

Perceptual measures of normal-hearing and hearing-impaired listeners across defined virtual acoustic scenes

Merle Gerken^a, Julia Schütze^a (shared first authors), Christoph Kirsch^a, Stephan D. Ewert^a, Bernhard U. Seeber^b, Jan Heeren^c, Tahereh Afghah^c, Kirsten C. Wagener^c, Birger Kollmeier^a, Anna Warzybok^a

^aMedizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

^bAudio Information Processing, School of Computation, Information and Technology, Technical University of Munich, Munich, Germany

^cHörzentrum Oldenburg gGmbH and Cluster of Excellence Hearing4all, Oldenburg, Germany

Keywords: speech intelligibility, listening effort, hearing impairment, hearing aid evaluation, dataset, virtual acoustic scenes

Abstract

Objective

To achieve a multi-dimensional characterization of normal-hearing and hearing-impaired listeners' hearing abilities, including standard audiological conditions and virtual acoustic scenes for the evaluation of hearing aids.

Design

Within-subjects design, including speech intelligibility and listening effort measurements in two standard conditions and four virtual acoustic scenes, and assessments of loudness scaling, tone-in-noise detection thresholds, and HEAR-COMMAND tool questionnaire.

Study Sample

76 age-matched listeners, including 20 normal-hearing, 25 hearing-impaired without hearing aids, and 31 hearing aid users.

Results

Speech intelligibility and hearing aid benefit in virtual acoustic scenes fell between the results of the standard audiological conditions *SON0* and *SON90*. The *SON0* and *UG_station* environment were the most challenging conditions regarding speech intelligibility and listening effort. The pure-tone average explained most of the differences between the listener groups in loudness perception and tone-in-noise detection thresholds. Moderate to strong correlations were found between the HEAR-COMMAND tool speech scores and speech intelligibility.

Conclusions

The study established a unique measurement database including complex virtual acoustic scenes and demonstrated a connection between speech intelligibility, hearing aid benefit, and other perceptual auditory measures. The database and findings provide a valuable foundation for advancing hearing aid evaluation and can serve as a benchmark in computational audiology.

1. Introduction

Methods in audiological diagnostics and in the assessment of hearing devices benefit vary considerably globally or even among hearing centers within the same region. A common shortcoming, however, is that most clinical tests involve low acoustic complexity and fail to reflect the challenging listening conditions individuals typically encounter in daily life. Several studies have reported a mismatch between laboratory-based measures of hearing aid benefit and real-life outcomes (Bentler, 2005; Cord et al., 2004).

The present article introduces a comprehensive audiological dataset specifically designed to address this gap. The test battery combines traditional and extended measures of hearing function. Standard clinical assessments, such as pure-tone audiometry and speech intelligibility in stationary noise, are complemented by additional measures that capture supra-threshold hearing deficits and listening-related demands. These include loudness perception, tone-in-noise detection, listening effort, speech intelligibility in acoustically complex conditions that are supposed to better simulate real-life listening environments, and subjective self-reports using the HEAR-COMMAND tool (Afghah et al., 2022, 2024), a comprehensive questionnaire assessing hearing, conversation, and communication abilities in daily-life contexts. Virtual acoustic scenes such as a living room, a noisy pub, and an underground station were created using room acoustics simulation and multichannel audio rendering. These provided realistic, yet reproducible, listening environments that approximate real-life challenges. Listening experiments were conducted with participants exhibiting different degrees of hearing loss, partially supplied with hearing aids (HAs).

A detailed technical description of the test battery structure and organization is provided in a companion article (Afghah, Heeren, et al., 2025), where the full test framework is made available as open-access resources to the research community. The companion article includes the full dataset, which is more comprehensive than the data presented here. In the present article, data are included only for those measurements in which all listeners were tested. For some of the remaining measures, only subgroups of listeners participated. The present article focuses on the conceptual scope, the audiological rationale, and detailed descriptions of measurements and outcomes.

The resulting dataset aims to contribute to a deeper understanding of the perceptual consequences of hearing loss and the individual benefit derived from hearing devices across a broad range of listening scenarios and tasks. It provides a valuable resource for the evaluation of hearing aid technologies, where access to varied and realistic test conditions is critical for evaluating signal processing strategies. Furthermore, this data set can be used as a benchmark for the development or validation of auditory computational models. By including a heterogeneous group of age-matched participants, with varying degrees of hearing loss, both aided and unaided, the dataset offers broad applicability for a range of research questions. The selection of test measures and listening scenarios was guided by the goal of enabling both basic auditory profiling and application-oriented research on hearing device performance.

Several existing studies have provided extensive audiological datasets with only a few of them available as open-source datasets (Gieseler et al., 2017; Kamberer et al., 2019; Rönnberg et al., 2016). One example is the HearCom project (Vlaming et al., 2011) which proposed a test battery covering seven domains: loudness perception, spectral and temporal resolution, speech intelligibility in quiet and in noise, spatial hearing, cognitive abilities, listening effort, and self-reported disability and handicap. This battery was evaluated in an international multi-center study involving over 100 participants from four countries. However, it remains accessible only through consortium agreements.

Another key initiative is the BEAR test battery (Sanchez-Lopez et al., 2021), which includes data from young normal-hearing reference listeners, older normal-hearing listeners and older hearing-impaired

listeners with mild to severe sensorineural hearing loss (unaided results only). The BEAR battery encompasses measurements of pure-tone threshold in quiet and in noise, loudness scaling, word intelligibility, speech-in-noise tests, spectro-temporal modulation and binaural audiometry, all aimed at improving data-driven auditory profiling. Most importantly, the full BEAR dataset is openly accessible on Zenodo.

Recently, the Oldenburg Hearing Health Record was published with data from 581 participants collected between 2013 and 2015 at the Hörzentrum Oldenburg, encompassing pure-tone audiometry, adaptive loudness scaling, speech reception tests, cognitive measures, and self-report questionnaires (Jafri et al., 2025). While audiometric tests were administered unaided, cognitive measures and self-report questionnaires were administered aided for hearing-impaired listeners who own hearing devices. This resource is anonymized and publicly released under a CC-BY 4.0 license via Zenodo.

Another open-access resource is the dataset accompanying Regev et al. (2025b), which includes measurements such as audiograms, speech reception thresholds, and amplitude modulation sensitivity for younger and older hearing-impaired listeners (unaided only). While this dataset is valuable for targeted investigations of specific suprathreshold deficits, it is limited in scope, as it only includes a narrow range of audiological tests. Additionally, the cohort size is relatively small, which may restrict its generalizability for broader applications in auditory profiling or hearing aid evaluation. The data is accessible via Regev et al. (2025a).

However, most of the existing databases are limited to rather simple and highly controlled listening scenarios that lack the acoustic complexity of daily environments. Measurement conditions involving reverberation, multiple competing talkers, and realistic background noise are rarely included. The test battery presented here addresses this gap by systematically incorporating more realistic everyday-life, complex acoustic conditions into the test protocol and by combining well-established audiological measures with complementary perceptual assessments resulting in a unique and comprehensive dataset. In addition, for hearing-impaired listeners who regularly use hearing aids in daily life, speech intelligibility and listening effort results are available in unaided and aided conditions. Furthermore, the HEAR-COMMAND tool captures self-reported daily listening experiences, with the aided condition reflecting typical real-life listening scenarios.

Traditional audiological testing, usually performed in quiet or in simple background noise, does not reflect the dynamic, multisource, and reverberant nature of real-world acoustic environments. This can lead to overestimation of both hearing ability and hearing aid benefit. Integrating complex acoustic scenes in measures like speech intelligibility and listening effort provides more ecologically valid insights into functional hearing. Additional perceptual measures such as loudness perception, tone-in-noise detection, and subjective self-reports assess supra-threshold deficits and individual variability that are not revealed by traditional audiological tests.

By providing this dataset along with open-access resources for reproduction and further research, this work aims to support the development of data-driven auditory profiling, computational models of hearing, and enable evidence-based evaluation of hearing device performance in conditions that more closely approximate the complexity of everyday listening.

This study is guided by the following research questions:

- How does speech intelligibility performance differ between age-matched normal-hearing listeners and unaided/aided hearing-impaired listeners across traditional audiological setups and complex acoustic scenes?
- How does hearing aid benefit vary with the complexity of the listening environment?

- How do measures such as loudness perception, tone-in-noise detection, and subjective self-reports contribute to a multidimensional characterization of hearing abilities, and how do speech intelligibility and listening effort vary with the complexity of the listening environment?

2. Methods

2.1 Listeners

In total, 76 listeners (35 w, 41 m) participated. They have been divided into 3 groups: 1. normal-hearing (NH, $N=20$), 2. hearing-impaired (HI) not wearing hearing aids in daily life (HI_noHA, $N=25$), and 3. hearing-impaired wearing bilateral, commercial hearing aids (HI_HA, $N=31$). For the latter group, measurements were conducted without (HI_HAu) and with hearing aids (HI_HAa). For aided measurements, hearing aids were used in their daily settings with no adjustments or deactivation of any system features (e.g. noise reduction), ensuring that aided measurements reflected natural, daily life listening conditions. The NH group was defined as listeners whose pure-tone threshold average across frequencies 0.5, 1, 2 and 4 kHz (PTA4) and ears was below 25 dB HL. According to a one-way ANOVA, the groups were matched in age [$F(2,75)=1.98$, $p=0.145$, $\eta^2=0.51$]. The age of the listeners ranged from 49 to 82 years in the NH group, 59 to 84 years in the HI_noHA group, and 60 to 84 years in the HI_HA group, with a mean age of 74.0 years (standard deviation $SD=7.1$ years) across all participants. In the following, the PTA4_M refers to a mean PTA4 across the ears. A one-way ANOVA showed that the listener groups significantly differed in the PTA4_M [$F(2, 75)=146.98$, $p<0.001$, $\eta^2=0.801$]. Pairwise comparisons (with Bonferroni correction) showed that all groups differ from each other in the PTA4_M (all comparisons with $p<0.001$). The absolute PTA4 difference between the right and left ears did not differ statistically across the groups [$F(2, 75)=0.957$, $p=0.389$, $\eta^2=0.026$] and averaged 4.7 dB HL ($SD=3.4$ dB HL), with a maximum difference of 15 dB HL. The NH group ranged in the PTA4_M from 1.3 to 24.4 dB HL, with a mean PTA4_M of 14.6 dB HL ($SD=6.3$ dB HL). For HI_noHA, the PTA4_M ranged from 26.9 to 53.8 dB HL, with a mean of 38.4 dB HL ($SD=6.7$ dB HL). Listeners of the HI_HAu group ranged in the PTA4_M from 33.8 to 76.3 dB HL, with a mean of 52 dB HL ($SD=8.9$ dB HL). Individual and group-specific pure-tone thresholds are shown in Figure 1.

All participants provided written informed consent in accordance with the approval of the pre-medical Ethics Committee of the University of Oldenburg. They received monetary compensation for their participation. All measurements for participants with hearing aids were conducted across three appointments, each lasting approximately 1.5-2 hours, resulting in a total testing time of about 5-6 hours. For participants without hearing aids, testing required approximately 3-4 hours in total. The first session included the general audiological and psychoacoustic measurements, the second session comprised speech intelligibility and listening effort tests in the unaided condition, and the third session included the corresponding aided measurements. The present paper reports data for a subset of these measurements that were conducted for all listener groups. The overall order of measurements was fixed, while the order of conditions within each measurement was randomized. Participants could take self-paced breaks at any time, and additional short breaks were scheduled approximately every 30 minutes to minimize fatigue.

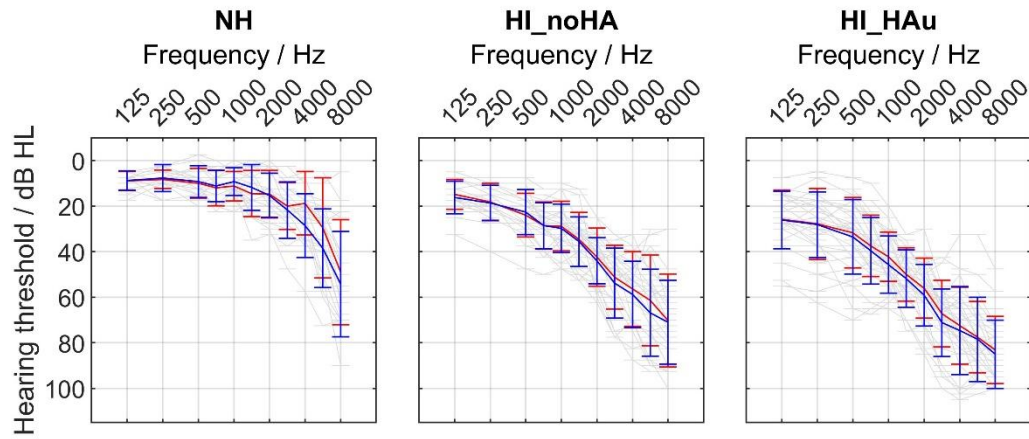


Figure 1: Mean pure-tone thresholds with corresponding standard deviation for the three different participant groups for left (blue) and right (red) ear. Individual audiograms averaged across the left and right ear are shown in grey.

2.2 Speech intelligibility measurements

The Oldenburg sentence test (Oldenburger Satztest, OLSA) (Wagener et al., 1999; Wagener & Brand, 2005) was used as speech material for speech intelligibility measurements. It consists of 5-word semantically unpredictable sentences of a fixed grammatical structure (name, verb, numeral, adjective, noun). Test lists of 20 sentences and an adaptive procedure converging to 50% speech intelligibility, defined as speech reception threshold (SRT), were used (Brand & Kollmeier, 2002). This holds for all acoustic conditions, including standard laboratory conditions and complex acoustic scenes. The first sentence was always presented with a signal-to-noise ratio (SNR) of 0 dB SNR. Prior to the actual measurements, listeners were trained with one test list to familiarize them with the speech material and task. The order of the acoustic conditions, the index of the test list as well as the order of the sentences within a test list were randomized across the listeners.

Speech intelligibility was measured in six acoustic conditions, including standard laboratory conditions and complex acoustic scenes. Standard laboratory conditions included two common spatial configurations: one with the target and masker co-located frontally (*SON0*), and another with the target presented frontally and the masker presented from a 90° angle (*SON90*). A stationary, speech-shaped masker (OLnoise) was used for both configurations.

Complex acoustic scenes involved different environments, which were originally introduced by van de Par et al. (2022) and include a living room, a pub, and an underground station. These scenes were selected to represent realistic everyday environments with varying levels of acoustic complexity, as typically experienced by hearing aid users. The simulations of the complex acoustic scenes used in this study were based on acoustic measurements conducted in real-world locations in Oldenburg and Munich. Those acoustic measurements have only been used to design the acoustical simulation parameters. The audio playback in the current experiments was taken from simulated audio renderings, and not from recordings. Visual representations of the environments, along with the measured room impulse responses, are available on Zenodo (Grimm et al., 2021; Hladek & Seeber, 2022; Schütze et al., 2021).

2.2.1 Living Room

The living room environment (van de Par et al., 2022; Schütze et al., 2021; Schütze, Kirsch, Kollmeier, et al., 2025) comprises a seating area and a television and a coupled kitchen room. The living room had the dimensions 4.97 m x 3.78 m x 2.71 m yielding a volume of 51 m³, and the kitchen had the dimensions

4.97 m x 2.00 m x 2.71 m with a volume of 27 m³. Both rooms were connected by a door. The reverberation time of the combined environment was $T_{30}=0.56$ s.

This environment was tested in two different sound source configurations. Figure 2 provides an overview of the spatial configuration of the sound sources in the living room scene for both a symmetric condition (referred to as *LR_sym*) and an asymmetric condition (referred to as *LR_asym*). In both configurations, the receiver (i.e., listener) was positioned on a sofa at location R1. Both configurations included masker sources in the kitchen room at positions S6, S7, and S8. In the symmetric condition *LR_sym*, further maskers were located at positions S4 and S5 to the left and right of the listener, while the target was placed directly in front of the listener at position S-TV. In the asymmetric condition *LR_asym*, the target source was placed at position S4 and a further masker was placed at position S-TV.

A detailed overview of the maskers is shown in Table A.1 in the Appendix. The maskers at positions S4 and S5 in the *LR_sym* configuration and at position S-TV in the *LR_asym* configuration used male transformed ISTS noise from Schubotz et al. (2016). This noise is based on the female International Speech Test Signal (ISTS) (Holube et al., 2010) and modified using the STRAIGHT algorithm (Kawahara et al., 2008), which included lengthening of the vocal tract and lowering of the fundamental frequency (F0) resulting in a mean F0 of 110 Hz matching that of the original male OLSA target speaker. In both configurations (i.e., *LR_asym* and *LR_sym*), a dishwasher signal was used as the masker at position S6 in the kitchen area. At position S7 and S8, a dialogue spoken by two female voices was played, which was a conversation of two female persons about the similarities and differences of two images (Bitzer et al., 2014).

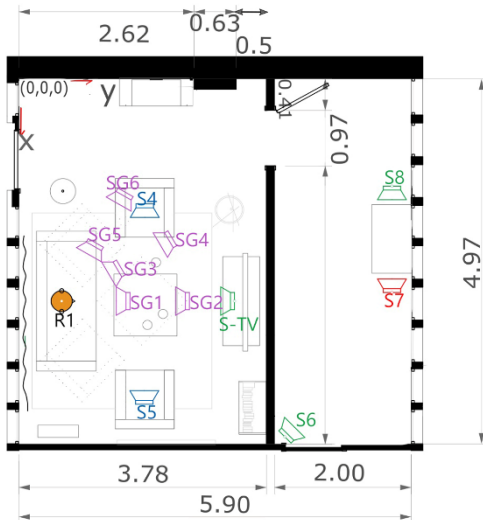


Figure 2: Living room scene map (view from the top). The receiver position R1 was located at the sofa, and the target position was located to the front (S-TV) in the symmetrical configuration *LR_sym*, or to the left of the receiver (S4) in *LR_asym*. Different masker positions are distributed throughout the scene. In both configurations, the maskers in the kitchen room were active (S6, S7, S8). In *LR_sym*, additional maskers were present at S4 and S5, and in *LR_asym*, an additional masker was present at position S-TV [Image is from van de Par et al., 2022].

2.2.2 Pub

The pub environment (Grimm et al., 2021; van de Par et al., 2022) resembles a restaurant setting, featuring different groups of guests engaged in conversations in background music and typical ambient sounds such as food being served. The room had dimensions of 15 m x 10 m x 2.95 m resulting in a volume of 442 m³. The reverberation time was $T_{30}=0.66$ s.

Figure 3 provides an overview of the spatial configuration of the sound sources in the *Pub* environment. The implemented acoustic scene simulates a group conversation at a table. The listener was positioned at receiver location R1 with the target speaker located directly across at position T2, at a distance of 0.97 m. In addition to the target, three interfering talkers were positioned at the same table at locations T1, T3, and T4; an overview of the various maskers in the *Pub* environment is given in Table A.1 in the Appendix. To generate a realistic babble noise background, multiple independent speech signals and conversations were distributed throughout the room (positions S1-S8, N1-N3, P1-P8). For ambient background noise, music was played using the professional audio system (PA). Additional typical pub sounds were incorporated, such as beer pouring (bartender), dishware handling (positions S4, S7), and clinking glasses (positions S2, P8).

The *Pub* environment was presented at a noise level of 74 dBA (see Table 1), which falls within the realistic range for restaurant environment. Hodgson et al. (2007) and Lebo et al. (1994) investigated typical background noise levels in occupied restaurants, reporting values ranging from 55.3-80 dB A.

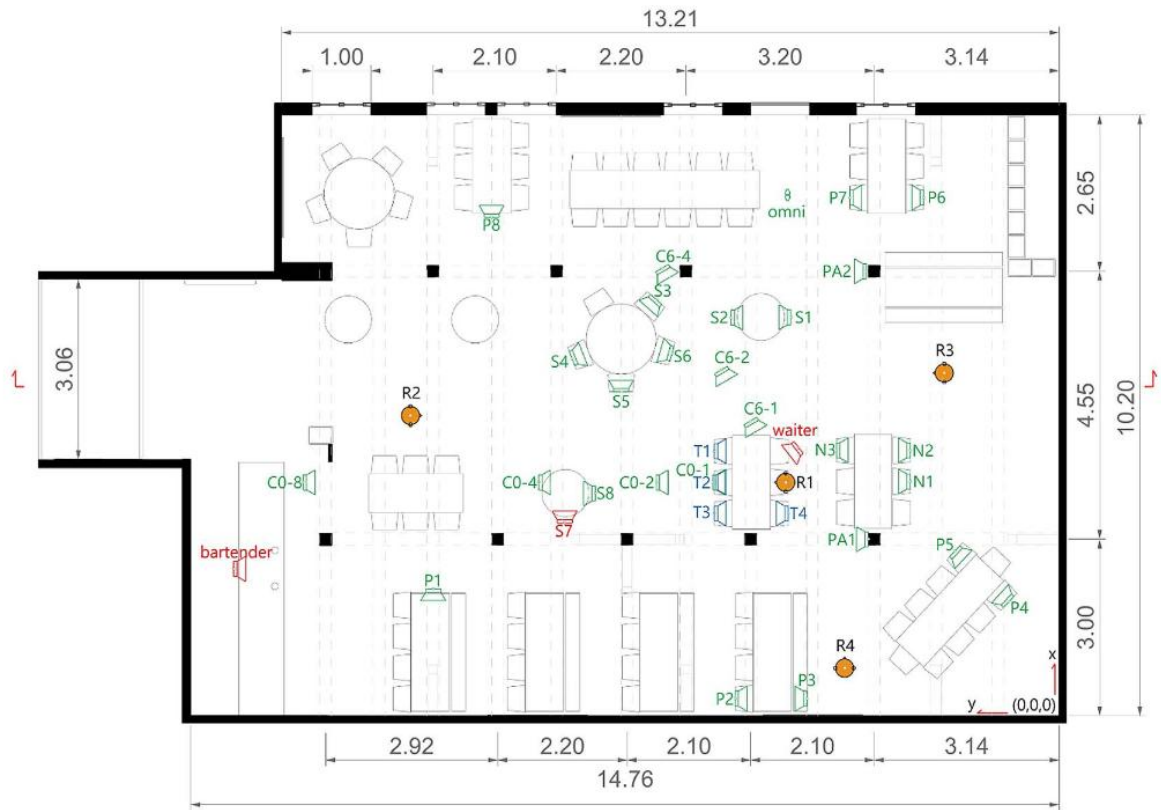


Figure 3: Floorplan of the acoustic scene *Pub*. The receiver position R1 was located at a table and is indicated by an orange head. The target position T2 was straight across of R1, indicated by a green loudspeaker. Three masker positions (T1, T3, and T4) were placed at the same table as the receiver and target. Additional masker positions are distributed throughout the scene [Image adapted from van de Par et al., 2022].

2.2.3 Underground Station

The underground station environment (Hladek & Seeber, 2022; van de Par et al., 2022), referred to as *UG_station*, represents a subway platform. The environment had dimensions of 120.00 m x 15.70 m with a ceiling height of 4.16 m at the platform level and a height of 11-54 m around the escalator area, resulting in a volume of 8555 m³. The reverberation time was $T_{30}=1.68$ s. Figure 4 shows a floorplan of the acoustic scene. The simulated scene resembled a conversation between two people on the platform.

The receiver R1 was located approximately in the middle between both train tracks. The target source was located 1.6 m in front of the receiver, in the direction of gaze. Two masker sources were also placed at a distance of 1.6 m from the receiver, symmetrically positioned at $\pm 60^\circ$ relative to the target. These maskers presented the male transformed ISTS signal, which were temporally shifted with respect to each other. A third masker, located 10.1 m from the receiver, represented the sound of an escalator. Additionally, an ambient noise recorded in the real underground station from Hladek et al. (2021) was incorporated into the scene. A detailed overview of the maskers is shown in Table A.1 in the Appendix.

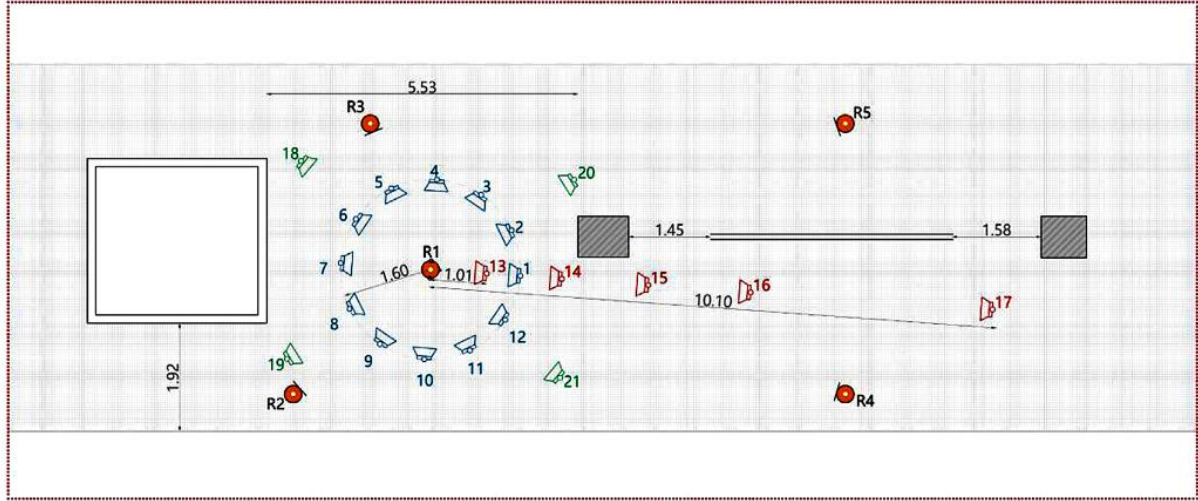


Figure 4: Floorplan of the acoustic scene underground station (UG_station). The receiver position R1 is indicated by a red head. The target position 1 is straight ahead of R1 and indicated by a blue loudspeaker. The two symmetrically positioned maskers 3 and 11 are indicated by blue loudspeaker symbols. The escalator sound source was positioned at the distant red loudspeaker with the number 17 [Image adapted from van de Par et al., 2022].

2.2.4 Acoustic simulation

Virtual acoustic representations of the acoustic scenes used in this study were generated using the Room Acoustics Simulator (RAZR) (Ewert et al., 2025; Kirsch et al., 2023; Wendt et al., 2014). RAZR uses a hybrid approach that combines discrete direct sound and early reflections up to the 3rd order from an image source model (ISM) (Allen & Berkley, 1979) with spatially distributed late reverberation generated by a feedback delay network (FDN) (Jot & Chaigne, 1991). A high degree of perceptual plausibility of the simulated room acoustics is supported by a favorable performance compared to other state-of-the-art methods (e.g., Brinkmann et al., 2019, see algorithm B in their Fig. 8; Starz et al., 2025).

As input to RAZR, geometrical (proxy) shoebox models were created based on the dimensions of the respective environments. The measured reverberation times reported by van de Par et al. (2022) were used to parametrize the frequency-dependent wall absorption filters and the FDN parameters. All audio stimuli of the virtual acoustic scenes in the current experiments were synthesized using RAZR, including background noise and room impulse responses (RIRs), which were later convolved with the target speech. In the *Pub* environment, a finite flat surface was incorporated in the simulation to account for the reflecting properties of the table located in front of the listener position (see also Kirsch et al., 2023; Kirsch & Ewert, 2022, 2024). Similarly, the coffee table in the living room environment was modeled. In the underground station and the living room with the adjacent kitchen, dual slope late reverberation was generated using a second FDN and spatially mapped according to the method described in Kirsch et al. (2023).

2.2.5 Stimulus playback and calibration

The measurements were conducted in a 24-channel horizontal loudspeaker array in a free-field laboratory of the Hörzentrum Oldenburg. The laboratory had dimensions of 5 m x 5.25 m x 2.5 m, and the reverberation time was approximately 0.2 s. The laboratory included carpet on the floor and a cylindrical metal frame covered in dark, heavy fabric reducing ambient impacts such as light and noise. 24 loudspeakers of type Genelec 8030B were arranged equally spaced on a horizontal circle. Starting at 0°, the loudspeakers were positioned every 15° at a radius of 2 m and a tweeter height of 1.25 m. Listeners were seated on a chair in the center of the setup. The measurements did not include any visual stimuli. For reproduction via a loudspeaker array, vector base amplitude panning (Pulkki, 1997) was applied to render 24-channel audio files using RAZR. The simulated RAZR scenes were rendered to 24-channel sound files and were played back at a sampling rate of 44.1 kHz using an RME HDSPe MADI sound card, Ferrofish Pulse 16 converters, and the Toolbox for Acoustic Scene Creation and Rendering (TASCAR) (Grimm et al., 2019) for multichannel playback.

Speech material and noise were calibrated omnidirectionally using a Brüel & Kjær (B&K) 2260 sound level meter at the listener position. In the two standard laboratory conditions, OLnoise was used to calibrate both the speech and masker signals to 65 dB SPL. For the complex acoustic scenes, the speech signal was calibrated with OLnoise convolved with the direct sound component of the respective RIR to match the noise level of the scene (see top row of Table 1), resulting in a broadband SNR of 0 dB. To estimate the direct sound component from the simulated RIR, a von-Hann window was applied, centered around the maximum of the direct sound and ending just before the first reflection. All data presented in this paper are based on the speech levels referenced to this direct-sound calibration. This calibration method was used to improve comparability with anechoic conditions (*SON0* and *SON90* in the present study) in which only the direct sound of the speech target is available by definition. For comparability with other studies using calibration to the reverberated speech target, Table 1 (second row) also provides calculated SNRs obtained for target speech convolved with the full RIR. These recalculations were performed using the speech signal recorded with the omnidirectional microphone at the listener position in the laboratory where the listening experiments were conducted. These recordings are available in Gerken et al. (2025). The resulting SNRs can be interpreted as correction factors that should be applied to the measured SRTs when comparing present results with studies using reverberant target calibration. For example, if the SNR for the speech signal convolved with the full RIR is +1 dB, the corresponding SRTs reported in the present study would need to be shifted by +1 dB to ensure directly comparability with SRTs from studies based on full-RIR calibration.

It should be noted that the temporal window applied to estimate the direct-sound component from the RIR affects its spectrum, depending on the window duration. To estimate the effect of the temporal windowing for speech, speech-weighted SNRs (swSNR) were computed using the method proposed by Greenberg (1993). These are provided in the third row of Table 1. The swSNR accounts for the relative contributions of different regions of the frequency spectrum to speech intelligibility and is thus more informative than broadband SNR, especially with regards to speech intelligibility and different acoustic conditions that differ in spectral properties. To compute it, speech and noise signals were first filtered into octave bands. The SNRs in each frequency band were then weighted according to their contribution to speech intelligibility using the frequency-weighting function from Table 3 of the Speech Intelligibility Index standard (ANSI, 1997). The noise levels including direct sound and reverberation were calibrated separately for every scene to the sound levels presented in Table 1 (top row “speech level (direct sound only) and noise level / dB SPL”). Sound levels for the complex scenes were chosen to reflect realistic conversational environments while maintaining experimental control. For the *LR* scenes, noise levels were defined based on a talker level of 65 dB SPL. *UG_station* was assigned the same overall noise level as *LR_sym* because both scenes were dominated by two ISTS talkers. The *Pub* scene contained three competing talkers in a lively environment that typically induces the Lombard effect, leading to

higher speech levels. Accordingly, the noise level in the *Pub* scene was set 4 dB higher than in the *LR* scenes. The choice of +4 dB is supported by Wagener et al. (2008), who reported comparable increases in conversational levels in the presence of background noise relative to quiet conditions. The selected levels are consistent with ranges reported in the literature for the respective environments (Hodgson et al., 2007; Lebo et al., 1994; Wagener et al., 2008).

Table 1: Speech levels for the target signal convolved with the windowed direct sound and noise levels in dB SPL (resulting in a broadband SNR of 0 dB between the direct sound component of target speech and reverberant noise) for the complex scenes (the first row), calculated broadband SNRs for the speech signal convolved with the full RIR (the second row), and calculated speech-weighted SNR difference between the SNRs in these two cases (the third row).

| Scene | LR_sym | LR_asym | Pub | UG_station |
|---|--------|---------|------|------------|
| Speech level (direct sound only) and noise level / dB SPL | 70 | 67 | 74 | 70 |
| Broadband SNR for noise and speech convolved with the full RIR / dB | 6.9 | 5.4 | 1.6 | 0.7 |
| Speech-weighted SNR difference between speech with the direct sound only and speech with the full RIR / dB | 1.8 | 2.0 | -3.2 | -0.9 |

2.3 Adaptive Categorical Listening Effort Scaling (ACALES)

The Adaptive Categorical Listening Effort Scaling (ACALES) measurement method (Krueger, Schulte, Brand, et al., 2017) allows for measuring the mental effort that a person must exert to understand speech in background noise. This represents an additional dimension beyond speech intelligibility. The listeners rate their subjective listening effort after each presentation of the test signal using a 13-point scale ranging from "no effort" to "extreme effort". In addition, a category labeled "only noise" was provided. Listeners were instructed to select this option if they perceived only noise without any speech signal. The categories are mapped to their numerical representation using the Effort Scale Categorical Unit (ESCU), which ranges from 1 to 14. A value of 1 represents "no effort" and 14 represents "only noise". The scale is anchored at 3 ESCU (very little effort), 5 ESCU (little effort), 7 ESCU (moderate effort), 9 ESCU (considerable effort), 11 ESCU (very much effort), and 13 ESCU (extreme effort). Intermediate even numbers between the anchors were also available to indicate intermediate levels of effort. The numbers of the ESCU were not visible to the listeners.

The test signal consisted of three randomly selected OLSA sentences, which offers a reasonable amount of time to listen to the stimuli and assess the listening effort. Speech material was presented in the same acoustic conditions as for the speech intelligibility measurements, i.e., in stationary speech-shaped noise (OLnoise) at 0° and 90°, and in all three complex environments (living room, pub, and underground station). The configuration of noise and speech sources in complex environments was the same as described in the speech intelligibility measurements. For the hearing aid users, the ACALES measurements were performed unaided and aided.

To familiarize participants with the task, a training session was conducted in the condition that was measured first in the actual measurements. Training data was excluded from the subsequent analyses.

2.4 Loudness scaling

Binaural broadband loudness perception was tested using a categorical loudness scaling according to ISO16832 (Brand & Hohmann, 2002). In this measurement, loudness was rated using the ACALOS scale (Brand & Hohmann, 2002) with 11 categories from "not heard" to "too loud", displayed on a touch screen. After each response, the presentation of the next stimulus started automatically. The verbal scale with categorical units (CUs) was mapped to numerical values between 0 ("not heard") and 50 ("too loud") and plotted as a function of the level of the test signal. A loudness function was fitted to the data according to the BTUX method proposed by Oetting et al. (2014). The fit is described by the parameters m_{low} , m_{high} , and l_{cut} , where m_{low} is the slope below 25 CU, m_{high} is the slope above 25 CU, and l_{cut} is the juncture at 25 CU. During the measurement, a narrowband loudness compensation based on individual hearing thresholds was applied to the stimuli, to compensate for deviations from loudness perception of normal hearing listeners (Suck et al., 2020; Zimmer et al., 2024).

According to the trueLOUDNESS fitting method (Oetting et al., 2018) restoring binaural broadband loudness perception, two test stimuli were investigated in this measurement, including a female speech-shaped noise (IFnoise, Holube et al., 2009) and uniform exciting noise (Fastl & Zwicker, 2006) with a bandwidth of 17 Bark (referred to as UEN17). IFnoise masker was derived from the ISTS (Holube et al., 2008), and its long-term average speech spectrum corresponds to international female speakers (Byrne et al., 1994). The UEN17 masker had a bandwidth of 5100 Hz and a center frequency of 10.5 Bark, corresponding to 1370 Hz. The stimuli were played back without further hearing aid amplification except for individual narrowband loudness compensation gains via Sennheiser HDA 200 headphones.

The procedure includes two measurement phases: the first phase estimated the individual dynamic range by an adaptive measurement interleaved loudness measurement until the responses "not heard" and "too loud" were obtained. The second phase presented stimuli in a random order within the individual dynamic range. The maximum presentation level was 100 eq. dB SPL (equivalent sound pressure level in free field), and stimulus duration was 1 s, with van Hann ramps of 50 ms at the start and end of each stimulus.

The HDA 200 headphones were free-field equalized according to ISO389-8 (2004) using IFnoise as a stimulus. They were calibrated using a Brüel & Kjær artificial ear type 4153, a 0.5-inch microphone type 4134, a microphone preamplifier type 2669, and a measuring amplifier type 2610. An RME Fireface UCX II at 44.1 kHz was used for signal presentation.

2.5 Tone-in-noise detection thresholds

Tone-in-noise detection thresholds were measured for the tone frequencies 0.5 kHz and 2 kHz, separately for the left and right ear using the single interval adjustment matrix (SIAM) (Kaernbach, 1990) procedure. Single intervals with a 50% chance level of including a tone were presented, and for each interval, the listeners were asked to respond if a tone was perceived. It was an adaptive procedure where the tone level was adapted steering towards a percentage of 87.5% correctly responded trials (Schädler et al., 2020), with an initial step size of 8 dB and a decreasing step size after reversals. The step size decreased to 4 dB, 2dB, and 1 dB. Listeners did not get feedback, and the initial tone level was 60 dB SPL. The duration of the tones was 200 ms flanked with 10 ms half-cosine ramps. The masker was a two-octave-wide white noise with a duration of 800 ms, centered on the target frequency (on a log scale), fixed at a level of 30 dB per Hertz. This resulted in a noise level of 49.0 dB SPL and 53.8 dB SPL per equivalent rectangular bandwidth at 500 and 2000 Hz, respectively. To prevent off-frequency tone detection, an f-flanking noise (blue noise, power density increase of 3 dB per octave, lowpass-filtered at the lower cutoff frequency of the bandpass masking noise) and a 1/f-flanking noise (pink noise, power density decrease of 3 dB per octave, highpass-filtered at the upper cutoff frequency of the bandpass making noise) were added to the masker. The spectral density levels of the flanking noises

were set to 6 dB below the level of the two-octave-wide white noise at its lower/upper cutoff frequency. The level of the broadband noise in a frequency range of 20 Hz to 8000 Hz was 69 dB SPL. Tone-in-noise recognition thresholds were calculated as a median of the levels at the reversals (discarding the first four reversals).

Presentation and calibration setups were the same as for the loudness measurements. Narrowband noise stimuli of different frequencies in the range of 123 Hz to 13943 Hz were calibrated to 80 dB SPL. The calibration was then checked for broadband stimuli as used in the measurement.

2.6 HEAR-COMMAND tool

The HEAR-COMMAND tool is a questionnaire that aims to provide a self-reported status of hearing loss, functioning, communication, and conversation disability (Afghah et al., 2022; Alfakir, Dunaway, et al., 2025). The questionnaire was developed and validated in English, German, Korean, and Arabic as a result of an international collaboration of experts in Germany, the United States, Egypt, and the Netherlands (Alfakir, Dunaway, et al., 2025; Alfakir, Hammady, et al., 2025). The tool consists of two groups of items (questions): 1) A set of 30 items inquiring about demographic information and hearing status, and 2) a set of 90 items developed based on utilizing the World Health Organization's International Classification of Functioning Disability and Health (ICF) framework, Core Sets for Hearing Loss (Danermark et al., 2013; Granberg, 2015; WHO, 2001). These items target the ICF domains, including impairments of body functions, difficulties with activities and participation, and facilitators or barriers in the environment. An overview of them is shown in Table A.2 in the Appendix. The response options were provided on a scale of 0 to 4 along with corresponding terminology, as the following example: H.51: "Do you have difficulty with socializing with people living in your community (e.g. classmates, co-workers)?" 0 (No difficulty) / 1 (Mild difficulty) / 2 (Moderate difficulty) / 3 (Severe difficulty) / 4 (Profound/Complete difficulty) / I don't know / Not applicable. As an outcome of the tool validation, ICF-based items were divided into two major groups, and a score was assigned to each group. 1) "Non-hearing-related items", encompassing 41 items that correspond to the evaluation of Interpersonal interaction functionality and infrastructure accessibility", "Social determinants and infrastructure compatibility", "Other sensory integration functionality", and "Cognitive functionality". 2) "Hearing-related items", which includes 37 items directly targeting "Auditory processing functionality", "Sound quality compatibility", and "Listening and communication functionality" (Afghah et al., 2024). To facilitate a direct comparison between the outcomes of laboratory-based speech intelligibility tests and the self-reported speech perception disability, a "Speech perception" score was formulated based on the responses to 32 of the "Hearing-related items" focusing on the perception of a sound in general or speech specifically. To calculate the score, the numeric responses to these 32 items were summed and then multiplied by a weighting coefficient of 0.0782, resulting in a numeric score out of 10, where a score of "0" represents "no" and "10" represents a "profound/complete" self-reported speech perception disability. The participants of this study received the printed German version of the questionnaire by mail at home and either mailed the filled-in questionnaire back or brought it with them during the follow-up visits. Versions in all languages, including German, are available at: <https://www.hz-ol.de/en/open-tools-for-science/hear-command-tool/>.

2.7 Statistical analysis

The experimental data was analyzed statistically using IBM SPSS Statistics. Listener group differences were tested using one-way ANOVAs for age, PTA_{4M}, and PTA_{4M} differences between the ears. Absolute and relative (benefit) speech intelligibility and listening effort data, as well as loudness fit parameter, loudness level, and tone-in-noise detection threshold data were analyzed using a general linear model with a repeated measures design. Bonferroni corrections were applied for multiple comparisons in post-

hoc tests, and a Greenhouse-Geisser correction was applied in cases where sphericity was violated. PTA4_M was used as a covariate, and listener group as a between subject factor. A generalized linear model is suitable for a repeated measures design with a between subject factor and a covariate, which is crucial for the current database. Relationships between different measures were analyzed using Pearson correlation coefficients for comparing two continuous variables: SRTs and SRT benefits across conditions, listening effort compared to speech intelligibility, and tone-in-noise detection thresholds compared to hearing thresholds and speech intelligibility. The relationship between the HEAR-COMMAND tool outcomes and speech intelligibility was analyzed using Spearman's rank correlation coefficient, because the HEAR-COMMAND tool data is ordinal. The ear effect of tone-in-noise detection thresholds was tested using a Wilcoxon rank-sum test due to non-normally distributed data. The HEAR-COMMAND tool outcome was only reported for descriptive statistics, because a more detailed analysis would have been out of the scope of this study.

3. Results

3.1 Speech intelligibility measurements

Measured SRTs in dB SNR are shown in Figure 5 for the different listener groups and different acoustic conditions. As described in Section 2.2.5, speech levels in the SNR calculation refer to the target level calibrated from the direct sound alone, and noise levels refer to the whole signal including reverberation. The results show that SRTs differed across measurement conditions [$F(5, 360)=28.5$, $p<0.001$, $\eta_p^2=0.283$]. Also, a statistically significant interaction between measurement conditions and PTA4_M was found [$F(5, 360)=7.7$, $p<0.001$, $\eta_p^2=0.097$] but not between acoustic condition and group [$F(10, 360)=1.8$, $p=0.064$, $\eta_p^2=0.047$]. Both, PTA4_M and group had a statistically significant effect on SRTs ([$F(1, 72)=69.2$, $p<0.001$, $\eta_p^2=0.049$], [$F(2, 72)=9.0$, $p<0.001$, $\eta_p^2=0.20$], respectively).

Pairwise comparisons across the listener groups (using Bonferroni corrections and PTA4_M as covariate) showed statistically significant differences between NH and HI_noHA ($p=0.03$) but no differences between NH and HI_HAu ($p=0.353$) and between HI_HAu and HI_noHA ($p=0.272$).

For the acoustic conditions, the post-hoc comparisons (using Bonferroni corrections and PTA4_M as covariate) revealed that *S0N0*, *Pub*, and *UG_station* resulted in SRTs that differed significantly from all remaining acoustic conditions (for all $p<0.001$). No statistical differences were found between *S0N90* and *LR_sym* and *LR_asym* as well as between *LR_sym* and *LR_asym*. The remaining comparisons showed statistically significant SRT differences (all with $p<0.001$).

Since the interaction between listener group and measurement condition was not statistically significant, the results can be generalized for all three groups: The lowest SRT was reached for *S0N90*, *LR_sym* and *LR_asym* with a mean SRT across listener groups of -7.5, -7.2, and -7.3 dB SNR, respectively. The *Pub* environment resulted in a mean SRT of -5.0 dB SNR, followed by the *S0N0* condition with a mean SRT of -2.9 dB SNR. The highest SRTs were observed in the *UG_station* environment with a mean SRT of -1.0 dB SNR.

The NH group had the lowest SRTs varying from -5.3 dB SNR for *S0N0* to -12.8 dB SNR for *LR_asym*. *LR_sym*, *Pub* and *UG_station* had mean NH SRTs of -10.7, -8.9, and -5.4 dB SNR, respectively. *S0N90* resulted in a mean NH SRT of -11.4 dB SNR, improving speech intelligibility by 6.1 dB by spatial separation of the noise from the target (compared to *S0N0*).

The mean SRTs for the group HI_noHA were statistically higher compared to the NH group. The differences between the HI_noHA and NH groups varied from 5.2 dB in *LR_asym* to 1.5 dB in *S0N0*. On average, the HI_noHA performed 3 dB worse in terms of SRT, considering all acoustic conditions. The differences could not be explained by the differences in absolute hearing threshold. Speech

intelligibility of HI_HAu showed on average 8 dB higher SRTs than the group of NH, and 5 dB higher SRTs than the group of HI_noHA. These differences, however, could be explained by the differences in average absolute hearing thresholds across the participants.

SRTs obtained with hearing aids (HI_HAa) were in the same range as for HI_noHA. The mean SRT with HA for *SON0* was -3.2 dB and for *SON90* -7.9 dB SNR. The mean SRTs in the complex acoustic scenes ranged between -0.2 (*UG_station*) and -6.4 dB SNR (*LR_asym*). The variability across listeners, as characterized by the standard deviation (see Figure 5), differed between groups. The NH group showed small interindividual differences, with standard deviations ranging from 0.8 dB (*SON0*) to 2.4 dB (*LR_sym*). The variability within the group HI_HAu was at least four times larger in the standard acoustic conditions and at least twice as large in the complex scenes. Hearing aids reduced the variability in the HI_HAu group, resulting in standard deviations that were approximately half the size of those observed without amplification. However, the variability remained higher than that observed in the NH group.

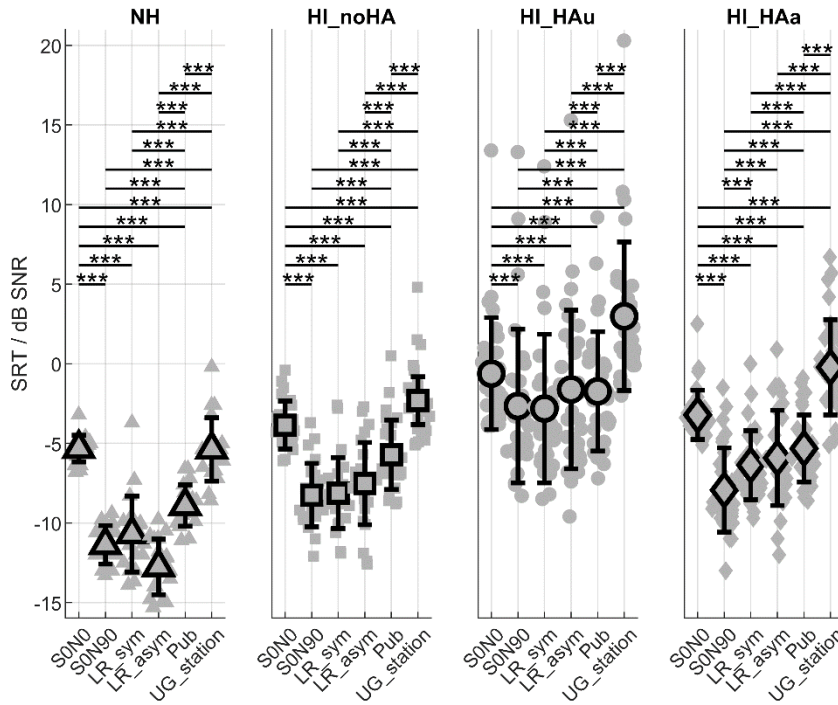


Figure 5: Speech reception thresholds (mean and standard deviation across listeners, and individual data points) for the different acoustic conditions *SON0*, *SON90*, *LR_sym*, *LR_asym*, *Pub* and *UG_station*. From left to right the different listener groups NH, HI_noHA, HI_HAu, HI_HAa are shown. Horizontal lines indicate statistically significant differences across conditions with *** corresponding to $p < 0.001$.

The individual benefit from hearing aids in different acoustic conditions as well as the benefit on the group level are shown in Figure 6. A general linear model revealed statistically significant differences in hearing aid benefit across the various acoustic conditions [$F(5, 150)=8.1$, $p < 0.001$, $\eta_p^2=0.212$]. Post-hoc comparisons with Bonferroni correction showed that the SRT benefit in *SON0* was significantly lower than in *SON90* ($p < 0.001$), *LR_asym* ($p=0.013$), and *Pub* ($p=0.045$). Additionally, the SRT benefit in *SON90* was significantly higher than in the *Pub* ($p=0.02$) and *UG_station* ($p=0.03$) conditions. No other pairwise comparisons showed statistically significant differences. The *SON0* condition resulted in a mean benefit of 2.6 dB, whereas the highest benefit (5.3 dB) was obtained in the *SON90* condition.

The benefit observed in the complex acoustic scenes fell between these two values (Δ SRT from 3.2 to 4.3 dB).

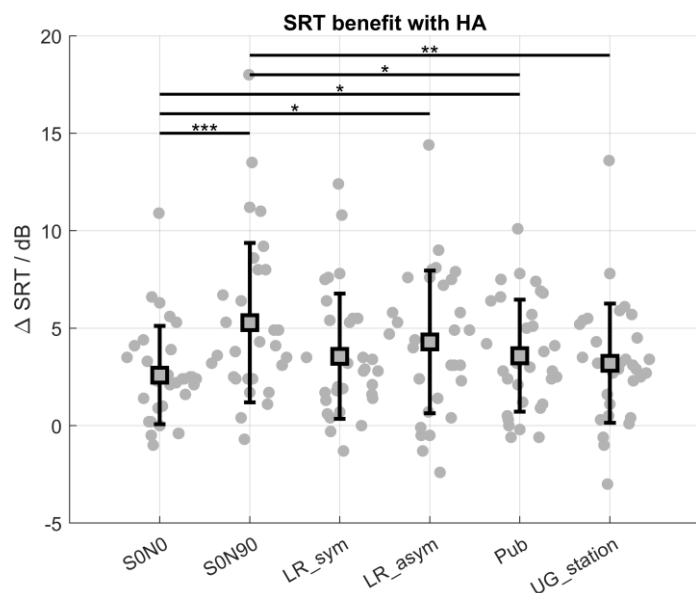


Figure 6: Benefit (mean and standard deviation across listeners, and individual data points) for HA users in the different acoustic conditions S0N0, S0N90, LR_sym, LR_asym, Pub and UG_station, calculated as the difference (Δ SRT) between the measurement with and without HA. Significance of pairwise comparisons is indicated by horizontal lines as follows: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

To examine the relationships of SRTs across conditions and of SRT benefits across conditions, linear regression analyses were performed. These analyses allowed to assess the extent to which performance in one condition predicts performance in other conditions, including whether SRTs and SRT benefits measured in a standard laboratory setting generalize to more complex experimental environments.

Table 2 summarizes the Pearson correlation coefficients (r) for all comparisons. The strength of the linear relationships between variables was categorized according to Evans (1996) as follows: 0.00-0.19 = very weak, 0.20-0.39 = weak, 0.40-0.59 = moderate, 0.60-0.79 = strong, and 0.80-1.00 = very strong. These correlations provide insight into the consistency and generalizability of response patterns across experimental conditions.

Table 2: Below the diagonal: correlation coefficients for SRT comparisons across different acoustic conditions; above the diagonal: correlation coefficients for SRT benefit from hearing aids comparisons across different acoustic conditions.

| | S0N0 | S0N90 | LR_sym | LR_asym | Pub | UG_station |
|------------|------|-------|--------|---------|------|------------|
| S0N0 | 1.00 | 0.68 | 0.75 | 0.71 | 0.81 | 0.75 |
| S0N90 | 0.88 | 1.00 | 0.65 | 0.71 | 0.76 | 0.67 |
| LR_sym | 0.90 | 0.92 | 1.00 | 0.67 | 0.71 | 0.60 |
| LR_asym | 0.88 | 0.93 | 0.92 | 1.00 | 0.80 | 0.69 |
| Pub | 0.89 | 0.89 | 0.91 | 0.92 | 1.00 | 0.76 |
| UG_station | 0.91 | 0.90 | 0.91 | 0.93 | 0.92 | 1.00 |

SRTs measured in the standard laboratory condition (S0N0) were very strongly correlated with SRTs in all complex acoustic conditions ($r = 0.88$ - 0.91). A similar pattern was observed for SRTs in the spatially separated standard condition (S0N90) ($r = 0.89$ - 0.93). These consistently high correlations suggest that both S0N0 and S0N90 provide robust predictive strength, with only minor variation depending on the

spatial complexity of the acoustic scene. For *SON0*, the highest correlation was observed for the *LR_sym* and *UG_station* conditions ($r = 0.90$ and 0.91 , respectively), while the *LR_asym* condition showed slightly lower, though still very strong, correlation ($r = 0.88$). Correlations among the complex acoustic conditions themselves were also very strong ($r = 0.91$ - 0.93), indicating a high degree of consistency in SRT performance across different complex listening environments.

When considering SRT benefits from hearing aids in the standard condition (*SON0*), correlations with the more complex acoustic conditions ranged from strong to very strong ($r = 0.68$ - 0.81), suggesting that the relative improvement observed in the simplest condition provides a useful, though not perfect, indication of benefit in more challenging environments. Across the complex conditions, the strongest associations were observed between the *LR_asym* and *Pub* conditions ($r = 0.80$), while the *LR_sym* condition exhibited the lowest, yet still moderate, correlation with other complex scenarios ($r = 0.60$ - 0.67). These findings indicate that, while the absolute magnitude of SRT benefit may differ depending on the acoustic scene, individuals who gain more benefit in one complex condition tend to gain more benefit in other complex conditions as well.

3.2 Adaptive Categorical Listening Effort Scaling (ACALES)

ACALES outcomes are analyzed by comparisons of three ESCUs values, namely 1, 7, and 13, corresponding to the categories of low, moderate, and extreme effort, respectively. Corresponding SNRs for all listeners groups and all conditions are shown in Figure 7.

Listening effort data were statistically compared using a general linear model with individual SNRs corresponding to the ESCU of 1, 7, and 13, as well as acoustic conditions as within-subject factors, listener group as a between-subject factor and PTA_{4M} as a covariate. The analysis revealed significant main effects of both acoustic condition [$F(4.15, 257.1)=6.99, p<0.001, \eta_p^2=0.101$] and ESCU [$F(1.22, 75.87)=50.32, p<0.001, \eta_p^2=0.448$] indicating that both factors independently influenced the outcome measure. However, no significant interactions were found between these factors ($p=0.397, \eta_p^2=0.017$) or with the group ($p=0.901, \eta_p^2=0.014$) and PTA_{4M} ($p=0.252, \eta_p^2=0.021$) suggesting that their effects were additive rather than interactive. The group factor approached significance [$F(2, 62)=3.10, p=0.052, \eta_p^2=0.091$], suggesting that PTA_{4M} cannot fully explain the variability across the groups. Significant pairwise differences were found between most acoustic condition combinations: The *LR_sym* condition consistently showed lowest SNRs, significantly lower than all other conditions (all $p<0.001$) suggesting that this environment is the least effortful. Contrary to that, the *UG_station* resulted in significantly higher listening effort than every other acoustic condition (all $p<0.001$). The standard condition *SON0* also differed significantly from all other acoustic conditions (all $p<0.001$) showing higher effort than all conditions except *UG_station*, but lower effort than *UG_station*. *SON90* showed comparable listening effort to *LR_asym* ($p=0.391$) and *Pub* ($p=0.085$) and differed significantly from the remaining environments (all $p<0.001$). A significant difference was also found between the *Pub* and *LR_asym* ($p<0.001$).

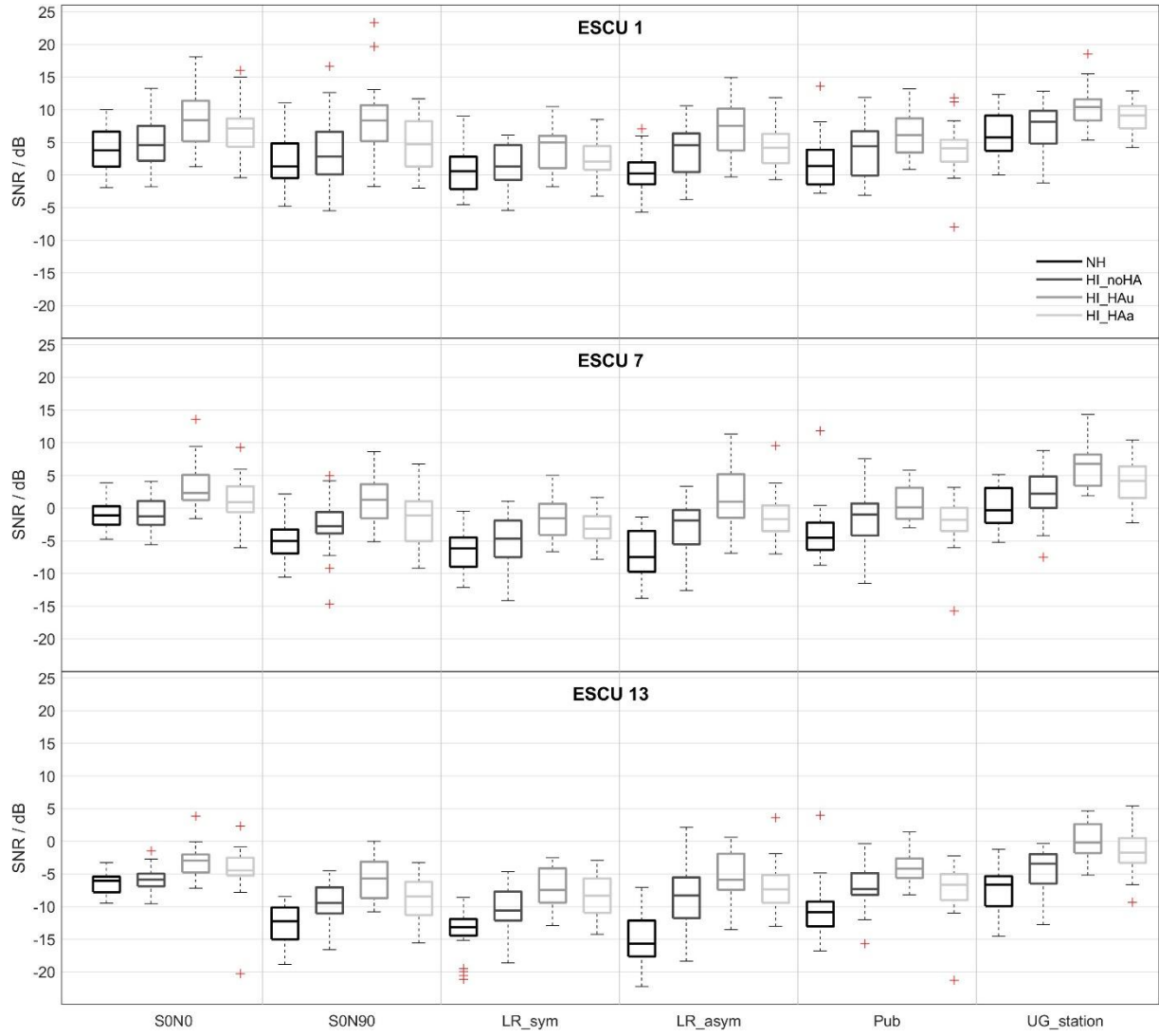


Figure 7: Grouped boxplots of SNR for ACALES test across six listening conditions for four listener groups (NH, HI_noHA, HI_HAu, HI_HAa) and ESCU values of 1, 7, and 13 corresponding to the categories no effort, moderate effort, and extreme effort, respectively. Boxes represent the interquartile range, horizontal lines indicate the median, whiskers show the range of non-outlier data, and plus symbols (+) denote outliers.

For hearing aid users, the benefit in terms of listening effort was calculated and analyzed for each listener and acoustic condition as shown in Figure 8. A general linear model for repeated measures was applied with two within-subject factors: acoustic condition and ESCU. Due to violations of the sphericity assumption for condition and interaction between condition and ESCU factors, degrees of freedom were corrected using the Greenhouse-Geisser method. The analysis revealed no statistically significant main effects or interactions on listening effort benefit. Specifically, neither the acoustic condition [$F(2.30, 41.5) = 0.560$, $\eta_p^2 = 0.034$], the ESCU factor [$F(2, 36) = 0.532$, $p = 0.592$, $\eta_p^2 = 0.029$], nor their interaction [$F(2.579, 49.4) = 1.84$, $p = 0.16$, $\eta_p^2 = 0.093$] reached statistical significance. As a result, listening effort benefits were averaged across listeners, conditions, and ESCUs. The mean benefit (averaged across ESCUs) ranged from 1.1 dB for *LR_sym* to 3.0 dB for *Pub*, with an overall average benefit of 2.1 dB across all acoustic environments.

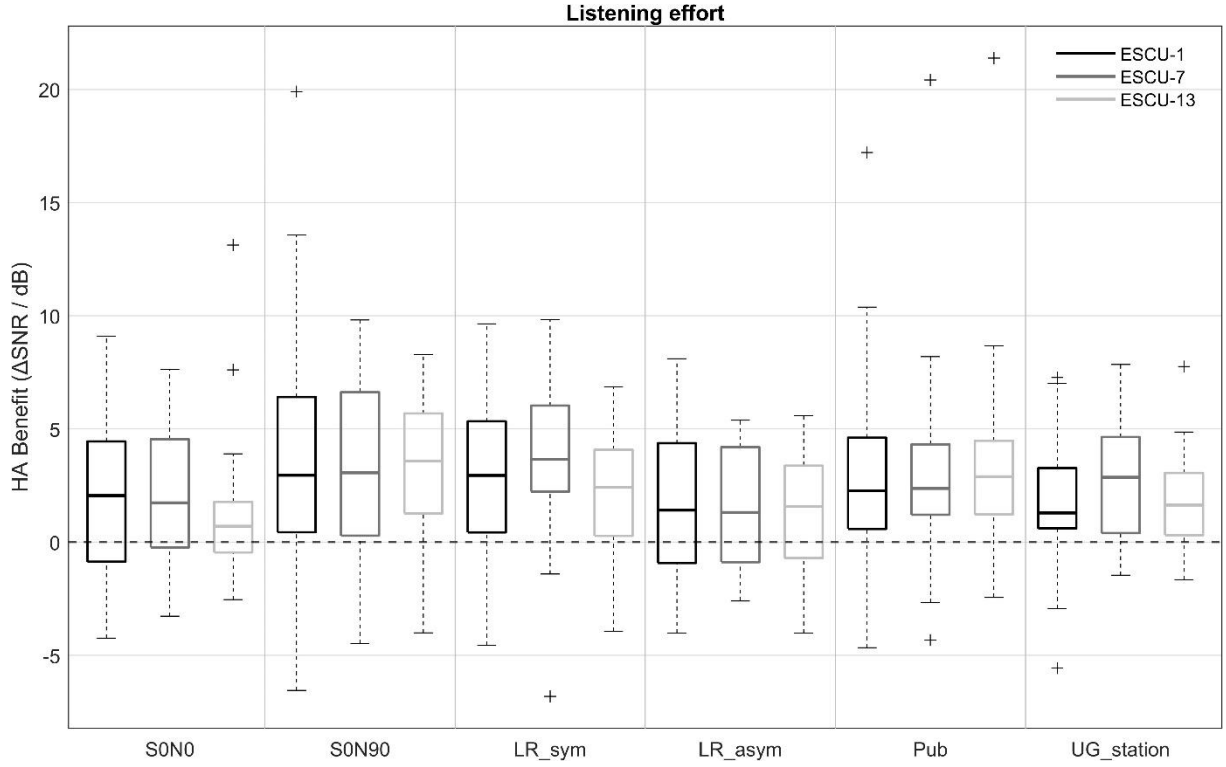


Figure 8: Box plots of SNR benefit in terms of listening effort across six listening conditions for ESCU values of 1, 7, and 13 corresponding to the categories of no effort, moderate effort, and extreme effort, respectively. Boxes represent the interquartile range, horizontal lines indicate the median, whiskers show the range of non-outlier data, and plus symbols (+) denote outliers.

3.3 Loudness scaling

Figure 9 shows loudness functions after narrowband loudness compensation. The functions were fitted with the procedure suggested by Brand & Hohmann (2002) for the signals IFNoise (left figure) and UEN17 (right figure). For each listener group and test signal, the loudness function for the median fit parameters across listeners was calculated. The shaded area visualizes the interquartile range across fit functions of the individual listeners per listener group and test signal.

The loudness fit parameters m_{low} , m_{high} , and l_{cut} have been originally derived from the fitted loudness functions. For each fit parameter, a separate general linear model with noise type as a within-subject factor, listener group as a between-subject factor, and PTA4_M as a covariate was conducted to statistically analyze the fit parameters across noises and listener groups.

For the slope below 25 CU, m_{low} , there was no significant effect of noise type [$F(1, 72)=0.001$, $p=0.981$, $\eta_p^2<0.001$], listener group [$F(2, 72)=0.18$, $p=0.834$, $\eta_p^2=0.005$], or the covariate PTA4_M [$F(1, 72)=0.14$, $p=0.713$, $\eta_p^2=0.002$]. None of the interactions were significant: noise * PTA4_M [$F(1, 72)=0.06$, $p=0.803$, $\eta_p^2=0.001$] and noise * listener group [$F(2, 72)=0.68$, $p=0.511$, $\eta_p^2=0.018$].

For the slope above 25 CU, m_{high} , there was a significant effect of the covariate PTA4_M [$F(1, 72)=4.17$, $p=0.045$, $\eta_p^2=0.055$]. There was no significant effect of noise type [$F(1, 72)=0.39$, $p=0.535$, $\eta_p^2=0.005$] or listener group [$F(2, 72)=2.32$, $p=0.106$, $\eta_p^2=0.060$]. None of the interactions were significant: noise * PTA4_M [$F(1, 72)=0.38$, $p=0.539$, $\eta_p^2=0.005$] and noise * listener group [$F(2, 72)=0.15$, $p=0.858$, $\eta_p^2=0.004$].

For the juncture at 25 CU, l_{cut} , there was a significant effect of the noise type [$F(1, 72)=5.23$, $p=0.025$, $\eta_p^2=0.068$]. The remaining parameters were not significant: listener group [$F(2, 72)=2.04$, $p=0.137$, $\eta_p^2=0.054$] and covariate PTA4_M [$F(1, 72)=0.27$, $p=0.608$, $\eta_p^2=0.004$]. None of the interactions were

significant: noise * PTA4M [$F(1, 72)=2.26, p=0.137, \eta_p^2=0.030$] and noise * listener group [$F(2, 72)=1.72, p=0.187, \eta_p^2=0.046$].

Across test signals, loudness functions started at approximately -5 to 0 dB SPL and reached 50 CU at levels of 85-95 dB SPL. At low signal levels, loudness was similar across listener groups, with increasing differences toward higher levels. Hearing-impaired listeners showed steeper loudness growth than normal-hearing listeners, and the HI_HAu group tended to report higher loudness than the HI_noHA group. Interquartile ranges indicated large interindividual variability. For low to intermediate levels (≤ 60 dB SPL), the slope of loudness growth was flatter than for higher levels. Across all groups, UEN17 elicited higher loudness ratings than IFNoise, with differences of up to approximately 2.5 CU at high levels.

Figure 10 shows the input sound level for the reported loudness levels of 15, 25, and 35 CU, which were calculated for each listener using their individual fitting parameters. These input sound levels are presented for the two test signals IFNoise (left figure) and UEN17 (right figure), and with the individual listeners assigned to the groups NH, HI_noHA, and HI_HAu. The mean and standard deviation are calculated for each listener group.

Similar statistical analyses as shown in Figure 9 were conducted for the data in Figure 10. In this analysis, the input levels corresponding to loudness levels of 15, 25, and 35 CU, as well as noise type, were considered as within-subject factors. Listener group was included as a between-subject factor and PTA4M as a covariate.

Perceived loudness differed significantly across noise types [$F(1, 71)=10.5, p=0.002, \eta_p^2=0.123$] and across CU values [$F(1.191, 84.5)=111.85, p<0.001, \eta_p^2=0.612$]. The interaction between noise and CU values was not significant [$F(1.36, 96.22)=0.2, p=0.817, \eta_p^2=0.003$]. The main effect of listener group was not statistically significant after controlling for PTA4M [$F(2, 71)=2.89, p=0.062, \eta_p^2=0.075$]. Pairwise comparisons with Bonferroni correction showed significant differences between all CU values ($p<0.001$). 15 CU yielded significantly lower mean values than 25 CU and 35 CU, and 25 CU yielded significantly lower mean values than 35 CU. All differences were large in magnitude, with mean differences ranging from 15.6 dB to 38.1 dB (between 15 CU and 35 CU).

Across all groups and signals, input level increased with CU. The same CU values were reached at lower input levels for UEN17 compared to IFNoise. Hearing-impaired listeners with and without hearing aids showed similar input levels. Interindividual variability in loudness perception was large.

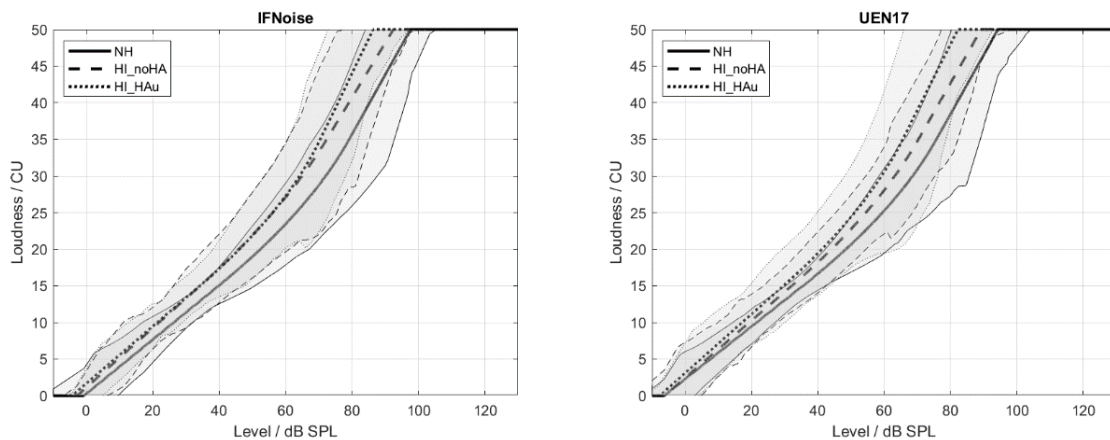


Figure 9: Loudness functions in CU of median fit parameters and interquartile range across loudness functions of the groups NH, HI_noHA, and HI_HAu as a function of the signal level in dB SPL for the signals IFNoise (left) and UEN17 (right), both with individual narrowband loudness compensation applied.

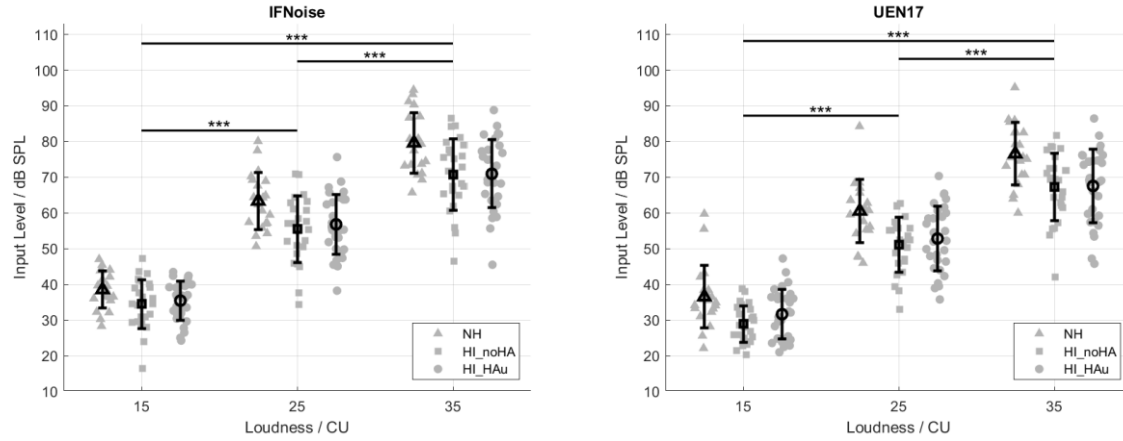


Figure 10: Input level in dB SPL as a function of the loudness at 15 CU (soft), 25 CU (medium), and 35 CU (loud) for the listener groups NH (triangles), HI_noHA (squares), HI_HAu (circles) for the test signals IFNoise (left) and UEN17 (right), both including individual narrowband loudness compensation. The graph shows the mean and standard deviation of the input levels along with individual listeners' levels. The horizontal lines indicate statistically significant differences between the CU groups with $p < 0.001$.

3.4 Tone-in-noise detection thresholds

Tone-in-noise detection thresholds were analyzed based on mean values over the left and right ear for each listener and test frequency. Prior to taking the mean, it was analyzed whether the ear had a significant effect on the outcome of the detection thresholds. Because the data were not normally distributed, a Wilcoxon rank-sum test was applied to test the significance of the ear at a level of $\alpha = 0.05$. The ear had no significant effect ($p=0.84$), so further analysis was computed with the mean values over the left and right ear.

Figure 11 shows tone-in-noise detection thresholds averaged across ears as a function of tone frequency for the different listener groups. Thresholds were statistically compared using a general linear model with tone frequency as a within-subject factor, listener group as a between-subject factor and PTA_{4M} as a covariate.

The analysis revealed a significant main effect of the listener group [$F(2, 68)=8.0, p<0.001, \eta_p^2=0.190$]. Descriptively, the detection thresholds for groups NH and HI_noHA were similar, whereas the HI_HAu group showed higher thresholds. However, Bonferroni-corrected post-hoc tests did not indicate significant differences between any pair of groups (all $p>0.05$). In the descriptive statistics, the mean values of the detection thresholds were higher at 2000 Hz compared to 500 Hz. At 500 Hz, thresholds were centered around 54 dB for NH and HI_noHA, and 58 dB for HI_HAu. At 2000 Hz, mean thresholds were approximately 60 dB for NH and HI_noHA, and 67 dB for HI_HAu. Interindividual variability increased across groups, from NH to HI_noHA and HI_HAu. For HI_HAu, thresholds ranged from 52 to 80 dB, while for NH they ranged from 52 to 58 dB at 500 Hz and from 54 to 74 dB at 2000 Hz. This effect of tone frequency was not significant [$F(1, 68)=0.002, p=0.966, \eta_p^2<0.001$]. The interactions tone frequency * PTA_{4M} [$F(1, 68)=2.3, p=0.136, \eta_p^2=0.032$] and tone frequency * listener group [$F(2, 68)=0.5, p=0.607, \eta_p^2=0.015$] were not significant.

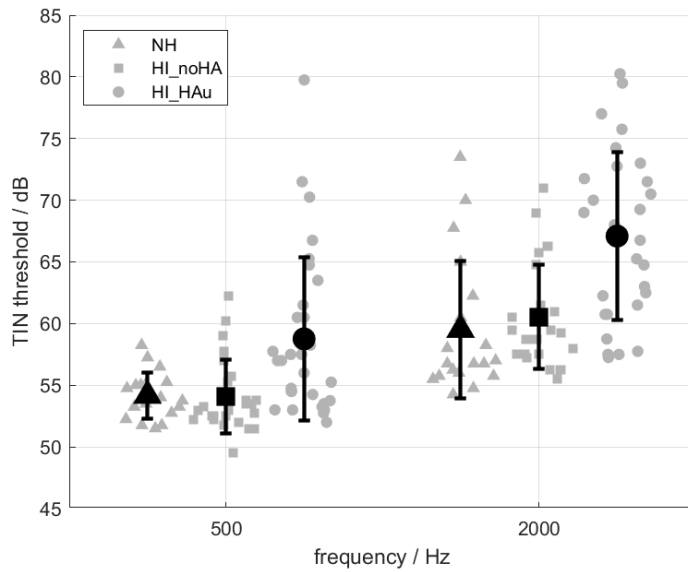


Figure 11: Tone-in-noise detection thresholds in dB as a function of the tone frequency in Hz. The graph shows the mean and standard deviation of the thresholds averaged over the left and right ear for the groups NH, HI_noHA, and HI_HAu along with individual listeners' thresholds. There were no statistically significant differences with $p < 0.05$.

3.5 HEAR-COMMAND tool outcome

Figure 12 shows the distribution of responses to the ICF-based items across all participants. Among the 90 ICF-based items, items H.41 to H.48, which assess speech production ability, are excluded from the figure, as they were applicable to only two participants. Following the ICF framework, environmental factors were categorized as facilitators and barriers. Across participants, most domains show low median scores, indicating generally low perceived effort or high functionality. In the interpersonal interactions domain, 12 out of 18 items had a median of 0, indicating that most participants did not report limitations in social relationships. Auditory processing, listening and communication, and cognitive functionality display some variability, with occasional higher scores reflecting increased effort for certain individuals. The domains hearing aid benefits and social determinants showed high median scores indicating strong satisfaction with support from services, systems, and devices. Items in the sound quality barriers domain had medians ranging from 2 to 3 indicating that, on average, participants perceived sound quality-related aspects as moderate barriers to effective listening. Variability and outliers were observed in nearly all domains and categories, suggesting that a subset of participants experienced substantially greater difficulties. Overall, the responses highlight specific challenges in certain functional areas, despite generally low effort ratings across most items.

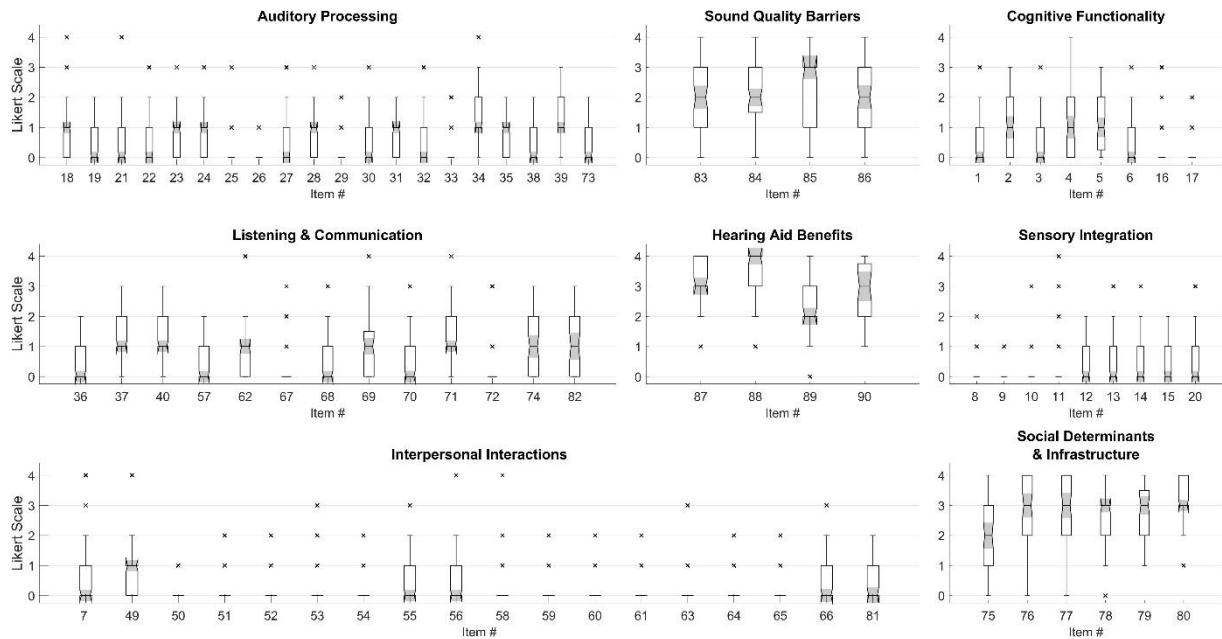


Figure 12: Overall distribution of participants' responses to the ICF-based items of the HEAR-COMMAND Tool. The X-axis shows the item number, and the Y-axis shows the Likert scale (Likert, 1932) from "0" (corresponding to "No") to "4" (corresponding to "profound/complete"). Items are categorized based on the construct factor addressed. For the complete list of items, refer to Appendix, Table A.2.

The descriptive statistics of speech perception scores (formulated based on the responses to 32 of the "Hearing-related items" focusing on the perception of a sound in general or speech specifically) with differentiation across the listener groups are given in Table 3. Speech perception scores were lowest in NH listeners and progressively higher in HI groups, with the highest scores observed in hearing aid users. Since dispersion and tail behavior clearly differ by hearing status (SD approximately doubles from NH (0.61) to HI_noHA (1.13) and HI_HAa (1.17), and the upper range extends from 2.10 (NH) to 4.37 (HI_noHA) and 5.62 (HI_HAa), full distributions of responses to speech perception-related items, categorized based on the participants' hearing status, distinguishing between NH, HI_noHA, and HI_HAa are given in Figure 13. Across individual items, NH participants consistently showed low ratings with narrow distributions, indicating minimal perceived difficulties. In contrast, HI_noHA reported higher and more variable scores across most items, suggesting greater heterogeneity of perceived disability. HI_HAa generally exhibited the highest median and spread of scores, particularly in Sound Quality Compatibility and Listening and Communication Functionality domains, with several items showing extended upper ranges and outliers, reflecting subgroups experiencing pronounced difficulties. Overall, the distributions highlight that while group means differ, the variability and skewness within HI groups provide crucial information about the diversity of individual experiences.

Table 3: Descriptive statistics of speech perception scores from the HEAR-COMMAND Tool across listener groups. The score is designed such that a score of "0" represents no and "10" represents a profound/complete self-reported speech perception disability. For details regarding the scoring method, see Afghah et al. (2024).

| Group | Median | Mean (\pm standard deviation) | Range |
|---------|--------|----------------------------------|-----------|
| NH | 1.17 | 1.20 \pm 0.61 | 0.07-2.10 |
| HI_noHA | 1.64 | 1.94 \pm 1.13 | 0.15-4.37 |
| HI_HAa | 2.58 | 2.81 \pm 1.17 | 1.32-5.62 |

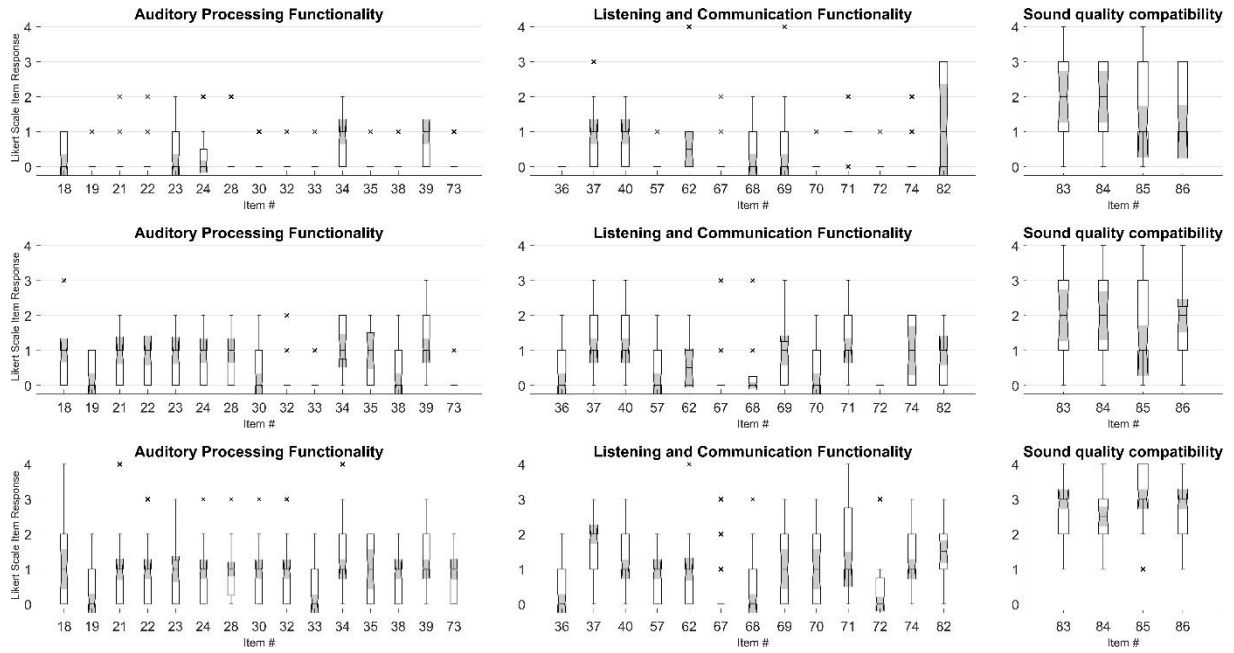


Figure 13: Distribution of responses to speech perception-related items of the HEAR-COMMAND Tool, categorized based on the participants' hearing status. Top: NH, Middle: HI_noHA, Bottom: HI_HAa. The X-axis shows the item number, and the Y-axis shows the Likert scale (Likert, 1932) from "0" (corresponding to "No") to "4" (corresponding to "profound/complete"). For the complete list of items, refer to Appendix, Table A.2.

3.6 Comparisons across different measures

To gain a comprehensive understanding of how different aspects of hearing relate to speech intelligibility, we examined associations between speech intelligibility and several complementary measures. Listening effort was compared to speech intelligibility to explore how the perceived task difficulty corresponds to actual speech intelligibility performance. Tone-in-noise detection thresholds were analyzed in relation to speech intelligibility to assess how suprathreshold tone detection abilities at low and high frequencies contribute to understanding speech in noise. In addition, self-reported hearing and communication abilities from the HEAR-COMMAND tool were compared with the auditory measures to evaluate how subjective experiences correspond to behavioral performance. Together, these comparisons provide insights into the multidimensional nature of hearing and the extent to which different measures capture overlapping or distinct aspects of auditory functioning.

3.6.1 Listening effort and speech intelligibility

To quantitatively compare listening effort and speech intelligibility, the relationship between SRTs and ACALES SNRs corresponding to an ESCU of 7 (moderate effort) was analyzed. Correlation analyses were conducted separately for each acoustic condition to examine whether the association between intelligibility and effort varied across environments. Comparisons were performed both at the level of absolute values (SRTs vs. SNRs at ESCU 7) and for derived benefit measures, where hearing aid benefit was expressed as the difference between aided and unaided values for each outcome (Δ SRT and Δ SNR). For the unaided analyses, the data of all listeners across groups were pooled, and the absolute aided and benefit analyses involved all hearing aid users. The benefit analysis allowed to evaluate whether amplification affected speech recognition and perceived effort in a comparable way across acoustic conditions. Table 4 summarizes the results, with Pearson correlation coefficients (r) for the absolute values and for the benefit comparisons in each acoustic condition.

Correlation analyses revealed a generally moderate to strong relationship between SRTs and ACALES SNRs at ESCU 7 across acoustic conditions. In the unaided condition, correlations ranged from $r = 0.54$ to $r = 0.74$ (all $p < 0.001$), indicating that listeners with poorer speech intelligibility (higher SRTs) tended

to reach moderate listening effort at more favorable SNRs (higher ACALES SNRs), reflecting a consistent association between speech recognition difficulty and perceived effort. Aided correlations were slightly lower ($r = 0.40$ - 0.57), suggesting that amplification generally improved speech perception but did not fully alter the relationship with perceived effort. Analyses of hearing aid benefit (Δ SRT vs. Δ SNR) showed moderate to strong correlations in most conditions ($r = 0.50$ – 0.68), with the strongest effects in spatially asymmetric conditions (*LR_asym*, *S0N90*) and weak relation in highly reverberant setting (*UG_station*, $r = 0.33$, $p=0.079$). Overall, these findings indicate that hearing aid amplification generally improves speech intelligibility and reduces listening effort, though the correspondence between the benefit in both domains varies across acoustic environments.

Table 4: Pearson correlation coefficients (r) and corresponding p -values between SRTs obtained from OLSA measurements and listening effort ratings at ESCU 7 for unaided, aided, and benefit data, across different conditions.

| Acoustic condition | Absolute correlation unaided (r) | Absolute correlation aided (r) | Correlation benefit from HA (r) |
|--------------------|--------------------------------------|------------------------------------|-------------------------------------|
| S0N0 | 0.68 ($p<0.001$) | 0.51 ($p=0.004$) | 0.50 ($p=0.006$) |
| S0N90 | 0.63 ($p<0.001$) | 0.51 ($p=0.003$) | 0.64 ($p<0.001$) |
| LR_sym | 0.65 ($p<0.001$) | 0.41 ($p=0.026$) | 0.54 ($p=0.007$) |
| LR_asym | 0.74 ($p<0.001$) | 0.57 ($p<0.001$) | 0.68 ($p<0.001$) |
| Pub | 0.54 ($p<0.001$) | 0.40 ($p=0.028$) | 0.54 ($p=0.003$) |
| UG_station | 0.70 ($p<0.001$) | 0.48 ($p=0.006$) | 0.33 ($p=0.079$) |

3.6.2 Tone-in-noise detection thresholds, hearing thresholds, and speech intelligibility

Tone-in-noise detection thresholds were compared to other measures (hearing thresholds, SRTs) in terms of the Pearson correlation coefficient r . To test if the tone-in-noise detection thresholds can be treated frequency-independent by using the mean over both frequencies 500 Hz and 2000 Hz, the correlation across both frequencies was tested. Their correlation was weak ($r=0.22$), so further correlation analyses were conducted frequency-dependent.

For the correlation analyses, the unaided data of all listeners across groups were pooled. The Pearson correlation coefficients r between frequency-dependent tone-in-noise detection thresholds and frequency-dependent hearing thresholds were in the range of $r=0.05$ - 0.67 , so there was a very weak to strong correlation. Figure 14 (left) shows a map of the coefficients of determination between the detection thresholds of separate frequencies. It was observed that there was a stronger relationship between tone-in-noise detection thresholds at low frequencies with hearing thresholds at low frequencies than with hearing thresholds at high frequencies. Vice versa, tone-in-noise detection thresholds at high frequencies correlated stronger with hearing thresholds at high frequencies. Generally, the detection thresholds cannot be inferred from each other, and tone-in-noise detection thresholds provide a large spread across listeners.

The relation between tone-in-noise detection thresholds and SRTs is shown in Figure 14 (right), in the same way as described for Figure 14 (left). This relation was in a similar range as with hearing thresholds, with coefficients of determination in the range of $r = 0.16$ - 0.64 . While the tone-in-noise detection thresholds at 500 Hz did not show a relation with SRTs, there was a much stronger relation for tone-in-noise detection thresholds at 2000 Hz. There was no clear pattern visible across different types of SRT conditions.

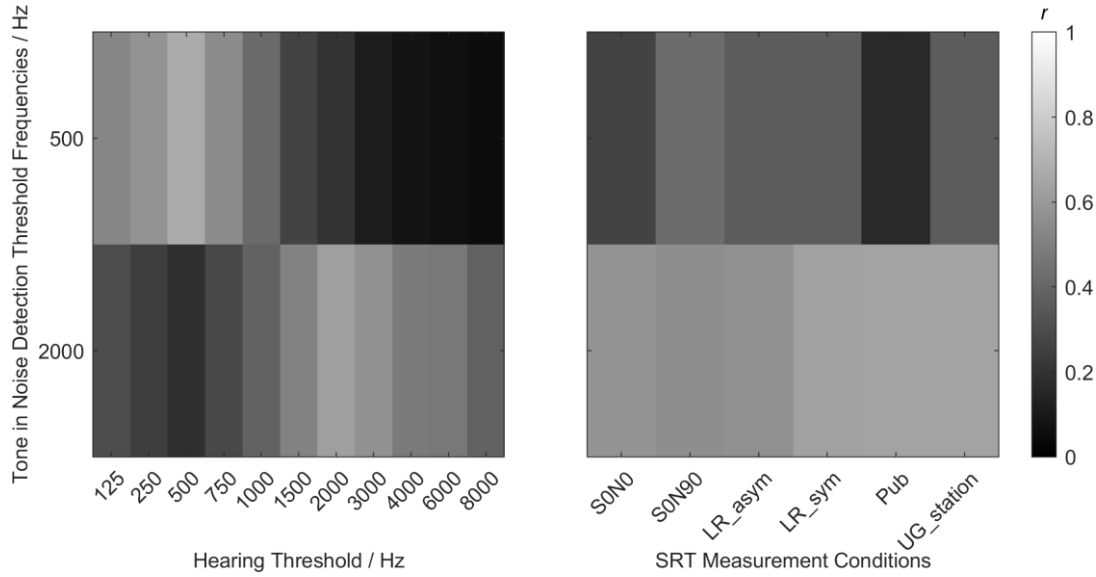


Figure 14: Heatmaps showing the Pearson correlation coefficient r between the tone-in-noise detection thresholds at 500 Hz and 2000 Hz and the frequency-dependent hearing thresholds (left figure), and between the tone-in-noise detection thresholds at 500 Hz and 2000 Hz and the speech reception thresholds in different conditions (right figure)

3.6.3 HEAR-COMMAND tool and speech intelligibility

The speech perception scores from the HEAR-COMMAND tool were compared to the SRT values obtained in speech intelligibility measurements across all acoustic conditions to evaluate their mutual consistency and potential complementarity. The correlation analyses were performed separately for the different listener groups, because there are no unaided HEAR-COMMAND tool results available for the hearing aid users. Table 5 presents the corresponding Spearman's rank correlation coefficients (r) (Spearman, 1961). For the HI_noHA group, moderate to strong correlations ($r = 0.42$ - 0.78) were observed across acoustic scenes reaching statistical significance in four conditions (*LR_asym*, *LR_sym*, *Pub*, and *UG_station*). For the NH and HI_HAa groups, correlations were weaker ($|r| \leq 0.39$ and $|r| \leq 0.33$, respectively) and did not reach statistical significance. This indicates that the relation between HEAR-COMMAND tool scores and SRTs is less pronounced for listeners with normal hearing and for hearing-impaired listeners with hearing aids. Findings highlight the potential utility of the HEAR-COMMAND tool as a complementary measure to traditional speech tests, especially in characterizing the unaided functional limitations of hearing-impaired listeners across complex acoustic environments.

Table 5: Spearman's rank correlation coefficients (r) between speech perception scores from the HEAR-COMMAND tool and SRT values obtained from OLSA measurements across different conditions.

| Condition | NH | HI_noHA | HI_HAa |
|------------|-----------------------|----------------------|----------------------|
| S0N0 | 0.25 ($p = 0.313$) | 0.55 ($p = 0.063$) | 0.22 ($p = 0.248$) |
| S0N90 | -0.39 ($p = 0.102$) | 0.42 ($p = 0.082$) | 0.24 ($p = 0.208$) |
| LR_sym | 0.15 ($p = 0.527$) | 0.53 ($p = 0.023$) | 0.33 ($p = 0.075$) |
| LR_asym | 0.37 ($p = 0.120$) | 0.78 ($p < 0.001$) | 0.15 ($p = 0.417$) |
| Pub | 0.30 ($p = 0.212$) | 0.60 ($p = 0.008$) | 0.31 ($p = 0.100$) |
| UG_station | -0.13 ($p = 0.588$) | 0.55 ($p = 0.016$) | 0.15 ($p = 0.431$) |

4. Discussion

4.1 Speech intelligibility measurements

The significant effect of acoustic condition on SRTs highlights the strong influence of listening environments on speech intelligibility. The particularly favorable performance in *SON90* and *Pub* reflects the benefit of spatial release from masking, underscoring the importance of spatial separation cues for everyday communication. Conversely, the reduced variability observed in *SON0* compared to *SON90* suggests that co-located speech and noise provide a more uniform, though generally more challenging, listening situation. The interaction between acoustic condition and hearing sensitivity suggests that reduced auditory thresholds exacerbate difficulties in complex listening environments, consistent with clinical expectations. As shown in Figure 5, NH listeners achieved substantially lower SRTs in the *LR_sym*, *LR_asym*, and *Pub* conditions compared to stationary co-located noise (*SON0*), whereas both hearing-impaired groups showed only small differences across these conditions. In the *UG_station*, hearing-impaired listeners performed even worse than in *SON0*, a pattern not observed in NH participants, highlighting the detrimental effect of reverberation for individuals with hearing loss. Overall, the rank order of performance followed the anticipated gradient from NH to HI_noHA to HI_HAu. The elevated SRTs in hearing-aid users under demanding conditions may further reflect limitations of current hearing-aid processing in resolving complex acoustic scenes.

The observed mean SRTs for *SON0* (-5.3 dB) and *SON90* (-11.4 dB) in NH listeners were slightly higher than the reference value for monaural speech and noise (OLnoise) presentation (-7.1 dB) (Wagner et al., 1999) and also higher than the -17.0 dB reported for *SON90* (OLnoise) in young NH participants (Schütze, Kirsch, Kollmeier, et al., 2025). Schütze, Ewert et al. (2025) measured SRTs for both young NH and older HI listeners and reported -8.0 dB (NH) and -5 dB (HI with moderate hearing loss) for *SON0* and -13.0 dB (NH) and -9.0 dB (HI with moderate hearing loss) for *SON90*. The elevated SRTs in the present study are likely attributable, at least in part, to the higher age of our age-matched NH group. For the HI_noHA group, SRTs in *SON0* (-3.9 dB) and *SON90* (-8.3 dB) fell within the range reported by Schütze, Ewert et al. (2025) in the same spatial configuration of speech and masker.

For *LR_asym*, the mean SRTs observed here (NH: -12.8 dB; HI_noHa: -7.5 dB; HI_HAu: -1.6 dB) closely match those of Schütze, Ewert et al. (2025) (NH -13.2 dB; HI: -9.5 dB), where the same living room environment was tested with S-TV as target and S5 as masker. However, unlike their study, our setup included a different target position (S4 instead of S-TV), and additional maskers from the coupled room, which may have contributed to the somewhat higher SRTs observed in the HI groups. Furthermore, the two studies differed in their calibration procedures and, by extension, in the definition of SNR: while the current analysis calculated speech level based on the direct sound, Schütze, Ewert et al. (2025) included the full speech signal with reverberation. Recalculation factors provided in Table 1 could be used in the future to examine the influence of different calibration schemes.

Hládek et al. (2021) examined speech intelligibility in an underground environment using the OLSA test with Fastl noise (a fluctuating speech noise) (Fastl, 1987) as the masker. With the masker fixed at source position 1 (as shown in Figure 4) and target speech co-located at 1.6 m, participants achieved mean SRTs of -15.7 dB. In contrast, our age-matched NH group reached a mean SRT of -5.4 dB when tested with two spatially separated, fluctuating maskers. This large difference is likely due to a combination of factors, including differences in listener characteristics, the type of background noise used, and the calibration to direct sound applied in the present study.

The significant effect of acoustic condition on hearing aid benefit underscores that amplification does not provide uniform support across environments. While some complex scenes, such as *LR_asym* and

Pub, revealed greater benefit than the standard *SONO* setup, the lack of consistent differences across all conditions suggests that *SONO* alone may underestimate the potential advantages of hearing aids in everyday listening. Interestingly, hearing aid benefit in *SON90* did not differ from that observed in the living room environments but exceeded that in the *Pub* and *UG_station*, indicating that spatially separated noise remains a favorable condition for aided listening. The mean hearing aid benefit ranged between 2.6 dB (*SONO*) and 5.3 dB (*SON90*), but importantly, not all HI users reached the 2-dB threshold typically considered clinically relevant (Gemeinsamer Bundesausschuss, 2012). This limited improvement may reflect the focus of current hearing-aid praxis, where fitting is still primarily optimized for speech intelligibility in quiet rather than in noisy environments. Consequently, a considerable number of participants did not achieve the expected 2-dB gain in the *SONO* condition. In addition, suboptimal fitting strategies or progressive hearing decline since the last adjustment may further reduce the effectiveness of amplification for some users.

With regards to the correlations across different acoustic conditions examined here, our findings indicate that SRTs measured in the standard laboratory condition (*SONO*) are very strongly predictive of performance in more complex acoustic environments. The very strong correlations across conditions ($r = 0.88\text{--}0.91$) suggest that simple laboratory measurements can serve as reliable proxies for real-world listening performance, supporting their continued use in both research and clinical assessments. The small variability in correlations across complex conditions indicates robust predictive power, although slightly lower correlations for the *LR_asym* condition suggest that environments with higher spatial complexity may introduce additional variability.

Correlations among the complex conditions themselves were also very strong, highlighting stable individual differences in SRT performance regardless of acoustic complexity. Listeners who perform well in one challenging scenario tend to perform well in others, reflecting consistent individual hearing profiles. Similarly, SRT benefits relative to the standard condition showed strong to very strong correlations across complex environments ($r = 0.67\text{--}0.81$). While the absolute magnitude of benefit can differ depending on the acoustic scene, individuals who gain more benefit in one condition tend to gain more in other conditions as well, indicating that relative improvements are somewhat predictable.

These findings have important implications for experimental design and clinical practice. Standard laboratory measurements appear to provide meaningful information about both absolute performance and benefit in real-world listening scenarios. Nonetheless, the slightly lower correlations observed for SRT benefits suggest that including some complex listening condition may still be valuable in studies aiming to capture real-world auditory performance. Future research should investigate a wider range of acoustic environments and larger participant samples to further validate the predictive strength of standard laboratory measures.

Finally, it is important to acknowledge that the comparability of the different acoustic environments is limited by differences in reverberation, noise spectra, and the spatial separation between target and receiver. Each of these factors can independently affect speech intelligibility, making it difficult to attribute observed differences solely to one acoustic characteristic. For example, in the pub condition, target-listener distance was shorter (1.0 m) than in the underground station (1.6 m) or living room (1.6 m). Moreover, the reverberation time was substantially longer in the underground station ($T_{30} = 1.68$ s) compared to the living room ($T_{30} = 0.56$ s) and pub ($T_{30} = 0.66$ s).

An important aspect of this study is the improved ecological validity compared to standard audiometric assessments. By incorporating complex virtual acoustic scenes, the design simulates typical conversational situations more closely than in highly controlled laboratory settings. However, the task

of the listener remains somewhat artificial, as the measurements were based on matrix test sentences pronounced in a clear and standardized speaking style. This type of material has been shown to yield a very high accuracy and, importantly, allows for reliable comparisons across languages (Kollmeier et al., 2015). Nevertheless, such speech material does not fully capture the variability and dynamics of everyday communication, which often involves spontaneous speech, reduced articulation, and overlapping talkers (see also Kothe et al., 2025; Schütze, Kirsch, Ewert, et al., 2025). Future work should therefore explore the use of more realistic speech material and dialog-based tasks to further bridge the gap between experimental conditions and everyday listening demands.

A further limitation in the sense of ecological validity may be the comparability of perceptual properties compared to real environments. To some extent, listeners in a virtual acoustic scene perceive some differences between a lab environment and a real environment (Cubick & Dau, 2016; Hendrikse et al., 2019; Oreinos & Buchholz, 2016). However, the use of virtual acoustic scenes allows for a reproducible and controllable way of assessing listeners' perception in realistic scenes. For the room acoustics simulator tool RAZR that was used to create the complex acoustic scenes in the current study, the adequate perceptual representation compared to real environments has been validated in previous studies (e.g., Brinkmann et al., 2019, see algorithm B in their Fig. 8; Stärz et al., 2025). The perception in the scenes of the current study has not been directly compared to the corresponding real environments, but the room simulation properties were based on measurements in the real environments. Furthermore, Schütze, Kirsch, Kollmeier et al. (2025) found a good agreement between the virtual acoustic representation of the living room scene and the real living room.

In terms of measurement reliability, it is important to note that the inclusion of realistic virtual scenes does not compromise the robustness of SRT assessment. Although such environments inherently introduce greater variability in SNR over time and between ears, previous studies have demonstrated that test-retest reliability remains high. For example, Kramer et al. (2020) reported root-mean-square errors below 1.7 dB in normal-hearing listeners for both short excerpts and full-length realistic scenes, indicating stable and repeatable SRT estimates even under acoustically complex conditions. Consistent with these findings, the reliability of SRTs measured in virtual scenes typically falls between that observed in the standard S0N0 and S0N90 configurations. Measures derived from SRT differences, such as hearing aid benefit, may be slightly more susceptible to cumulative measurement error, particularly between stationary and fluctuating maskers. However, the overall impact of this error is likely limited, given the high test-retest reliability reported.

4.2 Listening effort

The present study provides an insight into how listening effort varies across different acoustic environments and hearing status, as assessed using ACALES. The significant main effects of acoustic condition and ESCU indicate that both the listening environment and the subjective effort level independently shape perceived effort. Specifically, the *LR_sym* condition consistently produced the lowest SNRs across all ESCU levels, reflecting minimal listening effort, whereas the *UG_station* condition resulted in the highest SNRs, indicating maximal effort. This pattern highlights the pronounced influence of environmental characteristics, such as reverberation and background noise, on subjective listening effort.

The standard *S0N0* condition elicited higher effort than the *LR_sym* condition but remained lower than the *UG_station*, positioning it as an intermediate benchmark for listening effort. Interestingly, *S0N90* showed effort levels comparable to *LR_asym* and *Pub*, suggesting that spatial separation of target and masker may mitigate perceived effort in complex settings. These findings reinforce the notion that the

acoustic properties of the listening environment, including spatial cues and reverberation, can strongly modulate listening effort independently of hearing status.

Regarding group differences, the factor listener group approached significance, implying that hearing thresholds alone do not fully account for variability in listening effort. This suggests that individual differences, potentially related to cognitive factors or prior auditory experience, may also contribute to effort perception. The general trend of hearing impairment leading to greater listening effort than that experienced by normal-hearing listeners was consistent with the findings of Krueger et al. (2017).

The analysis of hearing aid benefit for listening effort revealed no significant effects of acoustic condition, ESCU, or their interaction. This indicates that the magnitude of effort reduction provided by hearing aids was relatively consistent across different acoustic environments and effort levels, rather than varying systematically with scene complexity or ESCU. Across all conditions, the mean benefit ranged from 1.1 dB to 3.0 dB, with an overall average of 2.1 dB, demonstrating a modest but relatively uniform improvement in perceived effort with amplification.

Overall, these results highlight that, like speech intelligibility, listening effort is highly sensitive to acoustic environment characteristics, less strongly modulated by hearing status, and only partially influenced by amplification. This emphasizes the importance of considering complex and ecologically valid listening scenarios when assessing subjective effort, rather than relying solely on standard laboratory conditions.

Comparisons of SNRs at ESCU 7 and SRTs provide insights into the relationship between speech intelligibility and listening effort. Moderate to strong correlations between SRTs and ACALES indicate a consistent association between speech perception and listening effort. With amplification, correlations were slightly lower ($r = 0.40-0.57$), suggesting that hearing aids improve speech intelligibility and reduce listening effort, but slightly attenuate the strength of the direct association between these measures. Importantly, the benefit correlations (Δ SRT vs. Δ SNR) showed that improvements in speech recognition and reductions in effort were related, but not in a one-to-one manner. This aligns with the descriptive results: the hearing aid benefit in SRT differed significantly across acoustic conditions, whereas the benefit in listening effort was more consistent. Thus, the correlation patterns suggest that amplification stabilizes perceived effort across environments, while intelligibility gains remain more sensitive to the specific acoustic scene. The relationship between speech intelligibility and listening effort was also evaluated by Krueger et al. (2017), where the analysis was conducted per rating category and for different measurement conditions than in the present study. They found that the strongest correlation between both measures was found at negative SNRs compared to positive SNRs. On that basis, one might expect the aided condition, where SRTs are shifted to lower SNRs, to show higher correlations. In the present data, however, the aided correlation was lower. A parsimonious explanation is precision, i.e. the unaided analysis included 76 listeners, whereas the aided analysis included only 20. With such disparity, correlation estimates in the aided group are substantially less stable and may be biased toward weaker or more variable effects.

Although the absolute ACALES values showed a similar pattern to the SRT results, ACALES provides complementary information because it is measured at substantially higher (positive) SNRs, where conventional speech intelligibility tests typically reach ceiling performance. This allows ACALES to capture differences in perceived effort and hearing-aid benefit under conditions closer to everyday listening, particularly when assessing specific hearing-aid features such as noise reduction or directional processing.

4.3 Loudness scaling

In line with previous studies (e.g., Oetting et al., 2016), the results show that hearing-impaired listeners perceive complex, i.e. binaural broadband signals, as louder than normal-hearing listeners at high input sound levels. This effect was more pronounced for the HI_HAu group, suggesting that elevated loudness perception increases with hearing loss. Notably, the data revealed a high interindividual variability of loudness perception, although individual narrowband loudness compensation was already applied to the stimuli, highlighting the importance of measuring loudness perception for characterizing listeners.

In the context of complex listening environments, it is particularly relevant to conduct loudness measurements in binaural and broadband conditions, as this study does. These conditions mirror real-life listening situations, in which sounds are usually both binaural and broadband. However, no direct conclusions can be drawn about the specific effect of scene complexity on loudness perception, since both test signals were binaural and broadband, and no simple control condition was included. Previous studies suggest that scene complexity does influence loudness perception (Fichna et al., 2021; Oetting et al., 2016). For instance, Fichna et al. (2021) found that reverberation significantly lowers loudness ratings, while Oetting et al. (2016) reported that individual differences in binaural broadband loudness summation cannot be explained by narrowband loudness perception alone. This is supported by the findings of the present study, because even after individual narrowband loudness compensation, there was a high interindividual variability. These previous findings emphasize the potential value of extending loudness assessment to complex and ecologically valid listening environments, which may provide a more complete characterization of loudness perception in hearing-impaired individuals.

4.4 Tone-in-noise detection thresholds

Measuring tone detection thresholds in noise is closely related to assessing a listener's speech intelligibility in noisy environments, so it is relevant in the context of complex, i.e., noisy listening environments. In this study, correlations between tone-in-noise detection thresholds and hearing thresholds in quiet were very weak to strong ($r = 0.05-0.67$), indicating a predominantly pronounced difference between a listener's ability to detect tones in quiet and in noise. Correlations with speech intelligibility were in a similar range, but no consistent pattern emerged between simple and more complex environments.

The expectation that hearing-impaired listeners show elevated tone-in-noise thresholds was partially supported by the data of this study. While the HI_HAu participants showed higher thresholds than normal-hearing listeners, HI_noHA listeners had thresholds similar to normal-hearing listeners. These findings suggest that elevated thresholds are more pronounced in listeners with greater hearing loss.

Previous research supports the relevance of these measures. Schädler et al. (2020) found that tone-in-noise detection thresholds can improve speech intelligibility model's predictions of individual performance, highlighting the contribution of tone-in-noise detection thresholds to individualization of hearing abilities in hearing-impaired listeners.

4.5 HEAR-COMMAND Tool

Consistent with expectations, NH participants reported the lowest disability scores with narrow distributions, reflecting minimal perceived challenges in everyday listening. In contrast, HI_noHA listeners reported higher and more variable scores, highlighting greater heterogeneity in perceived difficulties within this group. Interestingly, HI_HAa participants reported the highest median and spread of scores, particularly in domains related to sound quality and listening functionality. This suggests that while hearing aids may restore audibility, they do not necessarily reduce the subjective perception of

effort or difficulty. Instead, amplification may introduce additional challenges such as altered sound quality or processing artifacts, which are reflected in higher, i.e., worse ratings. The presence of outliers and extended ranges across multiple domains further reflect the importance of considering individual variability rather than group averages alone. For some participants, difficulties were considerably more pronounced than the central tendency suggests, pointing to the multifactorial nature of hearing disability. Importantly, the breadth of domains captured by HEAR-COMMAND extends beyond what can be quantified by clinical intelligibility tests, as it encompasses aspects of communication, auditory processing, and social participation. These results demonstrate the added value of integrating self-report tools to complement behavioral measurements, offering a more complete characterization of hearing-related challenges in both normal-hearing and hearing-impaired listeners.

The comparison between HEAR-COMMAND tool speech perception scores and SRT values obtained with the OLSA revealed important insights into the sensitivity of both measures across listener groups and acoustic conditions. Particularly in the HI_noHA group, the HEAR-COMMAND tool captured relevant aspects of speech perception in challenging listening environments. In contrast, correlations were weaker and non-significant in normal-hearing listeners and aided hearing-impaired listeners. One possible interpretation for the NH group is that the participants consistently provided low ratings with narrow distributions in HEAR-COMMAND tool scores (see Table 3), reflecting minimal perceived difficulties. This likely weakened the association between HEAR-COMMAND tool scores and speech intelligibility measurements, as limited variability in ratings reduces the potential for strong correlations. For the HI_HAa group, correlations with SRT were also weak despite larger variability and overall higher HEAR-COMMAND tool scores compared to NH or HI_noHA. One possible interpretation is that amplification improves audibility but does not directly align subjective ratings of perceived difficulty with objective speech intelligibility outcomes. A likely explanation is that the HEAR-COMMAND tool captures a broader range of factors related to speech intelligibility which may lead to differences in sensitivity, with HEAR-COMMAND tool revealing difficulties that are not fully represented in SRT outcomes. In summary, the HEAR-COMMAND tool is particularly valuable for detecting functional limitations in unaided hearing-impaired listeners, offering a complementary perspective to traditional speech-in-noise testing.

4.6 Outlook

The dataset generated in this study offers multiple avenues for future work. First, the availability of individual-level data (Afghah, Biermann, et al., 2025) together with recordings of all acoustic scenes and room impulse responses (Gerken et al., 2025) provides a valuable resource for auditory modeling. Models of speech intelligibility, listening effort, or loudness perception can be tested not only against averaged outcomes but also against detailed individual performance, thereby supporting the development of more personalized predictive frameworks. The repository of individual data also includes additional outcomes such as coupler measurements for hearing aid users, obtained with the International Speech Test Signal (ISTS) (Holube et al., 2010) at input levels of 50, 65, 80 dB SPL as test signals for the left and right hearing aids. Furthermore, speech intelligibility data are provided not only for matrix test sentences but also for everyday sentences, though the latter are available only under the standard conditions (*SON0*, *SON90*).

Beyond auditory-based modelling, statistical approaches to auditory profiling, such as those proposed by Saak et al. (2022), could be applied to our dataset to identify subgroups of listeners with distinct hearing profiles, potentially improving the individualization of rehabilitation strategies. Another promising direction would be the application of the open master hearing aid (openMHA) framework (Kayser et al., 2022), an open-source platform that enables flexible testing and controlled individualization of hearing aid algorithms.

Third, the data have clear potential for cross-study comparisons. Because the present study calibrated complex acoustic scenes to the direct sound, the absolute SNR values may differ from datasets calibrated differently. To facilitate comparability, recalculation factors are provided in Table 1, which allow researchers to align these results with alternative calibration schemes.

Taken together, this dataset enables a wide range of methodological, theoretical, and translational applications – from refining auditory models, to advancing statistical characterizations of hearing impairment, to ensuring comparability across experimental paradigms.

5. Conclusions

Age-matched normal-hearing listeners consistently outperformed both unaided and aided hearing-impaired listeners across all acoustic environments. Performance decreased from NH to HI_noHA to HI_HAu, with hearing-impaired listeners showing substantially greater variability than normal-hearing listeners.

SRTs measured in traditional laboratory setups (*SONO*) strongly predicted performance in more complex listening scenarios, supporting their continued use for research and clinical assessment. Nonetheless, complex environments revealed variability in hearing aid benefit, highlighting the value of including challenging acoustic conditions to fully capture real-world listening performance. These findings, however, apply to matrix-type sentences, which differ from everyday conversational speech.

Hearing aids provide modest improvements in speech intelligibility and listening effort, with benefit largely independent of environmental complexity. Individual differences, however, remain substantial, emphasizing the need for personalized assessment and amplification strategies.

Combining objective measures, such as tone-in-noise detection and loudness perception, with self-report tools like HEAR-COMMAND tool provides a more complete assessment of hearing-impaired listeners' abilities beyond standard speech intelligibility tests. This integrated approach supports an accurate, individualized evaluation and guides personalized rehabilitation strategies in clinical practice.

Acknowledgements

The authors thank Melanie Krüger for her input about listening effort, and Dirk Oetting for his input about loudness perception. The authors thank all participants who took part in the study.

Data availability statement

Empirical data is publicly accessible on Zenodo under the "SFB 1330 Hearing Acoustics" community: <https://doi.org/10.5281/zenodo.14864856>.

Recordings of noise signals and room impulse responses are publicly accessible on Zenodo under the "SFB 1330 Hearing Acoustics" community: <https://doi.org/10.5281/zenodo.16895975>.

Ethics statement

The studies involving humans were approved by the “Research Ethics Committee of the Carl von Ossietzky University of Oldenburg”, in German: “Kommission für Forschungsfolgenabschätzung und Ethik der Carl von Ossietzky Universität” (Drs.EK/2021/031, Drs.EK/2021/031-02, Drs.EK/2021/031-03, Drs.EK/2021/031-04).

Funding

This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID: 352015383 – SFB 1330 A5/C4/C5.

CRedit authorship contribution statement

Merle Gerken: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Julia Schütze:** Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Christoph Kirsch:** Methodology, Writing – original draft, Writing – review & editing. **Bernhard U. Seeber:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition. **Stephan D. Ewert:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Jan Heeren:** Conceptualization, Methodology, Data collection. **Tahereh Afghah:** Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Kirsten C. Wagener:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Birger Kollmeier:** Conceptualization, Writing – review & editing, Funding acquisition. **Anna Warzybok:** Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Conflict of interest

The authors report there are no competing interests to declare.

References

- Afghah, T., Alfakir, R., Meis, M., Hammady, M., Youssif, M., Abd Al-Ghaffar, M., Kramer, S. E., & Wagener, K. C. (2024). ICF-based hearing and functioning assessment: Validation and research outcomes of utilizing the HEAR-COMMAND tool for patients with mild to moderately severe hearing loss and individuals with normal hearing. *Frontiers in Rehabilitation Sciences*, 5, 1389653.
- Afghah, T., Alfakir, R., Meis, M., van Leeuwen, L., Kramer, S. E., Hammady, M., Youssif, M., & Wagener, K. C. (2022). The development of a Self-Rated ICF-based questionnaire (HEAR-COMMAND Tool) to evaluate Hearing, Communication, and Conversation disability: Multinational experts' and patients' perspectives. *Frontiers in Rehabilitation Sciences*, 3, 1005525.
- Afghah, T., Biermann, P., Warzybok, A., Heeren, J., Wulff, A., Gerken, M., Schütze, J., Kirsch, C., Seeber, B., Ewert, S. D., Kollmeier, B., & Wagener, K. C. (2025). *A FAIR and Open-Access Database of Audiological Perceptual Measures* [Dataset]. Zenodo.
<https://doi.org/10.5281/zenodo.14864856>

- Afghah, T., Heeren, J., Hartog, L., Biermann, P., Wulff, A., Warzybok, A., & Wagener, K. C. (2025). *An open access dataset of perceptual measures for individuals with normal hearing and hearing loss*. [Preprint] Zenodo. <https://doi.org/10.5281/zenodo.17085598>
- Alfakir, R., Dunaway, L., Hyun, J., Sagong, H., Afghah, T., & Kang, S. (2025). Translation, Cultural Adaptation, and Field Testing of the Korean Version of the HEAR-COMMAND Tool: A Self-Rated ICF-Based Questionnaire for Assessing Hearing, Communication, and Conversation Disabilities in Korean-Speaking Populations. *Journal of Audiology & Otology*, 29(3), 197.
- Alfakir, R., Hammady, M., Afghah, T., Abd Al-Ghaffar, M., & Youssif, M. (2025). Translation, cultural adaptation, and validation of the HEAR-COMMAND Tool – Arabic: A self-rated ICF-based questionnaire for assessing hearing, communication, and conversation disability in Arabic-speaking populations. *The Egyptian Journal of Otolaryngology*, 41(1), 101. <https://doi.org/10.1186/s43163-025-00832-4>
- Allen, J. B., & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4), 943–950. <https://doi.org/10.1121/1.382599>
- ANSI. (1997). S3. 5-1997, Methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*, 19, 90–119.
- Bentler, R. A. (2005). Effectiveness of Directional Microphones and Noise Reduction Schemes in Hearing Aids: A Systematic Review of the Evidence. *Journal of the American Academy of Audiology*, 16(07), 473–484. <https://doi.org/10.3766/jaaa.16.7.7>
- Bitzer, J., Rollwage, C., & Neumann, M. (2014, June 12). *Evaluation Results of Speaker Verification for VoIP Transmission with Packet Loss*. Audio Engineering Society Conference: 54th International Conference: Audio Forensics. <https://www.aes.org/e-lib/browse.cfm?elib=17329>
- Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, 112(4), 1597–1604. <https://doi.org/10.1121/1.1502902>

- Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America*, 111(6), 2801–2810. <https://doi.org/10.1121/1.1479152>
- Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., & Weinzierl, S. (2019). A round robin on room acoustical simulation and auralization. *The Journal of the Acoustical Society of America*, 145(4), 2746–2760. <https://doi.org/10.1121/1.5096178>
- Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., El Kholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., ... Ludvigsen, C. (1994). An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America*, 96(4), 2108–2120. <https://doi.org/10.1121/1.410152>
- Cord, M. T., Surr, R. K., Walden, B. E., & Dyrland, O. (2004). Relationship between Laboratory Measures of Directional Advantage and Everyday Success with Directional Microphone Hearing Aids. *Journal of the American Academy of Audiology*, 15(05), 353–364. <https://doi.org/10.3766/jaaa.15.5.3>
- Cubick, J., & Dau, T. (2016). Validation of a virtual sound environment system for testing hearing aids. *Acta Acustica United with Acustica*, 102(3), 547–557. <https://doi.org/10.3813/AAA.918972>
- Danermark, B., Granberg, S., Kramer, S. E., Selb, M., & Möller, C. (2013). The creation of a comprehensive and a brief core set for hearing loss using the international classification of functioning, disability and health. *American Journal of Audiology*, 22(2), 323–328.
- Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.
- Ewert, S. D., Gößling, N., Buttler, O., Par, S. van de, & Hu, H. (2025). Computationally-efficient rendering of diffuse reflections for geometrical acoustics based room simulation. *Acta Acustica*, 9, 9. <https://doi.org/10.1051/aacus/2024062>
- Fastl, H. (1987). Ein Störgeräusch für die Sprachaudiometrie. *Audiologische Akustik*, 26(1), 2–13.

- Fastl, H., & Zwicker, E. (2006). *Psychoacoustics: Facts and Models*. Springer Science & Business Media.
- Fichna, S., Biberger, T., Seeber, B. U., & Ewert, S. D. (2021). Effect of Acoustic Scene Complexity and Visual Scene Representation on Auditory Perception in Virtual Audio-Visual Environments. *2021 Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, 1–9. <https://doi.org/10.1109/I3DA48870.2021.9610916>
- Gemeinsamer Bundesausschuss. (2012). Richtlinie des Gemeinsamen Bundesausschusses über die Verordnung von Hilfsmitteln in der vertragsärztlichen Versorgung (Hilfsmittel-Richtlinie/HilfsM-RL). *BAnz AT, 10(2012)*, B2.
- Gerken, M., Schütze, J., Kirsch, C., Seeber, B. U., Ewert, S. D., Heeren, J., Wagener, K. C., & Warzybok, A. (2025). *Simulations and recordings of scenes used in “Perceptual measures of normal-hearing and hearing-impaired listeners across defined virtual acoustic scenes”* [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.16895975>
- Gieseler, A., Tahden, M. A. S., Thiel, C. M., Wagener, K. C., Meis, M., & Colonius, H. (2017). Auditory and Non-Auditory Contributions for Unaided Speech Recognition in Noise as a Function of Hearing Aid Use. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.00219>
- Granberg, S. (2015). *Functioning and disability in adults with hearing loss: The preparatory studies in the ICF core sets for hearing loss project*.
- Greenberg, J. E., Peterson, P. M., & Zurek, P. M. (1993). Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. *The Journal of the Acoustical Society of America*, 94(5), 3009–3010. <https://doi.org/10.1121/1.407334>
- Grimm, G., Hendrikse, M., & Hohmann, V. (2021). *Pub environment* [Dataset]. Zenodo. <https://zenodo.org/records/5886987>
- Grimm, G., Luberadzka, J., & Hohmann, V. (2019). A toolbox for rendering virtual acoustic environments in the context of audiology. *Acta Acustica United with Acustica*, 105(3), 566–578. <https://doi.org/10.3813/AAA.919337>

- Hendrikse, M. M. E., Llorach, G., Hohmann, V., & Grimm, G. (2019). Movement and Gaze Behavior in Virtual Audiovisual Listening Environments Resembling Everyday Life. *Trends in Hearing*, 23, 2331216519872362. <https://doi.org/10.1177/2331216519872362>
- Hládek, L., Ewert, S. D., & Seeber, B. U. (2021). Communication Conditions in Virtual Acoustic Scenes in an Underground Station. *2021 Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, 1–8. <https://doi.org/10.1109/I3DA48870.2021.9610843>
- Hladek, L., & Seeber, B. U. (2022). *Underground station environment* [Dataset]. Zenodo. <https://zenodo.org/records/6025631>
- Hodgson, M., Steininger, G., & Razavi, Z. (2007). Measurement and prediction of speech and noise levels and the Lombard effect in eating establishments. *The Journal of the Acoustical Society of America*, 121(4), 2023–2033. <https://doi.org/10.1121/1.2535571>
- Holube, I., Blab, S., Fürsen, K., Gürtler Grober, S., Meisenbacher, K., Nguyen, D., & Taesler, S. (2009). Einfluss des Maskierers und der Testmethode auf die Sprachverständlichkeit von jüngeren und älteren Normalhörenden. *Zeitschrift Für Audiologie (Audiological Acoustics)*, 48, 120–127.
- Holube, I., Blab, S., Fürsen, K., Gürtler, S., Meisenbacher, K., Nguyen, D., & Taesler, S. (2008). Einfluss des Störgeräuschs und der Testmethode auf die Sprachverständlichkeitsschwelle von jüngeren und älteren Normalhörenden. *Proceedings of the Annual Meeting of the German Association of Audiology (DGA)*.
- Holube, I., Fredelake, S., Vlaming, M., & Kollmeier, B. (2010). Development and analysis of an international speech test signal (ISTS). *International Journal of Audiology*, 49(12), 891–903.
- ISO 389-8. (2004). Acoustics—Reference zero for the calibration of audiometric equipment—Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones. *International Organization for Standardization, Geneva, Switzerland*.
- Jafri, S., Berg, D., Buhl, M., Vormann, M., Saak, S., Wagener, K. C., Thiel, C. M., Hildebrandt, A., & Kollmeier, B. (2025). *The Oldenburg Hearing Health Record (OHHR)* (p. 2025.03.30.25324761). medRxiv. <https://doi.org/10.1101/2025.03.30.25324761>

- Jot, J.-M., & Chaigne, A. (1991). *Digital delay networks for designing artificial reverberators*. Audio Engineering Society Convention 90.
- Kaernbach, C. (1990). A single-interval adjustment-matrix (SIAM) procedure for unbiased adaptive testing. *The Journal of the Acoustical Society of America*, 88(6), 2645–2655.
<https://doi.org/10.1121/1.399985>
- Kamerer, A. M., AuBuchon, A., Fultz, S. E., Kopun, J. G., Neely, S. T., & Rasetshwane, D. M. (2019). The Role of Cognition in Common Measures of Peripheral Synaptopathy and Hidden Hearing Loss. *American Journal of Audiology*, 28(4), 843–856.
https://doi.org/10.1044/2019_AJA-19-0063
- Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3933–3936.
<https://doi.org/10.1109/ICASSP.2008.4518514>
- Kayser, H., Herzke, T., Maanen, P., Zimmermann, M., Grimm, G., & Hohmann, V. (2022). Open community platform for hearing aid algorithm research: Open Master Hearing Aid (openMHA). *SoftwareX*, 17, 100953. <https://doi.org/10.1016/j.softx.2021.100953>
- Kirsch, C., & Ewert, S. D. (2022). Computationally efficient simulation of edge diffraction in virtual acoustic environments. *Proceedings of the 24th International Congress on Acoustics*.
- Kirsch, C., & Ewert, S. D. (2024). Effects of measured and simulated diffraction from a plate on sound source localization. *The Journal of the Acoustical Society of America*, 155(5), 3118–3131.
<https://doi.org/10.1121/10.0025922>
- Kirsch, C., Wendt, T., Van De Par, S., Hu, H., & Ewert, S. D. (2023). Computationally-Efficient Simulation of Late Reverberation for Inhomogeneous Boundary Conditions and Coupled Rooms. *Journal of the Audio Engineering Society*, 71(4), 186–201.
<https://doi.org/10.17743/jaes.2022.0053>
- Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across

- languages: A review. *International Journal of Audiology*, 54, 3–16.
<https://doi.org/10.3109/14992027.2015.1020971>
- Kothe, A., Hohmann, V., & Grimm, G. (2025). *Effect of Avatar Head Movement on Communication Behaviour, Experience of Presence and Conversation Success in Triadic Conversations* (No. arXiv:2504.20844). arXiv. <https://doi.org/10.48550/arXiv.2504.20844>
- Kramer, F., Schädler, M., Hohmann, V., Oetting, D., & Warzybok, A. (2020). *Speech intelligibility and loudness perception with the trueLOUDNESS fitting rule*. 46, 114–117.
- Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017). Development of an adaptive scaling method for subjective listening effort). *The Journal of the Acoustical Society of America*, 141(6), 4680–4693. <https://doi.org/10.1121/1.4986938>
- Krueger, M., Schulte, M., Zokoll, M. A., Wagener, K. C., Meis, M., Brand, T., & Holube, I. (2017). Relation Between Listening Effort and Speech Intelligibility in Noise. *American Journal of Audiology*, 26, 378–392. https://doi.org/10.1044/2017_AJA-16-0136
- Lebo, C. P., Smith, M. F., Mosher, E. R., Jelonek, S. J., Schwind, D. R., Decker, K. E., Krusemark, H. J., & Kurz, P. L. (1994). Restaurant noise, hearing loss, and hearing aids. *Western Journal of Medicine*, 161(1), 45–49.
- Oetting, D., Brand, T., & Ewert, S. D. (2014). Optimized loudness-function estimation for categorical loudness scaling data. *Hearing Research*, 316, 16–27.
<https://doi.org/10.1016/j.heares.2014.07.003>
- Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., & Ewert, S. D. (2016). Spectral and binaural loudness summation for hearing-impaired listeners. *Hearing Research*, 335, 179–192.
<https://doi.org/10.1016/j.heares.2016.03.010>
- Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., & Ewert, S. D. (2018). Restoring Perceived Loudness for Listeners With Hearing Loss. *Ear and Hearing*, 39(4), 664.
<https://doi.org/10.1097/AUD.0000000000000521>
- Oreinos, C., & Buchholz, J. M. (2016). Evaluation of Loudspeaker-Based Virtual Sound Environments for Testing Directional Hearing Aids. *Journal of the American Academy of Audiology*, 27(7), 541–556. <https://doi.org/10.3766/jaaa.15094>

- Par, S. van de, Ewert, S. D., Hladek, L., Kirsch, C., Schütze, J., Llorca-Bofi, J., Grimm, G., Hendrikse, M. M. E., Kollmeier, B., & Seeber, B. U. (2022). Auditory-visual scenes for hearing research. *Acta Acustica*, 6, 55. <https://doi.org/10.1051/aacus/2022032>
- Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6), 456–466.
- Regev, J., Zaar, J., Relaño-Iborra, H., & Dau, T. (2025a). *Dataset for: “Investigating the effects of age and hearing loss on speech intelligibility and amplitude modulation frequency selectivity”* [Dataset]. Technical University of Denmark. <https://doi.org/10.11583/DTU.25771884.v1>
- Regev, J., Zaar, J., Relaño-Iborra, H., & Dau, T. (2025b). Investigating the effects of age and hearing loss on speech intelligibility and amplitude modulation frequency selectivity. *The Journal of the Acoustical Society of America*, 157(3), 2077–2090. <https://doi.org/10.1121/10.0036220>
- Rönnerberg, J., Lunner, T., Ng, E. H. N., Lidestam, B., Zekveld, A. A., Sörqvist, P., Lyxell, B., Träff, U., Yumba, W., Classon, E., Hällgren, M., Larsby, B., Signoret, C., Pichora-Fuller, M. K., Rudner, M., Danielsson, H., & Stenfelt, S. (2016). Hearing impairment, cognition and speech understanding: Exploratory factor analyses of a comprehensive test battery for a group of hearing aid users, the n200 study. *International Journal of Audiology*, 55(11), 623–642. <https://doi.org/10.1080/14992027.2016.1219775>
- Saak, S., Huelsmeier, D., Kollmeier, B., & Buhl, M. (2022). A flexible data-driven audiological patient stratification method for deriving auditory profiles. *Frontiers in Neurology*, 13. <https://doi.org/10.3389/fneur.2022.959582>
- Sanchez-Lopez, R., Nielsen, S. G., El-Haj-Ali, M., Bianchi, F., Fereczkowski, M., Cañete, O. M., Wu, M., Neher, T., Dau, T., & Santurette, S. (2021). Auditory Tests for Characterizing Hearing Deficits in Listeners With Various Hearing Abilities: The BEAR Test Battery. *Frontiers in Neuroscience*, 15. <https://doi.org/10.3389/fnins.2021.724007>
- Schädler, M. R., Hülsmeier, D., Warzybok, A., & Kollmeier, B. (2020). Individual Aided Speech-Recognition Performance and Predictions of Benefit for Listeners With Impaired Hearing Employing FADE. *Trends in Hearing*, 24, 2331216520938929. <https://doi.org/10.1177/2331216520938929>

- Schubotz, W., Brand, T., Kollmeier, B., & Ewert, S. D. (2016). Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features. *The Journal of the Acoustical Society of America*, 140(1), 524–540. <https://doi.org/10.1121/1.4955079>
- Schütze, J., Ewert, S. D., Kirsch, C., & Kollmeier, B. (2025). Unaided and aided speech intelligibility in a real and virtual acoustic environment. *Trends in Hearing*, *in press*.
- Schütze, J., Kirsch, C., Ewert, S. D., Yadav, M., & Kollmeier, B. (2025). Effects of Background Noise on Communication Abilities and Subjective Effort in Listeners with Normal Hearing and Listeners with Impaired Hearing. *Proceedings of I3DA*.
- Schütze, J., Kirsch, C., Kollmeier, B., & Ewert, S. D. (2025). Comparison of speech intelligibility in a real and virtual living room using loudspeaker and headphone presentations. *Acta Acustica*, 9, 6. <https://doi.org/10.1051/aacus/2024068>
- Schütze, J., Kirsch, C., Wagener, K. C., Kollmeier, B., & Ewert, S. D. (2021). *Living room environment*. <https://doi.org/10.5281/ZENODO.5747753>
- Spearman, C. (1961). “ *General Intelligence* ” Objectively Determined and Measured.
- Stärz, F., Par, S. V. D., Roßkopf, S., Kroczeck, L. O. H., Mühlberger, A., & Blau, M. (2025). Comparison of binaural auralisations to a real loudspeaker in an audiovisual virtual classroom scenario: Effect of room acoustic simulation, HRTF dataset, and head-mounted display on room acoustic perception. *Acta Acustica*, 9, 31. <https://doi.org/10.1051/aacus/2025012>
- Suck, L., Hartog, L., Ewert, S., Hohmann, V., & Oetting, D. (2020). Verkürzung der trueLOUDNESS-Anpassmethode zur binauralen breitbandigen Lautheitsnormalisierung in Hörgeräten. *Deutsche Gesellschaft Für Audiologie eV, Editors*, 23.
- Vlaming, M. S. M. G., Kollmeier, B., Dreschler, W. A., Martin, R., Wouters, J., Grover, B., Mohammadh, Y., & Houtgast, T. (2011). HearCom: Hearing in the Communication Society. *Acta Acustica United with Acustica*, 97(2), 175–192. <https://doi.org/10.3813/AAA.918397>
- Wagener, K. C., & Brand, T. (2005). Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters La inteligibilidad de frases en silencio para sujetos con audición normal y con hipoacusia: la influencia del procedimiento de medición y de los parámetros de enmascaramiento.

International Journal of Audiology, 44(3), 144–156.

<https://doi.org/10.1080/14992020500057517>

Wagener, K. C., Brand, T., Kuehnel, V., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache I-III: Design, Optimierung und Evaluation des Oldenburger Satztest. *Zeitschrift Für Audiologie*, 38.

Wagener, K. C., Hansen, M., & Ludvigsen, C. (2008). Recording and classification of the acoustic environment of hearing aid users. *Journal of the American Academy of Audiology*, 19(04), 348–370.

Wendt, T., van de Par, S., & Ewert, S. D. (2014). A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation. *Journal of the Audio Engineering Society*, 62(11), 748–766.

WHO. (2001). ICF: International classification of functioning, disability and health. In *ICF: international classification of functioning, disability and health*.

Zimmer, J., Hartog, L., Oetting, D., Pöntynen, H., & Dietz, M. (2024). Loudness and lateralization of binaural broadband noise for subjects with asymmetric hearing loss. *GMS Zeitschrift Für Audiologie - Audiological Acoustics*, 6, Doc10. <https://doi.org/10.3205/zaud000045>

Appendix

A.1 Source levels in acoustic scenes

Table A.1: Overview of the various maskers used in the acoustic scenes. The source position, audio, and audio source are given.

| Scene | Source Position | Audio | Audio Source |
|---------|------------------|-----------------------|--|
| LR_sym | S4, | Male transformed ISTS | Schubotz et al., 2016 |
| | S5 | | |
| | S6 | Dishwasher | Freesound.org bm_dishwasher.wav |
| | S7, S8 | Female dialogue | Bitzer et al., 2014 NK06UL23_EN06RK01_spontan_44100_Hz.wav |
| LR_asym | S-TV | Male transformed ISTS | Schubotz et al., 2016 |
| | S6 | Dishwasher | Freesound.org bm_dishwasher.wav |
| | S7, S8 | Female dialogue | Bitzer et al., 2014 NK06UL23_EN06RK01_spontan_44100_Hz.wav |
| | | | |
| Pub | T1, T3, T4 | Conversation 1 | Gerken et al., 2020 (https://doi.org/10.5281/zenodo.4160499) |
| | S1, S2 | Conversation 2 | Bitzer et al., 2014 |
| | S3, S5 | Male transformed ISTS | Schubotz et al., 2016 |
| | S4, S6 | Female ISTS | Holube et al., 2010 |
| | S7, S8 | Conversation 3 | Bitzer et al., 2014 |
| | N1, N2, N3 | Conversation 4 | Gerken et al., 2020 (https://doi.org/10.5281/zenodo.4160499) |
| | | | |

| | | | |
|------------|-------------|-----------------------|--|
| | P1 | Conversation 5 | Bitzer et al., 2014 |
| | P2 | Conversation 6 | Bitzer et al., 2014 |
| | P3 | Conversation 7 | Bitzer et al., 2014 |
| | P4 | Conversation 8 | Gerken et al., 2020 (https://doi.org/10.5281/zenodo.4160499) |
| | P5 | Conversation 9 | Bitzer et al., 2014 |
| | P6, P7 | Conversation 10 | Gerken et al., 2020 (https://doi.org/10.5281/zenodo.4160499) |
| | P8 | Conversation 11 | Gerken et al., 2020 (https://doi.org/10.5281/zenodo.4160499) |
| | PA1, PA2 | Music Loop | Christoph Kirsch |
| | bartender | Pouring a beer | Freesound.org 428334__zembacraftworks__pouring-a-beer-from-the-tap.wav |
| | S4, S7 | Dishware | Freesound.org 219217__robinhood76__04812-laying-table-for-dish.wav |
| UG_station | S2, P8 | Glasses klinging | Freesound.org 166582__matucha__beerglasses-01.wav |
| | 3, 11 | Male transformed ISTS | Schubotz et al., 2016 |
| | 17 | Escalator | Freesound.org escalator(roachpowder-freesound).wav (https://freesound.org/s/170231/) |
| | | Ambient | Hladek et al. (2021) |

A.2 HEAR-COMMAND questionnaire overview

Table A.2: Summary of the targeted concepts covered by the 90 ICF-based items of the HEAR-COMMAND Tool. The complete questionnaire is available in multiple languages at: <https://www.hz-ol.de/en/open-tools-for-science/hear-command-tool/>.

| Body Functions | | Body Functions | | Activities and Participation | | Environmental Factors | |
|----------------|------------------------------------|----------------|---|------------------------------|--|-----------------------|---|
| # | Item content | # | Item content | # | Item content | # | Item content |
| H.1 | Mood swings | H.25 | Detecting noise in household | H.49 | Dealing with stressful situations | H.75 | Support received from society |
| H.2 | Sleeping | H.26 | Discriminating the sound of a car/bus | H.50 | Interacting with people in a socially appropriate manner | H.76 | Emotional support from family/friends |
| H.3 | Focusing attention | H.27 | Recognizing musical instruments | H.51 | Socializing with people in your community | H.77 | Support from family/friends in daily functioning |
| H.4 | Maintaining focus | H.28 | Detecting where a sound comes from | H.52 | Dealing with unknown people | H.78 | Support from health services/systems |
| H.5 | Remember information | H.29 | Telling a bus/truck is getting close or far | H.53 | Having formal relationships with people in authority | H.79 | Support from healthcare professional |
| H.6 | Recall new information | H.30 | Detecting corner of a room when one is talking | H.54 | Socializing with your family or friends | H.80 | Communication services/systems usefulness |
| H.7 | Sadness or depression | H.31 | Telling how far away a bus/truck is | H.55 | Making new friends | H.81 | Design of workplace as a barrier |
| H.8 | Seeing across the road | H.32 | Telling where a human is when he screams/dog barks | H.56 | Having an argument or debate | H.82 | Insufficient light as a barrier |
| H.9 | Seeing over an arm length | H.33 | Detecting whether the person on left/right starts talking | H.57 | Understanding a statement during communication | H.83 | Low volume of speech as a barrier |
| H.10 | Taste loss | H.34 | Hearing a single jumbled sound when hearing more than one sound | H.58 | Maintaining relationships with immediate family | H.84 | Background noise as a barrier |
| H.11 | Smell loss | H.35 | Understanding the speech over distance | H.59 | Joining in community activities | H.85 | Reverberant environment as a barrier |
| H.12 | Dizziness | H.36 | Understanding the speech in a quiet environment | H.60 | Engaging in any hobby or pleasurable activity | H.86 | Unclear sound considered a barrier |
| H.13 | Loss of balance | H.37 | Understanding the speech in a noisy environment | H.61 | Continuing relationships in an appropriate manner | H.87 | Hearing aid usefulness in normal daily routines |
| H.14 | Pain (general) | H.38 | Understanding news presenter on TV | H.62 | Performing communication techniques | H.88 | Hearing aid usefulness in conversation activities |
| H.15 | Pain (head & neck) | H.39 | Understanding what one is saying while the TV is on | H.63 | Your day-to-day tasks | H.89 | Hearing aid usefulness while using phone |
| H.16 | Understanding meaning of a message | H.40 | Understanding the news presenter and someone else | H.64 | Doing your most important tasks well | H.90 | Hearing aid usefulness while watching TV |
| H.17 | Producing a meaningful message | H.41 | Having health conditions causing speech impairment | H.65 | Getting done all the tasks | | |
| H.18 | Ringings/buzzing in ears | H.42 | Making sounds other than speech | H.66 | Getting your tasks done quickly | | |
| H.19 | Pressure in ear | H.43 | Changing pitch of sounds other than speech | H.67 | Conversation or speaking with someone | | |
| H.20 | Irritation in ear | H.44 | Changing volume of sounds other than speech | H.68 | Conversation or speaking with many people | | |
| H.21 | Distinguishing pitch | H.45 | Pronunciation | H.69 | Carrying on a conversation during a crowded meeting | | |
| H.22 | Distinguishing tone | H.46 | Regulating the volume of speech | H.70 | Carrying on a conversation in a bus or car | | |
| H.23 | Distinguishing volume | H.47 | Regulating the speed of speech | H.71 | Following a conversation in a busy restaurant | | |
| H.24 | Detecting a sound in environment | H.48 | Telling stories or reporting | H.72 | Carrying a phone call in a quiet room | | |
| | | | | H.73 | Telling what one is saying when conversation switches | | |
| | | | | H.74 | Listening to the TV/Radio/Music | | |