The Development and Validation of the Speech Quality Instrument

Stephanie Y. Chen, MD; Brianna M. Griffin, MD; Dean Mancuso, AuD; Stephanie Shiau, PhD; Michelle DiMattia, MA; Ilana Cellum, AuD; Kelly Harvey Boyd, MS; Charlotte Prevoteau, MD; Gavriel D. Kohlberg, MD; Jaclyn B. Spitzer, PhD; Anil K. Lalwani, MD

Objective: Although speech perception tests are available to evaluate hearing, there is no standardized validated tool to quantify speech quality. The objective of this study is to develop a validated tool to measure quality of speech heard.

Study Design: Prospective instrument validation study of 35 normal hearing adults recruited at a tertiary referral center.

Methods: Participants listened to 44 speech clips of male/female voices reciting the Rainbow Passage. Speech clips included original and manipulated excerpts capturing goal qualities such as mechanical and garbled. Listeners rated clips on a 10-point visual analog scale (VAS) of 18 characteristics (e.g. cartoonish, garbled).

Results: Skewed distribution analysis identified mean ratings in the upper and lower 2-point limits of the VAS (ratings of 8–10, 0–2, respectively); items with inconsistent responses were eliminated. The test was pruned to a final instrument of nine speech clips that clearly define qualities of interest: speech-like, male/female, cartoonish, echo-y, garbled, tinny, mechanical, rough, breathy, soothing, hoarse, like, pleasant, natural. Mean ratings were highest for original female clips (8.8) and lowest for not-speech manipulation (2.1). Factor analysis identified two subsets of characteristics: internal consistency demonstrated Cronbach’s alpha of 0.95 and 0.82 per subset. Test–retest reliability of total scores was high, with an intraclass correlation coefficient of 0.76.

Conclusion: The Speech Quality Instrument (SQI) is a concise, valid tool for assessing speech quality as an indicator for hearing performance. SQI may be a valuable outcome measure for cochlear implant recipients who, despite achieving excellent speech perception, often experience poor speech quality.

Key Words: Speech perception, speech quality, validated tool.

Level of Evidence: 2b.

INTRODUCTION

Speech quality is essential for normal hearing (NH) listeners’ experience of telecommunications, multimedia, and sound systems, contributing to speech’s intelligibility and a listener’s overall experience of sound. In addition, patient populations including hearing aid (HA) users, cochlear implant (CI) recipients, and hard-of-hearing patients experience changes in the quality of speech that they hear. For example, some CI patients complain of speech as sounding mechanical or tinny. However, although there are several outcome measures to evaluate hearing, there is currently no validated tool to assess the subjective quality of speech heard across multiple characteristics.

There are many objective measures for speech quality, including acoustic pressure, weighted spectral slope, articulation index, and speech transmission index for clean and degraded speech signals.1,2 Whereas these are useful measures for quality, they were originally developed to predict speech-coding distortions and do not necessarily correlate with subjective speech quality. On the other hand, the Speech, Spatial, and Qualities of Hearing Scale (SSQ) is a widely used, 49-item survey that assesses a listener’s subjective hearing ability.3 The SSQ includes an assessment of whether “other people’s voices sound clear and natural” on a visual analog scale (VAS) of 0 to 10. However, this survey addresses only two of many characteristics reported by patients with hearing loss. Skinder-Meredith et al. developed the 22-item Speech Characteristics Rating Form (2009) to assess speech qualities in childhood speech sound disorders, which includes a 5-point VAS of how hoarse, breathy, monotone, nasal, and resonant speech sound.4 However, this survey has not been formally validated and addresses only a minority of the characteristics.
In addition, speech perception scores themselves cannot be used as an indirect measure of quality of speech perception because it is unknown whether there is a direct association between speech perception and speech quality. For example, although many CI users are able to achieve excellent speech perception in quiet, they still complain of speech quality. This is similar to the disassociation between music perception in quiet, they still complain of speech quality. Thus, there is a critical need for a validated tool to measure speech quality that addresses multiple qualities reported by patients with hearing impairment. The objective of our study is to address this shortcoming by developing a valid, concise, self-administered, subjective instrument to assess perceived speech quality.

MATERIALS AND METHODS

The Speech Quality Instrument (SQI) was created in two phases: 1) instrument development and 2) instrument validation.1–11 (Fig. 1).

Subjects

This study was approved by the Institutional Review Board of Columbia University Medical Center (New York, NY). Using flyer, word-of-mouth, and online postings, 35 NH adults were recruited for the study. All participants were native English-speaking and at least 18 years old. Normal hearing subjects with a history of hearing loss or otologic surgery were excluded.

Phase 1: Instrument Development

Item Selection. The content of the proposed instrument was generated by a focus group of otolaryngologists, audiologists, and speech pathologists, who identified speech qualities of interest based on their extensive experience with patients with hearing loss and CI users. The initial alpha instrument included 18 speech quality-related items that were scored using a continuous, 10-point bipolar VAS. The 18 different qualities included: male/female, mechanical/not mechanical, cartoonish/not cartoonish, natural/not natural, pleasant/not pleasant, clear/garbled, like/dislike, breathy/not breathy, smooth/rough, echo-y/not echo-y, tinny/bassy, far/near, hollow/full, strong/weak, soothing/not soothing, monotone/animated, hoarse/not hoarse, and sounds like speech/does not sound like speech. Higher numeric scores corresponded to better speech quality.

Development of Speech Stimuli. To generate the speech stimuli, two male and two female speakers individually recorded the same excerpt from the Rainbow Passage. We chose to include two speakers from both sexes based on prior speech instrument validation studies12 for a broad selection of both male and female samples and without creating an unmanageable number of samples for the subjects. Speakers were seated in a soundproof audiology booth and used a Shure microphone (Shure Inc., Niles, IL) positioned 6 to 12 inches away. Each speaker was instructed to recite the passage at a normal conversational pace, volume, and degree of enunciation. Recordings were monitored by a speech pathologist; in the event of mispronunciation or misread words, the speaker was instructed to repeat the sentence. Speech stimuli were shortened to 5-second clips, or the first sentence of the Rainbow Passage. These stimuli were then manipulated by a sound engineer, as described below, to capture 10 goal qualities (bassy, cartoonish, echo-y, far, garbled, hollow, mechanical, not speech, rough, tinny) for a total of 44 speech stimuli in the alpha instrument (4 original and 40 sound-engineered samples).

Speech Engineering. Modifications to speech clips to achieve the desired characteristic were accomplished using Apple Logic 9 Pro recording software (Apple Inc., Cupertino, CA). Parameters including pitch, time, color, mutation, dialect, reflectivity, formant, tilt, and linear attenuation were altered to generate the desired speech qualities. For example, to generate the cartoonish characteristic, processing was achieved using the Logic Vocal Transformer (Apple Inc.) to raise the overall pitch and formant of the vocal recordings by +8 for male and +6 for female clips. A similar approach was applied to generate all 10 desired speech characteristics (bassy, cartoonish, echo-y, far, garbled, hollow, mechanical, not speech, rough, tinny; see Supporting Materials and Methods for all manipulations).

Preliminary Testing in Normal Hearing. The alpha instrument was administered to the first cohort of 20 NH listeners, who listened to speech stimuli and rated them using the 18-item scale. To reduce listeners’ fatigue, the test was divided into five parts (parts A–E), and listeners could take an optional 5-minute break between parts. Each part consisted of 44 speech clips and four items, except for part E, which contained two items. Altogether, the alpha instrument consisted of 220 clips and took approximately 1 hour to complete.

Pruning of the Alpha Instrument to Finalized Speech Quality Instrument. To enable maximum clinical utility, we sought to create a final instrument that was as concise as possible to minimize respondent burden and testing time. We specifically aimed to create a self-administered test that could be completed in half an hour. Mean quality scores for each clip were calculated for each of the 18 items. We then eliminated items and clips with inconsistent or redundant responses, and kept those with skewed distributions within the upper and lower 2 points. Final retained clips therefore included original and manipulated clips that uniquely characterized a number of speech qualities. Additionally, final speech quality-related items represented speech qualities that had similar meaning to most listeners. The test was pruned to a final SQI of nine clips and 14 speech quality-related items, which would take approximately 25 minutes to complete. Another cohort of 15 NH listeners was recruited to take the final SQI and was retested within 1 week.

Scoring of SQI. Two types of scores were calculated. First, a total score was calculated by adding the raw scores for each item and dividing by 10. Given 14 total speech-quality items and nine clips, the highest achievable total score would be 126, with the highest score indicating the best possible speech quality. Next, means with 1 standard deviation (SD) of
ratings were calculated for each characteristic and clip of the SQI. These ranges would serve as the normative data, or scoring rubric, for each item to evaluate SQI results of non-NH listeners.

**Phase 2: Instrument Validation**

We adhered to classical test theory to validate the finalized instrument by performing validity (content and construct validity) and reliability (internal consistency, test–retest reliability) analyses. Content validity was ensured during design of the instrument in phase 1. Absence of gold standard for assessing speech quality precluded achieving criterion validity. Construct validity was assessed using Spearman inter-item correlation (moderate if \( r \geq 0.35 \)) and factor analysis. Exploratory factor analysis with varimax rotation was performed on all 14 items; scree test was used to determine the factor structure. Clips with factor loadings \( \geq 0.5 \) were retained for both analyses.

Reliability was assessed by measuring internal consistency and calculating the Cronbach \( \alpha \) coefficient. Internal consistency reliability is considered adequate if Cronbach \( \alpha \geq 0.70 \).\(^{a}\) Test–retest reliability was also assessed by administering the final SQI to study participants who were retested within 1 week, and calculating intraclass correlation coefficients (ICC) of total scores using a mean-rating, two-way random effects model. ICC values between 0.75 and 0.9 indicate good reliability.\(^{13}\)

**Data Collection and Analysis.** Testing was administered in a soundproof room on a designated research laptop (Dell Inspiron 11 3000 Series, Dell Technologies, Round Rock, TX) that both administered the test and recorded testing data. Testing stimuli were presented via SONY MDR-CD60 headphones (Sony Corp., Minato, Tokyo, Japan) connected to the headphone jack of the laptop. Before beginning the test, participants were encouraged to adjust the volume to their preferred, comfortable listening level.

Participants listened to and rated speech clips using the 10-point VAS, as previously described, using MathWorks, Natick, MA). Speech clips were presented at a random order during each part of the test. No feedback was given during the test, and participants listened to each clip only once before submitting ratings. A screenshot of the computerized test is shown in Figure 2.

Data analysis was performed using StataCorp Stata Statistical Software version 14.1 (College Station, TX).

**RESULTS**

A total of 35 NH listeners participated in this study. The mean age was 27.9 ± 3.1 years (mean ± SD for all values), and 54.3% of participants were female. Twenty participants were administered the alpha instrument. The 15 remaining participants were administered the finalized, pruned instrument, all of whom completed the test–retest portion of the study. With the exception of test–retest reliability, data from all 35 NH participants were used for all of the validation statistics of the final SQI. Although there is no consensus on the number of subjects needed for psychometric validation studies, some studies suggest that a sample size of 25 to 50 is considered adequate for studies utilizing factor analysis.\(^{10}\)

**Phase 1: Instrument Development**

Mean ratings of all speech clips and items were calculated to identify the skewed distributions, or the items and clips with ratings in the upper and lower 2-point limits of the VAS (ratings of 8–10, 0–2 respectively). Mean ratings of speech clips in NH listeners ranged from 0 to 10 across all characteristics. Ratings were highest for original female (8.8 ± 0.7) and lowest for the not-speech manipulation (2.1 ± 0.6). Initial item review demonstrated that items far/near, full/hollow, strong/weak, and animated/monotone resulted in inconsistent responses and no skewed distributions; therefore, these were eliminated from the final instrument. To minimize patient burden and testing time, individual clips were also eliminated to identify one representative clip per manipulation, for a total of nine final speech clips recited by five male and four female voices.

The final SQI consisted of nine clips, with two original and seven engineered clips (original male, original female, not-speech female, bassy male, cartoonish female, far male, garbled male, mechanical female, rough male), and 14 speech quality-related items (male/female, mechanical/not mechanical, cartoonish/not cartoonish, natural/not natural, pleasant/not pleasant, like/dislike, breathy/not breathy, smooth/not smooth, echo-y/not echo-y, tinny/not tinny, soothing/not soothing, hoarse/not hoarse, and speech/not speech, clear/garbled). This final version takes approximately 25 minutes to complete.

**Scoring of the Speech Quality Instrument.** Total scores averaged 68.7 ± 7.0 (range 51.5–82.9). Normal hearing data was also used to develop a scoring rubric to
assess ratings of non-NH listeners, that is, HA or CI users for future studies. Means with 1 SD were calculated for each retained speech-quality item and clip of the SQI (Table I). Clips and items with mean ratings within the upper or lower 2-point VAS limits were identified as significant given that they represented the skewed distributions, or the clips that NH listeners agreed were good representations of specific speech qualities, and there was general consensus on the meaning of a specific quality. For example, for the not-cartoonish quality item, there were five significant clips: bassy, cartoonish, mechanical, original female, and original male. Given the low rating of the cartoonish clip (0.3 for the not-cartoonish quality), this represents agreement of the clip as sounding very cartoonish and also of the linguistic meaning of cartoonish. The only exception to this rule was the bassy clip, fitting its linguistic meaning of the speech quality, both linguistically and subjectively upon listening to the speech clip, among NH listeners. These highlighted clips are thus significantly associated with that characteristic or item. The only exception to this rule was the bassy/tinny characteristic for the bassy clip. We decided to include the bassy item in the final instrument because it matched the intended goal of the bassy manipulation and is also a common complaint in CI users. The last row tabulates the total number of significant clips associated with each item as a proposed scoring mechanism or gold standard to evaluate non-NH results.

### Phase 2: Instrument Validation

The pruned, finalized instrument was validated in NH listeners by assessing content validity, construct validity (Spearman inter-item correlation, factor analysis), and reliability (internal consistency, test–retest). As described above, content validity was achieved in stage 1 during the design of the instrument. To assess construct validity, item–item correlation analysis demonstrated moderate correlation of all items with at least one other retained item ($r \geq 0.35$, Spearman). Scree test identified two subsets of characteristics for factor analysis. Items with factor loadings $\geq 0.5$ were retained (Table II), and internal consistency demonstrated Cronbach alpha coefficients of 0.95 and 0.82 for factors 1 and 2, respectively. As shown in Table II, seven items loaded onto a primary factor, four items loaded onto a secondary factor, and two items (like/dislike and smooth/rough) cross-loaded onto both factors. Bassy/tinny was the only characteristic that did not significantly load onto either factor (loading of 0.48). The three items (like/dislike, smooth/rough, and bassy/tinny) were therefore marked for potential elimination. Test–retest reliability of total scores was high, with an ICC of 0.76 (95% confidence interval: 0.50–0.92, $P < 0.001$).

### DISCUSSION

In this study, we developed and validated a new tool, the SQI, to evaluate a listener’s perception of speech quality in NH adults. SQI addresses the critical need for a validated tool to measure commonly reported patient complaints of the quality of speech perceived using their auditory assistive devices. The SQI is brief, easy-to-complete, and self-administered on the computer, with minimal respondent burden. This makes the test clinically practical and well-suited for repeated prospective testing and broad implementation in multiple clinical settings.

There were significant relationships between specific items on the SQI, as evidenced by the moderate inter-item correlation of all retained items and strong internal consistency with high Cronbach’s alphas. We also demonstrated excellent test–retest reliability with high ICC, indicating that the test results were consistently reproducible in participants with no change in functional status. On factor analysis, all items except for the bassy/tinny quality loaded strongly onto either


factor. Although the bassy/tinny item has been marked for potential elimination, it demonstrated near-significant loading at 0.48. Given the relevance of the quality, along with the trend toward significance, we decided to include this quality on the final, pruned instrument for further testing in other patient populations. For example, many CI patients complain of the speech they hear as sounding bassy or tinny, and thus the item is highly relevant as a quality of interest. The item also demonstrated moderate correlation with the cartoonish/not cartoonish speech quality, suggesting that there is an association between how cartoonish and tinny speech sounds. In addition, the like/dislike and smooth/rough items loaded onto both factors, indicating redundancy of the items; however, these are characteristics often reported by patients with hearing loss. Future testing is thus needed to determine whether elimination of the bassy/tinny, like/dislike, and smooth/rough items is necessary for each patient cohort.

The SQI is designed to be flexible in both how it is administered and how the results are presented. For example, the clinician could choose to focus on assessing the cartoonish quality alone; tabulating the number of significant clips with ratings within 1 SD of the NH ratings would provide a metric to adjust programming settings. Alternatively, total scores may be calculated using the comprehensive panel of characteristics to assess overall speech-quality performance (calculated as a percentage by dividing the total score by the 68.7 mean total from NH data), or using specific speech clips to further reduce testing time. Other potential uses and outputs for the test are outlined in Table III. This ability to adaptively measure and present results on speech quality will facilitate patient-directed care to improve the appreciation of speech quality.

There are some limitations to the study. First, we did not assess response sensitivity of the test, or the ability of the test to detect clinical change, which is often included as a component of classical test theory. We also did not include concurrent testing using speech perception tests to assess the association of speech quality and speech perception. Given that the study is performed in NH subjects, the responsiveness and speech perception measures are irrelevant given the lack of clinical change or hearing loss in this cohort. Future studies should examine other relevant patient populations, such as patients with hearing loss and CI recipients, to assess the validity, reliability, and response sensitivity of the tool, in addition to its relationship with speech perception performance.

CONCLUSION

The SQI is a concise, computerized test that addresses a critical need for a validated tool to measure perception of speech quality. Based on our validity studies in NH listeners, we have demonstrated validity and reliability of the tool. Using NH data as the normative gold standard, further validity testing should be explored in hard-of-hearing cohorts, hearing aid users, and CI recipients to validate the tool in these patient populations. These populations are ideally suited for investigating the relationship between speech quality and speech perception. The availability of a valid, self-administered instrument to assess speech quality will aid in outcomes research and provide insight on optimizing care in patients with hearing abnormalities.

<p>| TABLE II. Mean Item Ratings and Factor Loadings for Final Items Retained |</p>
<table>
<thead>
<tr>
<th>Quality</th>
<th>Mean Rating</th>
<th>SD</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not cartoonish</td>
<td>5.7</td>
<td>4.0</td>
<td>0.665</td>
<td></td>
</tr>
<tr>
<td>Clear</td>
<td>4.5</td>
<td>4.0</td>
<td>0.680</td>
<td></td>
</tr>
<tr>
<td>Like</td>
<td>3.8</td>
<td>3.7</td>
<td>0.666</td>
<td>0.564</td>
</tr>
<tr>
<td>Not breathy</td>
<td>7.1</td>
<td>2.9</td>
<td></td>
<td>0.507</td>
</tr>
<tr>
<td>Smooth</td>
<td>5.2</td>
<td>3.9</td>
<td>0.643</td>
<td>0.544</td>
</tr>
<tr>
<td>Not echo-y</td>
<td>5.9</td>
<td>3.6</td>
<td></td>
<td>0.531</td>
</tr>
<tr>
<td>Bassy</td>
<td>4.9</td>
<td>2.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soothing</td>
<td>4.0</td>
<td>3.4</td>
<td>0.762</td>
<td></td>
</tr>
<tr>
<td>Natural</td>
<td>4.4</td>
<td>3.9</td>
<td>0.850</td>
<td></td>
</tr>
<tr>
<td>Not mechanical</td>
<td>5.1</td>
<td>3.9</td>
<td>0.734</td>
<td></td>
</tr>
<tr>
<td>Not hoarse</td>
<td>6.8</td>
<td>3.3</td>
<td></td>
<td>0.503</td>
</tr>
<tr>
<td>Pleasant</td>
<td>4.5</td>
<td>3.4</td>
<td>0.833</td>
<td></td>
</tr>
<tr>
<td>Sex ID</td>
<td>8.4</td>
<td>2.0</td>
<td>0.728</td>
<td></td>
</tr>
<tr>
<td>Speech-like</td>
<td>6.2</td>
<td>3.7</td>
<td>0.719</td>
<td></td>
</tr>
<tr>
<td>Cronbach alpha</td>
<td>0.95</td>
<td>0.82</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scree test identified two subsets of characteristics. Items with factor loadings ≥ 0.5 were retained and reported above; blanks represent factor loadings < 0.5. Internal consistency demonstrated Cronbach alpha of 0.95 and 0.82 for factors 1 and 2, respectively. All items demonstrated moderate interitem correlation (r ≥ 0.35, Spearman).

TABLE III. Multiple Applications of the Speech Quality Instrument to Meet Clinical Needs

| Comprehensive testing using the entire SQI for baseline speech quality performance |
| Benchmark assessment: comparison of comprehensive speech quality performance to the mean performance of other listeners with similar characteristics (i.e., HA or CI users with a similar duration of HA use or implantation). |
| Playing only the original clips of the SQI to assess how patients may perceive speech quality on a daily basis. |
| Focused assessment of one or multiple speech qualities in a listener with one or several complaints about quality: clinician plays only the clips significantly associated with the quality, then adjusts programming settings accordingly (i.e., for a HA or CI user) and retests the patient to assess improvement. |
| Variations on presentation of results: display results using one graph for each individual quality, assessed against the mean with 1 standard deviation of NH and CI user performance; display mean VAS scores using a summary table with individual scores per item and clip. |
| Alternative scoring: tabulating the number of speech qualities captured (with NH as gold standard) or generating a mean VAS rating per characteristic |

| CI = cochlear implant; HA = hearing aid; SQI = speech quality instrument; VAS = visual analog scale. |