# Validity of rating scale measures of voice quality

Jody Kreiman[a] and Bruce R. Gerratt
*Division of Head and Neck Surgery, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, California 90095-1794*

(Received 2 June 1997; revised 10 March 1998; accepted 5 May 1998)

The validity of perceptual measures of vocal quality has been neglected in studies of voice, which focus more commonly on rater reliability. Validity depends in part on reliability, because an unreliable test does not measure what it is intended to measure. However, traditional measures of rating reliability only partially represent interrater agreement, because they cannot reflect variations or patterns of agreement for specific voice samples. In this paper the likelihood that two raters would agree in their ratings of a single voice is examined, for each voice in five previously gathered data sets. Results do not support the continued assumption that traditional rating procedures produce useful indices of listeners' perceptions. Listeners agreed very poorly in the midrange of scales for breathiness and roughness, and mean ratings in the midrange of such scales did not represent the extent to which a voice possesses a quality, but served only to indicate that listeners disagreed. Techniques like analysis by synthesis or judgment of similarity avoid decomposing quality into constituent dimensions, and do not require a listener to compare an external stimulus to an unstable internal representation, thus decreasing the error in measures of quality. Modeling individual differences in perception can increase the variance accounted for in models of quality, further reducing the error in perceptual measures. Thus such techniques may provide valid alternatives to current approaches.  [S0001-4966(98)04708-0]

PACS numbers: 43.71.Bp, 43.71.Gv [WS]

## INTRODUCTION

Measurement validity—the extent to which a scale or instrument measures what it is intended to measure—is a central concern in the development and evaluation of any measurement system. Measures that are weakly or variably related to a concept are not useful indices of that concept (e.g., Kerlinger, 1973; Carmines and Zeller, 1979; Crocker and Algina, 1986). This paper examines the validity of traditional rating protocols that use scales like breathiness, roughness, hoarseness, or harshness as measures of vocal quality. Although a few authors have expressed doubt about the validity of such scales (Jensen, 1965; Perkins, 1971), issues of the validity of perceptual measures are typically neglected in studies of voice, which focus more commonly on rater reliability (see Kreiman *et al.*, 1993; Kreiman and Gerratt, 1998a, for review).

The validity of traditional protocols for rating vocal quality is important in part because perceptual methods are often used clinically to evaluate vocal disorders (Gerratt *et al.*, 1991). Perceptual ratings are also used to validate acoustic and other instrumental or ''objective'' measures of voice (e.g., Fritzell *et al.*, 1986; Hillenbrand *et al.*, 1994; de Krom, 1995; Martin *et al.*, 1995; Sodersten *et al.*, 1995). Voice quality is an interaction between an acoustic voice stimulus and a listener; the acoustic signal itself does not possess vocal quality, it evokes it in the listener. For this reason, acoustic measures are meaningful primarily to the extent that they correspond to what listeners hear (Gerratt and Kreiman, 1995; Kreiman and Gerratt, 1996).

Finally, validity is important because measurement protocols imply a model of the construct being measured. Therefore studies of the validity of rating scales for voice also serve to test the adequacy of the implied model of vocal quality. Because vocal quality is a perceptual response to an acoustic signal, rating protocols for vocal quality comprise a set of claims about both signals and listeners. When vocal quality is measured by means of ratings on scales for particular aspects of quality, this implies that the overall impression a listener receives from a voice can be decomposed into several perceptually distinct aspects corresponding to various terms such as breathiness and roughness. It is assumed that individual listeners can focus their attention on these different aspects of the stimuli, and can make the judgments required. Finally, and crucially, it is assumed that characteristics of the measurement tool remain constant across listeners and voices, so that different listeners use the scales in the same way and measurements of different voices can be meaningfully compared. This implies that quality is fairly constant across listeners, so that voice quality may be treated as an attribute of the voice signal itself, rather than as the product of a listener's perception. That is, traditional protocols for assessing voice quality necessarily treat individual differences in perception as noise, and do not model them explicitly. Because voice signals provide listeners with large amounts of information (for example, about the identity and physical, mental, and emotional state of the speaker; see Kreiman, 1997, for review), such claims about the perceptual process have interest beyond their clinical applications, and the validity or invalidity of voice assessment protocols has important implications for models of auditory pattern recognition and perception of complex signals in general.

[a]Electronic mail: jkreiman@ucla.edu

### A. Approaches to the study of scale validity

Quality is traditionally defined as ''that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar'' (ANSI Standard S1.1.12.9, p. 45, 1960; cf. Helmholtz, 1885). However, most authors avoid studying overall vocal quality, preferring instead to focus on single dimensions or specific aspects (for example, breathiness, harshness, or strain). One way to motivate scales for specific aspects of quality is with reference to overall quality: Individual scales or sets of scales may be valid to the extent that as a group they measure overall quality. For example, studies using multidimensional scaling (Murry *et al.*, 1977; Kreiman *et al.*, 1990, 1992, 1994; Kempster *et al.*, 1991; Kreiman and Gerratt, 1996) attempt to identify the perceptual dimensions that underlie listeners' judgments of the overall similarity of pairs of voices. Unfortunately, traditional scales have not generally emerged as perceptual dimensions from these studies, which in consequence provide little support for the validity of such scales as measures of overall vocal quality.

However, for clinical purposes, it may not be necessary to model overall quality in detail. Instead, it may be adequate to focus quality assessment on a limited number of clinically significant perceptual dimensions, while neglecting other irrelevant aspects of vocal quality. In this case, motivating and defining individual scales remain critical aspects of scale development, to specify what is being measured, to justify why those aspects of voice (and not others) are of interest, and to clarify the relationship among different scales. Few studies have investigated these issues. Individual scales are typically validated by appeals to consensual or face validity (Silverman, 1977; Allen and Yen, 1979), or by reference to their association with purported acoustic, aerodynamic, and/or physiological correlates. However, because appeals to face or consensual validity do not involve empirical examination of evidence or reference to theory, they are of little use in the assessment of measurement systems. Thus the literature on pathologic voice quality does not provide convincing or consistent evidence for the validity of traditional scales for vocal quality. (See, e.g., Colton and Estill, 1981; Kreiman and Gerratt, 1998a, for extensive review of these issues.)

### B. Reliability as a tool for assessing validity

Although the validity of traditional scales for voice quality has never been formally established, little evidence exists that such scales are invalid, largely due to lack of research. However, because the validity of perceptual measures depends on characteristics of both listeners and stimuli, validity is partially determined by reliability. That is, because quality is a function of both listeners and stimuli, an unreliable test cannot be a valid measure of quality, because it does not model listener behavior accurately (e.g., Young and Downs, 1968; Cone, 1977; Ventry and Schiavetti, 1980; Suen and Ary, 1989). Thus evidence about patterns of agreement and disagreement among listeners in their use of quality scales can provide evidence for or against the validity of the scales. If listeners cannot agree when making the required judgments, the critical assumption of listener equivalence is violated, and the validity of traditional protocols for quality assessment is not supported.

In this study we combined traditional approaches to reliability with new analyses designed to examine patterns of agreement and disagreement among listeners that bear upon issues of measurement validity. Conventional statistical analyses of reliability do not provide enough information to answer questions about scale validity. Such analyses produce a single number representing the overall reliability of a set of ratings, across all the voices and listeners in a study. This conventional approach derives from the literature on psychological test construction (Allen and Yen, 1979; Crocker and Algina, 1986), with listeners substituted for test items and voices substituted for examinees or subjects. Errors are assumed to be random in this model, so averaging together scores from a large number of raters will give the best estimate of the ''true'' score for a voice on a scale, and the mean rating approaches the true score as the number of raters increases. Thus reliability in classic theory is a function of both the average interrater correlation and the number of raters in a study.

In this traditional framework, reliability implies that another sample of listeners would produce the same mean ratings for the same test voices, but does not necessarily inform us of how the subjects would agree in their ratings of a new set of voices. Conventional reliability statistics are not informative about many other important aspects of listener performance. For example, they cannot indicate agreement for specific voice samples (Young and Downs, 1968), and they cannot capture information about systematic variations in reliability or agreement across raters or parts of the rating scale. Patterns of agreement and disagreement among listeners may provide evidence about the perceptual processes that underlie judgments of vocal quality. Such evidence may be helpful in establishing the validity of different scales for vocal quality, and may help determine why measurement protocols may fail. Finally, because the validity of measurement systems ultimately depends on the success of the underlying perceptual model, such detailed knowledge about listener agreement may guide the design of future protocols for quality assessment.

## I. METHOD

To determine if patterns of rater agreement support rating scale validity, we reevaluated existing data from experiments using unidimensional scales for different traditional vocal qualities or ratings of the similarity of pairs of voices. Data were drawn from four previously published studies (Kreiman *et al.*, 1993; Kreiman *et al.*, 1994; Rabinov *et al.*, 1995; Kreiman and Gerratt, 1996) and one unpublished study (Chhetri, 1997). Two of these studies (Kreiman *et al.*, 1993; Rabinov *et al.*, 1995) were specifically concerned with issues of rating reliability. Listeners in these studies judged the roughness of samples of pathologic voices, and recorded their responses on equal-appearing interval (EAI) or visual analog (VA) scales.

Three other studies (Kreiman *et al.*, 1994; Kreiman and Gerratt, 1996; Chhetri, 1997) used EAI scales to address

TABLE I. Characteristics of the data sets.[a]

| Study | Raters | Speakers | Scale(s) | Rating task |
|---|---|---|---|---|
| Kreiman *et al.* (1993) | 30 expert[b] | 30 (22 disordered, 8 normal) | seven-point EAI<br>100 mm VA | Judgments of roughness<br>Judgments of roughness |
| Kreiman *et al.* (1994) | 5 expert | 18 disordered | seven-point EAI | Paired comparison: Dissimilarity of pairs of voices with respect to breathiness |
| | | | seven-point EAI | Paired comparison: Dissimilarity of pairs of voices with respect to roughness |
| | 8 expert | 18 disordered | seven-point EAI | Judgments of breathiness |
| | | 18 disordered | seven-point EAI | Judgments of roughness |
| Rabinov *et al.* (1995) | 10 expert | 50 disordered | 75 mm VA | Judgments of roughness |
| Kreiman and Gerratt (1996) | 8 expert | 80 disordered (males) | seven-point EAI | Paired comparison: Overall dissimilarity of pairs of voices |
| | | 80 disordered (females) | seven-point EAI | Paired comparison: Overall dissimilarity of pairs of voices |
| Chhetri (1997) | 9 expert | 32 disordered (pre/post operative) | seven-point EAI | Judgments of severity of pathology |

[a]EAI=equal-appearing interval scale; VA=visual analog scale.
[b]Data from experiments 1 and 2 have been combined.

more general issues of the perception of pathologic voice quality. Two groups of raters participated in the studies reported in Kreiman *et al.* (1994). The first group judged the similarity of pairs of voices with respect to breathiness or roughness. The second directly rated the breathiness or roughness of the individual voices. Raters in Kreiman and Gerratt (1996) judged the overall similarity of pairs of voices. Raters in Chhetri (1997) rated the severity of vocal pathology for samples of voices gathered pre- and post-operatively. Further details are given in Table I.

For our current purposes, we calculated several traditional measures of overall intra- and interrater reliability and agreement for each data set. We also examined an additional measure, the empirical likelihood that two raters would agree in their ratings of a specific voice, for each voice in the data sets. These finer-grained analyses assessed how likely it was that individual raters would agree with one another for specific voice stimuli, rather than how well the population of raters agreed on average or how well the averaged data estimated the ''true mean rating.'' This approach also allowed us to capture detailed information about variations in agreement across voices and parts of the rating scale. Similar analyses of intrarater agreement were undertaken, comparing the first and second rating of a voice by a single listener, to determine whether individuals were more self-consistent for some voices than for others.

To simplify comparisons among studies using VA and EAI scales, differences between pairs of ratings on the VA scale were converted from mm to ''scale value equivalents.'' For example, a 100-mm VA scale was divided into seven intervals of 14.3 mm each, analogous to a seven point EAI scale. Pairs of ratings within 7.2 mm of each other were considered to agree exactly; ratings that differed by 21.5 mm (7.2+14.3) were considered to be within 1 scale value of each other, and so on. For the 75-mm VA scale, a scale interval was defined as 10.7 mm. Thus ratings differing by 5.4 mm or less were considered to agree exactly, and ratings differing by 16.1 mm or less were considered within 1 scale value of each other. Differences in mm and in scale value equivalents were highly correlated (data from Kreiman *et al.*, 1993: $r=0.98$; data from Rabinov *et al.*, 1995: $r=0.98$).

## II. RESULTS

### A. Intrarater agreement: How self-consistent were listeners?

Traditional analyses of intrarater agreement examine overall levels of listener self-consistency, summed across voices. In contrast, Table II shows the likelihood that a given voice would be rerated consistently, calculated across listeners. Numbers in this table represent the likelihood that a single rerating of a single voice would agree with the first rating by some amount (for example, exactly or within one scale value).

Listeners produced the same value when rerating a stimulus for 32%–50% of trials, depending on the study. Pooled across studies, a second rating agreed exactly with the first for 38.6% of repeated trials, and 76.8% of repeated ratings agreed with the first within 1 scale value. In comparison, across studies traditional test–retest agreement (calculated across voices for each listener, and then averaged across listeners) ranged from 72.5%–92.0% of ratings within $\pm 1$ scale value.

TABLE II. Likelihood that a single rerating of a single voice would differ from the first rating by a given amount.[a]

| Study/Scale | $N$[b] | Exact agreement | Ratings differ by 1 scale value | Ratings differ by 2 scale values | Ratings differ by 3 or more scale values |
|---|---|---|---|---|---|
| Kreiman *et al.* (1993) (EAI/Roughness) | 900 | 44.9% | 38.6% | 12.0% | 4.6% |
| Kreiman *et al.* (1993) (VA/Roughness) | 900 | 48.8% | 33.6% | 11.8% | 5.9% |
| Kreiman *et al.* (1994) (EAI/Roughness) | 144 | 38.9% | 43.8% | 11.1% | 6.3% |
| Kreiman *et al.* (1994) (EAI/Breathiness) | 144 | 47.2% | 38.2% | 11.1% | 3.5% |
| Kreiman *et al.* (1994) (Dissimilarity/Roughness) | 765 | 36.5% | 36.9% | 15.3% | 11.4% |
| Kreiman *et al.* (1994) (Dissimilarity/Breathiness) | 765 | 32.0% | 40.5% | 16.3% | 11.1% |
| Rabinov *et al.* (1995) (VA/Roughness) | 500 | 44.2% | 35.4% | 13.8% | 6.6% |
| Kreiman and Gerratt (1996) (Dissimilarity/Male voices) | 5056 | 36.6% | 38.1% | 16.4% | 9.0% |
| Kreiman and Gerratt (1996) (Dissimilarity/Female voices) | 5056 | 38.4% | 38.9% | 15.7% | 7.0% |
| Chhetri (1997) (EAI/Severity) | 66 | 50.0% | 42.4% | 6.1% | 1.5% |
| Pooled data | 14 296 | 38.6% | 38.2% | 15.3% | 7.8% |

[a]EAI=equal-appearing interval scale; VA=visual analog scale.
[b]$N$=(number of listeners)×(number of repeated trials/listener). Differences between VA ratings were converted to scale equivalents, as described in the text.

Figure 1 shows how test–retest agreement varied across listeners and voices for EAI ratings of roughness [Fig. 1(a); Kreiman *et al.*, 1993], VA ratings of roughness [Fig. 1(b); Rabinov *et al.*, 1995], and ratings of overall similarity of pairs of voices [Fig 1(c); Kreiman and Gerratt, 1996]. In this figure, each point represents a single stimulus presented to a single rater; the difference between the first and second rating that voice received from that rater is plotted against the mean of that individual's two ratings for that voice. Because agreement is plotted against the mean rating for a given voice, the probability of agreement must be high when mean ratings are near scale end points. However, agreement in the midrange of a scale may be high (if a listener consistently rates voices as moderately pathologic) or low (if a listener responds with a large scale value on one occasion and a small value on another occasion).

As Fig. 1 shows, for all three tasks individual listeners were often self-consistent in their use of these rating scales. In particular, individual listeners appeared to maintain stable standards for the midrange of a scale, so that many voices received ratings of 3, 4, or 5 both times they were rated.

Figure 2 summarizes the data from Fig. 1(a) and (b) (Kreiman *et al.*, 1993) by showing the overall probability of test-retest agreement for individual voices. Levels of self-consistency for individual stimuli were quite high overall, with most values above 0.8. This suggests that individual listeners are able to make reasonably consistent judgments of traditional vocal qualities.

## B. Pairwise agreement among raters

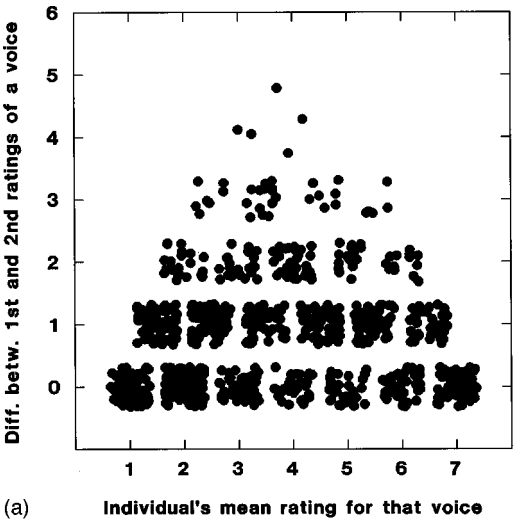
### 1. Overall likelihood of interrater agreement

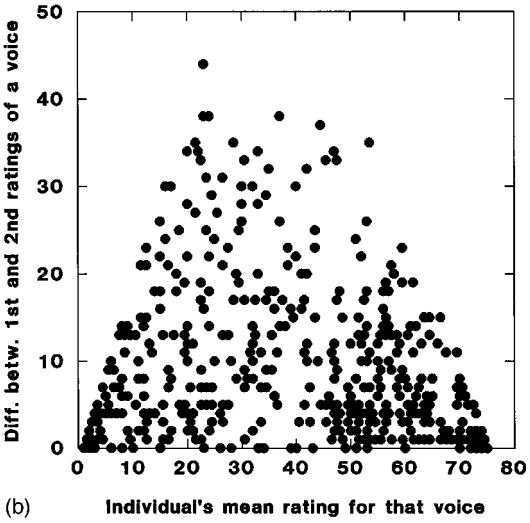
Measures of interrater agreement, like measures of intrarater agreement, usually sum across voices to provide a single measure of rater concordance. In contrast, the present analyses sum across listeners to provide a measure of the likelihood that two raters will agree in their ratings of individual stimuli. Table III lists the overall likelihood of raters agreeing exactly, within one scale value, and so on, in their ratings of a single voice or pair of voices. Across studies, pairs of listeners agreed exactly for 26.7% of trials (versus 38.6% test–retest agreement). Ratings differed by 1 scale value or less for 63.7% of trials (versus 76.8% test–retest agreement). Gross disagreements (ratings differing by 3 or more scale values on a seven-point scale) occurred for a total of 15.6% of trials (cf. Mackey *et al.*, 1997, who reported similar values for ratings of speech naturalness).

### 2. Patterns of interrater agreement for traditional rating scales

Patterns of interrater agreement depended on the listening task. For ratings of breathiness, roughness, and severity, interrater agreement levels were consistently poor in the midrange of the rating scales. Figure 3 shows the likelihood of two raters agreeing exactly [Fig. 3(a) and (c)] or within 1 scale value [Fig. 3(b) and (d)] for each voice in two representative data sets. Because we were interested in the extent to which mean ratings represent the underlying raw data, the probability of agreement is plotted against the group mean rating for each voice. As above, agreement near scale end points must be high in these plots, because average values can only approach scale end points when listeners agree. However, average values away from scale end points can result from agreement that voices are moderately pathologic, or from disagreement about the extent of pathology.

In the present data, the likelihood that two raters would agree exactly for voices with mean ratings between 2.5 and 5.5 on a seven-point EAI scale averaged 0.21 (range $=0.19–0.24$; chance agreement for independent ratings on a seven-point scale$=0.14$), despite the fact that individual listeners were self-consistent in the same scale range. The likelihood of agreement within 1 scale value averaged 0.57 (range$=0.50–0.61$; chance$=0.39$). Although these values



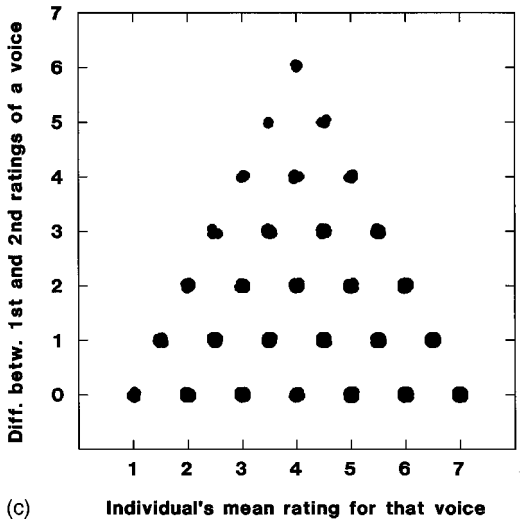
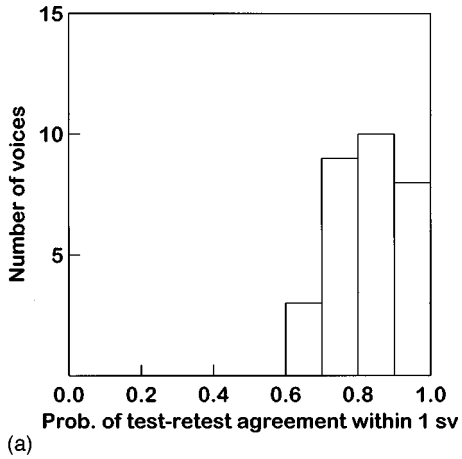




FIG. 1. Test–retest agreement for individual stimuli. A value of 0 on the *y* axis indicates that a rater gave that voice the same score both times it was rated (i.e., the difference between the first and second ratings was 0); a value of 1 indicates that the first and second ratings differed by 1; and so on. Values on the *x* axis represent the mean of a single individual's two ratings of that stimulus. Points have been jittered slightly to show overlapping values. (a) Test–retest agreement for EAI ratings of roughness (Kreiman *et al.*, 1993). (b) Test–retest agreement for visual analog ratings of roughness (Rabinov *et al.*, 1995). (c) Test–retest agreement for similarity ratings (Kreiman and Gerratt, 1996).



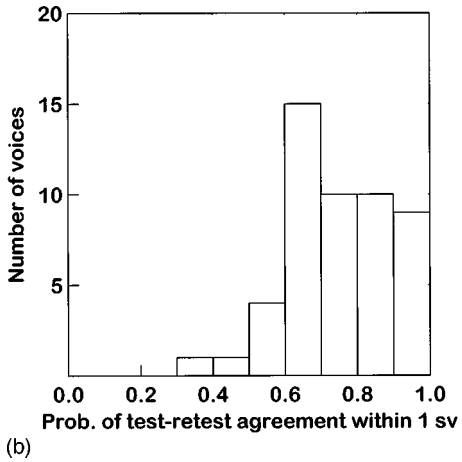




FIG. 2. The probability of observing test–retest agreement within one scale value (or scale value equivalent) for individual voices. Each column shows the number of voices for which overall test-retest agreement occurred with the given likelihood. (a) EAI ratings of roughness (Kreiman *et al.*, 1993). (b) Visual analog ratings of roughness (Rabinov *et al.*, 1995).

significantly exceed chance levels of agreement (one-sample t tests; $p<0.05$), they are very low. Further, across all the data examined here, we did not find a single voice that listeners consistently agreed was moderately deviant in quality. Thus the present data suggest that mean ratings in the midrange of the scale do not arise from a consensus among raters that the voice is moderately deviant, but indicate instead that raters disagreed about the extent of deviation on that scale.

Because a significant statistical result does not necessarily indicate the size of the effect (especially when *n* is large, as it is here), we also calculated the amount of variance in quality ratings that is attributable to differences among voices. Variance accounted for was estimated by one-way analyses of variance for the different sets of ratings (e.g., Young, 1993). The independent variable in these analyses was the voice being rated, and the dependent variable was the rating received; the error term reflects all other sources of variability in quality ratings, including (but not limited to) interrater variability and random error. Because agreement near scale end points is in part artifactual, analyses included only voices with mean ratings between 2.5 and 5.5 (inclusive).

Results are given in Table IV. Differences among voices

TABLE III. Pairwise agreement among raters.[a]

| Study/Scale | $N$[b] | Exact agreement | Ratings differ by 1 scale value | Ratings differ by 2 scale values | Ratings differ by 3 or more scale values |
|---|---|---|---|---|---|
| Kreiman *et al.* (1993) (EAI/Roughness) | 26 100 | 31.7% | 40.2% | 17.7% | 10.4% |
| Kreiman *et al.* (1994) (EAI/Roughness) | 1008 | 20.7% | 35.5% | 22.2% | 21.5% |
| Kreiman *et al.* (1994) (EAI/Breathiness) | 1008 | 25.4% | 41.7% | 21.0% | 11.9% |
| Kreiman *et al.* (1994) (Dissimilarity/Roughness) | 3060 | 24.4% | 34.9% | 21.3% | 19.3% |
| Kreiman *et al.* (1994) (Dissimilarity/Breathiness) | 3060 | 20.9% | 31.7% | 20.4% | 27.0% |
| Kreiman and Gerratt (1996) (Dissimilarity/Male voices) | 88 480 | 24.9% | 35.2% | 21.4% | 18.5% |
| Kreiman and Gerratt (1996) (Dissimilarity/Female voices) | 88 480 | 26.2% | 39.0% | 21.5% | 13.3% |
| Kreiman *et al.* (1993) (VA/Roughness) | 26 100 | 30.6% | 33.3% | 18.7% | 17.4% |
| Rabinov *et al.* (1995) (VA/Roughness) | 4500 | 27.0% | 37.9% | 17.6% | 17.4% |
| Chhetri (1997) (EAI/Severity) | 1152 | 32.2% | 35.3% | 22.3% | 10.2% |
| Pooled data | 242 948 | 26.7% | 37.0% | 20.7% | 15.6% |

[a]EAI=equal-appearing interval scale; VA=visual analog scale.
[b]$N$=(number of possible pairs of listeners)×(number of stimuli). Differences between VA ratings were converted to scale value equivalents, as described in the text.

with average ratings in the ''moderately pathologic'' range accounted for an average of 32% of the variance in ratings (range=22%–42%). In other words, for the midrange of the scales examined here, on average more than 60% (and as much as 78%) of the variance in ratings of voices was due to factors other than differences among voices in the quality being rated.

#### *3. Patterns of interrater agreement for similarity ratings*

The pattern of pairwise agreement among listeners for ratings of the similarity of pairs of voices was different than that for ratings of roughness, breathiness, and severity. Although agreement levels varied substantially across voice pairs, perfect or near-perfect agreement among raters was more common for ratings of overall similarity [Kreiman and Gerratt, 1996; Fig. 4(a) and (b)] than for ratings of traditional qualities (where the likelihood of two raters agreeing perfectly never exceeded 0.8). Good agreement occurred across the entire scale. In particular, listeners did agree that some pairs of voices were moderately similar.

Patterns of agreement for ratings of the similarity of voices with respect to specific vocal qualities [Fig. 4(c) and (d); Kreiman *et al.*, 1994] shared characteristics of both similarity ratings and ratings of specific qualities. Although levels of agreement were lower than for ratings of overall similarity, listeners did consistently agree in their ratings of at least some voices in the midrange of the scale.

### C. Conventional measures of rater reliability

Conventional measures of rating reliability, such as Cronbach's alpha (Cronbach, 1951) and the intraclass correlation for the reliability of mean ratings (Ebel, 1951; Berk, 1979; Shrout and Fleiss, 1979), do not reflect the variability that occurs in interrater agreement, because they cannot represent patterns of agreement among raters and they cannot indicate agreement for specific voice samples (Young and Downs, 1968). Table V lists values of these statistics for the data in Figs. 3 and 4. Summary reliability statistics were high overall, ranging from 0.68–0.99 across studies. Thus these data met conventional standards for reliability (see Kreiman *et al.*, 1993, for review), despite the great variability that appeared when agreement levels for specific voices were examined. In particular, values were very high for data sets where the likelihood of listener agreement was poor (Kreiman *et al.*, 1993) or variable (Kreiman and Gerratt, 1996), but $n$ was large.

## III. DISCUSSION

The experimental tasks examined here—ratings of breathiness, roughness and severity; similarity ratings; and ratings of similarity with respect to breathiness and roughness—showed varying patterns of agreement among listeners. For ratings of traditional vocal qualities (Kreiman *et al.*, 1993, 1994; Rabinov *et al.*, 1995; Chhetri, 1997), individual listeners were self-consistent in their use of EAI scales. However, there were relatively few voices about which listeners as a group consistently agreed. In particular, consistent agreement *never* occurred for voices with mean ratings in the midrange of a scale. In fact, only about 30% of the variance in quality ratings was related to differences among voices when average ratings were between 2.5 and 5.5 on a seven-point scale.

For ratings of the overall similarity of pairs of voices (Kreiman and Gerratt, 1996), listeners as a group did agree that some voices were moderately similar. Patterns of agreement for ratings of similarity with respect to breathiness or roughness (Kreiman *et al.*, 1994) shared characteristics of both traditional breathiness/roughness ratings and ratings of the overall similarity of pairs of voices. Unlike ratings of breathiness and roughness, listeners sometimes agreed in their ratings for voices with mean ratings in the midrange of the scale, but overall reliability was lower than for ratings of overall similarity.

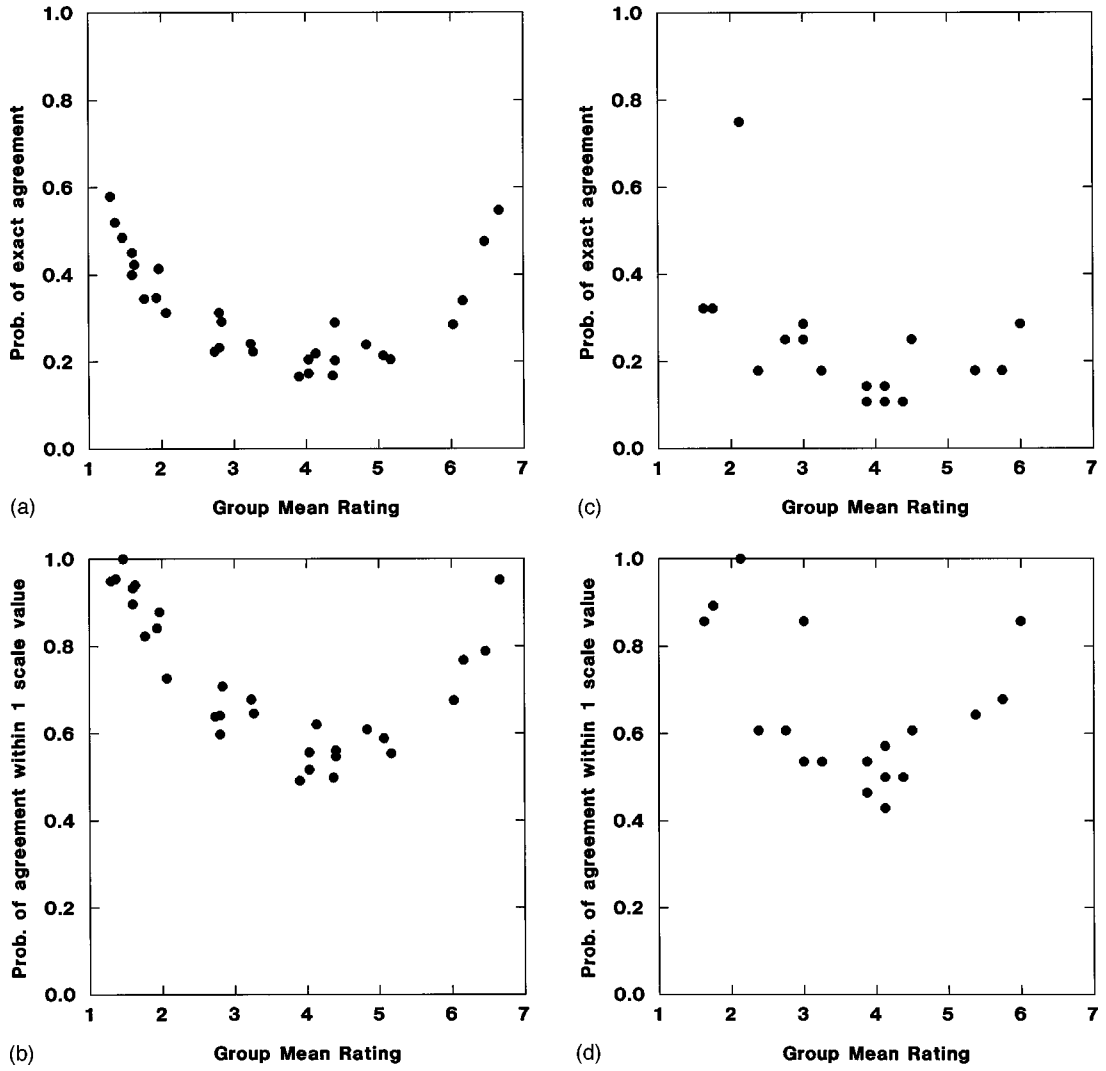



FIG. 3. For each voice in a data set, the probability that two raters agreed in their ratings of that voice, versus the overall mean rating for that voice. Results reflect only the first rating given each voice by each rater; the second rating was discarded. (a) The likelihood of exact agreement for EAI ratings of roughness; data from Kreiman *et al.* (1993). (b) The likelihood of agreement within 1 scale value for the same data. (c) The likelihood of exact agreement for EAI ratings of breathiness; data from Kreiman *et al.* (1994). (d) The likelihood of agreement within 1 scale value for the same data.

Several broad issues emerge from the patterns of agreement observed, and from observed differences among tasks. First, are patterns of results consistent with the assumption that traditional voice rating protocols provide valid measures of vocal quality? If these protocols are not sufficiently valid, how should vocal quality be measured? Finally, what measures of rating reliability are appropriate for evaluating data in studies of vocal quality?

TABLE IV. Variance in voice ratings accounted for by differences among voices with mean ratings in the midrange of a scale.[a]

| Study/Scale | $R^2$ |
|---|---|
| Kreiman *et al.* (1993) (EAI/Roughness) | 0.30 |
| Kreiman *et al.* (1993) (VA/Roughness) | 0.37 |
| Kreiman *et al.* (1994) (EAI/Breathiness) | 0.29 |
| Kreiman *et al.* (1994) (EAI/Roughness) | 0.22 |
| Rabinov *et al.* (1995) (VA/Roughness) | 0.42 |
| Chhetri (1997) (EAI/Severity) | 0.34 |

[a]EAI=equal-appearing interval scale; VA=visual analog scale. Midrange of a seven-point EAI scale is defined as the segment between 2.5 and 5.5, inclusive; VA scales were truncated proportionally.

### A. Validity of rating scale protocols

Paradigms for assessing vocal quality on traditional unidimensional scales like breathiness and roughness require the assumption that individual differences among listeners in ratings are noise or error, so that the ''true score'' for a voice on a scale is solely a function of the voice itself. Average ratings provide meaningful measures of quality only if this assumption holds. The present results are inconsistent with this assumption, and thus provide evidence against the validity of many protocols for assessing voice quality. Although listeners agreed at above-chance levels, most of the variance in quality ratings was due to factors other than differences among voices. The extent of variability in ratings received by voices away from scale end points indicates that mean ratings in the midrange of such scales poorly represent the

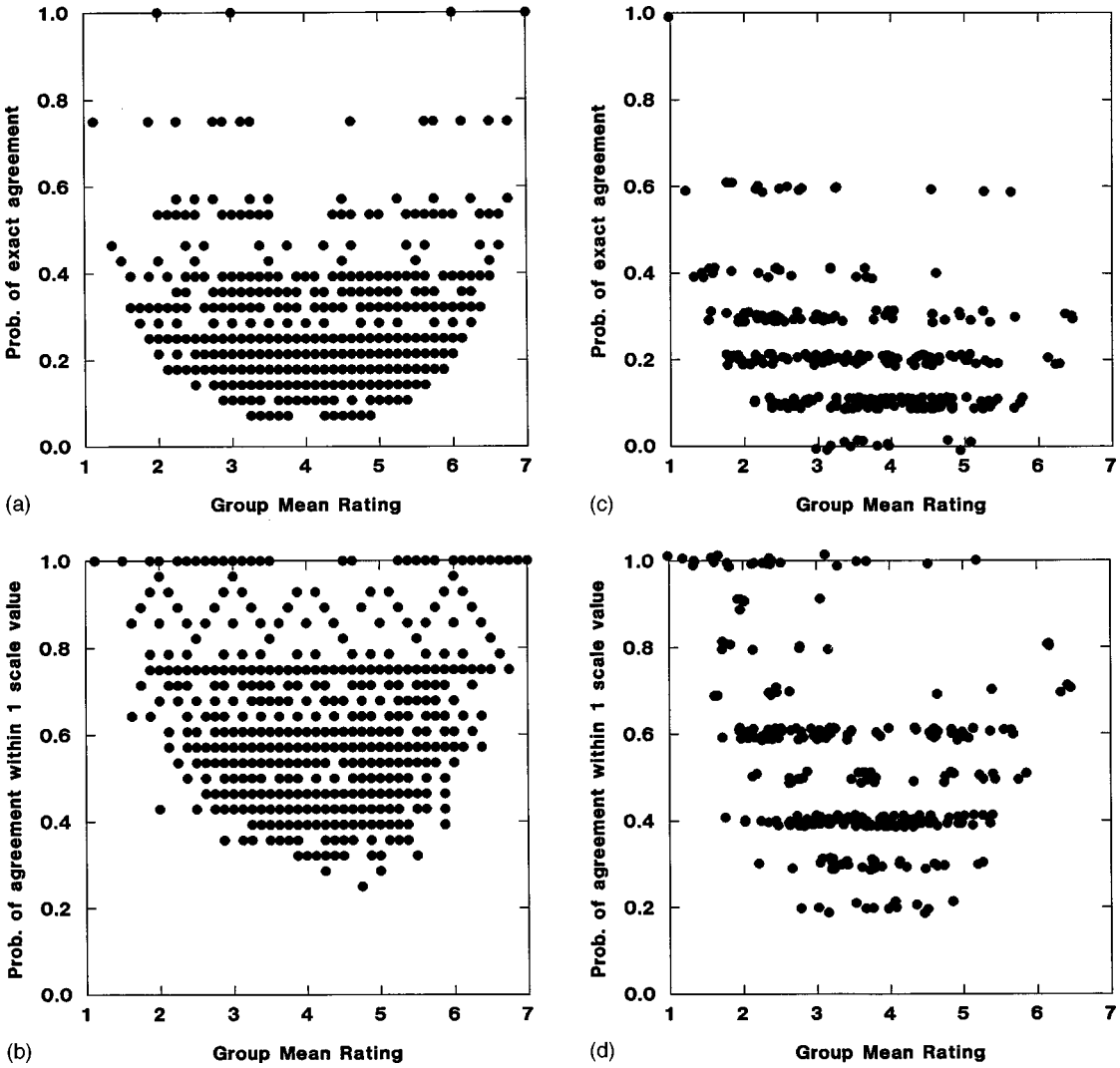



FIG. 4. For each pair of voices in a data set, the probability that two raters agreed in their ratings of the similarity of that pair of voices, versus the overall mean rating for that voice. Points have been jittered slightly to show overlapping values. (a) The likelihood of exact agreement for ratings of the similarity of pairs of female voices; data from Kreiman and Gerratt (1996). (b) The likelihood of agreement within 1 scale value for the same data. (c) The likelihood of exact agreement for ratings of the similarity of pairs of voices with respect to breathiness; data from Kreiman *et al.* (1994). (d) The likelihood of agreement within 1 scale value for the same data.

extent to which a voice possesses a quality. Instead, ironically, mean ratings in the middle of a scale serve primarily to indicate that listeners disagreed. If differences among voices are not consistently reflected by differences in ratings, then traditional rating scale protocols do not measure what they are intended to measure, and their validity is not supported.

Although traditional rating protocols do not appear to provide valid measures of the quality of a given voice, scalar

TABLE V. Traditional measures of rating reliability.[a]

| Study | Quality judged | Interrater agreement/reliability |
|---|---|---|
| Kreiman *et al.* (1993) | Roughness (EAI scale) | Reliability of mean rating (ICC) = 0.99<br>Cronbach's alpha = 0.99 |
| Kreiman *et al.* (1994) | Dissimilarity of pairs of voices with respect to breathiness (EAI scale) | Reliabity of mean rating (ICC) = 0.68<br>Cronbach's alpha = 0.74 |
| | Breathiness (EAI scale) | Reliability of mean rating (ICC) = 0.93<br>Cronbach's alpha = 0.97 |
| Kreiman and Gerratt (1996) | Dissimilarity of pairs of female voices (EAI scale) | Reliability of mean rating (ICC) = 0.89<br>Cronbach's alpha = 0.90 |

[a]ICC = intraclass correlation coefficient; EAI = equal-appearing interval scale.

ratings may still provide valuable information, if used to evaluate individual differences in perceptual strategy. For example, differences in patterns of disagreement that emerged from different rating tasks may provide insight into the mechanisms underlying the observed disagreements. Both traditional ratings of specific qualities and judgments of similarity with respect to specific qualities require listeners to compare observed voice stimuli to mental representations for the selected levels of that quality. This external-to-internal comparison introduces several sources of rating variability, including short- and long-term changes in mental representations, differences across listeners in how they define a quality or in standards for particular scale values, and variations in the importance of a cue in the context of variations in other cues (e.g., Kreiman *et al.*, 1992, 1993). In contrast, similarity rating tasks require listeners to compare stimuli globally and directly, without the need to refer to mental standards or assess particular attributes. Thus such tasks are not subject to error related to internal representations of a quality or drift in standards for particular levels of that quality. However, all tasks are subject to errors due to individual differences, perceptual biases, influences of perceptual context, mistakes, and changes over time in attention to these complex multidimensional stimuli.

Hypotheses regarding the effects of unstable internal standards for nonextreme levels of a quality are supported by data from a rating protocol using explicit anchors for each scale point (Gerratt *et al.*, 1993). When listeners made their ratings with reference to external ''anchor'' stimuli (instead of presumed internal criteria), good agreement occurred when stimuli were identical to the anchors. However, agreement dropped sharply between anchors, again suggesting that listeners cannot maintain internal standards for different levels of traditional vocal qualities. These results also demonstrate the major weakness of anchored protocols. The increase in agreement gained by including an external anchor was limited to the stimuli identical to the anchor, and listener agreement quickly decreased when stimuli fell between anchors. These data indicate that unless a protocol includes a large number of anchors spaced closely together, reference stimuli will not solve the problem of listener disagreements in ratings of particular voices. Further, providing anchors for a traditional quality scale circumvents the issue of the scale's validity, which must be established by some other means.

It remains possible that listener training may provide a partial solution to these difficulties. Although short-term training has not been shown to consistently improve overall listener agreement (see Kreiman *et al.*, 1993 for review), with extensive training listeners may learn to focus selectively on different aspects of complex auditory stimuli. Whether this is in fact the case, and whether the effects of training persist after training ceases, remain as issues for future research. In any case, the scales and stimuli with which listeners are trained must be viewed as arbitrarily chosen, unless independent evidence supports their validity.

### B. How should vocal quality be measured?

If traditional unidimensional rating scales are abandoned, a large gap in the conventional approach to clinical voice assessment will result. Obviously, much study is necessary to evaluate alternative strategies. Novel approaches to quality assessment should address the problems that appear to underlie listener disagreements. First, the present findings are consistent with the view that listener disagreements result in part from comparing external stimuli to idiosyncratic and/or unstable internal standards when attempting to use traditional rating scales. Second, it appears that listeners are unable to selectively attend to individual elements or dimensions of quality, as required by traditional voice assessment paradigms.

Measurement of overall vocal quality offers an alternative to traditional unidimensional ratings of specific vocal qualities. Many approaches to measurement of overall quality are possible. Techniques using analysis by synthesis and/or similarity ratings have long histories in psychometric research, and issues of their validity have been addressed in some detail (e.g., Gregson, 1975). Such tasks involve explicit comparisons between stimuli, rather than mappings between stimuli and internal standards, and they do not require listeners to focus attention on single dimensions of quality. Thus, in theory, they should eliminate the two causes of listener disagreement described above.

We have previously suggested that analysis by synthesis could be used to determine how listeners manipulate acoustic or other parameters to construct a synthetic token that matches the quality of a natural voice of interest (Kreiman and Gerratt, 1996). The values of these parameters would then directly represent a listener's perceptual response, rather than only having a statistical association with that response as in current correlative approaches. Although synthesizer parameters are manipulated individually, listeners still judge quality as a whole when evaluating the success of the synthesis. Thus analysis by synthesis combines unidimensional and overall approaches to quality.

Further, with the addition of multivariate or multidimensional statistical techniques, analysis by synthesis may allow development and testing of specific hypotheses about the nature and direction of changes in quality. For example, single acoustic parameters can be manipulated systematically and the resulting quality changes evaluated with similarity judgments. If the acoustic parameter in question predicts patterns of perceived similarity, a strong case for its importance to perception can be made. Note that this approach allows hypotheses about perceptual dimensions and their correlates to be investigated without the use of traditional scales for single qualities.

### C. Reliability and the measurement of vocal quality

Although the minimum ''acceptable'' level for listener agreement and reliability varies from study to study, a consensus exists that for most statistics, a value above 0.7 (or 49% variance in common) is ''good'' to ''excellent,'' but that a value above 0.5 is adequate (e.g., Kazdin, 1977; Fleiss, 1981; Hammarberg and Gauffin, 1995; de Bodt *et al.*, 1997). The present results highlight several difficulties with this view. Measures of overall reliability (such as intraclass correlations and Cronbach's alpha) can mask large and predictable differences in agreement levels for different voices. For

example, a data set for which Cronbach's alpha equals 0.9 or better may include individual voices for which agreement levels do not exceed chance. The presence or absence of normal and/or extremely severely pathologic voices in the stimulus set inflates or deflates these statistics (Kearns and Simmons, 1988). For example, ratings of roughness in the present data were more reliable overall for studies that included normal voices (Kreiman *et al.*, 1993) than for studies that did not (Rabinov *et al.*, 1995; Kreiman *et al.*, 1994). The number of raters in a study also affects overall reliability. For example, a mean interrater correlation of 0.4 will produce Cronbach's alpha of 0.87, given an $n$ of only ten raters (Carmines and Zeller, 1979). Thus such measures depend on differences in experimental design as well as differences in ratings.

If averaging ratings is inappropriate, as argued above, it follows that unaveraged data must be analyzed. In other words, individual differences in quality perception must be modeled (Kreiman and Gerratt, 1996). With individual differences models of perception there is no expectation that listeners will agree, so mean ratings are without interest. In this case, the reliability of the mean rating and the extent to which listeners agree in their ratings become moot points. Measures of variance accounted for may provide an alternative method of assessing the usefulness of a set of listener judgments. Such measures are particularly useful because they make explicit the factors being used to predict variance in ratings. In this way, the statistical model (and its fit to the data) are precisely specified, rather than implied (as they have been in the past).

For example, in the present data, overall reliability was slightly lower for similarity ratings than for ratings of traditional qualities. However, multidimensional scaling analyses accounted for much of this increased variability by quantifying the contributions of presentation order and/or individual differences in perceptual strategy to rating variability (Kreiman *et al.*, 1994; Kreiman and Gerratt, 1996). The $r^2$ values for individual listeners' data in Kreiman and Gerratt (1996) ranged from 0.56 to 0.83; $r^2$ due to differences between voices in the unidimensional rating tasks reviewed here ranged from 0.22 to 0.42.

Finally, patterns of listener agreement provide information not available from measures of overall reliability, and thus may serve as a useful supplement to measures of total variance accounted for. For example, understanding which voices listeners consistently agree about, and which they cannot agree about, may provide clues to the factors underlying judgments of vocal quality [see Kreiman and Gerratt (1998b) for an example of this kind of analysis].

## IV. CONCLUSIONS

Accurate modeling of voice perception is essential to the success of many endeavors, including development of instrumental measures of voice, refinement of speech synthesizers, and evaluation of the effectiveness of treatments for voice disorders. The low levels of listener agreement reported here indicate that traditional protocols for assessing qualities like breathiness and roughness are not useful for measuring perceived vocal quality. More detailed analyses of listeners' performance in voice evaluation tasks, and better quantification of the adequacy of models of voice perception, will contribute to improved measurement of voice quality.

## ACKNOWLEDGMENTS

We thank Ted Bell, James Hillenbrand, and Don Dirks for discussions of many statistical and nonstatistical aspects of this work. Dinesh Chhetri generously provided his severity rating data. Winifred Strange, Thomas Baer, and one anonymous reviewer provided many helpful comments on an earlier version of the manuscript. This research was supported by Grant No. DC01797 from the National Institute on Deafness and Other Communication Disorders. Please address correspondence to Jody Kreiman, Division of Head and Neck Surgery, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, CA 90095-1794 (Electronic mail: jkreiman@ucla.edu).

Allen, M. J., and Yen, W. M. (**1979**). *Introduction to Measurement Theory* (Brooks/Cole, Monterey, CA).

ANSI (**1960**). ANSI S1.1-1960, ''Acoustical Terminology'' (American National Standards Institute, New York).

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

Carmines, E. G., and Zeller, R. A. (**1979**). *Reliability and Validity Assessment*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-017 (Sage, Newbury Park, CA).

Chhetri, D. K. (**1997**). ''Treatment of voice disorders related to unilateral paralysis of the vocal cord,'' unpublished senior medical student thesis, University of California, Los Angeles.

Colton, R. H., and Estill, J. A. (**1981**). ''Elements of voice quality: Perceptual, acoustic, and physiologic aspects,'' in *Speech and Language: Advances in Basic Research and Practice*, edited by N. J. Lass (Academic, New York), Vol. 5, pp. 311–403.

Cone, J. D. (**1977**). ''The relevance of reliability and validity for behavioral assessment,'' Behav. Therapy **8**, 411–426.

Crocker, L., and Algina, J. (**1986**). *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, New York).

Cronbach, L. J. (**1951**). ''Coefficient alpha and the internal structure of tests,'' Psychometrika **16**, 297–334.

de Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., and Croux, C. (**1997**). ''Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality,'' J. Voice **11**, 74–80.

de Krom, G. (**1995**). ''Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments,'' J. Speech Hear. Res. **38**, 794–811.

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

Fleiss, J. L. (**1981**). *Statistical Methods for Rates and Proportions* (Wiley, New York).

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

Gerratt, B. R., and Kreiman, J. (**1995**). ''The utility of acoustic voice measures,'' in *Proceedings of the Workshop on Standardization in Acoustic Voice Analysis*, edited by D. Wong (National Center for Voice and Speech, Denver), pp. GER1–GER7.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., and Berke, G. S. (**1993**). ''Comparing internal and external standards in voice quality judgments,'' J. Speech Hear. Res. **36**, 14–20.

Gerratt, B. R., Till, J., Rosenbek, J. C., Wertz, R. T., and Boysen, A. E. (**1991**). ''Use and perceived value of perceptual and instrumental measures in dysarthria management,'' in *Dysarthria and Apraxia of Speech*, edited by C. A. Moore, K. M. Yorkston, and D. R. Beukelman (Brookes, Baltimore), pp. 77–93.

Gregson, R. A. (**1975**). *Psychometrics of Similarity* (Academic, New York).

Hammarberg, B., and Gauffin, J. (**1995**). ''Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects,'' in *Vocal Fold Physiology: Voice Quality Control*, edited by O. Fujimura and M. Hirano (Singular, San Diego), pp. 283–303.
Helmholtz, H. (**1885**; reprinted 1954). *On the Sensations of Tone* (Dover, New York).
Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). ''Acoustic correlates of breathy vocal quality,'' J. Speech Hear. Res. **37**, 769–778.
Jensen, P. J. (**1965**). ''Adequacy of terminology for clinical judgment of voice quality deviation,'' The Eye, Ear, Nose and Throat Monthly **44**, 77–82.
Kazdin, A. (**1977**). ''Artifact, bias, and complexity of assessment: The ABCs of reliability,'' J. Appl. Behav. Anal. **10**, 141–150.
Kearns, K., and Simmons, N. (**1988**). ''Interobserver reliability and perceptual ratings: More than meets the ear,'' J. Speech Hear. Res. **31**, 131–136.
Kempster, G. B., Kistler, D., and Hillenbrand, J. (**1991**). ''Multidimensional scaling analysis of dysphonia in two speaker groups,'' J. Speech Hear. Res. **34**, 534–543.
Kerlinger, F. N. (**1973**). *Foundations of Behavioral Research* (Holt, Rinehart, and Winston, New York), 2nd ed.
Kreiman, J. (**1997**). ''Listening to voices: Theory and practice in voice perception research,'' in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego), pp. 85–108.
Kreiman, J., and Gerratt, B. R. (**1996**). ''The perceptual structure of pathologic voice quality,'' J. Acoust. Soc. Am. **100**, 1787–1795.
Kreiman, J., and Gerratt, B. R. (**1998a**). ''Measuring vocal quality,'' to appear in *Handbook of Voice Quality Measurement*, edited by R. Kent and M. J. Ball (Singular, San Diego).
Kreiman, J., and Gerratt, B. R. (**1998b**). ''Sources of listener disagreement in voice quality assessment'' (in preparation).
Kreiman, J., Gerratt, B. R., and Berke, G. S. (**1994**). ''The multidimensional nature of pathologic vocal quality,'' J. Acoust. Soc. Am. **96**, 1291–1302.
Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., and Berke, G. S. (**1993**). ''Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research,'' J. Speech Hear. Res. **36**, 21–40.
Kreiman, J., Gerratt, B. R., and Precoda, K. (**1990**). ''Listener experience and perception of voice quality,'' J. Speech Hear. Res. **33**, 103–115.
Kreiman, J., Gerratt, B. R., Precoda, K., and Berke, G. S. (**1992**). ''Individual differences in voice quality perception,'' J. Speech Hear. Res. **35**, 512–520.
Mackey, L. S., Finn, P., and Ingham, R. J. (**1997**). ''Effect of speech dialect on speech naturalness ratings: A systematic replication of Martin, Haroldson, and Triden (1984),'' J. Speech Lang. Hear. Res. **40**, 349–360.
Martin, D., Fitch, J., and Wolfe, V. (**1995**). ''Pathologic voice type and the acoustic prediction of severity,'' J. Speech Hear. Res. **38**, 765–771.
Murry, T., Singh, S., and Sargent, M. (**1977**). ''Multidimensional classification of abnormal voice qualities,'' J. Acoust. Soc. Am. **61**, 1630–1635.
Perkins, W. (**1971**). ''Vocal function: A behavioral analysis,'' in *Handbook of Speech Pathology and Audiology*, edited by L. Travis (Appleton Century Croft, New York), pp. 481–504.
Rabinov, C. R., Kreiman, J., Gerratt, B. R., and Bielamowicz, S. (**1995**). ''Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter,'' J. Speech Hear. Res. **38**, 26–32.
Shrout, P., and Fleiss, J. (**1979**). ''Intraclass correlations: Uses in assessing rater reliability,'' Psychol. Bull. **86**, 420–428.
Silverman, F. H. (**1977**). *Research Design in Speech Pathology and Audiology* (Prentice-Hall, Englewood Cliffs, NJ).
Sodersten, M., Hertegard, S., and Hammarberg, B. (**1995**). ''Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women,'' J. Voice **9**, 182–204.
Suen, H. K., and Ary, D. (**1989**). *Analyzing Quantitative Behavioral Observation Data* (Erlbaum, Hillsdale, NJ).
Ventry, I. M., and Schiavetti, N. (**1980**). *Evaluating Research in Speech Pathology and Audiology* (Addison-Wesley, Reading, MA).
Young, M. A. (**1993**). ''Supplementing tests of statistical significance: Variation accounted for,'' J. Speech Hear. Res. **36**, 644–656.
Young, M. A., and Downs, T. D. (**1968**). ''Testing the significance of the agreement among observers,'' J. Speech Hear. Res. **11**, 5–17.