Reliability and Validity of Judgments of Sound Quality in Elderly Hearing Aid Wearers

M. Mini Narendran and Larry E. Humes

Objective: The aim of the present study was to evaluate a particular sound-quality rating procedure, referred to here as the Judgments of Sound Quality (JSQ) test, for potential use as an outcome measure with elderly hearing aid wearers. The validity and reliability of the JSQ as an outcome measure were determined for unaided and aided listening conditions.

Design: A repeated-measures design was employed with two primary independent variables, each with two values: 1) aided condition (aided versus unaided listening); and 2) trial (test versus retest). Sixteen elderly, hearing-impaired hearing aid users participated in the study. The participants provided ratings on eight dimensions of sound quality (softness, brightness, clarity, fullness, nearness, loudness, spaciousness, and total impression) under four stimulus conditions (speech at 65 dB SPL with a +8 dB signal-to-noise ratio, speech at 65 dB SPL in quiet, music at 90 dB SPL, and music at 75 dB SPL). Test and retest ratings were obtained in both unaided and aided conditions.

Results: No significant differences were found between median test and retest scores on the JSQ. The median ratings for each sound-quality dimension were found to have moderate test-retest reliability, with test-retest correlations \( r \) ranging from 0.20 to 0.73 (median \( r \) value = 0.58) for the unaided listening condition, and from 0.23 to 0.85 (median \( r \) value = 0.51) for the aided listening condition. Test validity was established through significant differences in JSQ ratings for various stimulus pairs (e.g., speech in quiet versus speech in noise). In addition, significant differences were observed between unaided and aided ratings for the dimensions of clarity, nearness, loudness, and total impression with aided JSQ ratings approaching normative “ideal” values established previously.

Conclusions: The JSQ appears to be a potentially useful measure of hearing aid outcome, especially when using group data to document the benefits of amplification. Additional efforts should be directed at improving the reliability of the JSQ, however, before application to hearing aid wearers on an individual basis.

Department of Speech and Hearing Sciences, Indiana University, Bloomington, Indiana.

DOI: 10.1097/01.AUD.000051745.69182.14

In recent years, along with technological advances in hearing aids, there has also been a greater focus on methods of evaluating the efficiency of these devices, and the benefit afforded to wearers of hearing aids. These methods have been collectively termed outcome measures, and many outcome measures are available today. Outcome measures for hearing aids have been classified into those measuring aided performance, benefit, satisfaction, or use (Bentler & Kramer, 2000; Gatehouse, 2000; Humes, 1999, 2001; Humes, Halling, & Coughlin, 1996). Measurement of sound quality offers one possible subjective measure of aided performance and benefit.

One of the most noticeable attributes of sound is its quality. Sounds can be described subjectively using adjectives such as “soft,” “shrill,” or “harsh.” Clinically, it is often one of the first indicators by which a hearing aid user provides the clinician with his or her impression of the hearing aid. Sound quality is often used by new wearers of hearing aids in making decisions regarding the continued use of their hearing aids. Ovegård, Lundberg, Hagerman, Gabriëlssoon, Bengtsson, and Brändström (1997), for example, suggested that sound quality might affect both the perceived helpfulness and the extent of hearing aid use. Therefore, measurement of sound quality in hearing aid wearers could be a potentially useful outcome measure.

The subjective nature of sound quality makes it more difficult to use as a structured measure of hearing aid outcome. One of the more carefully developed measures of sound quality was described by Gabriëlssoon, Schenman, and Hagerman (1988) and is referred to here as the Judgments of Sound Quality (JSQ) test. The JSQ makes use of eight dimensions, seven of which deal with separate qualities of sound and one that deals with the total impression of the sound. These dimensions were derived originally from subjective descriptions in the Swedish language, through multivariate analyses (Gabriëlssoon & Sjögren, 1979).

Though this measure was developed initially to assess the performance of different types of loudspeakers (Gabriëlssoon & Sjögren, 1979), it has also been applied to other devices, such as earphones (Gabriëlssoon, Hagerman, Beech-Kristensen, & Lundberg, 1990), earmolds (Lundberg, Ovegård, Hager-
man, Gabrielsson & Brändström, 1992), and hearing aids (Gabrielsson & Sjögren, 1979; Lundberg et al., 1992). Subjective sound-quality ratings provided by listeners were found to be very useful in assessing the performance of these devices. Some of these studies measured the inter-listener and intra-listener reliability of the ratings, and the ratings were found to be reasonably reliable. Gabrielsson et al. (1988) also found the JSQ to be reliable when used with hearing-impaired individuals, although their evaluation did not include the use of hearing aids.

Most prior studies of the JSQ, however, were conducted with normal-hearing listeners. As a result, the reliability measurements from these studies cannot be generalized to clinical populations, such as elderly hearing-impaired listeners and hearing aid wearers. These individuals may be more variable in their performance due to underlying variability in sensory or cognitive function associated with aging or hearing impairment. The hearing aid itself represents another potential source of variability in aided sound quality measures. The limited information on JSQ reliability restricts the usefulness of the test as a potential measure of outcome in this population. In addition, it is not clear how differences between aided and unaided JSQ ratings should be interpreted. It is not the case, for example, that numerically higher ratings on each dimension can be interpreted as “better” sound quality. For example, for the loudness dimension of the JSQ, a maximum numerical rating of 10 corresponds to “too loud” and is clearly not the desired or ideal loudness rating. Additional work is required, therefore, to validate the JSQ measures and facilitate their interpretation.

The aim of the present investigation was to assess the JSQ test as a potential outcome measure in elderly hearing aid users. Test-retest reliability was examined to provide an indication of the short-term stability of listeners’ ratings of sound quality. Additionally, test validity was evaluated by comparing group ratings for various stimulus pairs. For example, speech presented in a background of multitalker babble should be judged to have less clarity than that same speech signal presented in quiet. Further, if both speech signals are presented at the same sound pressure level (and the speech-to-babble ratio is positive), the JSQ ratings on the loudness dimension should not differ for these two stimuli. On the other hand, if identical musical selections, presented in quiet, serve as stimuli for the JSQ, but the stimuli differ in presentation level, one would expect differences to be confined primarily to the loudness dimension of the JSQ for this stimulus pair. In this fashion, JSQ results for various stimulus pairs will be examined in this study to validate interpretation of the JSQ ratings from elderly hearing aid wearers.

**Method**

**Participants**

Sixteen elderly hearing-impaired individuals with mild to profound sensorineural hearing loss were recruited as participants in the study. The participants included 10 men and 6 women and ranged in age from 60 to 81 yr (mean = 74.6 yr, SD = 6.2 yr). These participants were recruited from the patient database of the Speech and Hearing Clinic at Indiana University, Bloomington (IUB), and had all been wearing hearing aids for a period of at least 6 mo. This criterion was chosen to ensure that only participants with enough experience wearing hearing aids were included in the study. All participants had been fitted with hearing aids at IUB, and previous fitting information was available to the experimenter. Only individuals whose hearing aids were in working condition were included in the study. The ages and hearing thresholds of the participants, as well as the group means and the standard deviations, are presented in Table 1.

The participants wore their own hearing aids during the study, and these included behind-the-ear, in-the-ear, and in-the-canal hearing aids. No attempt was made to control the type and circuit of the hearing aids used by the participants. Eleven of the participants used binaural amplification, and five wore monaural hearing aids. All participants were tested individually.

**Materials**

Subjective impressions of sound quality were measured using the JSQ test (Gabrielsson et al., 1988). Eight dimensions were evaluated: softness, brightness, clarity, fullness, nearness, loudness, spaciousness and total impression. The dimensions and their descriptions were initially developed in Swedish and were later translated into English by Gabrielsson et al. (1988).

Each of the eight dimensions was rated on a 10-point continuous rating scale. The two ends of the rating scale were defined using specific adjectives. In addition, points 1, 3, 5, 7, and 9 on each rating scale were labeled for reference. For example, the dimension “softness” was defined as “Sounds soft and gentle, in opposition to sharp, hard, shrill, or harsh,” and on the rating scale for softness, 1 was defined as “very harsh,” 3 as “rather harsh,” 5 as “midway,” 7 as “rather soft,” and 9 as “very soft.”

The test was administered under four stimulus conditions: speech at 65 dB SPL at a +8 dB signal-
to-noise ratio (SPN), speech at 65 dB SPL in quiet (SPQ), Music at 75 dB SPL (Music75), and Music at 90 dB SPL (Music90). The two speech stimuli consisted of connected discourse by a male speaker from the Speech Intelligibility Rating (SIR) test (Cox & McDaniel, 1989). In case of the speech-in-noise condition, the speech was presented along with multi-speaker babble, also from the SIR test. The 50-second speech samples were prerecorded on a compact disc entitled “Speech Intelligibility Tests,” developed and distributed by the Hearing Aid Research Laboratory at the University of Memphis. The two music conditions included music from Ludvig van Beethoven’s Symphony Number 9 “Choral”; Allegro ma non troppo, un poco maestro and Molto vivace, performed by the Dresden Symphonic Orchestra and Choir. The 60-second music samples were equated in peak level (±2 dB) and the variation in rms sound levels within the samples was 30 dB. Thus, each of the Music75 samples and the Music90 samples varied in level from 45 to 75 dB, and from 60 to 90 dB, respectively. The presentation levels for the music samples were selected using a calibration noise matched to the peak level of the samples.

All materials were prerecorded and were presented to the listeners in sound field through loudspeakers. Speech materials were recorded on a compact disc, and the music samples on digital audiotape. All testing was done with the participant seated comfortably in a sound-treated booth that complied with ANSI-S3.1 (1991) standards regarding ambient sound levels. Except for the speech-in-noise stimulus condition, all stimuli were presented through a single loudspeaker (Realistic Minimus 7) located at an azimuth and elevation of 0°, at a distance of one meter from the center of the listener's head. For the speech-in-noise condition, the speech was presented from the loudspeaker located at 0°, while the noise was presented through an identical loudspeaker located at an azimuth of 180° and at a distance of 1 meter from the participant’s head.

**Procedure** • The study was carried out in two sessions that were scheduled within 1 to 3 days of each other. This test-retest interval was selected to avoid any change in perception of sound quality that might occur over longer periods of time, while still being long enough to minimize memorization of responses between sessions. Each session lasted approximately 120 minutes and both sessions were scheduled at the same time of the day for each participant.

In each session, participants completed the judgments of sound quality twice, once with (aided) and once without (unaided) their hearing aids. The order
of the two testing conditions (unaided/aided) was randomized across participants. Also, the order of the four stimulus conditions was randomized across participants. For a given participant, the same order of stimulus condition was used for the unaided and aided listening conditions. Within each participant, the order of listening conditions and the stimulus conditions was maintained for the test and retest conditions. For example, if a participant originally was tested in the aided condition first, and then in the unaided condition, he or she was tested in the same order at the retest session. Similarly, if the order of presentation of stimuli in the first session was Music90, Speech-in-quiet, Speech-in-noise, and Music75, the same order was followed for unaided and aided testing within the same session, and also at the subsequent retest session. Thus, during each session, the participant made a total of 64 ratings (8 dimensions × 4 stimulus conditions × 2 hearing aid conditions). The two sessions together yielded a total of 128 ratings per participant.

Before testing in the first session, the participant’s hearing aids were checked using a Frye-2000 Fonix Hearing Aid Analyzer. The electroacoustic measures obtained were compared with the specifications for each hearing aid. The participants were included in the study only if the hearing aids were in working condition at the time of testing and if the hearing aids met manufacturer’s specifications.

Once the hearing aids were checked, the participant’s air conduction hearing thresholds were re-established using a portable audiometer (Beltone 10 D) in a double-walled sound-treated booth. The thresholds obtained were compared with the thresholds from the latest audiogram on file in the participant’s clinic file. This was done to ascertain that the participant’s thresholds had not changed more than 15 dB at any frequency. All the participants met this criterion for inclusion into the study.

The participants were then given laminated sheets containing the descriptions of the quality dimensions, the rating scales, and written instructions about the sound-quality judgments. Each dimension was described on an individual page, so that the participant could focus on a single dimension at a time.

The instructions provided to the participants were as follows: “Your task is to judge the sound quality of the speech or music sounds you hear (while wearing your hearing aids), using the scales on the response forms. The scales refer to various properties of the sound. They are graded from 10 (maximum) to 0 (minimum). You decide yourself on the accuracy that you consider necessary. As you can see, it is also possible to use decimals. The integers 9, 7, 5, 3, and 1 are defined on the response form. For instance, in the scale for clarity, 10 means the maximum (highest possible) clarity, 9 means very clear, 7 rather clear, 5 midway, 3 rather unclear, 1 very unclear, and 0 minimum (lowest possible) clarity. The other scales work in similar ways.”

No further instructions were given to any participant. To minimize any bias in the responses, the aim of the study was not described to the participants until the completion of the study.

The participants were then presented with practice samples of the stimuli that were similar to the actual test materials. They were not required to provide ratings during the practice tasks. While listening to the practice samples, the participants were asked to set the volume controls of their hearing aids to the most comfortable level. This setting was marked on the hearing aid and the participants were not allowed to manipulate the hearing aid controls for the remainder of the session. The samples were played until the participants indicated that they were comfortable with the materials. For monaural hearing aid wearers, all aided measurements were made with the nontest ear unoccluded.

After the participants indicated that they were comfortable with the task and the rating procedure, the test samples were played. The participants, seated inside the sound-treated booth, communicated with the experimenter outside through a lapel microphone. The stimulus presentation was controlled by the tester. A separate 50- to 60-second sample was presented for each judgment. The participants were instructed to provide the ratings only after listening to the complete 50 to 60-second stimulus samples. The presentation was then paused, and the participants said the rating aloud. This was then recorded on the response sheet by the tester, and the next sample was played. Frequent breaks were provided during the testing.

The retest session involved the same routine as the test session, except for the hearing threshold measurements. The hearing aids were checked against manufacturer’s specifications at the beginning of the retest session also, and then the volume control was set to the position marked during the test session. The participants were then tested using the same method as described above, and the same order of testing was maintained. The participants were paid for their participation in the experiment.

Results and Discussion

Gabrielsson et al. (1988) suggested that sound-quality measures should be obtained for a range of stimuli, including speech in quiet, speech in noise, and music. Their data, including normative “ideal”
sound-quality ratings for each dimension, were compiled for a broad set of stimulus conditions. Consequently, for the present study, median ratings for each participant and each sound-quality dimension were established across the speech-in-noise (SPN), speech-in-quiet (SPQ) and music-75 dB (Music75) stimulus conditions. Doing so results in a more global or broad definition of “sound” quality that encompasses both speech and music, as well as ideal and poor listening conditions. The individual median JSQ values were then compiled and group medians calculated, along with interquartile ranges. Given the ordinal nature of the JSQ rating scales, medians and interquartile ranges were the most appropriate measures of central tendency and variability, respectively, for the group data. This, in turn, necessitated the use of nonparametric statistics for data analysis. For all such analyses, unless noted otherwise, a significance level of $p < 0.05$ was used with the criterion $p$ level adjusted conservatively for multiple comparisons simply by dividing $p$ by the number of pairwise comparisons made in a particular analysis.

Before additional data analysis, a series of Mann-Whitney U analyses were conducted to examine differences in median JSQ ratings between the monaurally and binaurally aided participants. None of these 32 comparisons (8 JSQ dimensions x 2 aid conditions x 2 trials) were significant ($p > 0.05$). Thus, for all subsequent presentations of the data, aggregate results from all 16 participants were used.

Figure 1 displays the median and interquartile JSQ ratings for each sound-quality dimension for the unaided (top) and aided (bottom) listening conditions. In each panel, data are presented for both test (filled squares and solid lines) and retest (open circles and dashed lines) conditions. Wilcoxon Signed-Ranks analyses of the data in each panel revealed no significant differences between the test and retest conditions.

In Figure 2, the median JSQ ratings from Figure 1 have been replotted to emphasize the differences between aided (unfilled circles) and unaided (filled squares) conditions for both test (top panel) and retest (bottom panel). Wilcoxon Signed-Ranks analyses showed that the aided ratings for nearness, loudness and total impression were significantly greater than the unaided ratings for the initial test. On retest, the aided JSQ ratings for clarity, nearness and total impression were found to be significantly greater than the corresponding ratings for unaided listening. Although the aided JSQ ratings were significantly greater than the unaided ratings for several sound-quality dimensions in Figure 2, as noted previously, it is not necessarily the case that “higher” is “better” on the JSQ. To assist in interpretation of the aided changes in JSQ ratings, the normal-hearing “ideals” established for the JSQ by Gabrielsson et al. (1988) are also presented in Figure 2. For “total impression,” Gabrielsson et al. (1988) simply note that a rating of 10 is the ideal, whereas the other ideal values were established directly through measurement. Clearly, for all of the significant improvements in JSQ ratings from unaided to aided listening, the aided ratings approximate the normal-hearing ideals for those sound-quality dimensions. Thus, for these listeners, amplification has improved the perceived sound quality significantly and along several dimensions. Moreover, although there are differences in the details regarding the specific sound-quality dimensions affected, this general observation was repli-
There are other ways in which the data from the present study can assist in the interpretation of changes in JSQ ratings or in the validation of changes in various stimulus dimensions. For example, consider the SPN and SPQ stimulus conditions of the present study. For these two stimulus conditions, the speech stimulus was identical, but was administered in two different background conditions. One would expect that the SPN stimuli would be judged to be of poorer sound quality than the SPQ stimuli, all other things being equal. Figure 3 presents median JSQ ratings for these two stimulus conditions for unaided (top) and aided (bottom) listening conditions. Only the data from the initial test session are shown here because the group data were shown previously to be reliable and a single test session is most representative of routine clinical practice or data collection. For unaided listening, only the difference in total impression was significant with the SPQ stimuli being judged as higher quality than the SPN stimuli. When the audibility of the speech signal (and the background babble) have been improved through amplification (aided condition), moreover, significant differences in JSQ ratings emerged between these two stimulus conditions for the dimensions of clarity, fullness, and total impression. All of the differences in the aided listening conditions were in the direction of higher JSQ ratings for the SPQ stimuli than for the SPN stimuli. Further, all of the differences resulted in aided JSQ ratings more closely approximating the normal-hearing ideals than the unaided ratings (see the "N" symbols in Fig. 2 for comparison). Thus, the differences in the group JSQ ratings from the SPQ stimuli

Figure 2. Aided (unfilled circles) and unaided (filled squares) median JSQ ratings for test (top panel) and retest (bottom panel) conditions. "N" symbols represent "normal ideal ratings" from Gabrielsson et al. (1988).

Figure 3. Median JSQ ratings for the speech in quiet (unfilled circles) and the speech in noise (filled squares) stimuli for unaided (top) and aided (bottom) listening conditions.
to the SPN stimuli are in line with a priori expectations and further validate the JSQ.

Another comparison of JSQ ratings across stimulus conditions used in this study could serve to further validate this instrument. In particular, for the two conditions using music as the stimulus, the stimuli were identical except for a 15-dB difference in overall presentation level. As such, one would expect differences in JSQ ratings for these two stimulus conditions to be confined primarily to those associated with perceived loudness. This includes the sound-quality dimensions of loudness and nearness. Figure 4 presents the median JSQ ratings for the initial test session for the music stimulus presented at peak levels of 75 dB SPL (filled squares) and 90 dB SPL (unfilled circles). For unaided listening (top panel) the Music90 stimuli had significantly higher JSQ ratings than the Music75 stimuli for the dimensions of fullness, nearness and loudness. It is important to note that the changes were such that the Music90 stimuli more closely approximated the normal-hearing ideals for fullness, nearness and loudness than did the Music75 stimuli in this unaided listening condition. For the aided listening condition (bottom panel), the Music90 stimuli yielded group ratings that were significantly higher than those for the Music75 stimuli for the JSQ dimensions of loudness and nearness, but significantly lower for the dimension of softness. For all three JSQ dimensions showing significant differences between the Music90 and Music75 stimuli for aided listening, comparison to the normative ideals in Figure 2 reveals that the increase in presentation level moved the perceived quality away from the ideals, rather than closer to them. This might be attributed to output limiting of the hearing aids at the higher music presentation level. Again, this reinforces the notion that “higher” is not necessarily “better” in the JSQ and ratings should be interpreted relative to the ideals established by Gabrielson et al. (1988). In general, however, the changes in group ratings from softer to louder music are consistent with a priori expectations, both in terms of the direction of the changes and the sound-quality dimensions affected.

Thus far, the focus has been on the group data obtained in this study. Although the lack of significant differences between test and retest sessions (Fig. 1) indicated stability of the group data, this does not address the reliability of the individual JSQ ratings. One way to assess the reliability of the JSQ is through test-retest correlations. Table 2 provides the Spearman Rho (rank order) correlation coefficients between the test and retest sessions for individual JSQ medians. Correlations are shown for each dimension of sound quality and separately for unaided and aided listening conditions.

### Table 2. Spearman Rho rank-order correlation coefficients for the unaided and aided conditions of this study.

<table>
<thead>
<tr>
<th>Stimulus Conditions</th>
<th>Unaided</th>
<th>Aided</th>
<th>Gabrielsson et al. (1988)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Softness</td>
<td>0.66**</td>
<td>0.65**</td>
<td>0.47</td>
</tr>
<tr>
<td>Brightness</td>
<td>0.73**</td>
<td>0.23</td>
<td>0.93</td>
</tr>
<tr>
<td>Clarity</td>
<td>0.60*</td>
<td>0.66**</td>
<td>0.87</td>
</tr>
<tr>
<td>Fullness</td>
<td>0.62*</td>
<td>0.48</td>
<td>0.87</td>
</tr>
<tr>
<td>Nearness</td>
<td>0.31</td>
<td>0.85**</td>
<td>0.78</td>
</tr>
<tr>
<td>Loudness</td>
<td>0.56*</td>
<td>0.43</td>
<td>0.52</td>
</tr>
<tr>
<td>Spaciousness</td>
<td>0.55*</td>
<td>0.55*</td>
<td>0.74</td>
</tr>
<tr>
<td>Total Impression</td>
<td>0.20</td>
<td>0.37</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Significant values are marked with * (p < 0.05) and ** (p < 0.01). (The p values for these correlations have not been adjusted for multiple comparisons.) For comparison, comparable values from Gabrielsson et al. (1988) for unaided listening have been provided in the far right column.
aided and unaided listening conditions. For the unaided listening conditions, all correlations are positive, of moderate strength and statistically significant, except those for nearness and total impression. For the aided listening conditions, only four of the eight test-retest correlations are significant (softness, clarity, nearness and spaciousness). In general, it is desirable for such test-retest correlations to be at least 0.80 and preferably >0.90. Only one correlation from this study approximated this criterion. For comparison, test-retest correlations established by Gabrielsson et al. (1988) for unaided listening have been provided in the far right column of Table 2. In many cases, the correlations reported for the 12 hearing-impaired subjects of Gabrielsson et al. (1988) are higher than those observed here. This could be due to a number of factors, including the use of elderly hearing-impaired participants in this study, but not in the study by Gabrielsson et al. (1988). Perhaps older listeners are less reliable in their subjective ratings of sound quality than younger adults. In addition, a wider range of acoustical test conditions were evaluated in the earlier study and this could have affected the test-retest correlations. Many of the test-retest correlations from Gabrielsson et al. (1988), however, are below the minimum desirable value of 0.80.

Clearly, a major limitation in the clinical use of the JSQ for individual hearing aid wearers is the lower than desired reliability. In addition, as administered here, each JSQ rating required 50 to 60 seconds to complete. Thus, to obtain median JSQ ratings for three different stimuli (SPQ, SPN, Music75) for all eight sound-quality dimensions required about 24 minutes of testing, in both the aided and unaided listening conditions. Alternative ways of collecting the ratings, like instructing participants to rate the stimulus on all eight sound-quality dimensions for a single 50- to 60-second sample, would both reduce the time of testing and permit repeated trials of the same conditions to enhance the reliability of the individual JSQ ratings. However, the reliability of the test with modifications in the methodology will need to be determined before use in clinical situations. This is especially crucial in case of the music stimuli, for which the physical attributes of the sound change considerably over the course of a single stimulus presentation. The use of group JSQ ratings, on the other hand, appears to be both a reliable and valid tool for use in the evaluation of hearing aid outcome.

Acknowledgments:

The authors would like to thank Melissa Coy-Branam for her assistance with data collection. This work was supported, in part, by a research grant to the second author from the National Institute on Aging.

Address for correspondence: Mini Narendran, M.Sc., Department of Speech and Hearing Sciences, Indiana University, 200 South Jordan Avenue, Bloomington, IN 47405-7002. E-mail: mnarendr@indiana.edu.

Received August 29, 2001; accepted June 28, 2002

References


