will approach maximum reliability and validity.

[5]Random error may also be reduced by careful attention to listening conditions, by motivating subjects adequately to pay attention, and by limiting the number of judgments made at any one session.

Acknowledgments

We thank our expert listeners Mary J. Bacon, Steven Bielamowicz, Robert Block, Nancy Brough, Hong-Shik Choi, S. Trey Fyte, Patricia Gomeztrejo, Susanne Hildebrand, Leona Hubatch, Ann E. Kalec, Daniel Kempler, Ming Ye, Julianne Morlock, Lee Nguyen, Janet Novak, Clare Anne Paskiet, Douglas Ross, Shimon Sapir, Joel Sercarz, Steven Sloan, Julie Trautmann, Diana Van Lancker, and Mary Walsh, for volunteering their time. This work was supported in part by grant # DC 00855-01 from the NIDCD and Merit Review funds from the DVA.

References

Anders, L., Hollien, H., Hurme, P., Sonninen, A., & Wendler, J. (1988). Perception of hoarseness by several classes of listeners. *Folia Phoniatrica, 40,* 91–100.

Arends, N., Povel, D-J., Os, E. van, & Speth, L. (1990). Predicting voice quality of deaf speakers on the basis of certain glottal characteristics. *Journal of Speech and Hearing Research, 33,* 116–122.

Arnold, K. S., & Emanuel, F. (1979). Spectral noise levels and roughness severity ratings for vowels produced by male children. *Journal of Speech and Hearing Research, 22,* 613–626.

Askenfelt, A. G., & Hammarberg, B. (1986). Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures. *Journal of Speech and Hearing Research, 29,* 50–64.

Baken, R. J. (1987). *Clinical measurement of speech and voice.* Boston: College Hill.

Bassich, C., & Ludlow, C. (1986). The use of perceptual methods by new clinicians for assessing voice quality. *Journal of Speech and Hearing Disorders, 51,* 125–133.

Berk, R. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83,* 460–472.

Berliner, J.E., Durlach, N. I., & Braida, L. D. (1978). Intensity perception IX: Effect of fixed standard on resolution in identification. *Journal of the Acoustical Society of America, 64,* 687–689.

Brancewicz, T. M., & Reich, A. R. (1989). Speech rate reduction and "nasality" in normal speakers. *Journal of Speech and Hearing Research, 32,* 837–848.

Coleman, R. F. (1969). Effect of median frequency levels upon the roughness of jittered stimuli. *Journal of Speech and Hearing Research, 12,* 330–336.

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

Coleman, F., & Wendahl, R. (1967). Vocal roughness and stimulus duration. *Speech Monographs, 34,* 85–92.

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.

Darley, F., Aronson, A., & Brown, J. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12,* 246–269.

Deal, R. E., & Belcher, R. A. (1990). Reliability of children's ratings of vocal roughness. *Language, Speech, and Hearing Services in Schools, 21,* 68–71.

Deal, R., & Emanuel, F. (1978). Some waveform and spectral

features of vowel roughness. *Journal of Speech and Hearing Research, 21,* 250–264.

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

Ebel, R. (1951). Estimation of the reliability of ratings. *Psychometrica,16,* 407–424.

Emanuel, F., Lively, M. A., & McCoy, J. (1973). Spectral noise levels and roughness ratings for vowels produced by males and females. *Folia Phoniatrica, 25,* 110–120.

Emanuel, F., & Sansone, F. (1969). Some spectral features of "normal" and simulated "rough" vowels. *Folia Phoniatrica, 21,* 401–415.

Emanuel, F., & Scarinzi, A. (1979). Vocal register effects on vowel spectral noise and roughness: Findings for adult females. *Journal of Communication Disorders, 12,* 263–272.

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

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

Fukazawa, T., & El-Assuooty, A. (1988). A new index for evaluation of the turbulent noise in pathological voice. *Journal of the Acoustical Society of America, 83,* 1189–1193.

Gelfer, Marylou P. (1988). Perceptual attributes of voice: Development and use of rating scales. *Journal of Voice, 2,* 320–326.

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.

Gerratt, B. R., Till, J. A., Rosenbek, J. C., Wertz, R. T., & Boysen, A. E. (1991). Use and perceived value of perceptual and instrumental measures in dysarthria management. In C. A. Moore, K. M. Yorkston, & D. R. Beukelman (Eds.), *Dysarthria and apraxia of speech* (pp. 77–93). Baltimore, MD: Brookes.

Hammarberg, B., Fritzell, B., Gauffin, J., & Sundberg, J. (1986). Acoustic and perceptual analysis of vocal dysfunction. *Journal of Phonetics, 14,* 533–547.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngologica, 90,* 441–451.

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.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart and Winston.

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.

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *Journal of the Acoustical Society of America, 83,* 2361–2371.

Hirano, M. (1981). *Clinical examination of voice.* Vienna: Springer.

Hopkins, B. L., & Hermann, R. J. (1977). Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis, 10,* 2141–2150.

Isshiki, N., Okamura, H., Tanabe, M., & Morimoto, M. (1969). Differential diagnosis of hoarseness. *Folia Phoniatrica, 21,* 9–19.

Isshiki, N., & Takeuchi, Y. (1970). Factor analysis of hoarseness. *Studia Phonologica, 5,* 37–44.

Jensen, P. J. (1965, December). Adequacy of terminology for clinical judgment of voice quality deviation. *The Eye, Ear, Nose and Throat Monthly, 44,* 77–82.

Kane, M., & Wellen, C. J. (1985). Acoustical measurements and clinical judgments of vocal quality in children with vocal nodules. *Folia Phoniatrica, 37,* 53–57.

Kearns, K., & Simmons, N. (1988). Interobserver reliability and perceptual ratings: More than meets the ear. *Journal of Speech and Hearing Research, 31,* 131–136.

Kempster, G. (1984). *A multidimensional analysis of vocal quality in two dysphonic groups.* Unpublished doctoral dissertation, Northwestern University, Chicago.

Kempster, G. (1987, November). *A comparison of two scales for measuring vocal quality.* Paper presented at the Annual Meeting of the American Speech-Language-Hearing Association, New Orleans, Louisiana.

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

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.

Klich, R. J. (1982). Relationships of vowel characteristics to listener ratings of breathiness. *Journal of Speech and Hearing Research, 25,* 574–580.

Kojima, H., Gould, W., Lambiase, A., & Isshiki, N. (1980). Computer analysis of hoarseness. *Acta Oto-Laryngologica, 89,* 547–554.

Kreiman, J., Gerratt, B. R., & Berke, G. S. (1992). The multidimensional nature of pathologic vocal quality. Unpublished manuscript.

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33,* 103–115.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512–520.

Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication, 10,* 265–275.

Kreul, E. J., & Hecker, M. H. L. (1971). Description of the speech of patients with cancer of the vocal fold. Part II: Judgments of age and voice quality. *Journal of the Acoustical Society of America, 49,* 1283–1287.

Ladefoged, P., Maddieson, I., & Jackson, M. (1988). Investigating phonation types in different languages. In O. Fujimura (Ed.), *Vocal fold physiology: Voice production, mechanisms and functions* (pp. 297–317). New York: Raven Press.

Lively, M., & Emanuel, F. (1970). Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult females. *Journal of Speech and Hearing Research, 13,* 503–517.

Martin, R. R., Haroldson, S. K., & Triden, K. A. (1984). Stuttering and speech naturalness. *Journal of Speech and Hearing Research, 49,* 53–58.

Monsen, R. B. (1979). Acoustic qualities of phonation in young hearing-impaired children. *Journal of Speech and Hearing Research, 22,* 270–288.

Montague, J. C., & Hollien, H. (1973). Perceived voice quality disorders in Down's Syndrome children. *Journal of Communication Disorders, 6,* 76–87.

Moody, D. K., Montague, J., & Bradley, B. (1979). Preliminary validity and reliability data on the Wilson Voice Profile System. *Language, Speech, and Hearing Services in Schools, 10,* 231–240.

Moran, M. J., & Gilbert, H. R. (1984). Relation between voice profile ratings and aerodynamic and acoustic parameters. *Journal of Communication Disorders, 17,* 245–260.

Nieboer, G. L., De Graaf, T., & Schutte, H. K. (1988). Esophageal voice quality judgments by means of the semantic differential. *Journal of Phonetics, 16,* 417–436.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America, 85,* 913–925.

Prosek, R. A., Montgomery, A. A., Walden, B. E., & Hawkins, D. B. (1987). An evaluation of residue features as correlates of voice disorders. *Journal of Communication Disorders, 20,* 105–117.

Ptacek, P. H., & Sander, E. K. (1963). Breathiness and phonation length. *Journal of Speech and Hearing Disorders, 28,* 267–272.

Rees, M. (1958). Some variables affecting perceived harshness. *Journal of Speech and Hearing Research, 1,* 155–168.

Reich, A., & Lerman, J. (1978). Teflon laryngoplasty: An acoustical and perceptual study. *Journal of Speech and Hearing Disorders, 43,* 496–505.

Rossi, M., Pavlovic, C., & Espesser, R. (1990, November). *Reduc-*

ing context effects in the subjective evaluation of speech quality. Paper presented at the 120th Meeting of the Acoustical Society of America, San Diego, California.

Samar, V., & Metz, D. (1988). Criterion validity of speech intelligibility rating-scale procedures for the hearing-impaired population. *Journal of Speech and Hearing Research, 31,* 307–316.

Sansone, F., Jr., & Emanuel, F. (1970). Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult males. *Journal of Speech and Hearing Research, 13,* 489–502.

Sapir, S., & Aronson, A. E. (1985). Clinician reliability in rating voice improvement after laryngeal nerve section for spastic dysphonia. *Laryngoscope, 95,* 200–202.

Sapir, S., Aronson, A. E., & Thomas, J. E. (1986). Judgment of voice improvement after recurrent laryngeal nerve section for spastic dysphonia: Clinicians versus patients. *Annals of Otology, Rhinology, and Laryngology, 95,* 137–141.

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

Schlavetti, 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.

Sheard, C., Adams, R. D., & Davis, P. J. (1991). Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *Journal of Speech and Hearing Research, 34,* 285–293.

Sherman, D., & Linke, E. (1952). The influence of certain vowel types on degree of harsh voice quality. *Journal of Speech and Hearing Disorders, 17,* 401–408.

Shipp, T., & Huntington, D. (1965). Some acoustic and perceptual factors in acute-laryngitic hoarseness. *Journal of Speech and Hearing Disorders, 30,* 350–359.

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

Smith, B., Weinberg, B., Feth, L., & Horii, Y. (1978). Vocal roughness and jitter characteristics of vowels produced by esophageal speakers. *Journal of Speech and Hearing Research, 21,* 240–249.

Stoicheff, M. L., Clampi, A., Passi, J. E., & Fredrickson, J. M. (1983). The irradiated larynx and voice: A perceptual study. *Journal of Speech and Hearing Research, 26,* 482–485.

Stone, R. E., & Sharf, D. J. (1973). Vocal change associated with the use of atypical pitch and intensity levels. *Folia Phoniatrica, 25,* 91–103.

Takahashi, H., & Koike, Y. (1975). Some perceptual dimensions and acoustic correlates of pathological voices. *Acta Otolaryngologica* (Suppl. 338), 2–24.

Tinsley, H., & Weiss, D. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22,* 358–376.

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.

Wendahl, R. W. (1966a). Some parameters of auditory roughness. *Folia Phoniatrica, 18,* 26–32.

Wendahl, R. W. (1966b). Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness. *Folia Phoniatrica, 18,* 98–108.

Wendler, J., Doherty, E. T., & Hollien, H. (1980). Voice classification by means of long-term speech spectra. *Folia Phoniatrica, 32,* 51–60.

Whitehead, R. L., & Emanuel, F. W. (1974). Some spectrographic and perceptual features of vocal fry, abnormally rough, and modal register vowel phonations. *Journal of Communication Disorders, 7,* 305–319.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Wolfe, V., & Ratusnik, D. (1988). Acoustic and perceptual measurements of roughness influencing judgments of pitch. *Journal of Speech and Hearing Disorders, 53,* 15–22.

Yumoto, E., Sasaski, Y., & Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *Journal of Speech and Hearing Research, 27,* 2–6.

Received May 12, 1992
Accepted July 10, 1992

Contact author: Jody Kreiman, PhD, VAMC, West Los Angeles, Audiology and Speech Pathology (126), Wilshire and Sawtelle Boulevards, Los Angeles, CA 90073.

Appendix

LITERATURE REVIEWED

Key to abbreviations: (Training) E = Extensive; O = Orientation (see text for definitions). (Scale) EAI = equal-appearing interval scale; DME = direct magnitude estimation; paired comp = paired comparison task. (Interrater reliability) ICC = Intraclass correlation.
[a]Not significant.

Study	Raters	Training	Scale	Intrarater reliability	Interrater reliability
Anders et al. (1988)	34 naive, 79 expert	not reported	4 categories hoarseness	40/40 rerated. t = 0.02–0.57[a]	not reported
Arends et al. (1990)	10 expert	"a few sessions" (E)	5 pt. EAI 4 scales	not reported	Pearson's r = .62–.89
Arnold & Emanuel (1979)	11 grad students	not reported	5 pt. EAI roughness	20/200 rerated. 55–100% ± 1 scale value	72–95% ± 1 scale value
Askenfelt & Hammarberg (1986)	6 expert	4 years min. (E)	5 pt. EAI 12 scales	8/41 rerated. Pearson's r = .86–.98; 86% ± 1 scale value	not reported
Bassich & Ludlow (1986)	4 grad students	8 hours (E)	*n* pt. EAI (varied with scale) 13 scales	4/21 rerated. 100% ± 1 scale value 75% exact agreement	Mean ICC = .8 across all scales for normal voices; = .71 for pathological voices
Brancewicz & Reich (1989)	10 expert	9 practice trials (O)	100 mm visual analog scale nasality	26/120 rerated. Pearson's r = .84–.90	ICC(3,10) = .66–.75
Coleman (1969)	32 psychology students	none	paired comp; 7 pt. EAI roughness	not reported	not reported
Coleman (1971)	15 undergrads	none	anchored DME roughness	not reported	not reported
Coleman & Wendahl (1967)	18 undergrads	not reported	paired comp roughness	not reported	not reported
Cullinan et al. (1963)	15–20 undergrads in each of 7 groups	samples of extremes provided (O)	5, 7, 9 pt. EAI; anchored DME severity of stuttering	27/27 rerated. "ANOVA technique"; reliability of 1 judge making 1 rating = .74–.80 across tasks	"ANOVA technique"; reliability of 1 judge making 1 rating = .48–.75 across tasks
Darley et al. (1969)	3 authors	discussed ratings during study (O)	7 pt. EAI 38 scales	30/212 rerated. 80–95% ± 1 scale value	84% ± 1 scale value
Deal & Belcher (1990)	30 children 1, 3, 5 grade	4 anchor stimuli; discussed 1st 5 ratings (O)	5 pt. EAI roughness	15/40 rerated. Pearson's r = .37–.95; 60–100% ± 1 scale value	Pearson's r = .56–.98; 75–100% ± 1 scale value
Emanuel et al. (1973)	11 grad students	20 item training tape heard 2× (O)	5 pt. EAI roughness	100/400 rerated. Pearson's r for medians of 1st & 2nd ratings = .98; > = 95% ± 1 scale value	> = 92% ± 1 scale value
Emanuel & Scarinzi (1979)	15 grad students	not reported	5 pt. EAI roughness	20/90 rerated. Pearson's r = .59–.95; 94% ± 1 scale value; 59% exact agreement	ICC adjusted for between-rater variance = .83 reliability of average rating = .98

LITERATURE REVIEWED

Key to abbreviations: (Training) E = Extensive; O = Orientation (see text for definitions). (Scale) EAI = equal-appearing interval scale; DME = direct magnitude estimation; paired comp = paired comparison task. (Interrater reliability) ICC = Intraclass correlation.
[a]Not significant.

Study	Raters	Training	Scale	Intrarater reliability	Interrater reliability
Emanuel & Smith (1974)	8 grad students	20 practice trials (O)	anchored DME roughness	20/80 rerated. Pearson's *r* for mean of 1st/2nd ratings = .80	not reported
Fritzell et al. (1986)	21 expert	not reported	5 pt. EAI breathiness	not reported	not reported
Gelfer (1988)	19 expert, 18 naive	none	9 pt. EAI 22 scales	not reported	Kendall's coefficient of concordance = .17–.69 for experts; = .14–.57 for naive
Hammarberg et al. (1986)	11 expert	terms discussed/ defined (O)	5 pt. EAI 25 scales	3/22 rerated. Pearson's *r* = .76–.98; 86% ± 1 scale value	not reported
Hammarberg et al. (1980)	14 expert	not reported	5 pt. EAI 28 scales	3/17 rerated. Pearson's *r* for means of 2 ratings on 28 scales = .93–.97	not reported
Hammarberg et al. (1984)	10 expert	not reported	5 pt. EAI 11 scales	6/16 rerated. Pearson's *r* = .68–.97	not reported
Heiberger & Horii (1982)	A: 25	not reported	anchored DME roughness	unspec. number rerated. Pearson's *r* for medians = .78–.96;	not reported
	B: 52 students (grad & undergrad)	2 demonstration trials; 20 practice trials (O)	paired comp roughness	20/138 rerated. Pearson's *r* = .96; exact agreement = 68–93%	not reported
Isshiki et al. (1969)	5 expert	not reported	4 pt. EAI 4 scales	not reported	not reported
Isshiki & Takeuchi (1970)	34 students, 6 expert	3 example stimuli repeated 2× (O)	7 pt. EAI 17 scales	not reported	not reported
Kane & Wellen (1985)	1 expert	not reported	7 pt. EAI severity	not reported	N/A (single rater)
Kearns & Simmons (1988)	5 expert	3 1-hour sessions (E)	7 pt. EAI 40 scales	not reported	82% ± 1 scale value overall; 68% ± 1 scale value for voices not rated as normal
Klatt & Klatt (1990)	5 expert	14 practice trials (O)	6 pt. EAI change in breathiness	12/12 repeated 5×. "rarely" exceeded ± 1 scale value	not reported
Klich (1982)	27 undergrads	10 minutes (O)	7 pt. EAI breathiness	40/40 rerated. ANOVA: no sig. presentation effect	ICC > .92 for all voice samples
Kojima et al. (1980)	5 expert	not reported	4 pt. EAI hoarseness	not reported	not reported

LITERATURE REVIEWED

Key to abbreviations: (Training) E = Extensive; O = Orientation (see text for definitions). (Scale) EAI = equal-appearing interval scale; DME = direct magnitude estimation; paired comp = paired comparison task. (Interrater reliability) ICC = Intraclass correlation.
[a]Not significant.

Study	Raters	Training	Scale	Intrarater reliability	Interrater reliability
Kreul & Hecker (1971)	22 undergrads	none	9 pt. EAI 5 scales	10/50 rerated. Kendall rank-order coefficients = .4–.95 for path voices; = .4–1.0 for normal voices	Friedman 2-way ANOVA: all F ratios significant
Lively & Emanuel (1970)	11 grad students	samples of scale extremes played 2× before each block of trials (O)	5 pt. EAI roughness	50/200 repeated. Pearson's *r* for medians of 1st/2nd ratings = .98; 96% ± 1 scale value	92% ± 1 scale value
Martin et al. (1984)	30 undergrads	none	9 pt. EAI naturalness	30/30 rerated. 85% ± 1 scale value	ICC: reliability of average rating = .98; reliability of single rater = .57–.74; 75% ± 1 scale value
Monsen (1979)	10 expert	short practice tape (O)	4 pt. EAI adequacy of voice quality	100% rerated. Pearson's *r* = .66–.74	not reported
Montague & Hollien (1973)	31 naive, 17 expert	definitions; example stimuli; 25 practice items (E)	7 pt. EAI 3 scales	40/40 rerated. "ANOVA technique" correlation = .70–.94	"ANOVA technique" correlation = .35–.96
Moody et al. (1979)	11 grad students	workshop plus 1 training session (E)	Wilson voice profile (7 scales)	5/11 judges; 20/40 rerated. Phi correlation = −.11–.96	Phi correlation = −.18–.97
Moran & Gilbert (1984)	4 expert, 3 grad students	not reported	Wilson voice profile (4 scales)	5/25 rerated. Pearson's *r* = .85–.95	"ANOVA technique" coefficient = .77–.93
Nieboer et al. (1988)	34 undergrad, 51 grad students	15 example stimuli; 2 practice trials (O)	7 pt. EAI 13 scales	not reported	Mean Pearson's *r* = .18–.66; ICC: reliability of average rating = .94–.99
Prosek et al. (1987)	9 expert	90 practice trials (E)	7 pt. EAI severity	90/90 rerated. Pearson's *r* = .86–.93	Cronbach's coefficient of generalizability = .82
Ptacek & Sander (1963)	8 expert	anchor stimuli, 16 practice trials (O)	7 pt. EAI breathiness	20/240 rerated. Pearson's *r* = .73	not reported
Rees (1958)	32 grad students	definitions; sample of each scale value repeated 3×; 30 practice (O)	7 pt. EAI harshness	100/1080 rerated. Pearson's *r* = .90 Difference between 2 means sig. by *t*-test	Q = .79
Reich & Lerman (1978)	20 grad students	trained to use EAI scale (O)	5 pt. EAI hoarseness 7 pt. EAI rough, pleasant	unspec. number rerated. Pearson's *r* = .73–.83	not reported
Samar & Metz (1988)	4 panels of 3 experts	not reported	5 pt. EAI intelligibility	84/84 rerated. Pearson's *r* = .975	Pearson's *r* = .95

LITERATURE REVIEWED

Key to abbreviations: (Training) E = Extensive; O = Orientation (see text for definitions). (Scale) EAI = equal-appearing interval scale; DME = direct magnitude estimation; paired comp = paired comparison task. (Interrater reliability) ICC = Intraclass correlation.
[a]Not significant.

Study	Raters	Training	Scale	Intrarater reliability	Interrater reliability
Sansone & Emanuel (1970)	11 grad students	4 scale extremes played several times (O)	5 pt. EAI roughness	50/200 rerated. Pearson's *r* for medians of 1st/2nd ratings = .96	>95% ± 1 scale value for 46 pairs of raters; lowest value = 80%
Sapir & Aronson (1985)	3 expert (inc. 2 authors)	not reported	8 pt. EAI severity of dysphonia	224/224 rerated. Pearson's *r* = .94–.96	Pearson's *r* = .83–.87
Sapir et al. (1986)	3 authors	not reported	7 pt. categorical scale	25/25 rerated. Pearson's *r* = .87–.95	Pearson's *r* = .80–.88
Schiavetti et al. (1981)	40 grad students	5 example stimuli (O)	7 pt EAI; anchored DME intelligibility	not reported	not reported
Schiavetti et al. (1983)	45 grad students	3 example stimuli (O)	7 pt EAI; anchored & unanchored DME severity of stuttering	not reported	ICC: reliability of mean ratings = .98 (EAI ratings); = .96 (anchored DME); = .97 (unanchored DME). Reliability of single rating = .75 (EAI scale); = .61 (anchored DME); = .65 (unanchored DME)
Sherman & Linke (1952)	35 students	examples of scale extremes played several times (O)	7 pt EAI harshness	30/90 rerated. Pearson's *r* = .97	mean Q = .77
Shipp & Huntington (1965)	4 expert	none	8 pt. EAI hoarseness, breathiness	not reported	Pearson's *r* = .17–.78
Smith et al. (1978)	9 grad students; 8 expert	stim. tape played 1× (O)	paired comp roughness	36/36 rerated. % agreement = .69–.89	not reported
Stoicheff et al. (1983)	8 expert	examples of scale extremes played several times (O)	7 pt. EAI dysphonia	50/150 rerated. "test-retest reliability" = .91	ICC(2,1) = .79
Stone & Sharf (1973)	5 grad students	60 actual stimuli (E)	7 pt. EAI "extent of change"	15/450 rerated. Pearson's *r* = .75	58% ± 1 scale value; 86% ± 2 scale values
Toner & Emanuel (1989)	20 grad students	10 practice trials (O)	5 pt. EAI; anchored DME roughness	60/60 rerated. mean Pearson's *r* = .76 (EAI); = .775 (DME)	ICC: reliability of mean rating = .975 (EAI); = .965 (DME). Reliability of single rating = .685 (EAI); = .575 (DME)
Wendahl (1966a)	461 undergrads	10 practice pairs (O)	paired comp roughness	3/15 rerated. .93 "by test-retest procedures"	not reported

LITERATURE REVIEWED

Key to abbreviations: (Training) E = Extensive; O = Orientation (see text for definitions). (Scale) EAI = equal-appearing interval scale; DME = direct magnitude estimation; paired comp = paired comparison task. (Interrater reliability) ICC = Intraclass correlation.
[a]Not significant.

Study	Raters	Training	Scale	Intrarater reliability	Interrater reliability
Wendahl (1966b)	97 undergrads	20 practice pairs (O)	paired comp roughness	not reported	not reported
Whitehead & Emanuel (1974)	11 grad students	15 practice trials (O)	5 pt. EAI roughness	50/300 rerated. Pearson's *r* for medians of 1st/2nd ratings = .98	> = 90% ± 1 scale value
Wolfe & Ratusnik (1988)	8 grad students	15 sample stimuli presented (O)	7 pt. EAI roughness	40/102 rerated. 89% ± 1 scale value	Cronbach's alpha = .95
Yumoto et al. (1984)	8 expert	Standard training tape (O)	4 pt. EAI hoarseness	2 judges rerated 87/87. Results not reported	Spearman's rho = .51–.79