Binaural hearing can be simply defined as listening with both ears. Binaural hearing improves speech understanding and sound source localization, particularly in background noise (Gantz et al., 2002; Müller et al., 2002; Schleich et al., 2004; Schön et al., 2002). Although binaural hearing has many advantages, defining and assessing it can be challenging. Moreover, clinicians are restricted to a few tests that provide a limited understanding into the true benefits of binaural hearing, and these tests are rarely used in clinical practice.

A wide range of tests has been used in the literature to assess the phenomena associated with binaural hearing, including the head shadow effect (HSE), binaural squelch (SQ), binaural summation (SU), spatial release from masking (SRM), and localization (LOC; Arsenault & Punch, 1999; Cox et al., 1981; Gifford et al., 2014; Tyler et al., 2002, 2003; Van Deun et al., 2010). There are studies showing these benefits in bilateral/bimodal cochlear implant (CI) users and in normal-hearing (NH) people (Buss et al., 2008; Eapen et al., 2009; Litovsky et al., 2006; Schleich et al., 2004; Verhaert et al., 2012).
Each of the aforementioned phenomena can be evaluated by psychoacoustic tests that can be performed in a clinical audiological setting. Psychoacoustic measures provide valuable information about the auditory perception of sound. These methods, however, are inevitably influenced by a conscious attentional process and, therefore, can be challenging to use in very young children and older adults (Bertoli et al., 2005; Köse et al., 2022; Luo et al., 2020; Maxon & Hochberg, 1982). Although binaural benefits have traditionally been evaluated utilizing psychoacoustic methods, some recent studies have investigated the binaural benefits also through objective methods (Balkenhol et al., 2020; Rawool, 2016; Rawool & Parrill, 2018). Rawool (2016) used acoustic reflex threshold measurements to assess SU in women. A follow-up study by Rawool and Parrill (2018) found similar results also in men. Balkenhol et al. (2020) assessed HSE, SU, SQ, and SRM in individuals with bimodal hearing using auditory evoked potentials. Thanks to their linguistic independence, objective methods might have the potential to reduce the high variance observed in the results of binaural benefit assessments. However, due to the lack of standards, there are neither universally accepted test setups nor normal values for either psychoacoustic or electrophysiologic methods. The standardized measures of binaural testing become particularly important in the context of CI. Cochlear implantation was traditionally performed unilaterally, which means that binaural hearing was not restored. This has changed in children since bilateral implantation has become a common treatment in this population, and it is also becoming more common in adults (Brown & Balkany, 2007; Kan & Litovsky, 2015; Litovsky et al., 2006).

The traditional methods to assess the binaural benefits include comparing speech understanding in the binaural listening condition with that in the monaural condition (Eapen et al., 2009; Gantz et al., 2002; Schleich et al., 2004). Here are many factors that have impact on the outcomes of these methods. These factors include both measurement-related variables such as the test setups, the type of noise, the reference ear tested in the monaural condition, the number and position of the loudspeakers as well as the microphone settings, and participant-related variables such as the individual performance and gender (Bronkhorst & Plomp, 1988; Carhart, 1965; Dieudonné & Francart, 2019; Dirks & Wilson, 1969; Kurz et al., 2021; Laszig et al., 2004; MacKeith & Coles, 1971; Müller et al., 2002; Pyschny et al., 2014; Rawool & Parrill, 2018; Schön et al., 2002; Sheffield et al., 2015). To the best of our knowledge, however, there are no studies evaluating and comparing the effectiveness of different speech materials in binaural benefit assessments. Furthermore, as long as the item lists used in the speech tests are relatively short, 10% or more test–retest differences can be observed (Dillon, 2012, as cited in Avan et al., 2015). However, in most studies in the literature, the test–retest reliability of the applied test material is not measured. In fact, when the individual results are examined, the test results in many conditions may not exceed the test–retest variability.

The first goal of this study was to investigate which speech material (monosyllabic words, spondees, or sentences) is more effective for assessing the HSE and SQ based on the criteria: (a) the largest effect size in NHs, (b) the highest test–retest reliability in both NHs and CI users, and (c) the smallest interindividual variability in NHs. The most appropriate material was then used to obtain normative values for both tests.

The second goal was to explore the correlation between HSE, SQ, and LOC in bilateral CI users and possible influencing factors, such as the interval between implantations, experience with binaural hearing, etc. Furthermore, a correlation analysis was performed between the results of binaural tests, and the Speech, Spatial, and Qualities of Hearing (SSQ) scale (Falzone et al., 2022).

Method

A prospective study was conducted with NH participants and bilateral CI recipients. The Ethics Committee of Antwerp University Hospital (UZA) approved the study with Project ID 2021–0551 – BUN B3002021000155. Informed consent was obtained from all participants. For participants under 18 years of age, parental consent was obtained in addition to their own.

Participants

A total of 64 people participated in this study. All participants spoke Flemish as their mother tongue. Thirty-four of them were CI users with more than 6 months of binaural listening experience who were being followed in our clinic, and 30 were in the NH group with hearing thresholds of ≤ 20 dB HL at all the octave frequencies tested between 125 and 8000 Hz. The NH group consisted of 14 (47%) male and 16 (53%) female participants, with a median age of 23.5 years (Q1: 22 and Q3: 28.5; range: 19–39). The mean pure-tone average (PTA) of the NH group was 7 ± 4 dB HL in the right ear and 8 ± 4 dB HL in the left ear.

The CI group consisted of 19 (56%) female and 15 (44%) male participants, with a median age of 21 years (Q1: 16 and Q3: 28; range: 14–80). Twelve participants (35%) were implanted postlingually (older than 4 years of age), whereas 22 participants (65%) were implanted before the age of 4 years. Nine (75%) of the 12 participants implanted after 4 years of age had congenital hearing loss
and had used hearing aids until implantation. The onset of hearing loss in the remaining three participants (25%) was in adulthood and showed a progressive course. These participants also had a history of hearing aids use before the implantation. Nineteen participants (56%) had their first implant on the right, and 15 participants (44%) had their first implant on the left. None of the participants was implanted simultaneously. The median ages at the first and second implantations were 3 years (Q1: 1 and Q3: 9.3) and 9.5 years (Q1: 4.8 and Q3: 23.3), respectively. The participants had an average of 125 ± 67 months of experience with their 2nd CI.

**Procedure**

This study consisted of three phases as follows.

1. Investigating the effect of speech material in the HSE and SQ tests:
   We compared three different speech materials: monosyllabic words (NVA: Nederlandse Vereniging voor Audiologie; Wouters et al., 1994), spondee words (BLU: Brugge-Leuven-Utrecht; Wouters et al., 1994), and sentences (LiCoS: linguistically controlled sentences; Coene et al., 2018). We envisaged that the optimal speech material should be characterized by three preset criteria: (a) large effect size in the NH group, (b) strong test–retest reliability in both the NH and CI groups, and (c) small interindividual variability in the NH group. This phase of the study was conducted with 11 CI users and six NH individuals.

2. Obtaining normative data of HSE and SQ:
   The speech material selected in the first phase was used to test a further 24 NH participants to obtain normative values for both tests.

3. Assessing the binaural benefits in bilateral CI users:
   Both tests were also administered to a further 23 bilateral CI users. Besides the HSE and SQ tests, the LOC test and the SSQ scale were also applied. The second and third phases were carried out simultaneously after the completion of the first phase.

**Outcome Measures**

All tests were conducted using the Auditory Speech Sounds Evaluation (A§E) psychoacoustics test suite (Otoconsult NV; Govaerts et al., 2006). A§E is an audiological evaluation tool for detection, discrimination, and identification tasks (Govaerts et al., 2006). The suite includes tests for spectral discrimination, intensity coding, temporal fine structure (TFS), speech audiometry, and binaural integration.

**Pure-Tone Audiometry**

Pure-tone thresholds at octave frequencies between 250 and 8000 Hz were obtained from all participants using the modified Hughson-Westlake down-up procedure (Carhart & Jerger, 1959; Hughson & Westlake, 1944). NHs were tested in a sound-treated booth using an Aurical audiometer (Natus Medical Incorporated) and a TDH-39P headphone (Telephonics Corporation).

CI users were tested in free-field condition using Otocube (Otoconsult NV). Otocube is a portable desktop box that replaces classic soundproof booths in the testing of CI patients. Otocube has a built-in loudspeaker that allows delivering the stimuli to the patient’s sound processor in isolation from the external environment. Before testing in the Otocube, a long coil cable was first connected to the sound processor. The sound processor was then placed in the Otocube, as shown in Figure 1, and the tests were conducted.

**Speech Audiometry in Quiet**

Before the HSE and SQ tests, speech in quiet test in CI users was performed with their everyday program settings using Flemish monosyllables (Wouters et al., 1994) in Otocube (Otoconsult NV). Speech recognition scores (SRSs) were obtained by presenting two lists of 12 words at four different levels (40, 55, 70, and 85 dB SPL). The weighted average was calculated using the following formula:

\[
    EaSI = \frac{SRS_{40} + SRS_{55} + (2 \times SRS_{70}) + SRS_{85}}{5},
\]

Figure 1. Portable desktop box (Otocube) used in audiometric examinations of cochlear implant users.
where EaSI stands for Eargroup Speech Index and SRS\text{\textsubscript{x}}
stands for phoneme score at the presentation level of “x” dB SPL.

**HSE and SQ**

The test setup was created using three Fostex 6301NB Personal
loudspeakers (Foster Electric Company Ltd). The speakers were placed at a
distance of 1 m from the participant, at 0°, +90°, and −90° azimuth (see Figure
2). The test room was untreated and had an ambient
background noise level of 30 dB (A). The reverberation
time in the test room varied between 0.40 and 0.52 s for
frequencies between 125 and 4000 Hz. The test parameters
were as follows.

**Speech material.** SQ and HSE scores were deter-
mined by comparing the results of the speech understand-
ing in noise tests performed in binaural and monaural lis-
tening conditions. Three different types of speech mate-
rials (NVA, BLU, and LiCoS) were used for the first
phase of the study.

NVA speech material consists of monosyllabic
consonant–vowel–consonant (CVC) words. The test con-
tains 15 lists of 12 words each. Each list has a similar set
of initial consonants, vowels, and final consonants. NVA
lists are ideal for evaluating speech understanding using phoneme scores based on the percentage of correctly iden-
tified phonemes (Wouters et al., 1994). Each correctly
repeated phoneme counts about 2.78% since each list has
36 phonemes. Participants had to repeat the words or at
least the phonemes they heard in the NVA test. For exam-
ple, a participant who repeated the word *kop as kot* was
given a 5.6% phoneme score and 0% word score.

BLU was developed in response to the need for a
Flemish speech test based on spondaic words (Wouters
et al., 1994). The test contains 15 lists of 10 spondee
words each. Therefore, each correctly identified spondee
word counts for 10%. Each word has the CVC–CVC
structure, and each syllable is a separate existing word.
BLU lists are suitable materials for evaluating speech
understanding in both quiet and noise based on word
scores. The test required participants to repeat every word
they heard. Since BLU spondees consist of two syllables
that form two separate words, the lists can also be scored
based on syllables (Bosman et al., 1995). In such a case,
for example, someone who repeated “brood-mes” as
“groot-mes” scored 5%.

As a more representative speech material of mod-
ern Dutch and Flemish, LiCoS was developed by Coene
et al. (2018). The test material consists of sentences artic-
ulated by one female and one male speaker. In LiCoS,
there are 12 lists of 30 sentences each with two keywords.
LiCoS lists comprise sentences with varying syntactic
complexity. Different lists are balanced in terms of vari-
ous linguistic parameters, including lexical, phonological,
morphological, and syntactic components of modern
Dutch and Flemish (Coene et al., 2018). The number of
words in each sentence varies between 6 and 10. The lists
are also balanced based on the length of the sentences.
However, none of the sentences comprises fixed expres-
sions or proverbs, ensuring low semantic predictability.
In LiCoS, the keywords are balanced for eight major
word classes (adjectives, adverbs, nouns, prepositions,
pronouns, quantifiers, verbs, and conjunctions; Coene
et al., 2018). The keywords within the same sentence
never belong to the same word class. The participants

![Figure 2](https://pubs.asha.org) Head shadow effect and binaural squelch test setup with three speakers at 0°, −90°, and +90° azimuth, each at a 1-m distance from the participant.
are given 1.67% for each correctly repeated keyword within the same list. The participants had no preknowledge of how the sentences were scored or about the keywords. Their task was to try to repeat all the words in the sentences they heard.

In the first phase of the study, six NH participants underwent both the HSE and the SQ tests. Eleven CI participants were also tested, either with the HSE test \((n = 5)\) or the SQ test \((n = 5)\) or both \((n = 1)\). The tests were done twice (to calculate the test-retest difference) with each material. All participants took a 10-min break between the test and retest sessions. Participants who were tested with HSE only, for example, had to do the test 6 times (twice with NVA, twice with BLU, and twice with LiCoS). In each participant’s test, the order of the presentation of the speech materials was determined randomly.

One list from each speech material was played to the participants before the test, and the answers to these lists were not included in the scoring. All subsequent tests used different lists in a random order to minimize learning effects. After the study’s first phase, only the speech material that provided the optimal results based on the predefined criteria for the HSE and SQ was used.

**Reference ear.** The reference ear was the ear that was tested in the monaural listening condition. In CI users, the reference ear was always the first implanted ear. However, the reference ear of each NH participant was randomly assigned. The right ear was tested in half of the participants, and the left ear in the other half.

**Locations of the signal and noise.** In the HSE test, noise was presented from the speaker at 0° azimuth, while signal was presented from the second ear (nonreference) side \(± 90°\) azimuth. In the SQ test, on the other hand, the signal was presented from 0° azimuth, whereas noise was presented from the second ear side.

**Noise.** Steady-state speech-shaped noise was used for the measurements. The noise used for each speech material \(\text{(NVA, BLU, and LiCoS)}\) was different and had the same long-term average spectrum as the target speech.

**Adaptive algorithm.** Both the SQ and HSE tests began by determining a speech recognition threshold \(\text{(SRT)}\) using an adaptive algorithm. The adaptive procedure that was followed in this study was based on the simple up-down staircase method. The parameters of the algorithm are listed in Table 1.

Once an SRT was determined using the adaptive algorithm in the binaural listening condition, the test proceeded to the next step. Here, individuals were tested in the monaural listening condition at the SNR determined in the previous step. For example, if a person had an SRT of +5 dB SPL in the first test, the second test was run at a fixed SNR of +5 dB SPL (signal at 70 dB SPL and noise at 65 dB SPL). Since the SRT determined in the first step corresponded to a 70% correct response, a 60% score in the second test, for example, would indicate a 10% binaural benefit compared with monaural listening. Figures 3 and 4 show examples of the display of the HSE and SQ test results.

CI users were asked to turn off their second CI in the monaural listening condition. The nonreference ears of the NH group were blocked with E-A-RTone 3A insert earphones (Aearo Technologies LLC) and masked via a Madson Itera II audiometer (Natus Medical Incorporated), with a 60-dB HL broadband speech noise.

**Azimuth LOC**

An azimuth LOC test was carried out using seven Fostex 6301B loudspeakers (Foster Electric Company Ltd). This test was performed in the same room used for HSE and SQ tests. The test stimulus was broadband speech noise presented at 70 dB SPL. A ± 3 dB level rove was applied, which means that the presentation level for each loudspeaker randomly varied between 67 and 73 dB SPL, to avoid any additional cues caused by the characteristics of the loudspeakers.

In the test, speakers were placed at 20° angle intervals from \(-60°\) (left side) to +60° (right side) and numbered from \(-3\) to +3. Figure 5 illustrates the test setup.

Participants received a brief training before starting the test with feedback on their answers. In the test, stimuli were presented from the seven speakers in random order, and individuals were asked to point to the speaker from which they heard the sound. The stimulus was presented 5 times from each speaker, and at the end of the test, root-mean-square \(\text{(RMS)}\) was calculated.

### Table 1. Parameters of the adaptive algorithm used in HSE and SQ tests.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial signal level</td>
<td>75 dB SPL</td>
</tr>
<tr>
<td>Initial SNR</td>
<td>10 dB SPL</td>
</tr>
<tr>
<td>Noise intensity</td>
<td>65 dB SPL (Fixed)</td>
</tr>
<tr>
<td>Target</td>
<td>70%</td>
</tr>
<tr>
<td>Initial step</td>
<td>10 dB</td>
</tr>
<tr>
<td>Minimum step size</td>
<td>2 dB</td>
</tr>
<tr>
<td>Step size calculation</td>
<td>(s = S_i \times (\frac{1}{2})^R)</td>
</tr>
<tr>
<td>Stop criterion</td>
<td>After eight reversals</td>
</tr>
<tr>
<td>Threshold estimation</td>
<td>Average of last six reversals</td>
</tr>
</tbody>
</table>

Note. HSE = head shadow effect; SQ = binaural squelch; SPL = sound pressure level; SNR = signal-to-noise ratio; \(s\) = step size; \(S_i\) = initial step size; \(R\) = number of reversals.
Figure 3. BIN: Binaural squelch (SQ) test with both ears available with the signal coming from the front and the noise coming from −90° azimuth when the right ear is selected as the first ear. The speech recognition threshold obtained at 70% correct response rate in this example is 17 dB SPL (82–65 dB SPL). MON: SQ test with only the right ear available. The score at 17 dB SNR is 50%, so the SQ = 20% (70%–50%). LiCoS = linguistically controlled sentences.

SSQ Scale

We used the 12-item version of the SSQ, which was translated into Flemish by KU Leuven University (Noble et al., 2013). The SSQ is a subjective self-assessment tool that mainly focuses on speech, spatial listening, and other qualities of hearing (Gatehouse & Noble, 2004). Since its development, SSQ has been translated into many languages (Falzone et al., 2022; Kılıç et al., 2021; Moulin et al., 2015) and has been widely used in many other studies (Jang et al., 2019; Laske et al., 2009; Mertens et al., 2015; Vermeire & Van de Heyning, 2009). In this scale, respondents rate their own performance on each item from 0 (not at all) to 10 (perfect). Higher score always indicates a greater ability since none of the items is worded in a negative direction.

Statistical Analyses

Descriptive statistics (minimum, first quartile, median, third quartile, maximum, and interquartile range) were used to summarize data. The independent-samples t test was used to compare the CI and NH groups. The paired-samples t test and the Wilcoxon signed-ranks test, depending on

Figure 4. BIN: Head shadow effect (HSE) test with both ears available with the signal coming from −90° azimuth and the noise coming from front when the right ear is selected as the first ear. The speech recognition threshold obtained at 70% correct response rate in this example is 9 dB SPL (74–65 dB SPL). MON: HSE test with only the right ear available. The score at 9 dB SNR is 30%, so the HSE = 40% (70%–30%). VL = Vlaams (English translation: Flemish).
whether parametric test assumptions were met, were used to compare the performance of the first and second implanted ears in the CI group. The Friedman test was used to compare the results of three different speech materials. During the analysis, only the first results from each speech test were used. Retest results were not taken into account in this analysis. The Spearman’s rank correlation coefficient was used for correlation analysis. The Shapiro–Wilks test was used to test for normality. The significance level was always set at 0.05. IBM SPSS Statistics for Windows, Version 20.0 (IBM Corp.) was used for data analysis.

At the end of the first phase, the following data were available for each speech material: (a) test results (= effect values), (b) test–retest results, and (c) interindividual variability. The performance of a diagnostic test is judged by how accurately the test result can identify a diseased (sensitivity) or a nondiseased (specificity) person (Hui & Zhou, 1998). To ensure adequate sensitivity and specificity in the HSE and SQ tests, the NH and CI groups must be distinguishable from each other. It is therefore reasonable to choose speech material that produces the largest difference between the groups. Interindividual variability is called random, unexplained, residual, or variation due to error. That’s why it is expected to be as low as possible in control groups (in this case the NH group; Drummond & Vowler, 2012). Then, finally, the test–retest reliability is an indication of the internal validity of the test and of its ability to produce stable and reproducible results over time. For these reasons, we determined that the optimal speech material would be characterized by (a) the largest effect size in NHs, (b) the highest test–retest reliability in NHs and CIs, and (c) the smallest interindividual variability in NHs.

Effect size calculations used Cohen’s $d$ for independent-samples and paired-samples $t$ tests, effect size $r$ for Wilcoxon signed-ranks test, and Kendall’s $W$ for the Friedman test.

Test–retest reliability was calculated using the two-way mixed intraclass correlation coefficient (ICC; Koo & Li, 2016; McGraw & Wong, 1996). ICC is a widely used reliability index in test–retest analysis (Koo & Li, 2016; McGraw & Wong, 1996). ICC has advantages over other methods that can be used for test–retest analysis such as paired $t$ test and Pearson correlation. For example, the paired $t$ test is a method for analyzing agreement between two measurements, and Pearson correlation is a way to measure correlation that does not take systematic differences into account. ICC, however, reflects both degree of correlation and agreement (Koo & Li, 2016). ICC values less than 0.5 indicate poor test–retest reliability, those between 0.5 and 0.75 indicate moderate reliability, those between 0.75 and 0.9 indicate good reliability, and those greater than 0.9 indicate excellent reliability (Koo & Li, 2016; McGraw & Wong, 1996; Portney & Watkins, 2009).

Interindividual variability was calculated using the coefficient of variation (CV). It is calculated by dividing the standard deviation by the mean.

Results

In the first implanted ears of the CI recipients, the mean aided PTA was $18 \pm 5$ dB HL, whereas in the second implanted ears, it was $22 \pm 6$ dB HL. The median aided EaSI score was 89% ($Q_1$: 79.8 and $Q_3$: 93) on the first implanted side and 82.5% ($Q_1$: 76.8 and $Q_3$: 89) on the second implanted side. The PTA and the EaSI scores on the first CI side were 4 ± 6 dB HL (parametric paired-samples $t$ test; $t[33] = -3.812$, $p = .001$, Cohen’s $d = 0.65$) and 4% ($Q_1$: –0.3 and $Q_3$: 8.3: nonparametric Wilcoxon signed-ranks test; $Z = -2.602$, $p = .009$, effect size $r = .32$) significantly better than those on the second CI side.

The Effect of Speech Material in the HSE and SQ Tests

Effect Size

Both for the HSE and SQ tests, the BLU and LiCoS test results in the NH group ($n = 6$) were higher than those of the NVA, although the differences were not statistically significant (nonparametric Friedman test for HSE: $\chi^2[2] = 4.364$, $p = .113$, Kendall’s $W = 0.36$, and for SQ: $\chi^2[2] = 4.333$, $p = .115$, Kendall’s $W = 0.36$). Table 2 presents the descriptive statistics for the results.

Test–Re test Reliability

For both tests, LiCoS achieved the highest ICC agreement in NH individuals (83.4% and 82.6%; see Table 3). The median test–retest differences in HSE for NH individuals
The interquartile range in the NH group in the HSE test was 37.8%, 21%, and 19.5% for NVA, BLU, and LiCoS, respectively. For the SQ, the values were 26.3%, 41%, and 22.3%.

In summary, for the criteria tested, there was no statistical difference between the three different speech materials for the HSE and SQ tests. However, on the basis of descriptive statistics, results indicated that BLU and LiCoS were more robust materials compared with NVA. Although BLU and LiCoS showed similar results in terms of effect size and interindividual variability, we still judged LiCoS to be more effective than BLU because of its higher test–retest reliability, especially in CI users. All the results of this phase are summarized in Table 3.

### Normative Data of HSE and SQ

Thirty NHs were tested to obtain normative data for the HSE and SQ tests with LiCoS sentences. The mean HSE (±SD) was 58 ± 14% (95% CI [53, 64]), and the mean SQ was 22 ± 11% (95% CI [17, 26]) in the NH group.

The initial phase of this study revealed that the median test–retest differences in NH individuals with the LiCoS sentence test for HSE and SQ were 8% (Q1: 2 and Q3: 10; range: 0–10) and 5% (Q1: 0 and Q3: 11; range: 0–13), respectively.
Binaural Benefits in Bilateral CI Users

The mean HSE was 49 ± 13% (n = 25, 95% CI [42, 54]), and the SQ was 13 ± 14% (n = 25, 95% CI [8, 20]) in the CI group. There were six out of 25 participants (24%) with a negative SQ in the CI group (see Figure 6 for the histogram). Furthermore, the HSE and SQ scores of the CI group were significantly lower than those of the NH group, as illustrated in Figure 7 (parametric independent-samples t test for HSE: t[53] = 2.73, p = .009, Cohen’s d = 0.74, and for SQ: t[53] = 2.64, p = .011, Cohen’s d = 2.48).

The median test–retest differences in the CI users with the LiCoS sentence test for HSE and SQ were 12% (Q1: 5 and Q3: 20; range: 2–25) and 10% (Q1: 6 and Q3: 18; range: 5–25), respectively. The average RMS error in the azimuth LOC test was 15° ± 5° (n = 21, 95% CI [13, 17]); see Figure 8.

The average score on the SSQ-12 (n = 16) was 5 ± 2 (95% CI [4, 6]). There were no statistically significant correlations between the HSE, SQ, LOC, and SSQ results (nonparametric Spearman’s rank correlation coefficient for HSE-LOC: r(17) = −.061, p = .809; HSE-SSQ: r(12) = .229, p = .452; SQ-LOC: r(17) = .120, p = .636; SQ-SSQ: r(12) = −.237, p = .435; LOC-SSQ: r(15) = −.162, p = .549).

Further analysis did also not reveal any significant correlations between the test results and the demographic variables such as the interval between the implantations, duration of binaural experience, and asymmetry between the ears (nonparametric Spearman’s rank correlation coefficient for HSE-Interval between the CI’s: r(24) = −.240, p = .248; HSE-Binaural experience: r(24) = −.063, p = .766; HSE-PTA asymmetry: r(24) = −.148, p = .480; HSE-EaSI asymmetry: r(24) = .032, p = .879; SQ-Interval between the CI’s: r(24) = −.081, p = .702; SQ-Binaural experience: r(24) = .132, p = .530; SQ-PTA asymmetry: r(24) = −.230, p = .268; SQ-EaSI asymmetry: r(24) = −.154, p = .462).

Discussion

The primary purpose of this study was to identify which speech material is most effective for the HSE and SQ tests. After that, normal values in NH individuals were obtained using the speech material decided to be optimal. In addition, LOC, SSQ, HSE, and SQ tests were used to test bilateral CI users.
The benefits of hearing with two ears compared to hearing with one ear can be divided into two groups: physical effects and central effects. HSE is defined as a purely physical effect (Brown & Balkany, 2007; Van Hoesel & Litovsky, 2011), whereas SU, SQ, and LOC result from binaural processing in the central auditory system (Tyler et al., 2002, 2003). SU does not originate from interaural cues and is solely dependent on binaural redundancy (Laszig et al., 2004; Van Hoesel, 2012). Hence, a more redundant material in the SU test could result in a larger effect size, which would result in lower test–retest differences. Since HSE and SQ do not rely on redundancy, but rather originate from interaural differences, it is not possible to make such an assumption for them. As for SRM, its origins are controversial. While some researchers suggest it is a combination of HSE and SQ (Aronoff et al., 2011; Dieudonné & Francart, 2019), others say it is only the result of HSE (Gifford et al., 2014; Sheffield et al., 2015). However, in any case, its origins are not different from HSE and/or SQ. For these reasons, only HSE and SQ were investigated in this study. Even though we believe that the results of this study can be generalized to SU and SRM, the fact they were not included in this study can be considered a shortcoming.

The Effect of Speech Material in the HSE and SQ Tests

The materials tested in the initial phase included monosyllabic NVA words, spondaic BLU words, and LiCoS sentences.

The small sample size in the first phase was the most likely reason for the lack of statistical differences between the speech materials. Due to the absence of statistically significant differences based on inferential statistics, descriptive statistics were used to interpret the results of the first phase. This might be considered one of the limitations of this study. However, descriptive statistics indicated that the sentences provided higher test–retest reliability than the monosyllables and spondees.

One might question the applicability of linguistically controlled sentences in children with CI. Existing research suggests that speech intelligibility measured by predictable linguistically controlled sentences varies mainly by age and cognitive abilities (Uslar et al., 2011, 2013). Uslar et al. (2011) showed that young listeners have much more difficulty with complex speech materials than with simple materials in comparison to adults. Pichora-Fuller (2008) and Uslar et al. (2011) attributed the advantageous effects
of older age to (a) greater experience with language and speech processing in difficult circumstances; (b) improved benefits from supportive context and redundancies, such as syntactic cues; and (c) their expert knowledge of structures when signal quality is poor. However, when the sentences are unpredictable with varying degrees of syntactic complexity (like in LiCoS), older adults have no advantage over young listeners anymore because the strategies used by experienced listeners do not work for unpredictable sentences (Uslar et al., 2011, 2013). Uslar et al. (2013) reported also that cognitive skills play a greater role in speech understanding of complex sentences in the presence of fluctuating noise than in the presence of stationary noise. In other words, listening in fluctuating noise requires a higher cognitive contribution than, for instance, listening in quiet or in stationary noise. It is important to note that the noise used in this study was stationary. Cognitive skills are, therefore, not expected to have a significant impact on the outcomes of the LiCoS test.

**Effect Size**

To the best of our knowledge, there are no other studies in the literature evaluating and comparing the effectiveness of different speech materials in HSE and SQ tests. Nonetheless, when the results of a variety of studies, each of which used a different material, are analyzed together, it becomes evident that the choice of speech material could influence the outcomes. As an example, monosyllabic words have the lowest SQ effect (1.9–3.7 dB SRT), whereas spondee words and closed-set sentences produce similar effects (4–7 dB and 4.5–7 dB SRT, respectively; Bronkhorst & Plomp, 1988; Carhart, 1965; Dirks & Wilson, 1969; MacKeith & Coles, 1971). In this study, we also found that spondee words and sentences had the largest effect sizes both in the HSE and SQ tests. Thus, it becomes apparent that binaural benefits increase with the redundancy of the speech material of interest. In the test setup, we create a difficult listening situation that creates gaps in the incoming acoustic signal. The test task is to fill those gaps by adding the second ear, and it turns out that this addition does so better if the acoustic material also contains redundant information.

**Test–Retest Reliability**

Studies conducted so far have used a variety of speech materials. Although many researchers used sentence stimuli, some also used other materials. For example, Morera et al. (2012) and Verhaert et al. (2012) used spondee words, and Arsenault and Punch (1999) used nonsense syllables. In their studies, Aronoff et al. (2011) and Devocht et al. (2017) evaluated test–retest variations of the sentences they used but made no comparison with other materials. Aronoff et al. (2011) found a statistically significant correlation between test and retest results with a correlation coefficient of .9. In this study, however, the ICC method was preferred over correlation analysis. In both the HSE and SQ tests in NHs, LiCoS had an ICC value of > 80%, indicating good reliability (Koo & Li, 2016). For CI users, the ICC values were 69% and 70%, respectively, indicating moderate reliability, which is also in line with the results reported by Devocht et al. The ICC values reported by Spyridakou et al. (2020) for the right and left ears in speech understanding in noise test using monosyllabic words in NH individuals were 25% and 39%, respectively. The ICC values for NH individuals obtained using monosyllabic words in this study were 82.6% and 55.3% for the HSE and SQ, respectively (for LiCoS sentences the values were even higher: 83.4% and 82.6%). However, the test–retest variability observed among NHs was lower than that among CI users, possibly due to ceiling effects as indicated by other studies demonstrating that test–retest variation decreases with the increasing SNR of the test, with the increasing performance level of the listeners, and with the decreasing linguistic complexity of the speech material (Grange, 2013; Hey et al., 2014; Kim et al., 2015; Uslar et al., 2011, 2013). The number of items in a speech list is another important factor affecting test–retest differences. Kim et al. (2015) demonstrated that there was no significant difference in test–retest differences between 25- and 50-item lists, but there was a significant difference when a 10-item list was included in the comparison. In each speech list, NVA and BLU had 12 items; LiCoS had 30. However, the participants in this study were presented with two lists of NVA and BLU and one list of LiCoS to establish an equivalence. As a result, 24 items from each of the NVA and BLU tests and 30 items from LiCoS were presented to the participants, ensuring that the number of items would not affect the test–retest scores.

Even though the test–retest reliability reported in this study is consistent with those reported in the literature, and ICC analysis shows that these tests have moderate-to-good test–retest reliability depending on the target population, the results should be interpreted with caution. As discussed above, test–retest differences in speech audiometry can reach up to 10% in NH individuals and up to 15% in CI users. This makes it particularly difficult or risky to interpret such results in individual test subjects. It is only by analyzing group results that the relatively large test–retest differences get averaged out. Therefore, the authors believe that interpreting speech audiometry results, in this case especially the binaural test results, would yield more reliable results on a group basis.

**Interindividual Variability**

In line with previous research, the results showed considerable variation among individuals. As expected,
the variability among individuals in the NH group was lower than that in the CI group. One might argue that the higher variation in the CI group was due to the wider age range of the individuals. However, there were no outliers in the data of the participants aged > 50 years or < 16 years (a total of seven participants). In HSE, five out of seven participants (71%), and in SQ, all seven participants had data within the mean ± 1 SD. Other possible reasons for the higher variation in the CI group's results include different CI signal processing strategies, the different spread of the electrical field generated by the implant, and the preservation of spiral ganglion cells (Williges et al., 2015). Future research can focus on the development of test methods (e.g., different test setups or implementation of electrophysiological methods) that show less variable results in the assessment of binaural benefits.

Normative Data of HSE and SQ

The mean HSE was 58 ± 14%, and the mean SQ was 22 ± 11% in the NH group. In the literature, HSE varies between 8.9 and 10.7 dB in individuals with NH (Arsenault & Punch, 1999; Bronkhorst & Plomp, 1988), and the SQ ranges between 1.9 and 4.9 dB (Arsenault & Punch, 1999; Bronkhorst & Plomp, 1988; Carhart, 1965; Cox et al., 1981; MacKeith & Coles, 1971). In test setups run at fixed SNRs, HSE ranges between 20% and 30% (Arsenault & Punch, 1999) and SQ ranges between 10% and 26% (Arsenault & Punch, 1999; Cox et al., 1981).

Binaural Benefits in Bilateral CI Users

Bilateral CI users in this study had 49 ± 13% HSE and 13 ± 14% SQ. These results are in line with previous studies in literature in which HSE ranged from 22% to 49% and SQ from 1.7% to 18% (Buss et al., 2008; Eapen et al., 2009; Gantz et al., 2002; Laszig et al., 2004; Tyler et al., 2002; Van Hoesel, 2012; Van Hoesel et al., 2002; Verhaert et al., 2012).

HSE had the most robust results. In this study, all CI users (n = 25) showed HSE, but 72% were able to have a positive SQ. Tyler et al. (2002) and Gantz et al. (2002) demonstrated that 80% of bilateral CI users had an HSE, whereas a significant SQ was reported only in three out of nine participants. Laszig et al. (2004), on the other hand, reported significant HSE and SU but no significant SQ even 6 months after the implantation. Litovsky et al. (2006) reported that 94% of 34 simultaneously implanted bilateral CI users had HSE, whereas only 47% had SQ. Laske et al. (2009) reported significant results after bilateral implantation only in HSE. Thus, in most studies, less than half of bilateral CI users have a measurable SQ benefit from their second CI. This finding seems important to us. This means that more than half of the participants do not experience a significant SQ after the second CI, and some even have a negative SQ, indicating a deterioration of speech understanding in noise with the second CI on compared with the first CI alone. In our sample, 24% of the participants also had a negative SQ. However, better speech understanding in noise is actually often cited as an important argument for a second CI. It turns out that we should be careful with this argument. Therefore, the experience that most bilateral CI users are satisfied with the second implant is mainly attributable to the HSE that is observed for a high proportion.

Overall, the binaural test results of CI users were significantly lower than those of NH individuals. Other studies also reported similar results (Arsenault & Punch, 1999; Bronkhorst & Plomp, 1988; Kokkinakis & Pak, 2014). A functional hearing system is able to process and integrate bilateral acoustic cues smoothly. However, it is not always possible to achieve a similar success in artificial hearing provided by electrical stimulation in CI users. CI users perform less well in binaural hearing tasks than NH people for a variety of reasons. Most notably, CI users have a reduced sensitivity for interaural time and level differences (ITD and ILD) in addition to inadequately encoded TFS information with current sound processing strategies (Brown & Balkany, 2007; Kokkinakis & Pak, 2014; Litovsky et al., 2006; Tyler et al., 2003). In their study, Litovsky et al. (2006) outlined three main reasons for reduced binaural advantages in CI users: “hardware- and software-related, surgical-based, and pathology-related.” In summary, CI users have two independent monaural hearing systems. Thus, the time base of each processor can differ slightly, resulting in random jitters in the ITD of the envelope and the carrier pulses, disrupting the ITD cues (Litovsky et al., 2006). Most CI systems process incoming signals by extracting only the temporal envelope and amplitude, modulating it to a fixed-rate pulsatile carrier (Ching et al., 2007). They do not provide the TFS information that is critically important for detecting ITDs. Despite the potential for the temporal envelope to convey timing information, ITDs would not be consistent because of the variations in detection thresholds across different electrodes (Ching et al., 2007). Furthermore, although CI users’ speech understanding improves with higher pulse rates, their ITD sensitivities drop significantly (Dunn et al., 2006; Thakkar et al., 2018; Van Hoesel et al., 2009). Additional factors that may distort binaural cues include different microphone characteristics, independent automatic gain control and compression algorithms, and different signal processing strategies between the two implants (Brown & Balkany, 2007; Ching et al., 2007; Litovsky et al., 2012; Snik et al., 2015; Tyler et al., 2002, 2003, 2006). For
instance, when one of the processors compresses its input more than the other, the brain perceives the sound as moving from one side to the other, which may negatively affect spatial hearing (Tyler et al., 2003). Another problem is the spectral mismatch between the electrodes due to the different surgical insertion depths. Accordingly, Yoon et al. (2013) found that increased spectral mismatch caused by different insertion depths affected SQ negatively but not HSE. A binaural CI may eliminate the aforementioned problems of two independent CI systems. A binaural CI has two different electrode arrays protruding from a single internal device, and these electrodes are placed in both cochleae. While there is a sound processor on the side of the internal device, the contralateral ear only has a microphone connected to the sound processor by a cable. Verhaert et al. (2012) investigated the effects of binaural cochlear implantation after 12 months of use in 14 adults with postlingual hearing loss. There was a significant difference between participants’ SRSs in silence and noise in the binaural condition compared with the unilateral condition. Significant binaural advantages were also present in HSE, SU, and SQ tests. In addition, a significant improvement of 35° RMS was observed in the LOC task. These results confirm that pseudosynchronous stimulation of binaural CIs might have positive effects on binaural hearing.

There was no significant correlation between HSE/SQ test results and LOC. Similarly, Schleich et al. (2004) and Tyler et al. (2006) also could not find a significant correlation between HSE/SQ and horizontal LOC in bilateral CI users. While HSE is not directly related to spatial listening, it is believed that SQ relies on the same binaural cues (ITD) that also allow the LOC of sound sources (Eapen et al., 2009). It was therefore expected that LOC and SQ results could have a significant correlation. Lack of correlation between them raises the question of whether they represent the ability to use the same cues. Similarly, Cox and Bisset (1984) concluded in their study that SQ tested with traditional methods does not reflect the ability to exploit interaural differences. These findings support the idea that future research should focus on the development of different test methods to assess SQ.

The SSQ was used in this study to examine how well the psychoacoustic test results matched patients’ daily life experiences. However, SSQ scores were not significantly correlated with HSE/SQ or LOC results. Using a similar sample of 34 bilateral CI users with at least 6 months of binaural experience, Laske et al. (2009) found that the LOC results were significantly correlated with the “spatial hearing” subcategory of the SSQ. Despite similar patient populations, there were several methodological differences between the two studies, including the type of noise (broadband vs. speech) and the total number of speakers (12 vs. 7) used in the LOC test, and the number of items in the SSQ (25-item vs. 12-item). These methodological differences may have led to different results in the studies.

The results of this study showed that none of the variables investigated correlated with the binaural hearing advantages. Similarly, Schleich et al. (2004), Tyler et al. (2006), and Laske et al. (2009) could not find a relationship between binaural benefits and other variables such as the duration of deafness of the ears or the asymmetry in the performance of the ears. These findings support the argument of Gantz et al. (2002) that there is no parameter for predicting the postoperative binaural benefits. We believe that the main reason for this is the high variation in both individual and group results observed in the HSE and SQ tests.

Conclusions

The primary purpose of this study was to identify which speech material is most effective for the HSE and SQ tests. The results showed that although there were no significant differences between spondee words and sentences in terms of effect size and interindividual variability, test-retest reliability was higher with sentence stimuli, especially in CI users. With the LiCoS sentence test as optimal speech material selected, normative data were obtained in NH individuals. All CI users in the study sample had an HSE, whereas 72% had SQ. We conclude that the benefit of the second CI is mainly attributable to HSE whereas SQ benefit is not achieved in all cases.

Data Availability Statement

All data generated and/or analyzed in this study are available from the corresponding author on reasonable request.

Author Contributions

Okan Öz: Conceptualization (Equal), Data curation (Lead), Formal analysis (Lead), Investigation (Lead), Methodology (Equal), Project administration (Equal), Validation (Equal), Visualization (Lead), Writing – original draft (Lead). Hilal Dinçer D’Alessandro: Conceptualization (Equal), Methodology (Equal), Project administration (Equal), Supervision (Equal), Validation (Equal), Writing – review & editing (Equal). Merve Özbah Batuk: Conceptualization (Equal), Methodology (Equal), Supervision (Supporting), Writing – review & editing (Supporting). Gonca Sennaroğlu: Conceptualization (Equal), Methodology (Equal), Supervision (Supporting), Writing – review & editing (Supporting). Öz et al.: Assessment of Binaural Benefits

13
(Supporting). Paul J. Govaerts: Conceptualization (Equal), Data curation (Supporting), Formal analysis (Supporting), Methodology (Equal), Project administration (Equal), Resources (Lead), Software (Lead), Supervision (Equal), Validation (Equal), Writing – review & editing (Equal).

References


Laske, R. D., Veraguth, D., Dillier, N., Binkert, A., Holzmann, K., Kan, A., & Litovsky, R. Y.


