HHS Public Access

Author manuscript

Ear Hear. Author manuscript; available in PMC 2020 July 01.

Published in final edited form as:

Ear Hear. 2019 ; 40(4): 805–822. doi:10.1097/AUD.0000000000000672.

Efficacy and Effectiveness of Advanced Hearing Aid Directional and Noise Reduction Technologies for Older Adults with Mild to Moderate Hearing Loss

Yu-Hsiang Wu[1], **Elizabeth Stangl**[1], **Octav Chipara**[2], **Syed Shabih Hasan**[2], **Sean DeVries**[3], and **Jacob Oleson**[3]

[1]Department of Communication Sciences and Disorders, The University of Iowa

[2]Department of Computer Science, The University of Iowa

[3]Department of Biostatistics, The University of Iowa

Abstract

Objectives: The purpose of the current study was to investigate the laboratory efficacy and real-world effectiveness of advanced directional microphones (DM) and digital noise reduction (NR) algorithms (i.e., premium DM/NR features) relative to basic-level DM/NR features of contemporary hearing aids (HAs). The study also examined the effect of premium HAs relative to basic HAs and the effect of DM/NR features relative to no features.

Design: Fifty-four older adults with mild-to-moderate hearing loss completed a single-blinded crossover trial. Two HA models, one a less-expensive, basic-level device (basic HA) and the other a more-expensive, advanced-level device (premium HA), were used. The DM/NR features of the basic HAs (i.e., basic features) were adaptive DMs and gain-reduction NR with fewer channels. In contrast, the DM/NR features of the premium HAs (i.e., premium features) included adaptive DMs and gain-reduction NR with more channels, bilateral beamformers, speech-seeking DMs, pinna-simulation directivity, reverberation reduction, impulse noise reduction, wind noise reduction, and spatial noise reduction. The trial consisted of four conditions, which were factorial combinations of HA model (premium vs. basic) and DM/NR feature status (on vs. off). In order to blind participants regarding the HA technology, no technology details were disclosed and minimal training on how to use the features was provided. In each condition participants wore bilateral HAs for five weeks. Outcomes regarding speech understanding, listening effort, sound quality, localization, and HA satisfaction were measured using laboratory tests, retrospective self-reports (i.e., standardized questionnaires), and in-situ self-reports (i.e., self-reports completed in the real world in real time). A smartphone-based ecological momentary assessment system was used to collect in-situ self-reports.

Address correspondence to Yu-Hsiang Wu, 125C SHC, Department of Communication Sciences and Disorders, The University of Iowa, Iowa City, IA 52242, USA., yu-hsiang-wu@uiowa.edu, Fax number: 319-335-8851, Telephone number: 319-335-8728.

AUTHOR CONTRIBUTIONS

Y.H.W. designed experiments, interpreted data, and wrote the paper; E.S. collected data; O.C. and S.S.H. developed EMA software and processed EMA data; S.D. and J.O. provided statistical analysis. All authors discussed the results and implications and commented on the manuscript at all stages.

Results: Laboratory efficacy data generally supported the benefit of premium DM/NR features relative to basic DM/NR, premium HAs relative to basic HAs, and DM/NR features relative to no DM/NR in improving speech understanding and localization performance. Laboratory data also indicated that DM/NR features could improve listening effort and sound quality compared to no features for both basic- and premium-level HAs. For real-world effectiveness, in-situ self-reports first indicated that noisy or very noisy situations did not occur very often in participants' daily lives (10.9% of the time). Although both retrospective and in-situ self-reports indicated that participants were more satisfied with HAs equipped with DM/NR features than without, there was no strong evidence to support the benefit of premium DM/NR features and premium HAs over basic DM/NR features and basic HAs, respectively.

Conclusions: Although premium DM/NR features and premium HAs outperformed their basic-level counterparts in well-controlled laboratory test conditions, the benefits were not observed in the real world. In contrast, the effect of DM/NR features relative to no features was robust both in the laboratory and in the real world. Therefore, the current study suggests that although both premium and basic DM/NR technologies evaluated in the study have the potential to improve HA outcomes, older adults with mild-to-moderate hearing loss are unlikely to perceive the additional benefits provided by the premium DM/NR features in their daily lives. Limitations concerning the study's generalizability (e.g., participant's lifestyle) are discussed.

Keywords

hearing loss; hearing aid; directional microphone; noise reduction algorithm

INTRODUCTION

Hearing aid (HA) amplification is the primary intervention for sensorineural hearing loss and its related psychosocial consequences in elderly adults (Chisolm et al. 2007). Of those who could benefit from HAs, only one in ten with mild hearing loss and less than four in ten with moderate-to-severe hearing loss use HAs (Kochkin 2009; Lin et al. 2011). One of the most commonly reported reasons for not seeking amplification intervention is its expense (Kochkin 2007). HA adoption rates are even lower for older adults with lower incomes and for those of racial and ethnic minorities (Bainbridge & Ramachandran 2014). Therefore, there is an urgent need to promote affordable and accessible hearing health care (Donahue et al. 2010). Toward this end, it is important to determine if the high cost of premium HAs, which are often equipped with more sophisticated (i.e., high-end or premium) signal processing technologies, can be justified by a demonstrable benefit additional to that obtained using less expensive and more basic technologies.[1]

Among various HA technologies, it is arguable that technologies designed to reduce noise are one of the most important as difficulty in listening in noisy environments is often reported by HA users (Takahashi et al. 2007). These technologies include directional microphones (DMs; or beamformers) and noise reduction (NR) algorithms. The former

[1]The definitions of low-end and high-end technologies are relative and evolve across time. A current low-end technology could have been considered high-end technology years ago. There is little consensus about the use of terminologies of high-end/low-end, complex/simple, and premium/basic to describe the difference in HA technology level. In the current paper, the terms *premium* and *basic* are used in accordance with the works published by Cox and her colleagues (Cox et al. 2014, 2016; Johnson et al. 2016, 2017).

utilizes multiple microphones to achieve spatially dependent sound sensitivity, thereby improving signal-to-noise ratio (SNR), while the latter analyzes the incoming signal and alters the gain/output characteristics to enhance speech and/or attenuate noise.

Over the past decades, both DM and NR technologies have been evolving from basic algorithms into more sophisticated and complicated designs. For DMs, the most basic design is a fixed DM that cannot alter its spatially-dependent directivity pattern (i.e., the polar pattern). In contrast, a more advanced design is a multichannel adaptive DM that can steer the least sensitive area of its directivity pattern (i.e., the null) to the direction of the noise source with each channel working independently. Multichannel DMs of high-end HAs typically have more channels than low-end HAs. The most advanced DM technologies implemented in contemporary HAs include bilateral beamformers and “speech-seeking DMs”. The former utilizes the microphones of paired HAs on two ears to achieve high directivity, and the latter can detect the location of speech and steer the most sensitive area of the directivity pattern (i.e., the lobe) to the direction of speech when speech is not in front of the listener. In addition to improving speech understanding in noise, DM technologies have been used to replicate the natural directivity provided by the pinna at high frequencies in order to improve front-to-back sound localization (denoted as “pinna-simulation directivity”). In terms of NR, advanced NR generally have more working channels and are faster than basic NR. Advanced NR can also detect and process a broader range of sounds to reduce their adverse effect, including reverberation, impulse noises, and wind noises. Finally, advanced NR can work in conjunction with DMs such that the NR system uses the spatial information provided by the DM system to decide the amount of gain reduction (i.e., “spatial noise reduction”). For more details about each DM/NR technology, see Holube et al. (2014) and Launer et al. (2016). In the current paper, more sophisticated and complicated designs of DM and NR technologies are referred to as premium DM/NR features.

Previous research has sought to determine the relative effect of DM and NR features compared to no features. In laboratory studies, DMs have consistently been shown to improve speech recognition performance in noise compared to their omnidirectional counterparts (e.g., Valente et al. 1995; Bentler et al. 2004; Walden et al. 2005; Wu & Bentler 2010a). Laboratory data also indicate that NR features could increase ease/comfort of listening, reduce listening effort, and are highly preferred by listeners (e.g., Ricketts & Hornsby 2005; Mueller et al. 2006; Bentler et al. 2008; Sarampalis et al. 2009; Ohlenforst et al. 2017; Wendt et al. 2017). However, compared to the strong evidence from laboratory research, the evidence supporting the contribution of DMs and NR to outcomes in the real world is quite limited (Bentler 2005; Bentler & Chiou 2006). Although some field studies have shown that DMs produce a perceived benefit (Preves et al. 1999; Ricketts et al. 2003), more studies have not (Walden et al. 2000; Cord et al. 2002; Gnewikow et al. 2009; Humes et al. 2009; Wu & Bentler 2010b). Field studies for NR are relatively scarce and results have been mixed (Boymans & Dreschler 2000; Bentler et al. 2008; Zakis et al. 2009).

Researchers also have tried to determine if premium DM and NR technologies deliver more benefit than basic technologies. Laboratory research has consistently shown that bilateral beamformers (Picou & Ricketts 2018) and speech-seeking DMs (Wu et al. 2013a; Wu et al. 2013b) outperform basic-level DMs. Pinna-simulation directivity can improve front-to-back

localization (Keidser et al. 2009). In terms of NR, previous laboratory studies have demonstrated the benefit of reverberation reduction (Fabry & Tehorz 2005), impulse noise reduction (Korhonen et al. 2013), and wind noise reduction (Latzel & Appleton 2013; Korhonen et al. 2017) algorithms. In contrast, evidence supporting the effect of premium DM/NR features in the real world is limited. Recently Cox and her colleagues conducted a clinical trial to examine the relative effect of premium-feature HAs compared with basic-feature HAs in improving speech understanding, listening effort, sound localization, and quality of life (Cox et al. 2014, 2016; Johnson et al. 2016, 2017). Results indicated that there were no statistically significant or clinically important differences in improvement between premium and basic HAs in most laboratory tests and in the real world. Although Cox and her colleagues did not directly compare the effect of premium and basic DM/NR technologies, their findings seem to suggest that premium and basic DM/NR features would yield similar outcomes.

The main purpose of the current study was to comprehensively determine the relative efficacy (how well a feature can work given the best possible scenario; Cox 2005) and effectiveness (how well a feature actually works in real-world settings; Cox 2005) of premium DM/NR features compared with basic DM/NR features. This effect is referred to as the *effect of premium DM/NR features* in this article. The current study also investigated the relative efficacy and effectiveness (1) of premium HAs compared to basic HAs and (2) of DM/NR features compared to no DM/NR features. Determining *the effect of premium HAs* would clarify that the results of the research by Cox and her colleagues could be replicated. Examining *the effect of DM/NR features* would help elucidate the mixed results observed in previous research.

In the current study, efficacy was assessed using laboratory test materials and conditions designed specifically for each DM/NR feature. In short, many premium DM/NR technologies have the potential to provide their maximum benefit and outperform their more low-end predecessors only in certain best-case situations. For example, premium multichannel adaptive DMs (more channels) would provide greater benefit than basic adaptive DMs (fewer channels) only in sound fields that have multiple discrete noises with different frequencies presented from various directions; speech-seeking DMs would surpass traditional adaptive DMs only when the talker is not located in front of the listener. The laboratory test conditions of the current study were designed to create the best-case situations for each DM/NR feature.

In terms of real-world effectiveness, retrospective self-reports (i.e., standardized questionnaires) were used. Although retrospective self-reports have been widely used in audiology research, they could be subject to recall bias (Bradburn et al. 1987) and therefore may not have the sensitivity to detect the difference between features. To address this issue, the ecological momentary assessment (EMA) was also used to assess real-world effectiveness. EMA, also known as experience sampling or ambulatory assessment, is a methodology involving repeated assessments/surveys to collect data describing respondents' current or very recent (i.e., momentary) experiences and related contexts in their natural (i.e., ecological) environments (Shiffman et al. 2008). Because experiences are recorded almost immediately in each assessment, EMA is less affected by recall bias. Research has shown

that EMA is a valid methodology in audiology research (Wu et al. 2015; Timmer et al. 2017). The assessments delivered using EMA are referred to as in-situ self-reports in this paper.

MATERIALS AND METHODS

Older adults with hearing loss were recruited and fitted with bilateral HAs. Two HA models, one a less-expensive, basic-level device (basic HA) and the other a more-expensive, advanced-level device (premium HA), were used. The DM and NR features of the HAs were turned on or off to create four HA conditions: basic-on, basic-off, premium-on, and premium-off (2 × 2 factorial design). A single-blinded, crossover repeated measures design was used. During the field trial of each HA condition, participants wore the devices in their daily lives for five weeks. HA outcomes were measured using laboratory tests, retrospective self-reports, and in-situ self-reports.

The current study was part of a larger study. One of the goals required participants to carry digital audio recorders to record environmental sounds with and without wearing HAs. Therefore, an unaided condition was included in the study in addition to the four HA conditions. Portions of the audio recording results (Klein et al. 2018; Wu et al. 2018) have already been published and will not be reported in the current paper. Furthermore, because the larger study also aimed to examine the test-retest reliability of the EMA methodology, participants repeated one of the four HA conditions (randomly selected) before the end of the study. The results of the re-test conditions will not be reported in the current paper either.

Participants

Fifty-four participants (26 males and 28 females) were recruited from cities, towns and farms around eastern Iowa and north western Illinois and completed the study. Their ages ranged from 65 to 88 years with a mean of 73.6 years. Participants were eligible for inclusion in the study if their hearing loss met the following criteria: (1) postlingual, bilateral, sensorineural hearing loss (air-bone gap < 10 dB); (2) pure-tone average across 0.5, 1, 2, and 4 kHz greater than 25 dB HL but not worse than 60 dB HL (ANSI 2010); and (3) hearing symmetry within 20 dB for all test frequencies. The study focused on mild-to-moderate hearing loss because of its high prevalence (Lin et al. 2011). The mean pure-tone thresholds are shown in Figure 1. All participants were native English speakers. Upon entering the study, 30 participants had previous HA experience for at least one year. While 54 participants completed the study, 6 participants withdrew from the study due to scheduling conflicts (n = 4) or unwillingness to record other people’s voices (n = 2).

The subject number of the current study was determined by a power analysis. To calculate the power, field studies that evaluated the effect of DM were reviewed (Ricketts et al. 2003; Palmer et al. 2006; Gnewikow et al. 2009). It was estimated that the outcome difference between DMs turned on and off measured using self-reports (converted to a scale ranging from 0 to 10) was from 0.3 (i.e., 3%) to 0.6 points (6%), which would be a clinically relevant difference. It was then assumed that the effect of premium DM/NR relative to basic DM/NR was similar to the effect of DM relative to no feature. Because in the current study EMA was used, the variations in EMA data within and between subjects were further

estimated from the literature (Wu & Bentler 2012a). With these estimations, Monte Carlo simulations were conducted using statistical software SAS. Results indicated that the research (2 × 2 factorial repeated measures design) required 54 subjects to detect the effect of premium DM/NR features, assuming $\alpha = 0.05$ and $\beta = 0.2$.

Hearing aids and fitting

Two commercially-available behind-the-ear HA models were used in the current study: basic (retail price per pair ≈ $1500) and premium (retail price per pair ≈ $5000) HAs. The HAs were released in 2013 and are still on the market at this writing. Both models were from the same manufacturer and shared the same hardware and chip. Participants were unable to recognize the HA model based on the device's case. Table 1 compares the DM/NR features of basic and premium HAs. In short, automatic adaptive DMs were available in both basic and premium HAs. The major difference was that the basic HAs had one beamforming channel while the premium HAs had 33 beamforming channels. Bilateral beamformers were only available in the premium HAs. This feature could be set to a fixed or an automatic mode. In the automatic mode, the beamformer would be activated when the speech is in front of the hearing instrument and the noise level is higher than 70 dB SPL. The speech-seeking DM, which is also a premium feature, could only be manually activated by users. After activation, the algorithm could detect the direction of the talker and automatically steer the directivity lobe to left, right, or back of the user. As for the NR, the only NR feature of the basic HAs was a 12-channel gain-reduction NR (Ricketts et al. 2018) which analyzes the variations of sound over time and reduces gain for different frequency bands whenever noise is detected. The premium HAs had the same NR feature, but with more (20) working channels. The reverberation reduction of the premium HAs looks for repeated signals of an original signal. It is designed to reduce reverberant signal components in speech pauses and works best in an environment with high SNRs (e.g., speech in quiet). The impulse noise reduction of the premium HAs detects the leading edge of a transient by sensing a rapidly rising level. The transient could be attenuated within microseconds of detection. As for the wind noise reduction, it uses information from both microphone ports on a hearing device (rather than from paired HAs on two ears) to detect wind noise. The algorithm then applies a reduction of gain below 1000 Hz to reduce the annoyance from wind noise. Finally, the spatial noise reduction of premium HAs uses the DM system to estimate the direction-of-arrival of the sounds. The algorithm then uses this spatial information to selectively reduce the gain for noise coming from behind, even when the noise contains speech. This feature works best when speech signals are from the front and noise are from the back of the listener.

HAs were coupled to the participant's ears bilaterally using slim tubes and custom canal earmolds with clinically-appropriate vent sizes. The devices were programmed to meet real-ear aided response (REAR) targets (+/− 3 dB) specified by the second version of the National Acoustic Laboratory nonlinear prescriptive formula (NAL-NL2, Keidser et al. 2011) and was verified using a probe-microphone HA analyzer (Audioscan Verifit; Dorchester, Ontario, Canada) with a 65 dB SPL speech signal presented from 0-degree azimuth in quiet. The status of DM/NR features (on vs. off) of basic and premium HAs were manipulated to create four HA conditions. The REARs were equalized across all conditions.

Note that none of the DM/NR features would be activated during the REAR measurement (speech presented from 0-degree azimuth in quiet in a low-reverberant laboratory). Therefore equalizing the REARs across the four HA conditions would not eliminate the effect of the DM/NR features. The frequency-lowering feature was disabled. All other features (e.g., wide dynamic range compression, adaptive feedback suppression and low-level expansion) remained active at default settings. Since the main purpose of the study was to determine the effect of premium DM/NR features, the volume control of the device was disabled.

Because HA users often do not switch between programs (or memories) of the device (Kuk 1996; Cord et al. 2002), it was desirable that all DM/NR features evaluated in the current study could be enabled or disabled automatically in the default program, minimizing the likelihood that participants would not use the features in the field trial. Therefore, in the two feature-on conditions (basic-on and premium-on), the default program was configured with all DM/NR features set to the automatic mode as recommended by the manufacturer. In the premium HA, the speech-seeking DM could not be enabled as part of the automatic DM/NR features in the default program. For this reason, the speech-seeking DM was added in a manual program for the premium-on condition, with the directivity steering mode (i.e., steering the directivity lobe to left, right, or back) set to automatic. Participants were instructed to use this program in situations where speech was not from in front and they were unable to turn their faces toward the talker of interest, such as in the car. In keeping with the premium-on condition, the basic-on condition included a manual program with the adaptive DM enabled. Participants were instructed to use this program in noisy situations where speech was from in front. Note that although HA features may not yield their maximum benefits unless users know how the features work and when to use them, extensive instruction and training on features could reveal the technology level to participants and compromise the blinding of the study. Therefore, for both premium and basic HAs, the instructions about the manual program were brief and did not involve any technology details. Questions regarding the exact programming of the devices were discouraged to minimize bias. The strength and sensitivity of DM/NR features was set to default as recommended by the manufacturer.

In the two feature-off conditions (basic-off and premium-off), all DM/NR features were turned off in the default program. Since all DM/NR features were disabled, no manual program was used. A fake manual program that was identical to the default program was not used. This is because should the participants have obtained an impression that there was no difference between the default and manual programs, this impression could reduce their willingness to switch between programs in the two feature-on conditions. Participants were told that the devices were fully automatic and did not have selectable programs. Again, the explanations about the (lack of) manual program were brief and did not involve any technology details.

Laboratory tests

Speech recognition test—Speech recognition test conditions were designed to examine the efficacy of premium DM features, including multichannel adaptive DM, bilateral

beamformer, and speech-seeking DM. The Hearing in Noise Test (HINT; Nilsson et al. 1994), which is an adaptive SNR sentence recognition test, was used. During testing, the participant was asked to repeat a block of 20 sentences against speech-shaped noise. The HINT noise was fixed at 65 dBA and was presented continuously during the testing. The speech level was adjusted adaptively depending on the listener's responses using the one-up-one-down procedure. The correct response for each sentence was based on the repetition of all the words in the sentence, with minor exceptions such as "a" and "the." The SNRs of the final 17 presentations were averaged to derive the HINT score, which was the SNR where the listeners could understand 50% of the speech. Lower HINT scores represent better performance.

The HINT was administered in a low-reverberant sound field (reverberation time = 0.21 sec) created using eight Tannoy (Coatbridge, Scotland) i5W loudspeakers. The loudspeakers were placed 1.2 m from the seated participant at 0, 45, 90, 135, 180, 225, 270, and 315-degree azimuth. Five listening conditions were created and each condition was designed for a specific DM feature. See Table 2 for the details of each condition. The name of the condition suggests the characteristics of the sound field. For example, in the S0Ndiffuse condition, speech (denoted by the capital S) was presented from 0-degree azimuth and uncorrelated noise (denoted by the capital N) was presented from all eight loudspeakers. This test condition served as a baseline because even DMs with the most basic of designs can provide benefit in this scenario. For the S0Ndiscrete condition which was designed to demonstrate the effect of multichannel adaptive DMs, two band-pass filtered HINT noises (between 1 to 3 kHz) were presented from 90 and 270-degree azimuth, respectively, and a band-stop filtered noise (cutoff frequencies 1 and 3 kHz) was presented from 180-degree azimuth. These three uncorrelated noises had the same sound level and the overall level was 65 dBA. The S0Ndiffuse-BBF condition had the same sound field configuration as the S0Ndiffuse condition, but the bilateral beamformer was enabled and set to a fixed mode (rather than automatic) in the premium-on condition. In the S90Ndiffuse and S180Ndiffuse conditions, speech was presented from 90 and 180-degree azimuth, respectively, to examine the effect of the speech-seeking DM. HAs were switched to the manual program in the premium-on and basic-on conditions to enable the associated DM feature. All five listening conditions of the HINT were administered after participants completed each HA field trial (see below), with participants wearing HAs configured for that trial condition.

Paired comparisons: Listening effort and sound quality—The main purpose of the paired comparisons was to examine the effect of premium NR features, including reverberation reduction, impulse noise reduction, wind noise reduction, and spatial noise reduction algorithms. In short, speech or speech in noise stimuli were recorded through the HAs. The recorded stimuli were presented to participants via earphones to assess their preference in terms of listening effort and sound quality using a paired comparison paradigm.

The stimuli were recorded using the basic and premium HAs with DM/NR features configured to the feature-on and feature-off conditions. HAs were programmed to fit NAL-NL2 targets for a bilateral, symmetrical sloping hearing loss (thresholds 30, 40, 50, 60, and 65 dB HL for octave frequencies from 250 to 4000 Hz). The recorded stimuli were then

processed to compensate for each participant's hearing loss using the procedures described by Wu et al. (2013b). In short, each participant's NAL-NL2 targets for a 65-dB SPL speech input were compared to the REAR targets for the sloping hearing loss used in the recording. The difference was compensated by applying a filter, one for each ear of each participant, to shape the spectrum of the recorded stimulus.

During stimuli recording, a pair of HAs were coupled to the two ears of a manikin (Knowles Electronics Manikin for Acoustic Research; KEMAR). The outputs of the HAs were recorded using a pair of G.R.A.S. (Holte, Denmark) Type RA0045 ear simulators. The recording was completed in a sound treated booth. Two Grason-Stadler (Eden Prairie, MN) sound booth loudspeakers, which were placed 1.2 m from the KEMAR at 0 and 180-degree azimuth, were used to present stimuli. Stimuli were recorded in five main listening conditions, with each condition designed to assess a specific NR feature (Table 3). Four of the five main conditions had two sub-conditions (A and B, see Table 3), resulting in a total of nine sub-conditions. For all listening conditions, sentences of the revised Speech Perception in Noise (SPIN) test (Bilger et al. 1984) were concatenated and presented from 0-degree azimuth. In the S0N0-babble condition, the babble noise of the SPIN was presented from 0-degree azimuth with the speech. The levels of speech and noise were varied to create two SNRs (6 and 0 dB) for the two sub-conditions. To create the reverberation in the S0-reverberation condition, the SPIN sentences were processed using the reverberation presets "church" (reverberation time = 2.45 sec) and "small club" (reverberation time = 1.25 sec) of the Adobe Audition version 1.0 software. Since most reverberation reduction algorithms implemented in commercially available HAs detect and suppress the reverberation tail after the offset of the sound source (Launer et al. 2016), the reverberation simulation used in the current study is likely to demonstrate the effect of this feature. For the S0N0-transient condition, the transient sounds of a can dropping on a tile floor and of a hammer hitting on a piece of wood were used. The transient sounds were first created and recorded using a Larson-Davis (Depew, NY) 2560 0.5 inch microphone. The recorded transient sounds were calibrated to 90 dBA (measured using the impulse detector of a Larson Davis 824 sound level meter) in the sound booth. The transient sounds were then presented to the KEMAR with the 60-dB SPIN sentences from 0-degree azimuth and recorded via the HAs. Each SPIN sentence was mixed with one transient sound. To generate wind for the S0N180-wind condition, an electric fan was placed at 180-degree azimuth of the KEMAR with the wind blowing at the back of the KEMAR's head. The wind speed measured at the center of the KEMAR's head was 3.8 m/s, which was lower than the average wind speed recorded in cities in United States on a typical non-windy day (4.5 m/s; Chung et al. 2010) but was high enough to activate the premium HA's wind noise reduction feature based on the manufacturer's document. The noise generated by the fan was 49.8 dBA. Finally, to demonstrate the effect of the spatial noise reduction feature, the SPIN babble noise was presented from 180-degree azimuth in the S0N180-babble conditions.

Paired comparisons were conducted in a sound treated booth. The processed stimuli were presented to participants via a pair of Sennheiser (Wedemark, Germany) IE8 insert earphones. Psychological testing software E-prime 2.0 was used to present the stimuli and collect the participant's responses. A forced-choice paired-comparison paradigm was used. In each comparison, two recordings of the same SPIN sentence, which were recorded

through two different feature conditions of a given HA model (e.g., basic-on vs. basic-off), were presented. In other words, the feature-on condition was compared to the feature-off condition within each of the basic and premium HAs. After listening to the stimuli, participants clicked the buttons on a computer monitor to express their preference based on one of the three criteria: listening effort, sound naturalness, and sound annoyance (Brons et al. 2013). For each judgment criterion, the comparison was repeated two times using different SPIN sentences. The presentation order of listening condition, HA model, feature status, and judgment criteria was randomized. The entire test consisted of 108 comparisons (9 listening sub-conditions x 2 HA models x 3 preference criteria x 2 presentations). The entire test was repeated four times, one at the end of each HA field trial condition, resulting in a total of 432 comparisons for each participant (108 comparisons x 4 trial conditions).

The probability of a participant selecting the feature-on condition was calculated for each preference criterion and each listening sub-conditions. To simplify data presentation and analysis, for each preference criterion, the results of the two sub-conditions in a given main listening condition (e.g., the A and B sub-conditions of the S0N0-babble; see Table 3) were averaged. The data of sound naturalness and annoyance were further averaged for each participant. The averaged results were referred to as laboratory sound quality results in this paper.

Scale rating: Listening effort and sound quality—Although paired comparison is a sensitive test paradigm, it did not take into account the possible adjustment or acclimatization effect that could have occurred during the field trial. Therefore, the effect of DM/NR features on listening effort and sound quality was also measured using a scale-rating paradigm. After the field trial of each HA condition, participants were seated in the same sound booth that was used to record the stimuli for paired comparisons, wearing the HAs configured for that trial condition. The speech and/or noise shown in Table 3 were presented to participants in the same manner as the paired comparison stimuli recording. For each of the nine listening sub-conditions (Table 3), participants listened to the stimuli for 20 sec and then used a 21-point scale ranging from 0 to 10 to report the perceived listening effort, sound naturalness and sound annoyance. The ratings of listening effort and sound annoyance were then reversed so that higher ratings represented better outcomes. Similar to paired comparisons, the ratings of the two sub-conditions of a given main listening condition were averaged for each participant. The results of sound naturalness and annoyance were also averaged.

Localization test—A front/back localization test was administered to examine the effect of pinna-simulation directivity. The stimulus was 3-kHz octave band-filtered pulsed pink noises, which consisted of a 750-ms train of four pulses with a 150-ms pulse duration, a 50-ms interpulse interval, and 10-ms rise/fall times. Research has shown that this high-frequency stimulus was able to demonstrate the benefit of pinna-simulation directivity (Keidser et al. 2009). The stimuli were presented from 0 or 180-degree azimuth of participants. The stimuli were presented at 70 dB SPL with a ±3 dB roving effect added to the base level. The participants were seated in the sound treated booth that was used for scale ratings. Participants wore HAs and were instructed not to move their heads during the

experiment. Before testing, each participant completed practice sessions where feedback was provided. During testing, the stimuli were presented randomly from the front or back loudspeakers with equal probability. The participants indicated whether the stimulus was coming from front or back. No feedback was provided. In total 56 trials were conducted. The localization test was administered at the end of each HA trial condition, with participants wearing HAs configured for that trial condition. Localization accuracy (percent correct) was calculated for each participant.

Retrospective self-reports

Abbreviated Profile of Hearing Aid Benefit (APHAB)—The APHAB (Cox & Alexander 1995) is a 24-item inventory designed to evaluate benefit experienced from HA use and to quantify the degree of communication difficulty experienced in various situations. The questionnaire consists of four subscales. The ease of communication, background noise, and reverberation subscales are focused on speech communication and therefore the global score of the APHAB is the mean of the scores of these three subscales. The aversiveness (AV) subscale evaluates the individual's response to unpleasant environmental sounds. The global score (referred to as APHAB-Global) and the AV subscale score (APHAB-AV) were used in data analysis.

Speech, Spatial, and Qualities (SSQ) hearing scale.—The 49-item SSQ (Gatehouse & Noble 2004) is a validated scale designed to measure a range of hearing disabilities across several domains. The SSQ consists of three subscales that measure the ability of an individual to understand speech (denoted as SSQ-Speech), to localize acoustic events (SSQ-Spatial), and to evaluate auditory experience including music perception and the clarity and naturalness of sound (SSQ-Qualities). The SSQ-Qualities also contains three items that directly assess listening effort (items 14, 15, and 18) and have been referred to as the listening effort subscale (Dawes et al. 2014; Alhanbali et al. 2017). This subscale was referred to as SSQ-Effort in this paper. The scores of these four subscales were used in analysis.

Satisfaction with Amplification in Daily Life (SADL)—The SADL (Cox & Alexander 1999) is a 15-item inventory designed to evaluate an individual's satisfaction with his/her HAs. The questionnaire is divided into four subscales. The positive effect subscale quantifies improved performance while using HAs, such as reduced communication disability. The personal image subscale evaluates the domain of self-image and stigma. The negative features subscale assesses undesirable aspects of HA use, such as feedback problems. The service and cost subscale measures the adequacy of service provided by the professional and the cost of the devices. The mean of the scores for all items (except for the item related to cost, as HAs were provided at no cost in the current study) formed the global score (SADL-Global) and was used in data analysis.

The three retrospective questionnaires were administered at the end of each HA trial condition. The participants were asked to retrospectively recall their listening experiences during the field trial. For all questionnaires, the original scores (global or subscale scores)

were linearly transformed so that the score ranged from 0 to 10, with higher scores representing better outcomes.

In-situ self-reports

The EMA (i.e., the ecological momentary assessment) methodology was used to collect the participant's real-time experience with HAs in real-world listening situations. In the current study, EMA was implemented using Samsung (Seoul, South Korea) Galaxy S3 smartphones (i.e., smartphone-based EMA). Smartphone application software (i.e., app) was developed to deliver electronic surveys (Hasan et al. 2013). The app prompted participants to complete surveys at randomized intervals approximately every two hours within a participant's specified time window. The 2-hr inter-prompt interval was selected because it seemed to be a reasonable balance between participant burden, compliance, and the amount of data that would be collected (Stone et al. 2003). Participants were instructed to answer survey questions based on their experiences during the past five minutes so that recall bias was minimized. Participants were also encouraged to initiate a survey whenever they had a new listening experience lasting at least 10 min.

In each survey, a series of questions regarding listening environments and experience were asked. The survey was designed for the larger study but only the questions that are relevant to the current study are reported in this paper. See Hasan et al. (2014) for the complete set of survey questions. The survey first asked "*Were you listening to speech?*" and provided two options for the participants to select (yes/no). If participants were not listening to speech, the survey then asked "*What were you listening to?*" and provided two options (non-speech sound listening/not actively listening). The survey then presented a question to assess the noisiness level ("*How noisy was it*?", quiet/somewhat noisy/noisy/very noisy). This question was included because the effect of DM/NR features is a function of noise level or SNR (Walden et al. 2005; Wu & Bentler 2010a). Participants also answered a question regarding HA use ("*Were you using hearing aids?*", yes/no). For these questions, participants tapped a button on the smartphone screen to indicate their responses.

Next, the app presented a series of questions to assess the participant's listening experience. The first question assessed the participant's speech understanding ("*How much speech did you understand?*"). The response was collected using a visual analog scale with two anchors (from "0%" to "100%"). This question is referred to as EMA-Speech in this paper. The app then presented questions to assess listening effort (EMA-Effort; "*How much effort was required to listen effectively?*", from "very easy" to "very effortful"), loudness satisfaction (EMA-Loudness, "*Were you satisfied with the loudness?*", from "not good at all" to "just right"), sound localization (EMA-Localization, "*Could you tell where sounds were coming from?*", from "not at all" to "perfectly"), and HA satisfaction (EMA-Satisfaction, "*Were you satisfied with your hearing aids?*", from "not at all" to "very satisfied"). The survey questions were presented adaptively such that certain answers determined whether follow-up questions would be elicited. For example, the EMA-Speech would be presented only when participants indicated that they were listening to speech in the beginning of the survey. The EMA-Effort would be presented when participants indicated that they were actively listening to either speech or non-speech sounds. The EMA-Satisfaction would be asked only when

participants indicated that they were using HAs. For all listening experience questions, participants used the sliding bar on the visual analog scale to mark their perception. The ratio of the distance between the left end of the scale and the participant's mark to the length of the entire scale defined the score. All scores were linearly transformed so that the score ranged from 0 to 10, with higher scores representing better outcomes. Finally, to obtain an insight about the overall HA outcomes in the real world, the scores of the five EMA questions that assessed listening experience (EMA-Speech, -Effort, -Loudness, -Localization, and -Satisfaction) of a given survey were averaged. This variable was referred to as EMA-Global.

Procedures

The study was approved by the Institutional Review Board at the University of Iowa. After signing the consent form, participants' hearing thresholds were measured using pure-tone audiometry. If participants met all of the inclusion criteria, earmold impressions were taken by an audiologist. Next, demonstrations of how to work and care for the smartphone, as well as taking and initiating EMA surveys on the phone, were provided. Participants were instructed to respond to the auditory/vibrotactile prompts to take surveys whenever it was possible and within reason (e.g., not while driving). Participants were also encouraged to initiate a survey during or immediately after they had a new listening experience lasting at least 10 min. Each participant was given a set of take-home written instructions detailing how to use and care for the phone, as well as when and how to take the surveys. Once all of the participants' questions had been answered and they demonstrated competence in the ability to perform all of the related tasks, they were sent home with one smartphone and began a three-day practice session. Participants returned to the laboratory after the practice session. If participants misunderstood any of the EMA/smartphone related tasks during the practice session, they were re-instructed on how to properly use the equipment or take the surveys.

Next, the HAs of all four conditions were fit and the first field trial condition began. The order of the four HA trial conditions was randomized across participants. Participants were allowed to return to the laboratory for HA gain adjustments during the first two weeks of the first field trial condition. If adjustments were needed, the same gain modifications were applied in the other three HA trial conditions so that the REARs were equalized across the four trial conditions. The settings of the DM/NR features (e.g., strength and sensitivity) were not adjusted in order to blind the participants. Twelve out of 54 participants requested gain adjustment. Their first HA conditions were basic-off (n = 2), basic-on (n = 3), premium-off (n = 4) and premium-on (n = 3). In each HA condition, participants familiarized themselves with the hearing instrument settings for four weeks. Participants then returned to the laboratory and the first part of the laboratory testing (paired comparisons and scale ratings) was conducted. Participants were then given smartphones and the assessment week in which participants carried smartphones to conduct EMA surveys began. Participants were encouraged to go about their normal daily routines during the week. One week later, participants brought the smartphones back to the laboratory and the second part of the laboratory testing (the HINT and the localization test) and retrospective questionnaires were administered. HAs were inspected, cleaned, and reprogrammed. The functionality of DM

and NR features was verified in the test box of the Audioscan Verifit. Then, the instructions about the device's manual program were provided and the field trial of the next HA condition began. Figure 2 shows the flow diagram for the study.

Recall that the larger study also included an unaided condition. Twenty-nine participants (5 experienced HA users and 24 new users) agreed to complete the unaided condition (four weeks without wearing HAs plus one week for EMA) in addition to the four HA conditions. For these participants, the order of the unaided condition was randomized with the four HA conditions. Monetary compensation was provided to the participants upon completion of the study.

Data organization and reduction

The rich set of laboratory tests, retrospective self-reports, and in-situ self-reports allowed HA outcomes to be assessed comprehensively from different perspectives, though at the potential expense of increased data complexity. To facilitate result presentation and to reduce the number of measures to a more manageable set, the outcome measures were grouped into five domains (Table 4). Specifically, the speech understanding domain included the HINT, the APHAB-Global, the SSQ-Speech, and the EMA-Speech. The APHAB-Global and the SSQ-Speech scores (transformed score, ranging from 0 to 10) were further averaged for each participant because they were significantly correlated ($r = 0.59$ to 0.76 across four HA conditions, all p-values <0.0001). The listening effort domain included paired comparisons and scale ratings in which listening effort was used as the judgment criterion. This domain also included the SSQ-Effort and the EMA-Effort. For the sound quality domain, the sound quality results, which were the average of sound naturalness and annoyance data, from the paired comparisons and scale ratings were included. This domain also included the APHAB-AV and the SSQ-Qualities; these two subscale scores were not averaged because they were not correlated ($r = 0.08$ to 0.22, $p = 0.10$ to 0.54). The localization domain included the front/back localization test, the SSQ-Spatial, and the EMA-Localization. Finally, the satisfaction domain consisted of the SADL-Global and the EMA-Satisfaction and did not include any laboratory tests. Note that the domains shown in Table 4 were created to facilitate data presentation and therefore the measures included in a given domain may not assess identical construct. For example, in the sound quality domain, paired comparisons and scale ratings evaluated sound naturalness and annoyance, the SSQ-Qualities contained items regarding sound segregation, while the EMA-Loudness assessed loudness satisfaction.

RESULTS

Recall that the current study had four HA conditions (basic-off, basic-on, premium-off, and premium-on). To evaluate the relative effect of premium DM/NR features compared with basic DM/NR features, linear mixed models were used. The models included a random intercept for subject to account for the repeated observations per participant, with fixed effects being HA model (basic vs. premium), feature status (off vs. on), and the interaction between HA model and feature status. A significant interaction indicates that the effect of premium DM/NR features differs from the effect of basic DM/NR features. To determine the effect of premium HA relative to basic HA and the effect of DM/NR features relative to no

features, a similar linear mixed model that used HA condition (four levels) as the fixed effect was used. The outcomes of the premium-on and basic-on conditions were compared to determine the effect of premium HAs; the outcomes of the feature-on and feature-off conditions were compared within each HA model to determine the effect of DM/NR features. Finally, the outcomes of basic-off and premium-off conditions were also compared; the outcomes of these two conditions should be very similar because the basic and premium HAs used the same hardware and chip. To adjust for multiple comparisons, a Tukey-Kramer *p*-value adjustment was used. Statistical software SAS version 9.4 was used for all analyses.

Laboratory testing

Speech understanding domain—The mean HINT score, the SNR at which the participants could understand 50% of the speech, of each HA condition as a function of five listening condition is shown in Figure 3A. Circles and triangles represent basic and premium HAs, respectively. Solid and open symbols represent the feature-off and feature-on conditions, respectively. The y-axis has been reversed so that the top of the figure represents better performance. Separate linear mixed models were created for each listening condition and the analysis results are shown in the figure. Significant interaction ($p < 0.05$) between HA model and feature status is indicated by "INT" and significant difference ($p < 0.05$) between HA conditions is indicated by bracket. Detailed statistics are available in the appendix.

Results from the models first indicated that the interactions in all five listening conditions were significant ($p = <0.0001$ to 0.02), suggesting that the effect of premium DM/NR features on improving speech recognition performance was larger (better) than the effect of basic DM/NR features. Linear mixed models further revealed that the HINT score of the premium-on condition was better (lower) than that of the basic-on condition in four out of five listening conditions (indicated by wide brackets in Figure 3A; adjusted $p = <0.0001$ to 0.016), supporting that the premium HAs outperformed the basic HAs. In seven out of ten comparisons between the feature-on and feature-off conditions, the feature-on conditions yielded better (lower) HINT scores (indicated by narrow brackets in Figure 3A; adjusted $p = <0.0001$ to 0.0007), suggesting the beneficial effect of DM/NR features relative to no features. However, basic DM/NR features had a detrimental effect on speech understanding in the S180Ndiffuse condition (adjusted $p < 0.0001$). Finally, none of the differences between the premium-off and basic-off conditions was statistically significant (adjusted $p = 0.06$ to 0.99).

The results of Figure 3A were summarized in Table 5. In short, because the interactions in all five listening conditions were statistically significant, "P > B" was used in Table 5 to indicate the effect of premium DM/NR features compared to the basic features (denoted by "Feature: P vs B" in Table 5, letters P and B denote premium and basic, respectively). Next, because the HINT scores of the premium-on conditions were either better (lower) than or equal to the basic-on conditions, "P ≥ B" was used to indicate the effect of the premium HA (denoted by "HA: P vs B" in Table 5). Finally, because the feature-on conditions were either better (lower) than or equal to the feature-off conditions except for the basic HA in the S180Ndiffuse condition, "On ≥ Off" was used to suggest the effect of DM/NR features

(denoted by “Feature: On vs Off” in Table 5) and a footnote was used to indicate the exception.

Listening effort domain—The laboratory tests included in this domain were paired comparisons and scale ratings based on the listening effort judgment criterion (Table 4). For paired comparisons, recall that the comparisons were made between the feature-on and feature-off conditions within each HA model. Figure 4A shows the probability of a participant selecting the feature-on condition of each HA model as a function of listening condition. The dashed line represents chance probability (0.5). Linear mixed models with a random intercept were used to determine if the probability of the premium-on condition being preferred (over premium-off) was different from the probability of the basic-on condition being preferred (over basic-off). Significant differences ($p < 0.05$) are labeled using brackets in the figure. Because comparisons were made within each HA model, paired comparison data could not determine the effect of premium HA relative to basic HA. Finally, to examine the effect of DM/NR features relative to no features, linear mixed models were implemented to determine if the probability of a participant preferring the feature-on condition over the feature-off condition was significantly different from the chance level. Significant differences ($p < 0.05$) are labeled using asterisks in Figure 4A. Detailed statistics are available in the appendix.

Model results suggest that the probability of the premium-on and basic-on conditions being selected were not different in all listening conditions ($p = 0.32$ to 0.79), except for the S0N180-wind condition (bracket in Figure 4A; $p < 0.0001$). However, in the S0N180-wind condition the probability of the premium-on condition being preferred was significantly lower than the chance level (asterisks in Figure 4A; $p < 0.0001$), indicating a detrimental effect of premium DM/NR features. For both basic and premium HAs in the S0N0-babble and S0N180-babble conditions, the probability of feature-on condition being preferred was significantly higher than the chance level (all $p < 0.0001$), suggesting the effect of DM/NR features relative to no features. A small beneficial effect was also observed for premium HAs in the S0N0-transient condition ($p = 0.037$). No effect in the S0-reverberation condition was statistically significant.

Figure 4B shows the results of scale ratings in the listening effort domain. Higher ratings represent better outcomes. The significant interaction of the S0N0-babble condition (“INT” in Figure 4B; $p = 0.044$) indicates that the effect of premium DM/NR features on improving listening effort was better than that of basic features. In the S0N180-babble condition, although the interaction between HA model and feature status was not significant at the 0.05 level ($p = 0.058$), the feature-on condition had higher ratings than the feature-off condition for both HA models (brackets in Figure 4B; adjusted $p = <0.0001$ and 0.0024). No other interactions and effects were statistically significant. The laboratory results of the listening effort domain, paired comparisons and scale ratings combined, are summarized in Table 5.

Sound quality domain—Results of paired comparisons and scale ratings in the sound quality domain are shown in Figures 4C and 4D, respectively. These results were similar to those observed in the listening effort domain in all listening conditions, except for the S0N0-transient and S0N0-babble conditions. Specifically, in the S0N0-transient condition of the

paired comparisons, the premium-on condition was preferred more often than the basic-on condition (bracket in Figure 4B; $p < 0.0001$) and the probability of preferring the basic-on condition was lower than the chance level ($p = 0.0051$). The sound quality results of the S0N0-babble condition of scale ratings (Figure 4D) differed from the listening effort results of the same condition (Figure 4B) in that none of the interaction and effects were significant. See Table 5 for the result summary.

Localization domain—Figure 3B shows the front/back localization accuracy (percent correct) of each HA condition. Linear mixed models indicated that the interaction between HA model and feature status was significant ("INT" in Figure 3B; $p = 0.033$). Mixed models further indicated that the difference between the premium-on and basic-on conditions was not statistically significant (adjusted $p = 0.14$). However, the localization accuracy of the feature-on condition was higher than that of the feature-off condition when participants were wearing premium HAs (bracket in Figure 3B; adjusted $p = 0.002$).

Retrospective self-reports

Figure 5 shows the mean score of retrospective self-reports. Recall that all scores have been linearly transformed so that the score ranged from 0 to 10, with higher scores representing better outcomes. Linear mixed models first revealed that none of the interactions between HA model and feature status was significant ($p = 0.058$ to 0.90), nor was the difference between the premium-on and basic-on conditions (adjusted $p = 0.32$ to 0.98). In contrast, the feature-on score was significantly higher (better) than the feature-off score for premium HAs in the speech understanding domain (i.e., the mean of the APHAB-Global and SSQ-Speech; $p < 0.0001$) and the sound quality domain (the APHAB-AV; adjusted $p = 0.021$). The APHAB-AV score of the premium-off condition was also found to be significantly lower (poorer) than that of the basic-off condition (adjusted $p = 0.036$). Finally, for both basic and premium HAs, participants were more satisfied with HAs when the DM/NR features were turned on (both adjusted $p = 0.007$). Results of Figure 5 are summarized in Table 5. Detailed statistics are available in the appendix.

In-situ self-reports

Across the four HA conditions, a total of 8608 EMA surveys were completed by the 54 participants. On average each participant completed 5.69 surveys per day. Since the main focus of the current study was the effect of HA features, the surveys that were completed when participants did not wear HAs were excluded. For the remaining 7579 surveys, 5000 (66%) were prompted by the EMA app and the 2579 (34%) were initiated by the participants. It was determined a priori that both the app-initiated and participant-initiated surveys would be pooled together for analysis.

Recall that in each EMA survey participants reported the noisiness level in four categories. Among the 7579 surveys used in analyses, "quiet," "somewhat noisy," "noisy," and "very noisy" were reported 56.4%, 32.8%, 8.1%, and 2.7% of the time, respectively. Because the effect of DM/NR features is a function of noise level (Walden et al. 2005), it was determined a priori that the EMA data would be analyzed separately in each noisiness category. However, comparing the EMA data across the four HA conditions within each noisiness

category would be less meaningful if participants did not report noisiness level consistently across the entire field trial or if HA condition had an effect on how participants reported nosiness level. Therefore, analysis was first conducted to determine the effect of HA condition on the distribution of the four noisiness levels. A linear mixed model with a random intercept was used to model the number of the surveys completed by a participant using the covariates of noisiness level, HA condition, and their interaction. The result revealed that the interaction between noisiness and HA condition was not significant (p = 0.83), suggesting that the distribution of the survey numbers across the noisiness levels were similar in the four HA conditions. Because HA condition was unlikely to affect how participants reported the four noisiness levels used in the current study, EMA data across the four HA conditions could be compared in each noisiness level.

Speech understanding domain—Recall that there were five EMA survey questions assessing a participant's listening experience, one for each outcome domain (Table 4). Figure 6A shows the mean score of the EMA-Speech as a function of noisiness level. The scores ranged from 0 to 10, with higher scores representing better outcomes. The number of surveys completed in each noisiness category is also shown in the figure (four HA conditions combined). Linear mixed models indicated that none of the interactions and effects were significant.

Listening effort, sound quality, and localization domains—Figures 6B to 6D show the mean outcome scores of the EMA-Effort, the EMA-Loudness, and the EMA-Localization. The significant interaction between HA model and feature status in the EMA-Loudness ("INT" in somewhat noisy category of Figure 6C; p = 0.002) indicated that premium DM/NR features yielded better outcomes relative to basic features. However, the significant interactions in the EMA-Effort (Figure 6B; quiet category; p = 0.01) and the EMA-Localization (Figure 6D; noisy category; p = 0.0009) suggested the opposite. In terms of the effect of premium HA, the premium-on condition yielded higher scores than the basic-on condition in the somewhat noisy category of the EMA-Loudness (wide bracket in Figure 6C; adjusted p = 0.013). As for the comparison between the feature-on and feature-off conditions, all significant differences (feature-on better than feature-off) stemmed from the basic HAs (adjusted p = 0.0009 to 0.03), except for the premium HA in the somewhat noisy category of the EMA-Loudness (Figure 6C; adjusted p = 0.014).

Satisfaction domain—The EMA-Satisfaction results shown in Figure 6E revealed that the interaction between HA model and feature status was significant in the somewhat noisy category (p = 0.015), suggesting that the effect of premium DM/NR features on improving user's satisfaction was larger than that of basic features. Although none of the differences between the premium-on and basic-on conditions was significant (adjusted p = 0.14 to 0.88), participants were more satisfied with HAs in the feature-on than the feature-off conditions (adjusted p = < 0.0001 to 0.035). The exceptions were the basic HAs in the somewhat noisy category (adjusted p = 0.094) and the premium HAs in the very noisy category (adjusted p = 0.051). The results of in-situ self-reports in each outcome domain are summarized in Table 5. Detailed statistics are available in the appendix.

EMA-Global—Figure 6F shows the mean EMA-Global score as a function of noisiness level. The interaction in the noisy category was significant ($p = 0.042$), suggesting that the participants perceived basic DM/NR features to be better than premium features. None of the differences between the premium-on and basic-on conditions was significant (adjusted $p = 0.28$ to 0.98). Finally, three out of the four significant differences between the feature-on and feature-off conditions stemmed from the basic HA (adjusted $p = 0.0004$ to 0.0012).

DISCUSSION

The purpose of the current study was to determine the effect of premium DM/NR features relative to basic DM/NR features, the effect of premium HAs relative to basic HAs, and the effect of DM/NR features relative to no DM/NR features. Outcomes in five domains were measured using laboratory tests, retrospective self-reports, and in-situ self-reports.

Laboratory tests

For the speech understanding domain, the HINT results consistently supported the beneficial effect of premium DM/NR features, premium HAs, and DM/NR features (Figure 3A). Although in the current study the effects of DM and NR features were not examined separately, it is likely that the benefit observed in the HINT was mainly from DM technologies. The robust findings across all five listening conditions agree with previous literature regarding the efficacy of multi-channel adaptive DMs (Blamey et al. 2006), bilateral beamformers (Picou & Ricketts 2018), and speech-seeking DMs (Wu et al. 2013b).

For the listening effort and sound quality domains, the results were mixed (Figure 4). For example, although the significant interaction in the S0N0-babble condition of scale ratings supported the effect of premium DM/NR features on reducing listening effort (Figure 4B), this effect was not observed in the paired comparisons (Figure 4A) and in the sound quality domain (Figures 4C and 4D). Regardless, the paired comparison data obtained from the S0N0-babble condition were consistent with the literature (e.g., Ricketts & Hornsby 2005; Sarampalis et al. 2009; Wendt et al. 2017), suggesting that NR features could reduce listening effort and improve sound quality. For the S0-reverberation condition, none of the interactions and comparisons were significant, suggesting that the reverberation reduction algorithm of the premium HA had a minimal effect in the test conditions used in the current study. Recall that the reverberation was created using software simulation. The reverberation reduction algorithm might work differently in a true reverberant sound field. For the S0N0-transient condition, the paired comparison results were in line with the literature (Korhonen et al. 2013), supporting the effect of impulse noise reduction algorithms in improving sound quality.

Unexpectedly, paired comparisons of the S0N180-wind condition suggested that the premium wind noise reduction algorithm had a detrimental effect on listening effort and sound quality (Figures 4A and 4C). After examining the recorded stimuli used in the test, it was found that although the premium feature effectively attenuated the level of wind noise, it also generated artifacts and highly degraded the speech. It has been suggested that the wind noise reduction strategy could reduce wind noise at the cost of worsening low-frequency speech audibility, resulting in limited benefit (Ricketts et al. 2018). Newer wind noise

reduction strategies that wirelessly route the signal from the HA with less wind noise to the other HA to replace the signal with more wind noise (Latzel & Appleton 2013) and that use least mean squares filtering to attenuate wind noise (rather than simply reduce low frequency gain) (Korhonen et al. 2017) could be more effective in reducing wind noise while preserving the quality or audibility of speech. Of note, the results of the S0N180-wind condition shown in Figure 4 should be viewed as the combined effect of the wind noise reduction and gain-reduction NR. This is because the noise of the fan used to generate wind (49.8 dBA) would trigger the gain-reduction NR. A wind tunnel should be used to record stimuli for paired comparisons in order to more precisely assess the effect of the wind noise reduction feature. Finally, although the data obtained from the S0N180-babble condition strongly supported the effect of DM/NR features relative to no features, there was no evidence to support the benefit of the spatial noise reduction algorithm.

For the localization domain (Figure 3B), premium DM/NR features improved front/back localization accuracy more than basic features did. This is consistent with the literature (Keidser et al. 2009) supporting the effect of pinna-simulation directivity. On the other hand, although the localization accuracy of the premium HA was higher than the basic HA by 8.5%, this difference was not statistically significant (adjusted $p = 0.14$).

Table 5 summarizes all laboratory tests of the current study. In short, although premium DM/NR features examined in the study could have a detrimental effect in windy situations wherein speech is presented, the laboratory tests of the current study generally supported the beneficial effect of premium DM/NR features, premium HAs, and DM/NR features, especially in the speech understanding and localization domains.

Retrospective self-reports

Contrary to laboratory tests, retrospective self-reports did not demonstrate any differential effect of premium DM/NR features compared with basic features across all outcome domains (Figure 5). Furthermore, consistent with the research by Cox and her colleagues (Cox et al. 2014, 2016; Johnson et al. 2016, 2017), premium HAs and basic HAs yielded similar real-world outcomes. Retrospective self-reports, however, did indicate that DM/NR features significantly improved speech understanding (by 0.33 points, or 3.3%, for premium HAs) and satisfaction (by 4.4% and 4.2% for premium and basic HAs, respectively), although the small amount of improvement might not be clinically important. Retrospective self-reports also indicated that DM/NR features significantly improved APHAB-AV scores for premium HAs. However, this effect was due to the poor APHAB-AV scores in the premium-off condition. The reason for this poor APHAB-AV score is unclear.

In short, although the effect might not be considered clinically important, the retrospective self-report results supported the effect of DM/NR features in the real world, especially in the satisfaction domain. However, the evidence supporting the benefit of premium DM/NR features and premium HAs was limited (Table 5).

In-situ self-reports

Among the 7579 surveys used in analyses, only 10.9% were completed in noisy or very noisy situations. This is consistent with previous research showing that older adults with

hearing loss spent most of their time in quiet (Wu & Bentler 2012a). In quiet and somewhat noisy situations, the mean EMA-Speech scores were quite high (i.e., > 8 points, or 80%). Therefore, it is not surprising that the EMA-Speech scores were essentially the same across the four HA conditions in the quiet, the somewhat noisy and even in the noisy categories (Figure 6A). Research has shown that the benefit of HA features would be minimal when speech understanding is at the ceiling level (Walden et al. 2005; Wu & Bentler 2010a, 2012b). In the very noisy category, the feature-on scores were higher than feature-off scores by 14% (i.e., 1.4 points) and 9% for basic and premium HAs, respectively. However, the difference was not statistically significant which may be due to the small number of survey responses in this noisiness category.

The statistical results of the EMA-Effort, the EMA-Loudness, and the EMA-Localization were not consistent across noisiness levels (Figures 6B to 6D). This was in part due to the substantial variation in sample size (number of completed surveys) across the noisiness categories. The large sample size in the quiet and somewhat noisy categories allowed the analysis to detect very small outcome variation across the HA conditions, while the large outcome difference in noisier situations might not be statistically significant due to the small sample size. Since statistical results are confounded by sample size, the data pattern across HA conditions and across noisiness levels was inspected to obtain insight regarding the effect of features and HAs. For the EMA-Loudness (Figure 6C), there seemed to be a pattern of higher scores for feature-on than feature-off conditions, except for the somewhat noisy category. For the EMA-Localization (Figure 6D), the outcome scores of the four HA conditions were almost identical in the quiet, somewhat noisy, and noisy categories, while in the very noisy category the DM/NR features seemed to provide some benefit relative to no features. In contrast, the data of the EMA-Effort (Figure 6B) did not show a clear pattern, especially across the noisy and very noisy categories. This could be due to that the direction of the visual analog scale of the EMA-Effort (right side indicating more effortful) was opposite to other EMA questions. Participants might accidentally answer this question in the wrong direction, resulting in less reliable data. To summarize, the data pattern of the EMA-Effort, the EMA-Loudness, and the EMA-Localization seemed to suggest that the feature-on condition had better outcomes than the feature-off condition. However, because statistical analyses results were mixed, there was no strong evidence to support any effects of features or HAs on listening effort, sound quality, and localization.

In contrast, the results of the EMA-Satisfaction were more consistent and robust (Figure 6E). Participants were more satisfied with HAs in the feature-on conditions than the feature-off conditions across all noisiness levels. In very noisy situations, the DM/NR features could improve satisfaction by 1.2 points (or 12%, basic HAs), which is considered clinically important. Finally, the results of the EMA-Global (Figure 6F) were in line with the EMA-Satisfaction supporting the effect of DM/NR features. Neither the EMA-Satisfaction nor the EMA-Global provided robust evidence to support the effect of premium DM/NR features and the effect of premium HAs.

In short, in-situ self-reports of the current study supported the effect of DM/NR features relative to no features. The evidence supporting the effect of premium DM/NR features and premium HAs, however, was weak (Table 5).

Discrepancy between laboratory and real-world outcomes

Similar to previous studies that examined the effect of HA features (e.g., Gnewikow et al. 2009), the current study indicated that the effect of premium DM/NR features and premium HAs observed in the laboratory did not translate to the real world. Several reasons could explain the discrepancy between laboratory and real-world outcomes. First, although statistically significant in the laboratory, the benefit of premium features might not be large enough to be noticed in the real world. For example, the just-noticeable difference of SNR measured in well-controlled listening conditions in the laboratory is 3 dB (McShefferty et al. 2015). Because most differences in HINT scores across the four HA conditions were smaller than 3 dB (Figure 3A), participants might not notice the difference in the real world and therefor did not report it in retrospective and in-situ self-reports.

Second, the listening situations wherein premium DM/NR features and premium HAs could outperform their basic-level counterparts might not occur very often in the real world. In a study designed to characterize the behavior of automatic features in the real world, Banerjee (2011) found that the DM and NR features of a mid-level HA model (compared to the current technology level) were enabled only approximately 10% and 20% of the time, respectively. Therefore, it is unlikely that the premium DM/NR features would be enabled very often in the real world. If the benefit of premium features rarely occurred, participants might not remember or notice it enough to report it in retrospective questionnaires. In terms of in-situ self-reports, the number of the surveys that recorded the benefit of premium features would be too small for statistical analysis to detect significant differences. A smartphone-based EMA system that allows smartphones to wirelessly retrieve the feature status information from HAs and use this information to trigger the delivery of surveys (e.g., surveys are triggered when DM is activated) could address this issue in future research.

Third, the benefit of premium features could be offset by the negative effect, such as distortions and artifacts, generated by these features. For example, although the wind noise reduction feature could attenuate wind noise level, it could also degrade the speech (Ricketts et al. 2018 and Figures 4A and4B). In the current study, several participants reported that they did not appreciate premium HAs in the feature-on condition because they could hear HAs switching between processing schemes, which generated unpleasant distortions and/or artifacts.

Fourth, the benefit of premium DM/NR features and premium HAs observed in the laboratory might not be realized in the real world. For example, research has suggested that many real-world factors such as visual cues (Wu & Bentler 2010a) could decrease the benefit of DM technologies. Further, although HA's automatic algorithms may work well in simple and static sound fields in the laboratory, they may not enable or disable features appropriately in the complex and dynamic real-world. For example, research has demonstrated that HA users' voices could affect how automatic algorithms select microphone modes (omnidirectional vs. speech-seeking DM; Wu et al. 2013a). Ricketts et al. (2017) further suggested that the microphone mode (DM vs. omnidirectional) selected by HA's automatic algorithms was inappropriate 38% of the time in real-world listening situations.

EMA methodology

Although in the current study the results of the retrospective and in-situ self-reports were generally consistent, the EMA data were more informative than the retrospective questionnaires. For example, the noisiness information collected in each in-situ assessment allowed the effect of features and HAs to be examined at different noisiness levels. More specifically, the data shown in Figure 6 indicated that the difference between feature-on and feature-off scores increased as noisiness level increased. In the very noisy category, the differences (especially in the satisfaction domain) could be as large as 10% (or 1 point), which is considered clinically important and is larger than those reported by retrospective questionnaires (3% to 4%). The EMA data further suggested that although DM/NR features could yield clinically important benefit, older HA users would rarely perceive this benefit because the very noisy category occurred only 2.7% of the time in the current study. Such context-specific information is not available from retrospective questionnaires.

Although EMA seems to be a useful methodology in HA outcome research, the EMA system used in the current study has room for improvement. First, the EMA questions used in the study were created specifically for the study and, therefore, their wordings and response formats were not vigorously validated. It would be beneficial to establish and validate a set of standardized questions that can be used in future EMA research. Second, although a visual analog scale with a sliding bar allowed fine-grained data to be collected, some participants reported difficultly using them on the small touchscreens of smartphones. A five-point or seven-point scale with buttons could be a better response format for smartphone-based EMA. Third, very few surveys (11%) were completed in the noisy and very noisy categories in which DM/NR features are supposed to have a larger effect, precluding statistical analysis to detect significant differences in these situations. A context-sensitive EMA system that can use smartphones or other sensors to characterize the environments and deliver surveys more evenly across different situations could address this issue.

Limitations

The current study has several limitations. First, the volume controls of the HAs were disabled. Although this would ensure that the study results are not confounded by volume control setting, it limits the generalizability to a clinical setting. Previous research has shown that the effect of HA signal processing could be minimized if users can adjust the volume of the device (Souza & Kitch 2001). Second, the study procedures designed to blind the participants may have affected the results. For example, the DM/NR feature settings were set to the default and were not fine-tuned during the field trial. Because the premium DM/NR features had more flexibility in setting adjustment (e.g., strength and sensitivity) than the basic features, the effect of the premium DM/NR features might be underestimated in the current study. In order to blind the participants regarding the HA technology, no technology details were disclosed and minimal training on how to use the manual program was provided. However, some features (especially DM technologies) may not yield their maximum effect if users do not know how the features work and when to use them. Further, the outcomes of the feature-off conditions could be slightly inflated because the participants

were told that HAs were fully automatic (and therefore did not have selectable programs), which could lead to an impression of high-end technologies.

Third, all real-world outcomes of the current study were measured using self-reports. It is well known that self-reports are not always consistent with behavioral or physiological measures. For example, previous studies have suggested that listening effort measured using self-reported ratings are not necessarily correlated with the effort assessed using pupillometry (e.g., Zekveld et al. 2010) and dual-task paradigms (e.g., Wu et al. 2016). Therefore, even though the perceived effort is not affected by DM/NR features or HAs in the real world (Figures 5 and 6), changes in listening effort might still occur.

Fourth, the older participants in the current study, who lived in eastern Iowa and north western Illinois, could have quiet lifestyles (Figure 6). This might preclude them from perceiving the benefit of premium DM/NR features. It is unknown if the results of the current study could generalize to older adults who lived in urban areas and have very active lifestyles. Finally, the current study examined only two HA models from one major manufacturer. Therefore, although the current study (HAs released in 2013) and the research by Cox and her colleagues (Cox et al. 2014; HAs released in 2011) used different devices but generated similar results regarding the effect of premium HAs, the generalizability of the study results to other devices is unknown.

Future data analysis

Recall the current study was part of a larger study. The data collected for the larger study may help explain why the effect of premium features and HAs was observed in the laboratory but not in the real world. For example, the EMA survey designed for the larger study contained the questions that assessed the location of the listening activity (indoor, outdoor, or traffic) and the location of the talker of interest (from the listener's front, side, or back). In the HA conditions wherein the HAs had a manual program (i.e., the basic-on and premium-on conditions), the EMA survey also asked if the participant was using the default or manual program. With these data, it is possible to compare HA outcomes in specific real-world listening situations. Figure 7 shows the mean EMA-Satisfaction scores of the basic-on and premium-on conditions (1) in outdoor listening situations and (2) in the situations wherein the speech was from the side or back of the listener and the HAs were in the manual program (refer to as the "side/back speech" situations in the figure). See the figure legend for the details of these situations. It is hypothesized that the premium HA's wind noise reduction feature and speech-seeking DM (activated in the manual program of the premium HAs) could improve user's satisfaction in outdoor listening situations which are often windy and in situations wherein speech was not from the front of the listener, respectively. Figure 7 shows that the mean EMA-Satisfaction score of the premium-on condition is higher (better) than that of the basic-on condition by 7% and 9% in the two situations, respectively, consistent with the hypothesis. However, likely due to the small number of the surveys completed in these situations, none of the differences between the basic-on and premium-on conditions shown in Figure 7 are statistically significant. These results seem to suggest that the premium HAs could outperform their basic-level counterparts, despite the relatively low occurrence of these situations in the real world. More analyses on the EMA data to explore

the relationship between HA features and listening situations are needed. Further, the audio recordings collected for the larger study could be used to characterize the listening environments (Klein et al. 2018; Wu et al. 2018), which might be useful in explaining the effect of HAs and features in the real world.

CONCLUSIONS

The current study investigated the laboratory efficacy and real-world effectiveness of premium DM/NR features relative to basic features, of premium HAs relative to basic HAs, and of DM/NR features relative to no features. Outcomes regarding speech understanding, listening effort, sound quality, localization, and satisfaction were measured using laboratory tests, retrospective self-reports, and in-situ self-reports. Results of laboratory tests supported the effect of premium DM/NR features, premium HAs, and DM/NR features on improving speech understanding and localization accuracy. Laboratory data also suggested that DM/NR features could improve listening effort and sound quality compared with no features. However, although both retrospective and in-situ self-reports demonstrated that participants were more satisfied with HAs when the DM/NR features were turned on than turned off, there was no strong evidence to support the effectiveness of premium DM/NR features and premium HAs in the real world. The study has limitations that concern its generalizability, including disabled HA volume controls (which could overestimate the effect of features), minimal participant training on features, and participants' quiet lifestyles (which could underestimate the effect of features). Despite these limitations, the current study suggests that although both premium and basic DM/NR technologies evaluated in the study have the potential to improve HA outcomes, older adults with mild-to-moderate hearing loss are unlikely to perceive the additional benefits provided by the premium DM/NR features in their daily lives.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The current research was supported by NIH/NIDCD R03DC012551 (title: "Minimal Technologies for Hearing Aid Success in Elderly Adults") and the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR) 90RE5020–01-00 (title: "RERC on Improving the Accessibility, Usability, and Performance of Technology for Individuals Who are Deaf or Hard of Hearing"). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this paper do not necessarily represent the policy of NIDILRR, ACL, HHS, and the reader should not assume endorsement by the Federal Government. Portions of this paper were presented at the annual conference of the American Auditory Society, March 3, 2016, Scottsdale, AZ, USA.

Conflicts of Interest and Source of Funding:

Yu-Hsiang Wu is currently receiving grants from the National Institute on Deafness and Other Communication Disorders, the National Institute on Disability, Independent Living, and Rehabilitation Research, and the Retirement Research Foundation. Octav Chipara is currently receiving grants from the National Institute on Disability, Independent Living, and Rehabilitation Research and the National Science Foundation. Jacob Oleson is currently receiving grants from the National Institute on Deafness and Other Communication Disorders, National Heart, Lung, and Blood Institute, Department of Defense, Centers for Disease Control and Prevention, Fogarty International Center, and the Iowa Department of Public Health. The current research was supported by National

Institute on Deafness and Other Communication Disorders (R03DC012551) and the National Institute on Disability, Independent Living, and Rehabilitation Research (90RE5020–01-00).

REFERENCES

Alhanbali S, Dawes P, Lloyd S, et al. (2017). Hearing handicap and speech recognition correlate with self-reported listening effort and fatigue. Ear Hear, Published before print.

ANSI. (2010). Specification for audiometers (ANSI S3.6). New York: American national standards institute.

Bainbridge KE, & Ramachandran V (2014). Hearing Aid Use among Older United States Adults: The National Health and Nutrition Examination Survey, 2005–2006 and 2009–2010. Ear Hear, 35, 289–294. [PubMed: 24521924]

Banerjee S (2011). Hearing aids in the real world: Typical automatic behavior of expansion, directionality, and noise management. J Am Acad Audiol, 22, 34–48. [PubMed: 21419068]

Bentler RA (2005). Effectiveness of directional microphones and noise reduction schemes in hearing aids: a systematic review of the evidence. J Am Acad Audiol, 16, 473–484. [PubMed: 16295234]

Bentler RA, & Chiou LK (2006). Digital noise reduction: an overview. Trends Amplif, 10, 67–82. [PubMed: 16959731]

Bentler RA, Palmer C, & Dittberner AB (2004). Hearing-in-Noise: comparison of listeners with normal and (aided) impaired hearing. J Am Acad Audiol, 15, 216–225. [PubMed: 15119462]

Bentler RA, Wu YH, Kettel J, et al. (2008). Digital noise reduction: outcomes from laboratory and field studies. Int J Audiol, 47, 447–460. [PubMed: 18698521]

Bilger RC, Nuetzel JM, Rabinowitz WM, et al. (1984). Standardization of a test of speech perception in noise. J Speech Lang Hear Res, 27, 32–48.

Blamey PJ, Fiket HJ, & Steele BR (2006). Improving speech intelligibility in background noise with an adaptive directional microphone. J Am Acad Audiol, 17, 519–530. [PubMed: 16927516]

Boymans M, & Dreschler WA (2000). Field trials using a digital hearing aid with active noise reduction and dual-microphone directionality. Audiology, 39, 260–268. [PubMed: 11093610]

Bradburn NM, Rips LJ, & Shevell SK (1987). Answering autobiographical questions: the impact of memory and inference on surveys. Science, 236, 157–161. [PubMed: 3563494]

Brons I, Houben R, & Dreschler WA (2013). Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. Ear Hear, 34, 29–41. [PubMed: 22874643]

Chisolm TH, Johnson CE, Danhauer JL, et al. (2007). A systematic review of health-related quality of life and hearing aids: final report of the American Academy of Audiology Task Force On the Health-Related Quality of Life Benefits of Amplification in Adults. J Am Acad Audiol, 18, 151–183. [PubMed: 17402301]

Chung K, McKibben N, & Mongeau L (2010). Wind noise in hearing aids with directional and omnidirectional microphones: polar characteristics of custom-made hearing aids. J Acoust Soc Am, 127, 2529–2542. [PubMed: 20370035]

Cord MT, Surr RK, Walden BE, et al. (2002). Performance of directional microphone hearing aids in everyday life. J Am Acad Audiol, 13, 295–307. [PubMed: 12141387]

Cox RM (2005). Evidence-based practice in provision of amplification. J Am Acad Audiol, 16, 419–438. [PubMed: 16295230]

Cox RM, & Alexander GC (1995). The abbreviated profile of hearing aid benefit. Ear Hear, 16, 176–186. [PubMed: 7789669]

Cox RM, & Alexander GC (1999). Measuring Satisfaction with Amplification in Daily Life: the SADL scale. Ear Hear, 20, 306–320. [PubMed: 10466567]

Cox RM, Johnson JA, & Xu J (2014). Impact of advanced hearing aid technology on speech understanding for older listeners with mild to moderate, adult-onset, sensorineural hearing loss. Gerontology, 60, 557–568. [PubMed: 25139516]

Cox RM, Johnson JA, & Xu J (2016). Impact of hearing aid technology on outcomes in daily life I: the patients' perspective. Ear Hear, 37, e224–e237. [PubMed: 26881981]

Dawes P, Munro KJ, Kalluri S, et al. (2014). Acclimatization to hearing aids. Ear Hear, 35, 203–212. [PubMed: 24351612]

Donahue A, Dubno JR, & Beck L (2010). Guest editorial: accessible and affordable hearing health care for adults with mild to moderate hearing loss. Ear Hear, 31, 2–6. [PubMed: 20040828]

Fabry DA, & Tehorz J (2005). A hearing system that can bound back from reverberation. Hear Rev, 12, 48, 50.

Gatehouse S, & Noble W (2004). The Speech, Spatial and Qualities of Hearing Scale (SSQ). Int J Audiol, 43, 85–99. [PubMed: 15035561]

Gnewikow D, Ricketts T, Bratt GW, et al. (2009). Real-world benefit from directional microphone hearing aids. J Rehabil Res Dev, 46, 603–618. [PubMed: 19882494]

Hasan SS, Chipara O, Wu YH, et al. (2014). Evaluating auditory contexts and their impacts on hearing aid outcomes with mobile phones. In Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare (pp. 126–133). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Oldenburg, Germany.

Hasan SS, Lai F, Chipara O, et al. (2013). AudioSense: Enabling real-time evaluation of hearing aid technology in-situ. In Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (pp. 167–172). IEEE Porto, Portugal.

Holube I, Puder H, & Velde TM (2014). DSP Hearing Instruments In Metz MJ (Ed.), Sandlin's Textbook of Hearing Aid Amplification: Technical and Clinical Considerations. San Diego, CA: Plural Publishing.

Humes LE, Ahlstrom JB, Bratt GW, et al. (2009). Studies of hearing aid outcome measures in older adults: A comparison of technologies and an examination of individual differences. Seminars in Hearing, 30, 112–128.

Johnson JA, Xu J, & Cox RM (2016). Impact of hearing aid technology on outcomes in daily life II: Speech understanding and listening effort. Ear Hear, 37, 529–540. [PubMed: 27556363]

Johnson JA, Xu J, & Cox RM (2017). Impact of hearing aid technology on outcomes in daily life III: Localization. Ear Hear, 38, 746–759. [PubMed: 28700447]

Keidser G, Dillon H, Flax M, et al. (2011). The NAL-NL2 prescription procedure. Audiology Research, 1, 88–90.

Keidser G, O'Brien A, Hain JU, et al. (2009). The effect of frequency-dependent microphone directionality on horizontal localization performance in hearing-aid users. Int J Audiol, 48, 789–803. [PubMed: 19951147]

Klein KE, Wu YH, Stangl E, et al. (2018). Using a digital language processor to quantify the auditory environment and the effect of hearing aids for adults with hearing loss. J Am Acad Audiol, 29, 279–291. [PubMed: 29664722]

Kochkin S (2007). MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids. Hear J, 60, 24–50.

Kochkin S (2009). MarkeTrak VIII: 25-year trends in the hearing health market. Hear Rev, 16, 12–31.

Korhonen P, Kuk F, Lau C, et al. (2013). Effects of a transient noise reduction algorithm on speech understanding, subjective preference, and preferred gain. J Am Acad Audiol, 24, 845–858. [PubMed: 24224991]

Korhonen P, Kuk F, Seper E, et al. (2017). Evaluation of a wind noise attenuation algorithm on subjective annoyance and speech-in-wind performance. J Am Acad Audiol,, 28, 46–57. [PubMed: 28054911]

Kuk F (1996). Subjective preference for microphone types in daily listening environments. Hear J, 49, 29–35.

Latzel M, & Appleton J (2013). Evaluation of a binaural speech in wind feature, Part 2: Validation and real-life benefit. Hear Rev, 20, 36, 38, 43–44.

Launer S, Zakis JA, & Moore BCJ (2016). Hearing aid signal processing In Popelka GR, Moore BCJ, Fay RR & Popper AN (Eds.), Hearing aids (pp. 93–130). Switzerland: Springer.

Lin FR, Thorpe R, Gordon-Salant S, et al. (2011). Hearing loss prevalence and risk factors among older adults in the United States. J Gerontol A Biol Sci Med Sci, 66, 582–590. [PubMed: 21357188]

McShefferty D, Whitmer WM, & Akeroyd MA (2015). The just-noticeable difference in speech-to-noise ratio. Trends in hearing, 19, 1–9.

Mueller HG, Weber J, & Hornsby BW (2006). The effects of digital noise reduction on the acceptance of background noise. Trends Amplif, 10, 83–93. [PubMed: 16959732]

Nilsson M, Soli SD, & Sullivan JA (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. J Acoust Soc Am, 95, 1085–1099. [PubMed: 8132902]

Ohlenforst B, Zekveld AA, Jansma EP, et al. (2017). Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review. Ear Hear, 38, 267–281. [PubMed: 28234670]

Palmer C, Bentler R, & Mueller HG (2006). Evaluation of a second-order directional microphone hearing aid: II. Self-report outcomes. J Am Acad Audiol, 17, 190–201. [PubMed: 16646279]

Picou EM, & Ricketts TA (2018). An Evaluation of Hearing Aid Beamforming Microphone Arrays in a Noisy Laboratory Setting. J Am Acad Audiol, Published before print.

Preves DA, Sammeth CA, & Wynne MK (1999). Field trial evaluations of a switched directional/omnidirectional in-the-ear hearing instrument. J Am Acad Audiol, 10, 273–284. [PubMed: 10331619]

Ricketts TA, Bentler RA, & Mueller HG (2018). Signal classification and sound cleaning technologies *Essentials of Modern Hearing Aids: Selection, Fitting, and Verification* (pp. 383–426). San Diego, CA: Plural Publishing.

Ricketts TA, Henry P, & Gnewikow D (2003). Full time directional versus user selectable microphone modes in hearing aids. Ear Hear, 24, 424–439. [PubMed: 14534412]

Ricketts TA, & Hornsby BW (2005). Sound quality measures for speech in noise through a commercial hearing aid implementing digital noise reduction. J Am Acad Audiol, 16, 270–277. [PubMed: 16119254]

Ricketts TA, Picou EM, & Galster J (2017). Directional microphone hearing aids in school environments: Working toward optimization. J Speech Lang Hear Res, 60, 263–275. [PubMed: 28114614]

Sarampalis A, Kalluri S, Edwards B, et al. (2009). Objective measures of listening effort: effects of background noise and noise reduction. J Speech Lang Hear Res, 52, 1230–1240. [PubMed: 19380604]

Shiffman S, Stone AA, & Hufford MR (2008). Ecological Momentary Assessment. Annu Rev Clin Psycho, 4, 1–32.

Souza PE, & Kitch VJ (2001). Effect of preferred volume setting on speech audibility for linear peak clipping, compression limiting, and wide dynamic range compression amplification. J Am Acad Audiol, 12, 415–422. [PubMed: 11599876]

Stone AA, Broderick JE, Schwartz JE, et al. (2003). Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. Pain, 104, 343–351. [PubMed: 12855344]

Takahashi G, Martinez CD, Beamer S, et al. (2007). Subjective measures of hearing aid benefit and satisfaction in the NIDCD/VA follow-up study. J Am Acad Audiol, 18, 323–349. [PubMed: 17580727]

Timmer BH, Hickson L, & Launer S (2017). Ecological momentary assessment: Feasibility, construct validity, and future applications. Am J Audiol, 26, 436–442. [PubMed: 29049626]

Valente M, Fabry DA, & Potts LG (1995). Recognition of speech in noise with hearing aids using dual microphones. J Am Acad Audiol, 6, 440–449. [PubMed: 8580504]

Walden BE, Surr RK, Cord MT, et al. (2000). Comparison of benefits provided by different hearing aid technologies. J Am Acad Audiol, 11, 540–560. [PubMed: 11198072]

Walden BE, Surr RK, Grant KW, et al. (2005). Effect of signal-to-noise ratio on directional microphone benefit and preference. J Am Acad Audiol, 16, 662–676. [PubMed: 16515138]

Wendt D, Hietkamp RK, & Lunner T (2017). Impact of noise and noise reduction on processing effort: A pupillometry study. Ear Hear, 38, 690–700. [PubMed: 28640038]

Wu YH, & Bentler RA (2010a). Impact of visual cues on directional benefit and preference: Part I--laboratory tests. Ear Hear, 31, 22–34. [PubMed: 19864954]

Wu YH, & Bentler RA (2010b). Impact of visual cues on directional benefit and preference: Part II--field tests. Ear Hear, 31, 35–46. [PubMed: 19773657]

Wu YH, & Bentler RA (2012a). Do older adults have social lifestyles that place fewer demands on hearing? J Am Acad Audiol, 23, 697–711. [PubMed: 23072962]

Wu YH, & Bentler RA (2012b). The influence of audiovisual ceiling performance on the relationship between reverberation and directional benefit: perception and prediction. Ear Hear, 33, 604–614. [PubMed: 22677815]

Wu YH, Stangl E, & Bentler R (2013a). Hearing-aid users' voices: A factor that could affect directional benefit. Int J Audiol, 52, 789–794. [PubMed: 23777478]

Wu YH, Stangl E, Bentler R, et al. (2013b). The effect of hearing aid technologies on listening in an automobile J Am Acad Audiol, 24, 474–485. [PubMed: 23886425]

Wu YH, Stangl E, Chipara O, et al. (2018). Characteristics of real-world signal-to-noise ratios and speech listening situations of older adults with mild to moderate hearing loss. Ear Hear, 39, 293–304. [PubMed: 29466265]

Wu YH, Stangl E, Zhang X, et al. (2015). Construct Validity of the Ecological Momentary Assessment in Audiology Research. J Am Acad Audiol, 26, 872–884. [PubMed: 26554491]

Wu YH, Stangl E, Zhang X, et al. (2016). Psychometric Functions of Dual-Task Paradigms for Measuring Listening Effort. Ear Hear, 37, 660–670. [PubMed: 27438866]

Zakis JA, Hau J, & Blamey PJ (2009). Environmental noise reduction configuration: Effects on preferences, satisfaction, and speech understanding. Int J Audiol, 48, 853–867. [PubMed: 20017682]

Zekveld AA, Kramer SE, & Festen JM (2010). Pupil response as an indication of effortful listening: the influence of sentence intelligibility. Ear Hear, 31, 480–490. [PubMed: 20588118]

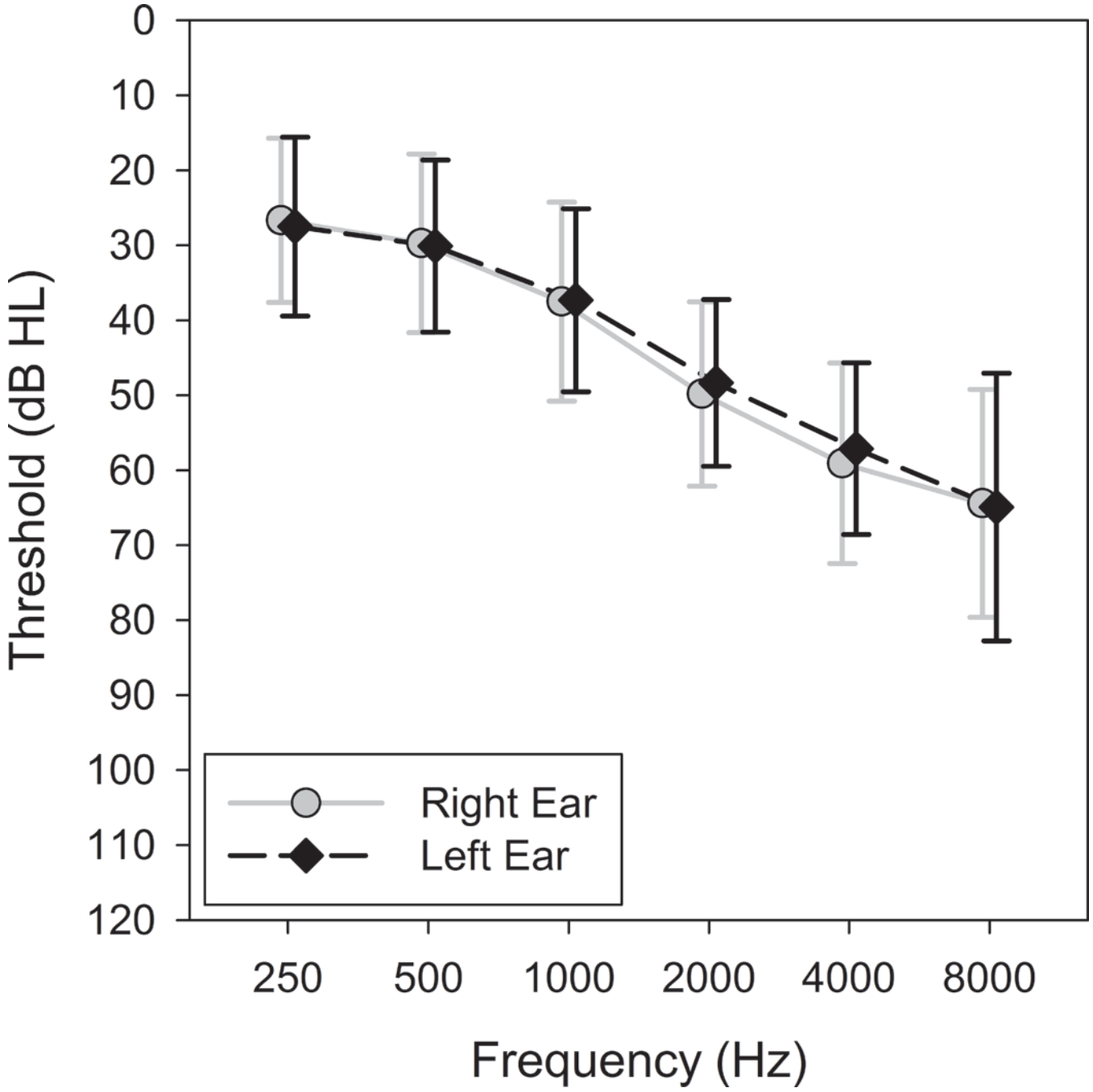

Figure 1.
Average audiograms for left and right ears of study participants. Error bars = 1 SD.

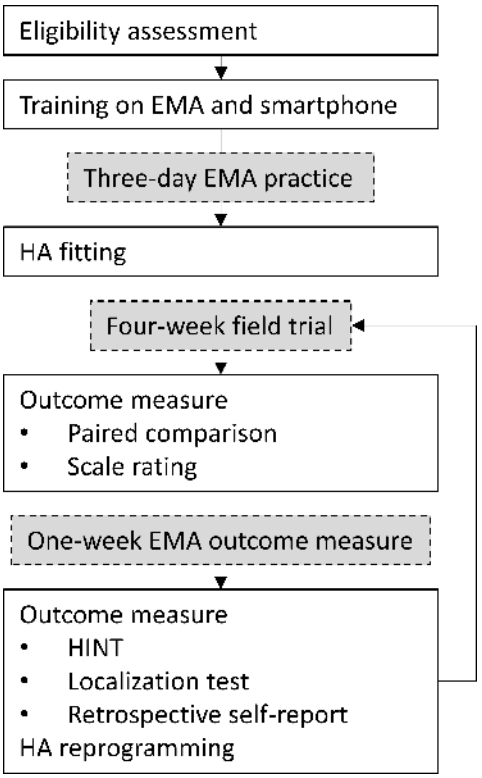

Figure 2.
Flow chart for the study. EMA: ecological momentary assessment; HA: hearing aid; HINT: the Hearing in Noise Test.

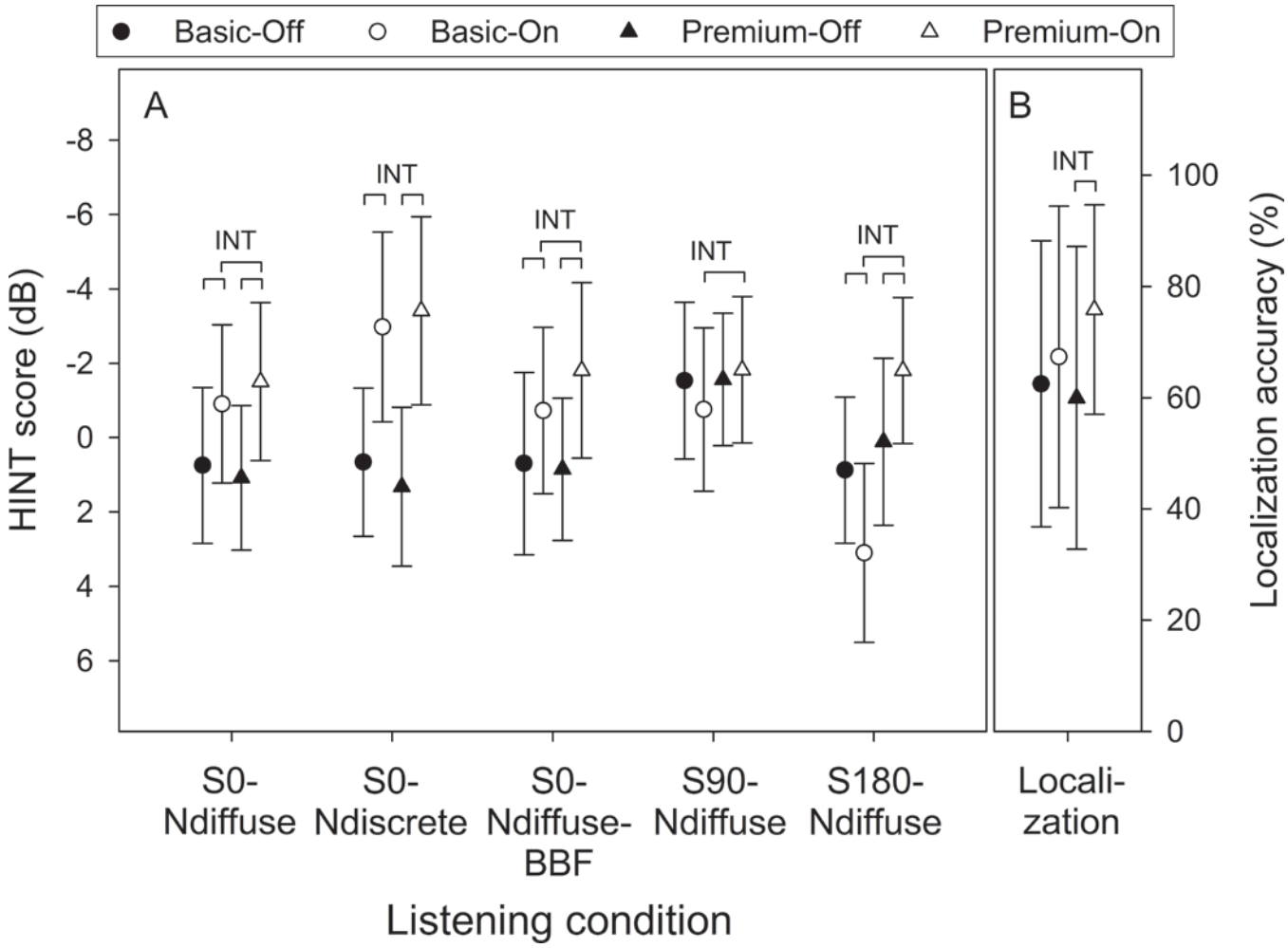

Figure 3.

A. Mean score of the Hearing in Noise Test (HINT) of each hearing aid condition as a function of listening condition. Lower scores represent better performance. The y-axis has been reversed so that the top of the figure represents better performance. B. Localization accuracy of each hearing aid condition as a function of listening condition. “INT” represents significant interaction between hearing aid model and feature status. Bracket represents significant difference. Error bars = 1 SD.

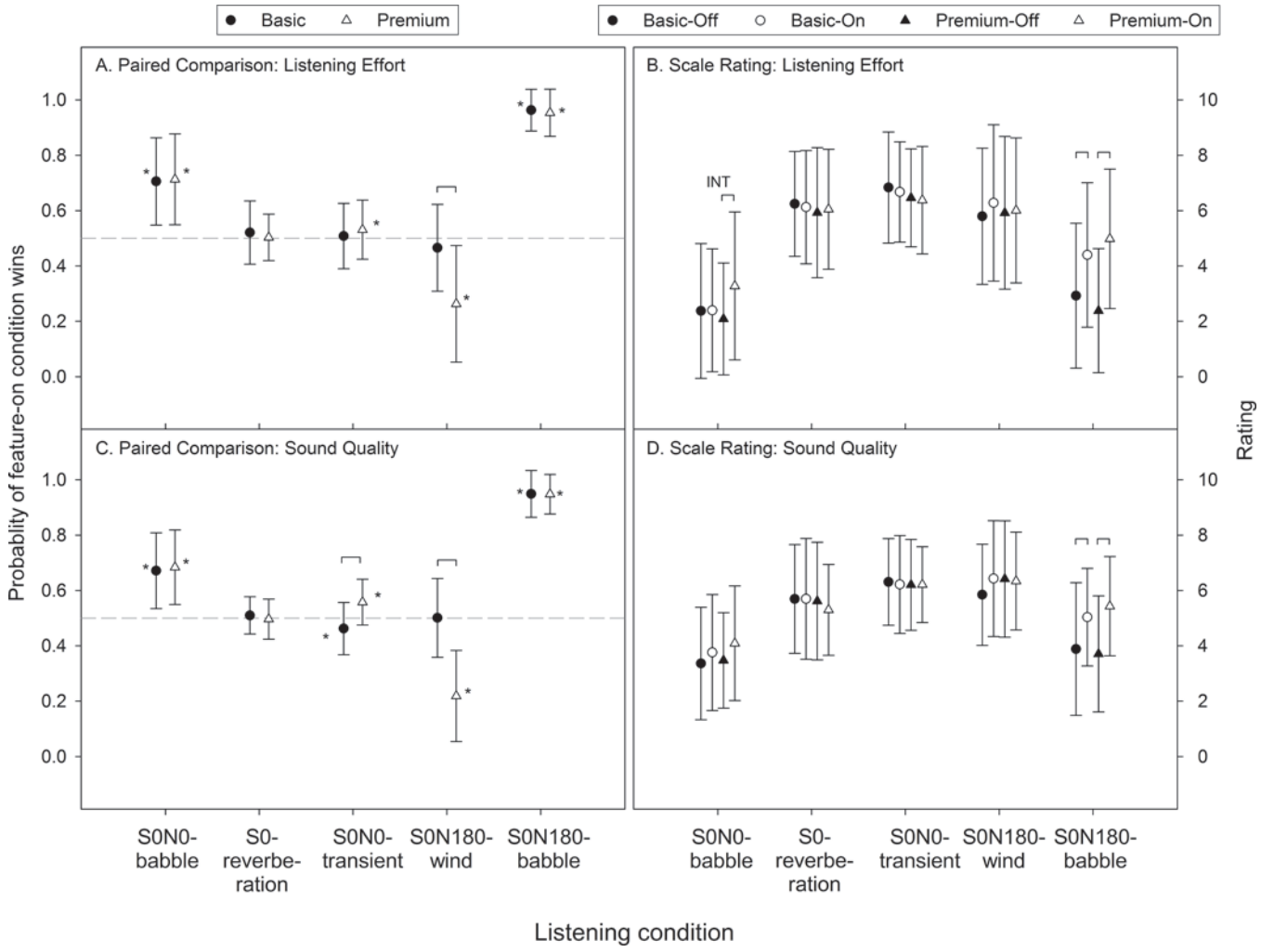

Figure 4.

Listening effort (A and B) and sound quality (C and D) results of paired comparisons (A and C) and scale ratings (B and D) as a function of listening condition. For B and D, higher ratings represent better outcomes. Brackets and asterisks represent significant difference. "INT" represents significant interaction between hearing aid model and feature status. Error bars = 1 SD.

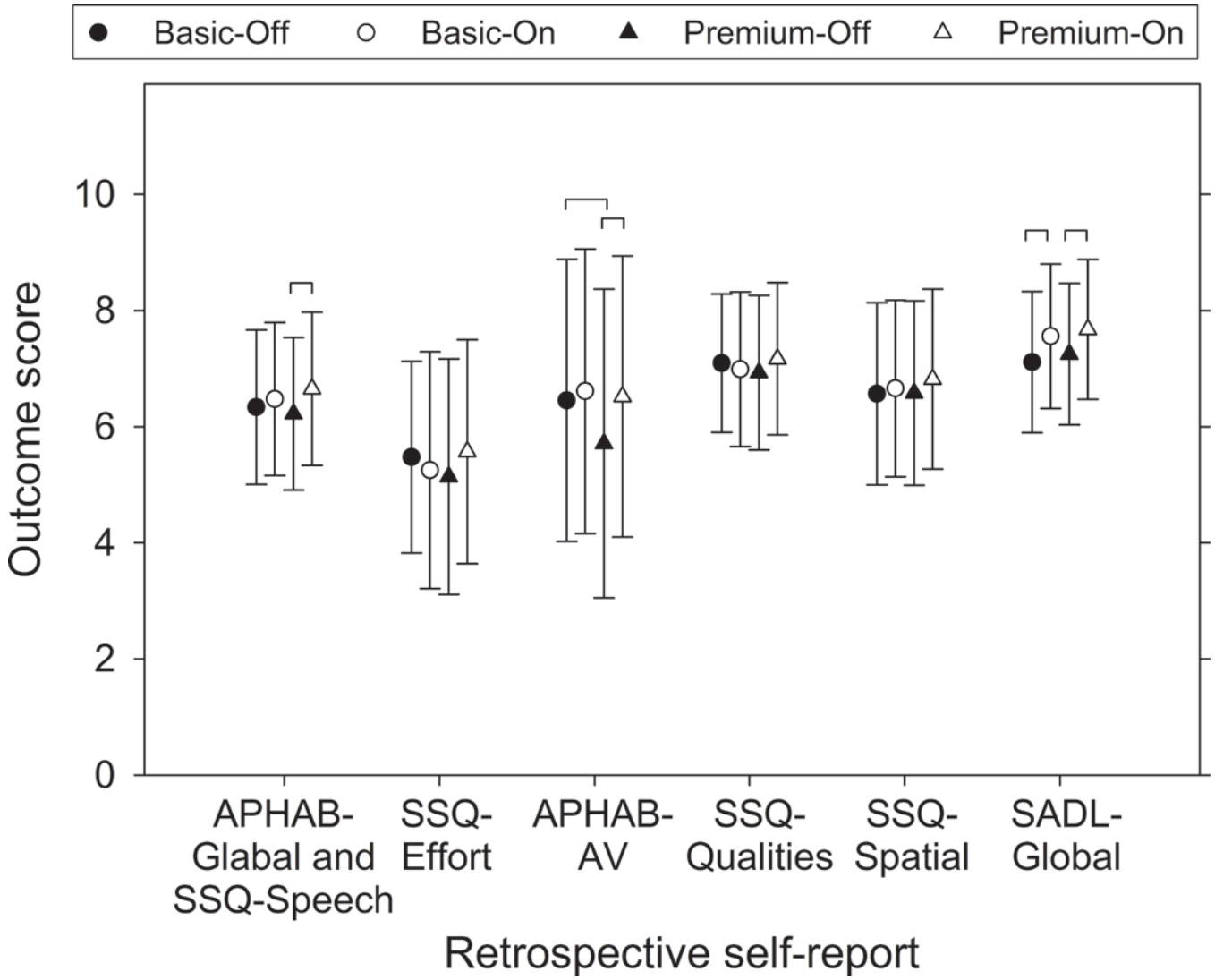

Figure 5.

Mean outcome scores of retrospective self-reports of each hearing aid condition. Higher scores represent better outcomes. “INT” represents significant interaction between hearing aid model and feature status. Brackets represent significant difference. Error bars = 1 SD.

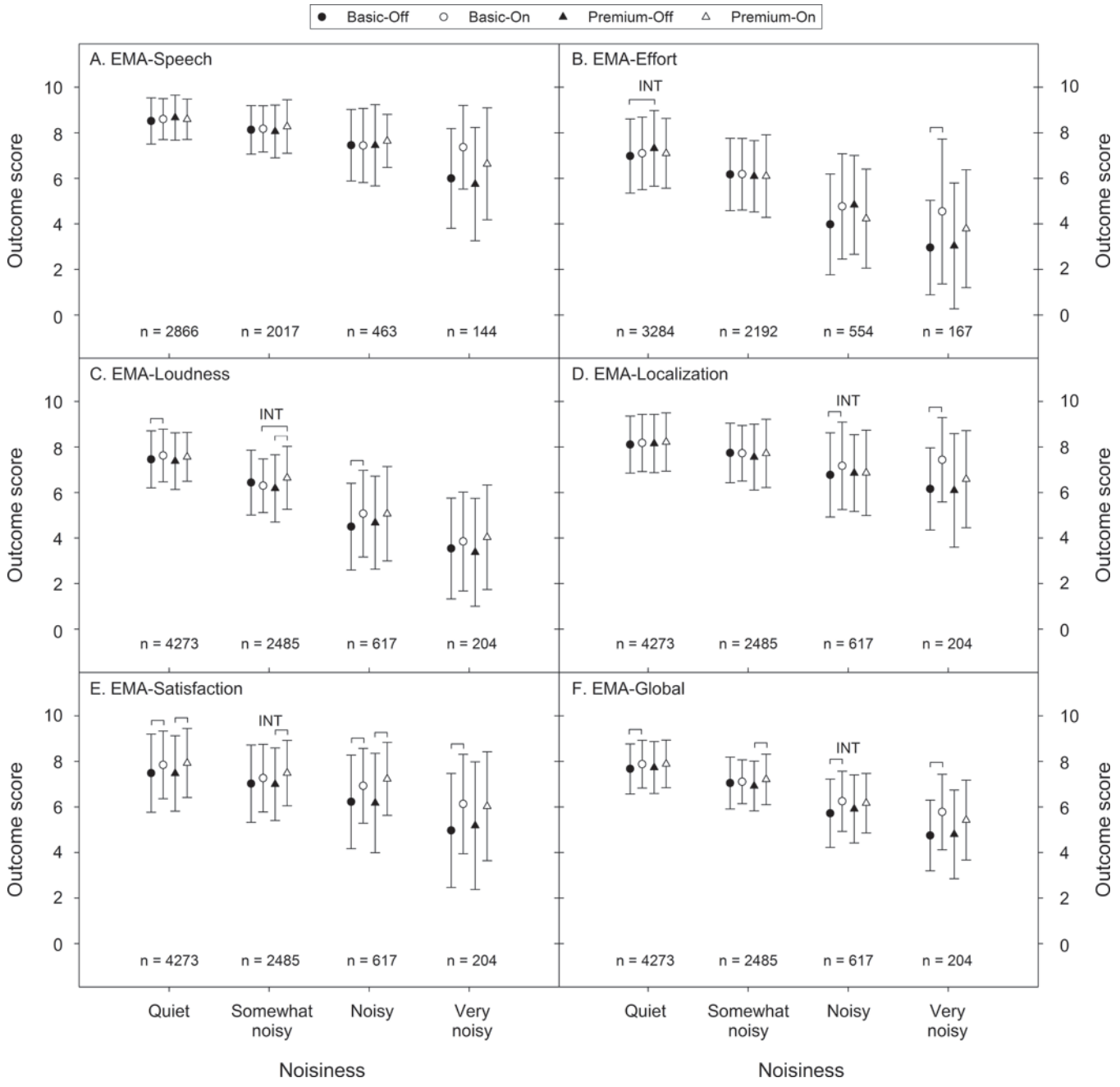

Figure 6.

Mean outcome scores of in-situ self-reports of each hearing aid condition. Higher scores represent better outcomes. “INT” represents significant interaction between hearing aid model and feature status. Brackets represent significant difference. Error bars = 1 SD.

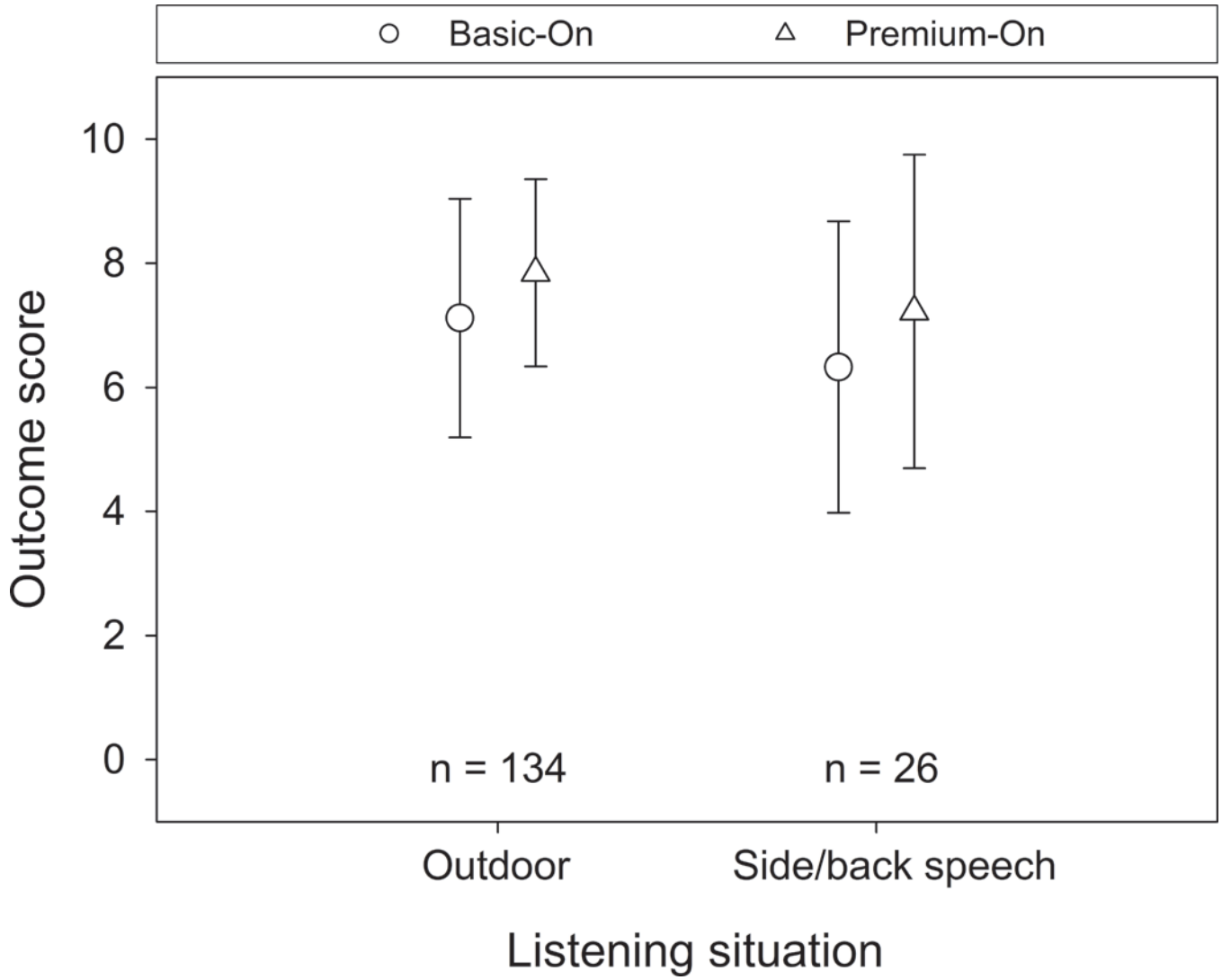

Figure 7.

Mean EMA-Satisfaction scores of the basic-on and premium-on conditions in two types of listening situations. For the “outdoor” situations, only the surveys that meet the following criteria are included: location of the activity = outdoor; noisiness = somewhat noisy, noisy, or very noisy. For the “side/back speech” situations, only the surveys that meet the following criteria are included: speech location = side or back; noisiness = somewhat noisy, noisy, or very noisy; hearing aid program = manual program. For both listening situations, only the participants who had survey data from both of the basic-on and premium-on conditions are included (19 and 7 participants in the outdoor and side/back speech situations, respectively). The numbers of the survey completed in each type of the situation are shown at the bottom of the figure (basic-on and premium-on combined).

Table 1.

Differences, as described by the manufacturer, between basic and premium hearing aids used in the study. DM: directional microphone

	Hearing Aid	
Feature	**Basic**	**Premium**
Automatic adaptive DM	Yes (1 channel)	Yes (33 channels)
Bilateral beamformer	No	Yes
Speech-seeking DM	No	Yes
Pinna-simulation directivity	No	Yes
Noise reduction	Yes (12 channels)	Yes (20 channels)
Reverberation reduction	No	Yes
Impulse noise reduction	No	Yes
Wind noise reduction	No	Yes
Spatial noise reduction	No	Yes

Table 2.

Listening conditions for the speech recognition test. DM: directional microphone; BBF: bilateral beamformer

	Speech location	Noise level and location	Designed for
S0Ndiffuse	0 degree	65 dBA; 8 loudspeakers	Baseline
S0Ndiscrete	0 degree	65 dBA; 90, 180, and 270 degree	Multi-channel adaptive DM
S0Ndiffuse-BBF	0 degree	65 dBA; 8 loudspeakers	Bilateral beamformer
S90Ndiffuse	90 degree	65 dBA; 8 loudspeakers	Speech-seeking DM
S180Ndiffuse	180 degree	65 dBA; 8 loudspeakers	Speech-seeking DM

Table 3.

Listening conditions for paired comparisons and scale ratings

	Speech level and location	Noise level and location	Designed for
S0N0-babble	A. 60 dBA; 0 degree B. 70 dBA; 0 degree	A. Babble; 54 dBA; 0 degree B. Babble; 70 dBA; 0 degree	Baseline
S0-reverberation	A. 60 dBA; 0 degree; church B. 60 dBA; 0 degree; small club	No noise	Reverberation reduction
S0N0-transient	60 dBA; 0 degree	A. Can drop (90 dBA); 0 degree B. Hammer hit (90 dBA); 0 degree	Impulse noise reduction
S0N180-wind	60 dBA; 0 degree	Wind, 3.8 m/sec; 180 degree	Wind noise reduction
S0N180-babble	A. 60 dBA; 0 degree B. 70 dBA; 0 degree	A. Babble; 54 dBA; 180 degree B. Babble; 70 dBA; 180 degree	Spatial noise reduction

Table 4.

Summary of outcome measures. HINT: Hearing in Noise Test; APHAB: Abbreviated Profile of Hearing Aid Benefit; SSQ: Speech, Spatial, and Qualities hearing scale; SADL: Satisfaction with Amplification in Daily Life; AV: aversiveness subscale of the APHAB; EMA: ecological momentary assessment

		Outcome Domain				
		Speech understanding	**Listening effort**	**Sound quality**	**Localization**	**Satisfaction**
Laboratory tests	Speech recognition test	HINT	-	-	-	-
	Paired comparison	-	Listening effort	Naturalness and Annoyance	-	-
	Scale rating	-	Listening effort	Naturalness and Annoyance	-	-
	Localization test	-	-	-	Front/back localization	-
Retrospective self-reports	APHAB	APHAB-Global and SSQ-Speech	-	APHAB-AV	-	-
	SSQ		SSQ-Effort	SSQ-Qualities	SSQ-Spatial	-
	SADL	-	-	-	-	SADL-Global
In-situ self-reports	EMA	EMA-Speech	EMA-Effort	EMA-Loudness EMA-Global	EMA-Localization	EMA-Satisfaction

Table 5.

Summary of results. HA: hearing aid; P: premium; B: basic.

		Outcome Domain				
		Speech understanding	**Listening effort**	**Sound quality**	**Localization**	**Satisfaction**
Laboratory tests	Feature: P vs B HA: P vs B Feature: On vs Off	P > B P ≥ B On ≥ Off [1]	P = B [2,3] P = B On ≥ Off [4]	P = B [2,5] P = B On ≥ Off [4,6]	P > B P = B On ≥ Off	- - -
Retrospective self-reports	Feature: P vs B HA: P vs B Feature: On vs Off	P = B P = B On ≥ Off	P = B P = B On = Off	P = B P = B On ≥ Off	P = B P = B On = Off	P = B P = B On > Off
In-situ self-reports	Feature: P vs B HA: P vs B Feature: On vs Off	P = B P = B On = Off	P ≤ B P = B On ≥ Off	P ≥ B P ≥ B On ≥ Off	P ≤ B P = B On ≥ Off	P ≥ B P = B On ≥ Off

[1]: On < Off for basic HAs in the S180Ndiffuse condition

[2]: P < B in the S0N180-wind condition of paired comparisons

[3]: P > B in the S0N0-babble condition of scale ratings

[4]: On < Off for premium HAs in the S0N180-wind condition of paired comparisons

[5]: P > B in the S0N0-transient condition of paired comparisons

[6]: On < Off for basic HAs in the S0N0-transient condition of paired comparisons