Programmable hearing aids offer a wide variety of adjustable electroacoustic parameters, including low- and high-cutoff frequencies, frequency-response slope, and crossover frequency, as well as single- or multiband gain, compression threshold, compression ratio, and attack and release times. Some of these aids also have multiple memories, providing the capability of storing the composite of these characteristics in separate memories that can be accessed for use in different acoustic environments. Although many of these electroacoustic parameters may be individually fitted to patients, procedural guidelines necessary to fit most of these parameters on an individual basis are lacking. Clearly, our ability to fit hearing aids to individual needs has not kept pace with the rapid advances in hearing aid technology.

The increased flexibility available with programmable hearing aids dictates that new methods and procedures be developed to fit these instruments efficiently. For many years, the potential of the paired-comparison method for individualizing hearing aid settings has been recognized (Levitt and White, 1978; Kuk, 1994). In contrast to traditional linear and nonlinear fitting formulas, the method of paired comparison allows a listener to compare directly various hearing aids, electroacoustic characteristics, and/or memory settings under everyday acoustic conditions such as listening in noise, listening to music, or listening to one's spouse (Kuk and Lau, 1995). When given a choice between these traditional fitting formulas and other alternatives, some listeners choose frequency-gain settings that differ from those settings determined by the formulas (Punch et al, 1994; Kuk and Lau, 1995; Preminger et al, 2000). Furthermore, the reliability and validity of paired-comparison judgments in hearing aid selection have been amply demonstrated by studies of both speech quality (Jeffers, 1960; Witter and Goldstein, 1971; Punch, 1978; Harris and Goldstein, 1979; Schwartz et al, 1979; Punch and Beck, 1980; Punch et al, 1980; Punch and Parker, 1981) and speech intelligibility (Zerlin, 1962; Punch and
comparison trials are likely to consist of a pair as the same or different with respect to tener may judge the characteristics of a stimulus of the two stimuli being compared. When lis-
there is frequently not a preference for either
experience in those settings, as well as in hear-
tive power of corrective lenses. Based on our
diopter measurements to establish the refrac-
paired comparisons are used routinely to make
fessions of optometry and ophthalmology. There,
additional paired-comparison method were modi-
tournament strategy or psychophysically adap-
t determination whether reliability might be improved if
the forced-choice paradigm inherent in the tra-
tioned to allow listeners a third choice of No
the first. Similarly, Leijon et al (1991) asked
their listeners, after deciding which of two stim-
uli was preferred, to indicate if the selected
for this study, we hypothesized that allowing a
listener to forego the choice of A or B, when
there is no clear perceptual choice, would
increase the data's reliability.

Some investigators have dealt with the issue
of the strength of paired-comparison preference
judgments by using confidence ratings to sup-
plement the judgments. For example, McGee
(1964) applied frequency distortions to spoken
sentences and asked listeners to indicate
whether the quality of the second of a pair of sen-
tences was better or worse than the quality of
the first, and, by use of a rating scale, how much
the second of the pair was better or worse than
the first. Similarly, Leijon et al (1991) asked
their listeners, after deciding which of two stim-
uli was preferred, to indicate if the selected
stimulus was slightly better, better, or much
better than its counterpart. These response cat-
gories were regarded as confidence ratings of
the subjective choices and were weighted for
analysis as subjective distance d', as defined by
signal detection theory (Corso, 1967).

More recently, Kuk and Lau (1995) advoc-
atized a different strategy to increase the reli-
bility of paired-comparison judgments. They
applied binomial-probability theory to calculate
the likelihood of acceptable outcomes in paired-
comparison hearing aid judgments. Essentially,
they derived cumulative probability distribution
curves that can be used to determine the
odds of choosing a particular frequency response
N times in M trials and in determining the like-
lihood of error in accepting the result when one
alternative wins over another in N of M com-
parisons. In their approach, increased reliabil-
ity of paired-comparison preferences is
dependent on increasing the number of trials.
This means effectively that the round-robin
tournament, which is inherently lengthy (Neu-
man et al, 1987), must be supplemented by addi-
tional trials. It would also still be possible that
no clear choice would emerge from the addi-
tional trials, particularly if differences were not
easily discriminable. To avoid this problem, Kuk

Howard, 1978; Punch and Parker, 1981; Stude-
baker et al, 1982; Kuk and Pape, 1992; Stel-
machowicz et al, 1994). To facilitate comparisons,
microprocessor control of the fitting method and
electroacoustic parameters is readily achiev-
able. Also, if desired, the electroacoustic fea-
tures can be simulated by signal-processing
schemes, as opposed to requiring that the listener
physically wear the aid.

Despite the promise offered by the method
of paired comparison, it remains to be imple-
mented as a formal approach to the clinical fit-
ing of hearing aids. Although studies of the
reliability and validity of this method amply
support clinical applications of the basic
approach, these issues require and deserve fur-
ther study before its widespread use can be
expected. Resolution of reliability and validity
issues may ultimately determine the actual form
taken by the method in its clinical implementa-
tion. For example, it may take the form of a
tournament strategy or psychophysically adap-
tive strategy such as the modified simplex
method (Neuman et al, 1987; Kuk, 1994).

In its classic form, the method of paired
comparison involves the presentation of paired
stimuli and a forced response from the listener.
As Stevens (1951) noted, "Stimuli are presented
in pairs. Each stimulus is paired with each other
stimulus (now commonly referred to as a round
robin). The observer indicates which of each
pair is greater in respect of a given attribute" (p.
43). We were interested, in this study, in deter-
mining whether reliability might be improved if
the forced-choice paradigm inherent in the tra-
tional paired-comparison method were modi-
ified to allow listeners a third choice of No
Preference in those instances in which a definite
preference was not apparent.

To appreciate the rationale for inclusion of
a No Preference response in paired-compari-
sion judgments, one needs not to look at the pro-
fessions of optometry and ophthalmology. There,
paired comparisons are used routinely to make
diopter measurements to establish the refrac-
tive power of corrective lenses. Based on our
experience in those settings, as well as in hear-
ing aid research studies, we have observed that
there is frequently not a preference for either
of the two stimuli being compared. When listen-
ing to hearing aid-processed speech, a lis-
tener may judge the characteristics of a stimulus
pair as the same or different with respect to
some attribute, but both may be equally desir-
able or equally undesirable. As a result, paired-
comparison trials are likely to consist of a
subset of data that is not helpful in determin-
ing the preferred electroacoustic characteristics
for a listener. At best, such data could represent
random guesses that reduce efficiency and are
not useful in determining the preferred value,
or winner. At worst, the forced-choice nature of
the paired-comparison method could substan-
tially limit the data's overall reliability and
may weaken a listener's confidence in the abil-
ity of the method to establish an optimal set of
values for his or her hearing aid. As a premise
for this study, we hypothesized that allowing a
listener to forego the choice of A or B, when
there is no clear perceptual choice, would
increase the data's reliability.

Some investigators have dealt with the issue
of the strength of paired-comparison preference
judgments by using confidence ratings to sup-
plement the judgments. For example, McGee
(1964) applied frequency distortions to spoken
sentences and asked listeners to indicate
whether the quality of the second of a pair of sen-
tences was better or worse than the quality of
the first, and, by use of a rating scale, how much
the second of the pair was better or worse than
the first. Similarly, Leijon et al (1991) asked
their listeners, after deciding which of two stim-
uli was preferred, to indicate if the selected
stimulus was slightly better, better, or much
better than its counterpart. These response cat-
gories were regarded as confidence ratings of
the subjective choices and were weighted for
analysis as subjective distance d', as defined by
signal detection theory (Corso, 1967).

More recently, Kuk and Lau (1995) advoc-
atized a different strategy to increase the reli-
bility of paired-comparison judgments. They
applied binomial-probability theory to calculate
the likelihood of acceptable outcomes in paired-
comparison hearing aid judgments. Essentially,
they derived cumulative probability distribution
curves that can be used to determine the
odds of choosing a particular frequency response
N times in M trials and in determining the like-
lihood of error in accepting the result when one
alternative wins over another in N of M com-
parisons. In their approach, increased reliabil-
ity of paired-comparison preferences is
dependent on increasing the number of trials.
This means effectively that the round-robin
tournament, which is inherently lengthy (Neu-
man et al, 1987), must be supplemented by addi-
tional trials. It would also still be possible that
no clear choice would emerge from the addi-
tional trials, particularly if differences were not
easily discriminable. To avoid this problem, Kuk
and Lau (1995) suggested obtaining preliminary judgments from listeners indicating whether stimulus conditions are different or not, and then comparing only those conditions that are perceived as different. The binomial-probability approach is generally more applicable to round-robin tournaments than other tournament or adaptive strategies because the probability distributions were generated under the assumption that paired stimuli have an equal chance of being selected for comparison, a condition not easily met by these other test strategies.

Our goal in this study was to determine whether a response strategy that incorporates a No Preference option might increase the reliability of paired-comparison data. We investigated, in normal-hearing listeners, forced- and unforced-choice response paradigms in determining the preferred frequency-response slope characteristics of a simulated programmable hearing aid when speech was presented against a background of multitalker babble. These respective paradigms are referred to as AB and ABN, where N refers to No Preference. Our goal was to answer the following research questions:

1. What is the intra- and intersession reliability of the forced- and unforced-choice paired-comparison paradigms overall?
2. With regard to the most preferred frequency-response slope characteristics, (a) Do the AB and ABN paired-comparison paradigms result in the same or different listener choices? and (b) Do the two paradigms demonstrate the same or a different degree of intra- and intersession reliability?
3. Under what stimulus conditions do listeners tend to select A or B versus N? Specifically, is there evidence of a relationship between frequency-response slope and the tendency of listeners to indicate a specific stimulus preference (A or B) versus a non-preference (N)?

Although we used a round-robin tournament in seeking answers to these questions, we expected the results to be generally applicable to a variety of hearing aid fitting strategies, including adaptive strategies.

**METHOD**

**Participants**

Fifteen listeners ranging in age from 20 to 56 years (mean = 34.5 years; SD = 10.1) participated in the experiment. Each was required to pass a hearing screening at 15 dB HL at all audiometric frequencies from 250 to 6000 Hz bilaterally.

**Stimuli**

Stimuli were the 36 most homogeneous passages of the Revised Speech Intelligibility Rating (RSIR) test. The 72-passage SIR was originally developed by Cox and McDaniel (1989) and revised by Speaks et al (1994), who used a 1/3-octave band-averaging technique to specify signal levels in determining those passages producing essentially equivalent intelligibility in normal listeners. The test, as used by Speaks et al and by us in this study, consisted of speech passages and multitalker babble (cafeteria noise) that were low-pass filtered at 5000 Hz and digitized with 12-bit resolution at a 12,800-Hz sampling rate. Stimuli used in this study consisted of a computer-stored version of the middle 36, most homogeneous passages from the RSIR (Table 1 in Speaks et al, 1994), digitally mixed with the competing noise at a signal-to-noise ratio (SNR) of +3 dB. A 1000-Hz calibration tone was used in specifying levels of the separate RSIR passages and babble. The speech-in-noise passages were delivered to listeners at an overall level of 80 dB SPL, as measured with a sound level meter (Larson-Davis 800B), using speech noise as a calibration signal. This level was determined by preliminary measurements on three normal listeners to be comfortable and sufficiently loud to produce subjective impressions of speech intelligibility (clarity) of at least 50 percent under the nine experimental slope conditions, which are described below.

The speech and noise, combined at a +3-dB SNR, were processed by nine frequency responses differing in low- and high-frequency slope characteristics. The nine frequency-response slopes (n) were administered in a round-robin tournament, resulting in 36 comparisons (n [n-1]/2) per run. The experimental frequency responses, shown in Figure 1, were based on a 3 × 3 (nine-cell) matrix whose X and Y axes were slope values of 3 dB/octave on respective sides of a fulcrum frequency of 1000 Hz (represented by the dots in Fig. 1). Use of frequency slopes that differed in 3 dB/octave steps ensured that the various stimulus conditions would be discriminable (Byrne, 1992). The reference cell, cell 5, represented a flat signal spectrum in which no frequency shaping was applied. A programmable 1/3-octave spectrum equalizer (Acoustic Research IEQ-801) produced the
required frequency-response slopes. The filtered signals were then amplified and delivered to the listener's preferred ear via an ER-3A insert earphone.

**Procedures**

In each of two experimental sessions scheduled 2 to 4 weeks apart, listeners were asked to judge the clarity of speech, against the babble background, in both a traditional paired-comparison paradigm (AB) and a modified paired-comparison paradigm in which they were also allowed a third choice of No Preference (ABN). Listener instructions for both tasks are provided in the Appendix. Essentially, listeners were given 30 seconds to compare each of the paired (A or B) conditions and responded in accord with the specific task under study. For both tasks, listener preferences were to be based on which signal, A or B, produced better speech intelligibility. In the ABN task, a No Preference option was allowed when intelligibility was perceived as equal. Throughout this article, preferences obtained in the AB paradigm, or task, will be referred to simply as AB. Preferences obtained in the ABN task will be referred to as ABN Preference, whereas nonpreferences, or N responses, in the ABN task will be referred to as ABN No Preference.

Two runs of the AB and ABN paradigms were administered per session, to allow for an analysis of both intra- and intersession reliability. A practice run on each paradigm was included in the first session prior to the first experimental run. The AB and ABN paradigms were counterbalanced across listeners and test and retest sessions. Paired stimuli were randomized without replacement during each run, based on a computer-controlled randomization sequence.

It is likely that changes in loudness occurred under the various slope conditions of the experiment (see Punch and Rakerd, 1993, for a discussion). We made no adjustments in the overall intensities of the stimuli to compensate for such changes, for three reasons. First, the most readily applicable data (Punch and Rakerd, 1993; Rakerd et al, 1999) were based on loudness differences occurring in the Parameter Adjustment and Selection procedure (Punch and Robb, 1992), and those data pertained to a limited number of the frequency-response slopes used in this study (cells 1, 5, and 9 only). Second, the magnitude of any such changes was estimated to be small here (Punch and Rakerd, 1993). Third, and most important, our primary interest was in a relative comparison of winners in the AB and ABN paradigms. Because any existing loudness differences would be present in both tasks, they were not considered critical to the experiment.

The computer program stored the data by run and listener in 9 X 9 matrices. All values from the generated matrices were transferred to similarly formatted computer spreadsheets for analysis. Sample (spreadsheet) matrices for listener 1, obtained in the first test session, are shown in Tables 1 to 3. For the AB task (see Table 1), there were four matrices per listener (AB test and AB retest for each of two test sessions). For the ABN task, data were stored in an AB Preference matrix (see Table 2) and an ABN No Preference matrix (see Table 3) for test and retest during each of the two test sessions, for a total of eight matrices. Each of the nine frequency-response slopes (see Fig. 1) was represented in the columns and rows of each table. In the AB paradigm (see Tables 1 and 2), a preference for A or B was stored as a 1 to indicate the row number as the winner over the column number. For example, referring to Table 1, frequency-response slope 2 beat slope 3, so a 1 was stored in the cell intersecting the second row and third column, and a 0 was stored in the counterpart cell intersecting the third row and second column. The blanks on the diagonal reflect the fact that cells were not compared with themselves. Winners, based on counts of preferred-cell responses (1s), were rank ordered for

![Figure 1](image-url)
Table 1  Example Case (Listener 1, Day 1 Test) for AB Paradigm Showing Total Number of Wins as Preferred Cells (Rows) over Compared Cells (Columns) and Rank Order of Preferences

<table>
<thead>
<tr>
<th>Compared Cell</th>
<th>Preferred Cell 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Wins</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>3.5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>3.5</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5.5</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>36</td>
<td></td>
</tr>
</tbody>
</table>

In the ABN paradigm, when listeners chose A or B, ABN Preference results were stored in the same way as for the AB task. In Table 2, for example, there was only one value of 1 stored in the entire upper diagonal, indicating that slope 2 beat slope 7. Generally, this listener preferred higher-numbered slopes to lower-numbered slopes, as shown in the lower diagonal. A value of 0 was assigned in Table 2 either when a row slope was beaten by a column slope value (e.g., row 7, column 2) or when there was no preference for either slope (e.g., row 3, column 2). Note that a winner emerged 23 times out of a total of 36 comparisons in the ABN paradigm. The remaining comparisons are accounted for in Table 3, where judgments of No Preference (in the ABN paradigm) are indicated by a 1 stored in the cell intersecting the two slopes and by a 0 stored in any cell intersecting two frequency-response slopes for which there was a preference for either A or B. In the ABN No Preference matrix (see Table 3), 1s and 0s were stored only in the upper diagonal of the matrix. Of the 36 paired comparisons, listener 1 indicated no preference for 13 comparisons. Such ABN No Preference responses were considered ties.

RESULTS AND DISCUSSION

Intra- and Intersession Reliability

Intra- and intersession reliability of the forced- and unforced-choice paired-comparison methods was evaluated, for test and retest on each of the 2 test days, by tabulating across all 15 listeners the number of times each frequency-response slope was preferred for all possible (36) comparisons. The total wins for preferred cells, established by summing 1s across all listeners for each respective condition, were then converted

Table 2  Example Case (Listener 1, Day 1 Test) for ABN Paradigm Using AB Data Only (ABN Preference), Showing Total Number of Wins as Preferred Cells (Rows) over Compared Cells (Columns) and Rank Order of Preferences

<table>
<thead>
<tr>
<th>Compared Cell</th>
<th>Preferred Cell 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Wins</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7.5</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7.5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td></td>
</tr>
</tbody>
</table>
Unforced-Choice Paired Comparisons/Punch et al

Table 3 Example Case (Listener 1, Day 1 Test) for ABN Paradigm Using N Data Only (ABN No Preference), Showing Total Number of Ties When Reference Cells (Rows) and Compared Cells (Columns) Were Judged as Nonpreferred

<table>
<thead>
<tr>
<th>Compared Cell</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Cell</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>--</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>--</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>--</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>13</td>
</tr>
</tbody>
</table>

Results of this analysis are shown in Table 4. Rho values ranged from .83 to .97, all significant at the .01 level. These findings indicate high reliability and demonstrate that the rank-ordered preferences on both the AB and ABN tasks were highly repeatable overall, both within and across experimental sessions. In answer to our first research question, therefore, the intrasession reliability of both the forced- and unforced-choice paradigms was high. There was no evidence in this first analysis to suggest that the ABN procedure should be expected to offer a clinical advantage over the traditional paired-comparison procedure in differentiating preferences for various frequency-response slopes.

Listeners' Most-Preferred Choices

Most-Preferred Slopes

Our concern in asking the first of our two-part research question 2 was whether the same preference outcomes, or frequency-response slopes, would be selected as most preferred in the two paradigms. As our interest was in the over-
all capacity of the different paradigms to produce the same or different outcomes, we summed preferences across experimental days and test–retest runs to generate separate overall matrices for the AB and AB Preference data. When preferences were summed across each of the nine cells within each matrix, the overall respective ranks of cells 1 to 9 (where 1 indicates most preferred) for AB were 9, 8, 6, 7, 3, 4, 5, 2, and 1. For AB Preference, the respective ranks were 9, 8, 6, 7, 5, 2, 4, 3, and 1. That is, cell 9, with a rank of 1, was most preferred overall in both paradigms, indicating the strong preference of listeners for signal processing in which low-frequency gain was minimized and high-frequency gain was maximized (as seen in Fig. 1). Incidentally, listeners’ least-preferred choices (ranks of 9, 8, 6, and 7, as applied to cells 1 to 4, respectively) were also the same in both paradigms. These findings indicate that the two paradigms converged overall on the same results with respect to the most-preferred frequency-response slope, as well as the several least-preferred slopes.

Reliability of Top Choices

The analysis related to the second part of our question 2, whether AB and ABN demonstrate essentially equivalent intra- and intersession reliability, was addressed by tabulating the number of times preferences agreed on an intra- and intersession basis across listeners and within paradigms. As in the preceding analysis, only the AB and ABN Preference data were considered here. For this analysis, we determined the percentage of times in which the most-preferred responses of the 15 individual listeners were in agreement with one another. A cell (in the $3 \times 3$ matrix) was counted as most preferred if it was ranked as the first choice or tied for first choice, and most-preferred responses were considered repeatable if they were selected in both conditions of a given intra- or intersession comparison.

Results are reported in Tables 5 and 6, where a plus sign indicates agreement of most-preferred choices and a minus sign indicates disagreement. For the AB condition (see Table 5), 8 of the 15 listeners (53%) showed intra-session agreement on day 1, whereas 5 of the 15 (33%) showed intra-session agreement on day 2. Seven and four listeners, respectively, showed intersession agreement on test and retest data. Four listeners (5, 9, 12, and 13) had no instances, on the four comparisons, in which their most-preferred slope values agreed. Of the possible 60 opportunities for agreement (4 comparisons $\times$ 15 listeners), agreement of most-preferred values occurred 40 percent of the time overall on the AB task.

For the ABN Preference condition (see Table 6), a total of 10, 12, 6, and 9 (of 15 listeners) showed intra- and intersession agreement on the respective conditions. All 15 listeners had at least one instance in which there was agreement with regard to their most-preferred slope value. In only one comparison (day 1 test versus day 2 test) did agreement on AB exceed that on ABN Preference. The overall agreement was 62 percent for ABN, which was significantly greater than the 40 percent agreement for AB ($z = 2.20$, $p < .05$). In answer to the first part of question 2, therefore, AB and ABN Preference data produced different outcomes with respect to reliability of the most-preferred frequency-response slope characteristics. The fact that the ABN Preference data produced significantly better overall agreement than AB with respect to most-preferred responses provides tangible evidence of the potential clinical utility of the ABN paradigm. This is true particularly because a listener’s most-preferred responses represent the information that is most likely to be useful in

Table 5  Intra- and Intersession Agreement for Most-Preferred Frequency-Response Slope for Individual Listeners (n = 15) in the AB Paradigm

<table>
<thead>
<tr>
<th>Listener</th>
<th>Intrasession</th>
<th>Intersession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day 1 T/R</td>
<td>Day 2 T/R</td>
</tr>
<tr>
<td></td>
<td>Day 1 T/ Day 2 T</td>
<td>Day 1 R/ Day 2 R</td>
</tr>
<tr>
<td>1</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>15</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Overall (%) 53 33 47 27

T = test; R = retest.
Table 6  Intra- and Intersession Agreement for Most-Preferred Frequency-Response Slope for Individual Listeners (n = 15) in the ABN Paradigm

<table>
<thead>
<tr>
<th>Listener</th>
<th>Intrasession</th>
<th>Intersession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day 1 T/R</td>
<td>Day 2 T/R</td>
</tr>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>14</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Overall (%)</td>
<td>67</td>
<td>80</td>
</tr>
</tbody>
</table>

T = test; R = retest.

determining the optimal electroacoustic characteristics of a hearing aid.

Preference versus Nonpreference Behavior as a Function of Slope

Our third research question was aimed at determining the existence of a link between frequency-slope characteristics and listeners' preference behavior in the AB and ABN tasks. We examined this issue in an attempt to identify the underlying motivation of listeners to choose N, or No Preference, in the ABN task, as opposed to a specific cell (A or B).

Three levels of low-frequency shaping—a cut of −3 dB/octave, a flat setting of 0 dB/octave, and a boost of +3 dB/octave—were orthogonally crossed with three levels of high-frequency shaping—again, cut, flat, and boost—to derive the nine frequency-response settings tested in the experiment (see Fig. 1). The impact of each of these frequency-shaping factors on AB and ABN preferences was assessed.

AB

The preference data obtained in the AB task showed statistically significant effects of both low- (F[2, 28] = 22.90, p < .001) and high-frequency shaping (F[2, 28] = 11.57, p < .001). The results are shown in Figure 2. The figure reports the mean number of wins (times preferred) that a particular shaping type earned when paired against the other types in the round-robin tournament. The mean preferences for AB are plotted with open symbols in the figure. The low- and high-frequency shaping results are given in the left- and right-hand panels of the figure, respectively. For low-frequency shaping, listeners showed a marked preference for the spectral cut and a marked dislike for the boost. Results for the flat setting were intermediate. With high-frequency shaping, listeners preferred a boost, disliked a cut, and were again intermediate regarding the flat setting. The whole pattern of responses indicates a clear listener preference for settings that enhanced speech intelligibility by minimizing the masking of higher-frequency speech cues by low-frequency components of the background babble.

ABN Preference

Mean preferences for ABN—plotted in Figure 2 with filled symbols—showed a pattern very similar to that for the AB data. Again, there were significant effects of both low- (F[2, 28] = 43.03, p < .001) and high-frequency shaping (F[2, 28] = 38.61, p < .001), and, again, listeners showed a marked preference for cutting the low frequencies and boosting the high ones. The F-ratio found for low-frequency shaping in this ABN analysis was nearly twice as great as the corresponding F-ratio for AB (43.03 vs 22.90);
the F-ratio for high-frequency shaping was more than three times greater (38.61 vs 11.57). This suggests an increased sensitivity with the ABN approach, which is most likely due to the fact that arbitrary or difficult choices in the AB paradigm that functioned largely as statistical noise were effectively removed by the availability of the N option.

The one difference between the ABN and AB results that is clearly visible in Figure 2 is that there were fewer wins overall with ABN for every condition, due to listeners’ use of the N response. In the AB paradigm, on average, listeners used the N option on 39 percent of all paired comparisons (SD = 19%). There were instances of N use for pairs involving every type of frequency-response setting, but the listeners chose N most frequently when either low- or high-frequency shaping was in a flat setting. In other words, the N response was used most often in cases where frequency-response changes were least extreme. An analysis of variance on the N results showed statistically significant effects of both low- (F [2, 28] = 5.57, p < .009) and high-frequency shaping (F [2, 28] = 14.34, p < .001).

Task Time for AB versus ABN

In comparing methods, we were interested in the general issue of whether the time spent on the ABN task might have been different from that spent on the AB task. Our interest stemmed from the notion that listeners might choose the N option decisively and quickly, when available, in cases when the choice between A and B was not readily apparent. If that were the case, ABN might require less time than AB to complete. Because we did not pose a specific research question related to the time spent by participants in listening to the paired comparisons and in indicating their preferences, a discrete measure of time was not part of the experimental design. Our computer program, however, did monitor the time at which each experimental run ended. This allowed us to estimate the time spent between runs of the AB and ABN tasks. For each paradigm, the time intervals, in minutes, between the completion of test and retest sessions on each of the 2 days were averaged (and rounded to the nearest tenth of a minute). These time periods essentially included the 36 trials encompassing the nine conditions in the retest run on each of the 2 days. They did not include time spent on breaks; when breaks did occur (for three of the participants), those times were excluded from the analysis. Times spent on the AB task ranged widely from 3.5 to 14 minutes across the 15 participants, and times spent on the ABN task ranged from 3 to 18 minutes. The respective means were 9.5 and 9.7 minutes, and respective standard deviations were 4.8 and 5.5 minutes, suggesting no difference in the time spent on the two tasks.

The total number of N responses for the ABN task was also highly variable across listeners, the range being from 1 to 97, of 144 opportunities (36 trials × 4 runs). The mean number of N responses was 56.5, with a standard deviation of 27.4. (Only one listener responded with as few as 1 N response; the next smallest number was 20.) To examine whether a relationship existed between the total number of N responses and the time to complete each task, Pearson product-moment correlation coefficients (r̂s) were computed between average times to complete an AB run and number of N responses and between average times to complete an ABN run and number of N responses. Correlations of .09 and .04, respectively, were obtained, indicating virtually no relationship between listeners’ use of the No Preference option and the time they needed to complete the paired-comparison procedure in either paradigm. Furthermore, when listeners were categorized based on the number of N responses, no evidence of a relationship could be found between the use of N responses and agreement on most-preferred responses across runs. We interpret these findings to mean that the times needed to complete the AB and ABN tasks did not differ substantially from one another, and thus that the ABN paradigm offered no time savings over the AB paradigm. Furthermore, the number of N responses on the ABN task does not seem to be associated with, or predictive of, the amount of time needed to complete either paradigm.

CONCLUSION

Overall, results showed a high degree of reliability for both paradigms, based on the overall rank ordering of preferences among the nine different frequency-response settings. Both paradigms tended to converge on the same most-preferred frequency-response slope. Most important, results specifying a listener’s most-preferred frequency response setting were significantly more reliable for the ABN paradigm than for AB. The salience of this latter finding lies in the clinical importance of the most-preferred setting for determining the
optimal value of the electroacoustic characteristics of a hearing aid. The finding supports our general view that the efficiency of the paired-comparison method may be reduced when listeners are forced to choose between stimuli that are perceived as equally desirable or undesirable.

When listening to speech against a competing multitalker babble (+3-dB SNR), normal-hearing listeners preferred frequency shaping that either cut low frequencies or boosted high frequencies, based on a criterion of relative speech intelligibility. This was the case when judgments were made using either a traditional AB paradigm or the experimental ABN paradigm. A response of N (No Preference) occurred most commonly (in the ABN paradigm) when either low- or high-frequency shaping was flat or neutral.

To our knowledge, this study is the first to report audiologic data based on an ABN paradigm. Although tentative, the results are encouraging with respect to the paradigm's potential applicability to the fitting of multimemory hearing aids. It is entirely reasonable to expect that use of unforced choices could be incorporated into adaptive hearing aid fitting protocols such as the modified simplex method (Neuman et al, 1987). The test sequence, for example, could incorporate a rule wherein a response of No Preference would be followed by a reversal in the direction of the next comparison (in a $3 \times 3$ or $5 \times 5$ matrix) or by an increased step size of the measured parameter in the same direction (in a $5 \times 5$ matrix). Responses of No Preference, if used in such a manner, could improve the efficiency of paired-comparison judgments in optimizing a hearing aid's characteristics. Further investigations of this prospect are warranted, especially in light of the time savings offered by a modified simplex method. Neuman et al (1987), using a 25-condition ($5 \times 5$) matrix, found that such a procedure could be completed in one-tenth of the time required for an iterative round robin, which was used in this study. The efficiency of the modified simplex procedure makes it extremely attractive for determining optimal settings for multimemory hearing aids, and our data suggest that the test–retest reliability of individual listeners would measurably benefit from use of an ABN paradigm.

Aside from potentially enhancing the reliability of most-preferred choices, the data from an ABN paradigm may have other properties that differentiate it from a traditional paired-comparison method. For example, the extent to which logical relationships exist in listeners' judgments may differ for the two paradigms, and this difference could provide an improved understanding of how to develop the method for clinical implementation. Thus, the study of transitivity may yield important additional information about the relative utility of AB and ABN paradigms in paired-comparison testing. This study has provided a voluminous database for such an analysis, which is currently under way.

Acknowledgment. The authors wish to acknowledge Timothy Trine of Starkey Labs, Inc., for providing a CD-ROM version of the Revised Speech Intelligibility Rating test. Scott Turner and Karrie Slominski were helpful in ensuring the accuracy of the spreadsheet data and in performing selected data analyses.

Portions of this article were presented at the Convention of the American Academy of Audiology, Chicago, IL, March 2000.

REFERENCES


**APPENDIX**

**AB Paradigm**

In this task, you are to listen to 30-second passages of speech, along with background noise, as if amplified by pairs of hearing aids. The pairs are referred to as A and B. You are to indicate which of the passages results in greater speech intelligibility. Think of intelligibility as the percentage of spoken words you can understand. Try to ignore the loudness of the speech, any unpleasantness in sound quality (i.e., tininess or too much bass), and the background cafeteria noise, and concentrate only on which of the pair, A or B, results in better speech understanding.

When you are ready to listen, flip the toggle switch to ON. All lights will come on for a couple of seconds to signal you to listen carefully. The listening task will randomly begin with either A or B, and after listening a few seconds to that condition, you should press the other button and listen to that condition for a few seconds. The light above A or B will indicate the current listening condition. Press A and B alternately as many times as you wish during the 30-second time period.

When you are ready to respond, flip the toggle switch to OFF, and then press A or B to indicate your preference. In the event that you need all 30 seconds to make a decision, the passage of speech and noise will automatically stop, and all of the lights will come on. This indicates the need for you to respond. Flip the toggle switch to OFF, and then press A or B to indicate your preference. Remember that you will need to toggle the switch to ON to listen and OFF to respond.

**ABN Paradigm**

In this task, you are to listen to 30-second passages of speech, along with background noise, as if amplified by pairs of hearing aids. The pairs are referred to as A and B. You are to indicate which of the passages results in greater speech intelligibility or that you have no preference, which is denoted by the letter N. Think of intelligibility as the percentage of spoken...
words you can understand. Try to ignore the loudness of the speech, any unpleasantness in sound quality (i.e., tinniness or too much bass), and the background cafeteria noise, and concentrate only on which of the pair, A or B, results in better speech understanding.

When you are ready to listen, flip the toggle switch to ON. All lights will come on for a couple of seconds to signal you to listen carefully. The listening task will randomly begin with either A or B, and after listening a few seconds to that condition, you should press the other button and listen to that condition for a few seconds. The light above A or B will indicate the current listening condition. Press A and B alternately as many times as you wish during the 30-second time period.

When you are ready to respond, flip the toggle switch to OFF and then press A or B to indicate your preference. If A and B sound the same to you, or if you perceive a difference, but don't have a preference for either, press N. In the event that you need all 30 seconds to make a decision, the passage of speech and noise will automatically stop, and all of the lights will come on. This indicates the need for you to respond. Flip the toggle switch to OFF and then press A, B, or N to indicate your preference. Remember that you will need to toggle the switch to ON to listen and OFF to respond.