Fitting Hearing Aids Using Clinical Prefitting Speech Measures: An Evidence-Based Review

Mead C. Killion*
Gail I. Gudmundsen*

Abstract
A systematic review of the literature addressed the question "Is there evidence of a good correlation between unaided prefitting speech measures and aided satisfaction on self-report measures?" This restricted question is only one of several possible questions related to speech measures and hearing aid fittings. The levels of evidence that were accepted included meta-analyses, randomized controlled trials, and nonrandomized intervention studies. Nearly 300 articles and book chapters were identified during the initial search; 220 were eliminated on the basis of their abstracts; and 80 papers and book chapters were reviewed in depth. Five studies met the criteria set forth in this review. No significant correlation between traditional unaided prefitting speech measures and aided satisfaction was found in any of the five studies. One of the studies showed a correlation between the results of a prefitting speech-in-noise test and self-reported aided satisfaction.

Key Words: Evidence-based review, hearing aids, self-reported satisfaction, speech discrimination, systematic review, word-recognition scores

Abbreviations: CID = Central Institute for the Deaf; IU = Indiana University; NIDCD/VA = National Institute on Deafness and Other Communicative Disorders/Veterans Administration; NU = Northwestern University; PB = phonetically balanced; QuickSIN = quick speech-in-noise test; SI = sentence intelligibility; SIN = speech-in-noise test; SNR = signal-to-noise ratio

Sumario
Una revisión sistemática de la literatura buscó respuesta a la pregunta “¿Existe evidencia de una buena correlación entre las mediciones de lenguaje no amplificado antes de la adaptación y las mediciones auto-reportadas de satisfacción con la amplificación?” Esta pregunta restringida es sólo una de las varias posibles preguntas que se relacionan con mediciones de lenguaje y la adaptación de auxiliares auditivos. Los niveles de evidencia que se aceptaron incluyeron meta-análisis, estudios aleatorios controlados, y estudios de intervención no aleatorios. Se identificaron alrededor de 300 artículos y capítulos de libros durante la búsqueda inicial; 220 fueron eliminados con base en sus sumarios; y 80 artículos y capítulos se revisaron a profundidad. Cinco estudios cumplieron los criterios establecidos en esta revisión. No se encontró una correlación significativa entre las mediciones de lenguaje no amplificado antes de la adaptación y la satisfacción en la amplificación en ninguno de los cinco estudios. Uno de ellos mostró una correlación entre los resultados de una prueba de lenguaje en ruido previa a la adaptación y el auto-reporte de satisfacción con dicha amplificación.

*Etymotic Research, Elk Grove Village, IL

Mead Killion, Etymotic Research, 61 Martin Lane, Elk Grove Village, IL; Phone: 847-228-0006; Fax: 847-228-6836; E-mail: m_killion@etymotic.com

In the introduction to a special issue of *JAAA* on outcome measures, Jerger commented that the era in hearing aid fitting from 1946 to about 1970 “might be described as the era of the ‘PB [phonetically balanced] list’” (2000, p. 350). The second era might be described as “the engineering solution,” and the third era, “which is just beginning, might be described as the era of the outcome questionnaire. A fitting is successful when the hearing aid user says it is successful (the ultimate statement of accountability)” (Jerger, 2000, p. 350). Self-reported satisfaction was the primary outcome measure we studied.

Clinical measurements of speech intelligibility have been part of hearing aid fitting protocols since the 1940s. In this paper, we discuss the evidence in support of the use of speech-based measures—word recognition tests, sentence tests, and speech-in-noise tests—as unaided prefitting measures on the basis of which it is possible to improve patient satisfaction with hearing aids as measured on self-report satisfaction scales. In order to put these things in proper perspective, we first review the historical use of speech measures. The reader who wishes to skip over the historical account is invited to go directly to the “Results” section.

Our search for evidence used (1) Web search engines PubMed, Ingenta, CINAHL (Cumulative Index to Nursing and Allied Health Literature), and ComDisDome, (2) a search through any file folder with related titles among some 2400 articles in our files, (3) a review of 22 books on hearing and hearing aids that included Skinner’s (1988) exhaustive review of evidence on all aspects of hearing aid fitting, and (4) telephone inquiries to Bill Carver, Robyn Cox, Ira Hirsh, Larry Humes, Sergei Kochkin, Wayne Olsen, Margo Skinner, Brian Walden, and Laura Ann Wilber. Finally, we checked numerous references in each journal article uncovered during the above search if an article included information on speech testing and hearing aid fitting. As an aside, what we now correctly call word-recognition scores was earlier called speech discrimination scores until Robert Bilger enthusiastically reminded all of us that we were using words and not speech, and that the task was recognition and not discrimination.

Prefitting speech measures can be (1) used to guide the hearing aid fitting, (2) used in patient counseling to set realistic expectations, or (3) simply noted for reference or for research purposes. In this study, we searched for any significant correlation between prefitting speech measures and satisfaction, under the assumption that the absence of any strong correlation warranted no further investigation. Thus, we searched for studies in which prefitting speech measures and outcome measures—specifically self-reported satisfaction—were reported on the same subjects.

**BRIEF HISTORY**

The earliest formal measurement of speech used in hearing aid fitting appears to date back to Bryant (1904), who described the use of the phonograph for hearing tests. Wharry recommended that “the patient first be tested by using the 4A audiometer with disk recording” (1932, p. 574). The choice of which ear to fit with the hearing aid was “based on the hearing loss for speech” (Berger, 1970, p. 130). In his book *Hearing and Deafness*, Davis (1947) reproduced many of the word lists available at the time, including...
Spondee lists, Larsen Sound-Discrimination lists, Rhyme lists, Question-Answer lists, Harvard PAL (Psychoacoustics Lab) PB-50 word lists, and SI (sentence intelligibility) Sentence lists. The SI Sentence list later came to be called the “IEEE (Institute of Electrical and Electronics Engineers) sentences,” selected subsets of which were used in the SIN (speech-in-noise) test and QuickSIN (quick speech-in-noise) test (Killion et al, 2004).

Watson and Tolan (1949) recommended aided speech measures but did not discuss the possible value of speech as a prefitting measure.

Carhart (1946) said that a measure of the patient’s unaided speech reception threshold was the first step in his protocol for selecting hearing aids, but its only apparent use was in conjunction with the second step—measuring the aided speech reception threshold—to determine the effective gain of the hearing aid. Carhart is often credited with recommending that the hearing aid corresponding to the highest speech discrimination score in noise should be chosen, but he also stated that in most cases patients will obtain excellent scores with each of several hearing aids. He suggested, however, that a score difference of eight percentage points was sufficient to indicate selecting the hearing aid that yielded the superior score. Despite the fact that it did not receive experimental validation in terms of benefit or satisfaction, the “Carhart method” of selecting one aid from among three hearing aids found its way into clinical practice.

The Harvard Report (Davis et al, 1947) used PB-50 speech scores at ten different presentation levels (obtained by varying the gain in a low-distortion master hearing aid) and three signal-to-noise ratios (SNRs) of 0, 5, and 10 dB. The summary of the Harvard Report, based on extensive speech testing on each subject, concluded that “the idea of individual selective amplification is fallacious” (Davis et al, 1946, p. 68).

Fletcher (1953) gave the first strong evidence that the conclusions of the Harvard Report were seriously in error. He demonstrated that the various speech scores of the Harvard Report could be predicted quite well for each subject from the subject’s threshold audiogram, once the erroneous assumptions about the real-ear headphone response in the master hearing aid used in the Harvard study were taken into account. His calculations used the Bell Labs’ articulation index (AI) reported by Fletcher (1929) and later by French and Steinberg (1947).

Harford and Dodds (1966) and Pascoe (1975) gave conclusive experimental evidence that choosing the frequency response to compensate for the audiogram gave significantly better word-recognition scores than the Harvard Report’s recommendation to use the same frequency response for everyone. In the Harford and Dodds and the Pascoe studies, however, speech testing was used as an outcome measure, not as a prefitting measure. The frequency responses of the experimental hearing aids were chosen on the basis of the subject’s prefitting audiogram, not any speech measures.

Indeed, Fletcher’s (1953) analysis implied that the hearing aid frequency response and gain that would provide best speech intelligibility for a given subject could be predicted from the audiogram without need for speech measures. Three comments are in order. First, in addition to the audiometric data, Fletcher used the average speech test data in his calculations relative to the scores in the Harvard Report. Use of the speech data resulted in a better prediction of each subject’s average performance but did not affect the calculated best frequency response. Second, the “best” speech intelligibility identified in Fletcher’s analysis did not necessarily mean “good,” only better than the intelligibility with alternative frequency responses. Thus, the best frequency response for a given subject might still yield poor intelligibility in noise or even in quiet. Third, many of these subjects had conductive hearing losses that did not require individualized adjustment of the frequency response.

**RELIABILITY**

Although Carhart (1946) suggested that speech discrimination scores could be used to select hearing aids, Shore et al (1960) reported that they could not reliably distinguish among intentionally created “good” and “bad” experimental hearing aid fittings on the basis of the 25-word speech discrimination scores commonly used.

We now know that the score for one of two
hearing aids will exceed that of the other nearly 30% of the time purely by chance for scores in the 30–70% range, using the 25-word lists often used today. Even with 50-word lists, an 8% difference would occur by chance 21% of the time between equally performing hearing aids.

The statistical arguments in the preceding paragraph were based on the binomial distribution. Boothroyd (1968) first showed that the binomial distribution gave a good prediction of the variability of speech scores. Thorton and Raffin (1978) made a major contribution by demonstrating that some 4120 scores obtained on 50-word CID (Central Institute for the Deaf) W-22 lists, compared for reliability across lists and half lists, gave an excellent fit to the binomial prediction (Figure 1). An enjoyable debate subsequently ensued between Dillon (1982) and Thorton and Raffin (1982) over their paper. What was particularly enjoyable about the debate, besides the fact that it brought to light several important fine points in the use of statistics, was that, when the dust had cleared, the error in standard deviation being debated (a factor of 1.05) amounted to the difference between a 95% and a 95.3% confidence level.

Note: A simple mnemonic, useful for estimating the approximate standard deviation of N-word scores between 30% and 70%, is \( sdev = 50\%/\sqrt{N} \). For example, 25-word lists give an estimated standard deviation of 10% for a single test score, and thus the critical difference between two scores at the 95% confidence level is approximately 28% (1.96*1.41*10%). If a subject scores 40% wearing one aid and 64% wearing a different hearing aid, this is not a statistically significant difference using a 25-word list.

Schwartz and Walden (1980) and Walden et al (1982) conducted a series of experiments using 50- and 100-word NU-6 lists. These experiments further demonstrated the futility of using speech scores to choose among hearing aids. In an experiment using 50-word NU-6 lists, Schwartz and Walden showed on retest two weeks later that the aid originally selected as “best” was lower, on the average, than the aid that was “second best” on the original test. This is a nice example of the statistical regression to the mean. Mead C. Killion (unpublished data, 1984) used a two-digit random number table to generate fictitious “speech scores” for eight subjects, creating a table remarkably similar to that of Schwartz and Walden, consistent with their explanation that the hearing aid was usually selected by chance. Studebaker (1982) calculated that to identify a true difference of 8 percentage points between two aids at the 95% confidence level would require 211 words in each condition, or a total of 422 words: eight 50-word lists (Table 1)!

![Figure 1](image-url)  

**Table 1. Number of Words Required in Speech Lists to Reliably Recognize a True Difference at Various Significance Levels**

<table>
<thead>
<tr>
<th>“True Difference” (in %)</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>33%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6785</td>
<td>3381</td>
<td>2048</td>
<td>822</td>
<td>241</td>
</tr>
<tr>
<td>4</td>
<td>1696</td>
<td>845</td>
<td>511</td>
<td>221</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>753</td>
<td>376</td>
<td>227</td>
<td>98</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>424</td>
<td>211</td>
<td>128</td>
<td>55</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>271</td>
<td>135</td>
<td>82</td>
<td>35</td>
<td>10</td>
</tr>
</tbody>
</table>

*Note: From Studebaker (1982), used with permission.*
The conclusion has been clear for some time: Within the time constraints of clinical practice, evidence-based practice can not justify choosing hearing aids on the basis of speech-test performance, certainly not based on an 8% difference. Skinner (1976) and Pascoe (1975) and the Harvard Report (Davis et al, 1947) all used hundreds of words with each subject to derive reliable results.

**VALIDITY**

Another pitfall in the use of speech tests in the prefitting of hearing aids arises from the fact that the intelligibility of a word may be controlled more by the talker than by the apparent difficulty of the word. Figure 2 illustrates Hood and Poole’s most dramatic comparison. Only two of their test subjects missed the test word “wedge” spoken by Speaker #2, while ten missed that word when spoken by Speaker #1. On the other hand, 12 subjects missed “our” spoken by Speaker #2, while only two missed “our” spoken by Speaker #1.

In a recent study of the equivalence of NU-6 lists, Stuart (2004) concluded that the NU-6 lists are not equivalent when used in noise. The results of Hood and Poole suggest that Stuart’s findings may have resulted from his use of a female talker whose relative clarity of pronunciation for NU-6 words differed from that of Tillman, whose recorded voice was used in the original NU-6 (male) equivalence studies. Hood and Poole’s study suggested that no general conclusion can be drawn about the equivalence of any lists of words per se without specifying the talker.

**SPEECH-IN-NOISE TESTS FOR SNR LOSS**

Although speech testing in noise goes back almost as far as speech testing in quiet, the common measure was percentage of words correct. Dirks et al (1982) suggested reporting the signal-to-noise ratio required for 50% words correct, rather than the traditional word-recognition scores in percent. They reported the results of their method on a large group of hearing-impaired subjects. Subsequently, the HINT (hearing-in-noise) test was introduced (Nilsson et al, 1994), and more recently the SIN test and QuickSIN test have been made available (Killion et al, 2004). The signal-to-noise ratio required to understand speech can now be measured by one of these methods.

**RESULTS**

Our present literature search produced five studies that correlated prefitting speech measures and satisfaction on self-report measures. Kapteyn (1977) gave a questionnaire to 160 patients six months after the prescription of the hearing aid. He found a minimal correlation (0.15) between the Dutch “speech discrimination score” and the subjective overall satisfaction score (“the subjects’ appreciation of the aid”) based on 38 questions scored on a scale of one to ten. The questions included “Do you use your hearing aid at home?” “Are you troubled by loud noises?” and “How does the sound of your hearing aid appeal to you?”

Parving (1991) reported that the aided Danish speech recognition score in noise (SRSN) of the 124 patients in their study could be predicted from their unaided SRSN score but that it was not possible to predict either hearing aid use or self-assessed benefit.
Table 2. Summary of Research Meeting Search Criteria

<table>
<thead>
<tr>
<th>Reference</th>
<th>Design</th>
<th>Intervention</th>
<th>Outcomes</th>
<th>Results</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kapteyn TS. (1977)</td>
<td>N = 160</td>
<td>Hearing aids</td>
<td>Dutch speech discrimination.</td>
<td>No correlation</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>Nonrandomized intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parving A. (1991)</td>
<td>N = 124</td>
<td>All fitted with ITC, same circuit: 67 binaural, 57 monaural</td>
<td>Unaided/Aided 25-word monosyllables (Danish) at +10 dB SNR.</td>
<td>No correlation</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>Nonrandomized intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humes et al. (2003)</td>
<td>N = 76</td>
<td>2 of 3 groups received bin HAs; linear, Class D-OL</td>
<td>NU-6, Time-compressed Quiet&amp;SNR+12; HHIE, CPHI</td>
<td>No correlation; no differences among groups; could not predict use/nonuse</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nonrandomized intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humes LE. (2003)</td>
<td>N = 173 NU-1, N = 53 NU-2</td>
<td>IU-1, single channel linear ITE; IU-2, 2Ch WDRC ITEs;</td>
<td>NST-U/A; CST-U/A; CST-U50Q; CST-U80+8SNR; CST-U80-OSNR; HAPI, HDABI, GHABP, MarkeTrak IV</td>
<td>No correlation</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td>Meta-analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NIDCD/VA multicenter study</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N = 338; double-blind</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>crossover design</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walden TC,</td>
<td>N = 50</td>
<td>Hearing aids</td>
<td>U-QS; A-QS; U-AI; A-AI; NU-6</td>
<td>Best predictor = (+)</td>
<td></td>
</tr>
</tbody>
</table>

Note: “++” = strongly support; “+” = support; “” = equivalent findings; “-” = not support; “—” = strongly not support; A-AI = aided articulation index; A-QS = aided QuickSIN; Class D-OL = class D output with output limiting; CPHI = Communication Profile for the Hearing Impaired; CST-U/A = Connected Sentence Test—unaided/aided; GHABP = Glasgow Hearing Aid Benefit Profile; HA = hearing aid; HAPI = Hearing Aid Performance Inventory; HAUS = Hearing Aid Usefulness Scale; HDABI = Hearing Disability and Benefit Inventory; HHIE = Hering handicap Inventory for the Elderly; ITC = in-the-canal hearing aid; IU = Indiana University; NIDCD/VA = National Institute on Deafness and Other Communicative Disorders/Veterans Administration; NST/U/A = nonsense syllable test—unaided/aided; NU = Northwestern University; SADL = Satisfaction with Amplification in Daily Life; SNR = signal-to-noise ratio; U-AI = unaided articulation index; U-QS = unaided QuickSIN; WDRC = wide dynamic range compression.
Humes et al (2003) studied three groups of hearing aid candidates, approximately 26 patients in each group: Group 1 purchased and kept their aids, Group 2 purchased but returned their aids, and Group 3 declined to purchase aids. The subjects were matched for audiometric loss. Humes et al found no differences in average speech scores among the three groups using various tests including 50-item NU-6 monosyllables in quiet and noise. If we assume that those who kept their aids were satisfied and those who returned their aids were dissatisfied, this study provides further evidence of a minimal correlation between satisfaction and prefitting speech measures.

Walden and Walden (2004) found that the best predictor of success with hearing aids in daily living was the SNR loss on the QuickSIN test, but they observed that the high correlation between success and SNR loss can be partly explained by the high correlation between SNR loss and age.

Finally, Humes (2003) reported a meta-analysis of three published studies, each of which contained large numbers of subjects: IU-1 (173 subjects), IU-2 (53 subjects), and the NIDCD/VA (National Institute on Deafness and Other Communicative Disorders/Veterans Administration) study (333 to 338 subjects). He found an excellent correlation within groups. For example, four unaided speech measures (nonsense syllable test [Levitt and Resnick, 1978], and Connected Sentence Test in quiet and noise [Cox et al, 1988]) were well correlated with each other. Similarly, three usage measures (hours used per day, etc.) were well correlated with each other, and two satisfaction measures and five subjective benefit measures were well correlated with each other, depending on the individual study. Table 2 summarizes these studies.

Table 3 shows the within-group correlation Humes found. Note, however, that he found no significant correlations among items in different groups. In particular, self-reported satisfaction and benefit could not be predicted from any of the speech measures.

**CONCLUSIONS**

Despite what many of us believe to be true and recommend to our students and colleagues, we found little evidence to support the hypothesis that there is a significant correlation between self-reported satisfaction with hearing aids and prefitting speech measures.

Although not the subject of this report, the authors believe that the use of prefitting SNR measures can provide an improved basis for providing realistic expectations during counseling and lead to use of features that can bring more than half of hearing aid wearers back to normal ability to understand speech in noise (Killion, 2004). Kochkin (1996) found a strong positive correlation between self-reported user satisfaction and the use of advanced features in hearing aids.

Specific evidence in support of this belief could be obtained in a nonrandom controlled intervention study in which one-half of the subjects are randomly assigned to “blind” hearing aid fittings, where prefitting SNR loss is measured but the person selecting the hearing aid is not informed of the results, and the other half receives counseling or intervention appropriate to their SNR loss. If further blinding is desired, the counseling

---

**Table 3. Percentage of Variance in the Overall Hearing Aid Outcome That Is Explained by Each of the Three Important Measures Found in Humes’s Two Studies and in the NIDCD-VA Study**

<table>
<thead>
<tr>
<th>Study</th>
<th>IU-2</th>
<th>IU-1</th>
<th>NIDCD-VA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech-test performance</td>
<td>28%</td>
<td>45%</td>
<td>41%</td>
</tr>
<tr>
<td>Speech-in-noise performance</td>
<td></td>
<td></td>
<td>7%</td>
</tr>
<tr>
<td>Usage</td>
<td>21%</td>
<td>15%</td>
<td>15%</td>
</tr>
<tr>
<td>Subjective Benefit</td>
<td>19%</td>
<td>7%</td>
<td>21%</td>
</tr>
<tr>
<td>Satisfaction (Separated from Subjective Benefit)</td>
<td></td>
<td></td>
<td>6%</td>
</tr>
</tbody>
</table>

*Note: From Humes (2003), used with permission.*
or intervention could be done on a random basis for the whole group, and the self-reported satisfaction after several months would be correlated to the prefitting SNR loss and random counseling and intervention given the subjects in each group.

Acknowledgments. We are grateful to Karen Briskey for her extensive help with the references, to three anonymous reviewers, and to Larry Humes, Robyn Cox, and Harry Levitt, who helped clarify several points in the paper.

REFERENCES


