Predicting Individual Hearing-Aid Preference From Self-Reported Listening Experiences in Daily Life

Jeppe H. Christensen,[1] Johanne Rumley,[2] Juan Camilo Gil-Carvajal,[2] Helen Whiston,[3] Melanie Lough,[3] and Gabrielle H. Saunders[3]

Objectives: The study compared the utility of two approaches for collecting real-world listening experiences to predict hearing-aid preference: a retrospective questionnaire (Speech, Spatial, and Qualities of Hearing Scale [SSQ]) and in-situ Ecological Momentary Assessment (EMA). The rationale being that each approach likely provides different and yet complementary information. In addition, it was examined how self-reported listening activity and hearing-aid data-logging can augment EMAs for individualized and contextualized hearing outcome assessments.

Design: Experienced hearing-aid users (N = 40) with mild-to-moderate symmetrical sensorineural hearing loss completed the SSQ questionnaire and gave repeated EMAs for two wear periods of 2-weeks each with two different hearing-aid models that differed mainly in their noise reduction technology. The EMAs were linked to a self-reported listening activity and sound environment parameters (from hearing-aid data-logging) recorded at the time of EMA completion. Wear order was randomized by hearing-aid model. Linear mixed-effects models and Random Forest models with five-fold cross-validation were used to assess the statistical associations between listening experiences and end-of-trial preferences, and to evaluate how accurately EMAs predicted preference within individuals.

Results: Only 6 of the 49 SSQ items significantly discriminated between responses made for the end-of-trial preferred versus nonpreferred hearing-aid model. For the EMAs, questions related to perception of the sound from the hearing aids were all significantly associated with preference, and these associations were strongest in EMAs completed in sound environments with predominantly low SNR and listening activities related to television, people talking, nonspecific listening, and music listening. Mean differences in listening experiences from SSQ and EMA correctly predicted preference in 71.8% and 72.5% of included participants, respectively. However, a prognostic classification of single EMAs into end-of-trial preference with a Random Forest model achieved a 93.8% accuracy when contextual information was included.

Conclusions: SSQ and EMA predicted preference equally well when considering mean differences, however, EMAs had a high prognostic classifications accuracy due to the repeated-measures nature, which make them ideal for individualized hearing outcome investigations, especially when responses are combined with contextual information about the sound environment.

Key words: Data-logging, Ecological Momentary Assessment, Hearing aid, Random forest, Retrospective hearing outcomes.

(Ear & Hearing 2024;XX;00–00)

INTRODUCTION

Ecological Momentary Assessment (EMA) and retrospective hearing questionnaires like the Speech Spatial Qualities of Hearing Scale (SSQ) provide insights into real-world hearing-aid use and outcomes that are unavailable from each in isolation (Wu et al. 2020). Specifically, EMA responses provide in-situ information about real-life listening experiences, and immediate opinions about outcome and listening intent at a specific moment in time (Shiffman et al. 2008; Wu et al. 2015; Timmer et al. 2018; Holube et al. 2020), while retrospective questionnaires provide insights into the way a user recalls hearing-aid function over time (Robinson & Clore 2002) over a wide variety of listening situations (Gatehouse & Noble 2004). Each approach has advantages and disadvantages (see later) but likewise, each can play a different role in understanding opinions about a hearing aid and assist in the choice of hearing-aid settings or model for a particular individual (Lelic et al. 2024). For clinical purposes, it is therefore valuable to learn what combination of information from retrospective questionnaires and EMAs provide optimal clinical insights, face validity, and generalizability of outcomes so that time spent on clinical investigations and patient burden can be minimized. In addition to this, today's technology allows for seamless recording of objective environmental data, such as ambient acoustic characteristics, which can potentially help contextualize EMA outcomes (Williger & Lang 2015; Christensen et al. 2021; Yellamsetty et al. 2021; Hart et al. 2022). However, EMA can be problematic because the process is invasive and burdensome to participants, necessitates a mobile phone, and can lead to biased sampling of situations due to a lack of willingness or it being inappropriate to respond to EMAs in certain situations (Schinkel-Bielefeld et al. 2020). Retrospective questionnaires, on the other hand are less invasive and burdensome, but they rely on recollection and thus encounter recency biases and recall issues (Bradburn et al. 1987; Stone & Shiffman 2002).

Retrospective self-reports have been directly compared with in-situ EMAs to learn about the sensitivity of each for assessing hearing-aid outcomes (Wu et al. 2020). The participants of Wu et al. (2020) evaluated experiences via an adapted version of the Glasgow Hearing Aid Benefit Profile questionnaire both retrospectively and with EMA with two different hearing-aid models. The in-situ EMA responses were more sensitive to differences in experiences with the two models of hearing aids compared with the retrospectively collected responses, as reflected in more subscales showing significant differences between the hearing-aid models. However, the study did not compare how well the EMAs or retrospectively collected responses aligned with participants' overall hearing-aid preference which is not necessarily the same as that rated most highly via EMA. This was illustrated in the work of Christensen et al. (2024), who found a within-participant

[1] Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark; [2] Oticon A/S, Centre for Applied Audiology Research; and Clinical Audiological Development, Smoerum, Denmark [3] Manchester Centre for Audiology and Deafness, School of Health Sciences, University of Manchester, Manchester, United Kingdom.

<zdoi; 10.1097/AUD.0000000000001520>

significant difference in EMAs from 40 participants regarding sound satisfaction with two different hearing-aid models, but when the participants were asked about their preferred choice between the models, the split was equivocal (17 versus 23).

It is also of interest to understand how subjective responses of listening experiences interacts with the contexts in which they were given to shape preference. Studies have shown that listening intentions or conditions can influence preferences for hearing-aid settings in daily life (Walden et al. 2004; Pasta et al. 2022) and it is therefore important to consider the way users' needs and experiences are impacted by hearing-aid settings across various real-life situations (Andersson et al. 2021; Bosman et al. 2021; Christensen et al. 2021). For example, Andersson et al. (2021) determined from EMA responses that real-life benefit from a hearing-aid noise management system was only present when listening in speech-dominated real-life environments, while data from the SSQ-12 showed no differences between hearing aids with and without noise reduction activated. In other words, this contextual benefit was not identified using a retrospective questionnaire and would not have been understood properly had the EMA data not been combined with data-logging of the encountered listening conditions collected by the hearing aid itself.

Besides revealing contextual effects, data-logging can provide insights regarding the face validity of field trials by giving an account of the listening environments encountered by the tested population, their equivalence across participants, and whether they change over time or interventions (Schinkel-Bielefeld 2020; Christensen et al. 2021). Thus, EMAs combined with contextual data-logging potentially allow for an even more precise account of real-world listening experiences as compared with retrospective questionnaires or EMAs alone (i.e., without data-logging).

Considering the uncertainties above and with a view to minimizing participant burden by for example enabling a choice between methods in future investigations, the aim of this study was to assess the utility of a retrospective questionnaire—here, the SSQ (Noble et al. 2013), and in-situ EMA with and without contextual data-logging for predicting preference between two different hearing-aid models. To this end, we collected EMAs and retrospective data (SSQ responses) from individuals who wore two models of hearing aid in a cross-over trial, along with contextual information from self-reports and hearing-aid data-logging during wear. At the end of the study, participants were permitted to keep their preferred hearing aid. The main aim of the study was to determine how well EMAs, and retrospective questionnaire responses could predict hearing-aid preference using both diagnostic and prognostic analytical approaches (Saeb et al. 2017). While the distinction between diagnostic and prognostic prediction methods is typically relevant in the context of medical interventions (van Smeden et al. 2021), we apply these approaches here to evaluate the accuracy of predicting an individual's hearing-aid preference based on listening experiences from SSQ responses or EMAs (diagnostic), and to assess how well changes over time in an individual's preference for a hearing-aid model can be predicted with listening experiences from EMAs (prognostic). Both diagnostic and prognostic prediction can be clinically relevant when for example, deciding between two hearing-aid models for a new user (the former) or when evaluating if a new feature or hearing-aid setting is preferred over an existing one for the same user (the latter) (Collins et al. 2015).

MATERIALS AND METHODS

Key aspects of the methods are described later. Additional details can be found in Christensen et al. (2024) which analyses data from the same sample.

Participants

Participants were recruited between January 2021 and June 2021 via social media, word of mouth, and Manchester Centre for Audiology and Deafness Hearing Health research volunteer database. Forty-one individuals met the study criteria of being aged 25 to 85 years, having mild-to-moderate symmetrical sensorineural hearing loss confirmed by a clinical audiogram completed within the prior 3 years, being an experienced hearing-aid user (used bilateral hearing aids daily for >1 year), owning an iPhone with iOS 13 or newer (necessary for the study apps), being proficient in the English language, and not having ever worn either of the hearing-aid models used in the study. Data from 1 participant were excluded due to missing hearing-aid data-logging data.

Study Design

A crossover trial was conducted that consisted of two 14-day wear periods. The order of hearing-aid wear was randomized across participants. Participants were not informed as to which hearing-aid model they were wearing and identifying labels were removed from the devices; however, 2 participants spontaneously reported having retrieved this information from the iOS settings information.

Study Hearing Aids

Both hearing-aid models (HA1 and HA2) were mini "receiver in the ear" (miniRITEs) made by Oticon A/S (Smoerum, Denmark). HA1 was the Oticon Opn S and HA2 was the Oticon MORE. Both have features that include wind noise management, feedback management strategies, and they have equal fitting bandwidths and number of frequency bands (64) for signal processing. Both hearing-aid models were programmed with a single general listening program (i.e., the participants did not have access to other programs for specific listening situations). The key difference between the two models is the noise reduction system. HA1 uses a 16-band noise reduction system with a fast-acting combination of a Minimum Variance Distortionless Response beamformer (Souden et al. 2010) and a single-band Wiener postfilter (Uwe Simmer et al. 2001; Kjems & Jensen 2012), while HA2 uses a fast-acting 24-band noise reduction system with a Minimum Variance Distortionless Response beamformer combined with the processing of a Deep Neural Network-based postfilter trained to enhance the contrast between speech and noise using across-band information from 12 million real-life sound scenes (Andersen et al. 2021). Both models of hearing aids could be controlled remotely via a smartphone app (Oticon, ON), however, participants were not informed of this possibility.

Data-Logging

Timestamped data were logged every 20 sec and were sent via Bluetooth for storage to an app (see Ecological Momentary Assessment) on the participant's iPhone. The data collected consisted of total time the hearing aid was switched on, and the ambient sound pressure levels (SPLs in dB) and signal to noise ratios (SNRs in dB). For both models of hearing aid, SPL is estimated from a low-pass infinite impulse-response filter with a time constant of 63 msec. SNR is then estimated as the difference between the lower envelope and the immediate level of the SPL estimation (see Figs. 3–10 in Kates 2008). Thus, a "low" SNR can indicate both a noisy environment (if the SPL is relatively high) and a quieter environment (if the SPL is relatively low). The frequency weighting of the SPL estimates from the two hearing-aid models differed. HA2 estimated an A-weighted SPL (Fletcher & Munson 1933), while HA1 estimated an un-weighted SPL across four frequency bands (independent of the 64 bands for signal processing) with center-frequencies 313, 1250, 2578, and 5547 Hz. These differences yield an offset between SPL and SNR levels estimated by HA1 and HA2. Also, see Supplementary Information in Christensen et al. (2024) for a detailed investigation into the level differences with the two hearing-aid models.

Ecological Momentary Assessment

Participants completed EMA surveys via a smartphone app, the same app that stored SPL and SNR information. The questions in the EMA app were designed specifically for the study. Surveys were self-initiated or initiated via a phone prompt up to a maximum of eight times a day, with a minimum time of one hour (pseudo-random) between prompts and no prompts earlier than 8 A.M. or later than 8 P.M. Self-initiated EMAs could be completed at any time. Responses to the EMA survey could only be submitted if the associated smartphone had an active Bluetooth connection to the study hearing aids.

The EMA survey consisted of eight questions. The first asked about the listening situation with eight predefined choices (menu-selection): "Television"; "Sounds around me"; "People talking"; "One person talking"; "Nothing in particular"; "Music (streaming)"; "Music (live)"; "Streaming broadcast." Note that the "Music (live)" category refers to any music that is not streamed. So, this could be a live performance or music from a "live" sound system (e.g., a radio). The second question was a simple radio button for the participants to indicate if the evaluated listening experience was still occurring. The last six questions (Q1 to Q6) asked participants to rate their hearing aids using continuous sliders from 0 to 10 with differing anchor points. Specifically, the questions (and anchor points) were:

- Q1—"How noisy was it?" ("Quiet" to "Very noisy"),
- Q2—"How satisfied were you with the sound from your hearing aids?" ("Not satisfied" to "Very satisfied"),
- Q3—"How was it to focus on the sounds you wanted to hear?" ("Difficult" to "Easy"),
- Q4—"How was it to ignore sounds you didn't want to hear?" ("Difficult" to "Easy"),
- Q5—"How was it to work out where different sounds were coming from?" ("Difficult" to "Easy"),
- Q6—"How well could you hear what was going on around you?" ("Not very well" to "Very well").

Although it was technically possible to answer Q3 before Q2 as the sliders are "active" at the same time, there is no reason to assume the questions were not answered in the order in which they were presented on the screen.

Protocol

The study was conducted during the Coronovirus Disease-19 (COVID-19) pandemic thus all interactions with participants took place remotely. Before enrollment, participants sent the study team a recent (within 3 years of enrollment) copy of their audiogram, which was used to program the hearing aids. The need to use an existing audiogram is why we elected to enroll only experienced hearing-aid users. At their initial encounter, participants were asked if they thought their hearing had changed since their last hearing test and they completed the Consumer Ear Disease Risk Assessment tool (Klyn et al. 2019) to screen for ear-related diseases. Participants reporting a change would have been ineligible to participate and would have been advised to go for a hearing test; this situation did not arise in practice. If following discussion with the research audiologist, the Consumer Ear Disease Risk Assessment identified a possible ear pathology that required a medical opinion, the participant was directed to contact their GP. They were still eligible to take part in the study, so long as the pathology did not contra-indicate hearing-aid use. Following enrollment, the participants were assigned randomly to a wear order. The appropriate pair of hearing aids were then programmed using the manufacturer's proprietary prescription generated from the pure-tone audiogram and personal details (age, sex, and personalized questions regarding how they like to listen to sounds). The programmed hearing aids were then sent by mail to the participants. Following receipt of the hearing aids, the participants met via Zoom with a research audiologist who provided comprehensive study procedure instructions followed by remote fine-tuning based on the participant's subjective feedback using the Oticon RemoteCare app. Gain was adjusted when fine-tuning the hearing aids but specific features such as noise reduction settings were not. Voiced Ling sounds (with the mouth obscured to prevent lip-reading cues) over video call were used to test audibility as part of the verification process (Ling 1976). The hearing-aid settings from the first fitting were copied to the second pair of hearing aids. The second pair of hearing aids were mailed a few days before the second wear period was due to start, so that participants had time to fully charge the hearing aids for the second wear period on the correct day. Participants then had a remote appointment (as they did for the first fitting), and fine-tuning was performed where required. The first pair of hearing aids were returned by mail to the laboratory once the swap over had occurred. Hearing-aid boxes were clearly labeled so that participants could distinguish between first and second pairs.

During pilot testing, it was learned that participants often missed or did not receive the automated EMA prompts, thus, we emphasized that, in addition to responding to automated prompts, participants should self-initiate 4 to 6 EMAs each day when they encountered a self-defined "interesting" listening situation.

One week into each wear period participants were sent a link to a secure online platform (RedCAP) to complete the SSQ. On average, participants completed the SSQ 8.2 days (SD = 1.0, range = 7.4 to 12.5) after the fine-tuning appointment. Once both trial periods had been completed participants selected which of the two models of hearing aid they wanted to keep. The selected model was considered to be the "preferred hearing aid" and was used as such in the subsequent analyses.

Analyses

We carried out three main analyses. Initially, we explored the differences between the responses given with the preferred and nonpreferred hearing aids. We also investigated the extent to which these differences were moderated by SSQ scales and items, and EMA items and usage contexts (both self-reported and logged).

In our second analysis, we examined how effectively the SSQ and EMA responses could predict the preferred hearing aid using data from unseen individuals. This is referred to as a diagnostic classification.

Lastly, we studied how accurately EMAs could predict hearing-aid preferences using unseen ratings from the same individuals, a process known as a prognostic classification. We also investigated how the prediction accuracies were influenced by contextual data associated with EMAs, including self-reported listening activity and data-logged sound environment.

Contextual Data • EMAs were linked to the ambient sound environment by time-averaging SPL and SNR from hearing-aid data-logging in the 5-min preceding EMA completion. The choice of a 5-min time-window was based on prior work with similar hearing-aid data-logging (Andersson et al. 2021, 2023; Bosman et al. 2021).

Next, the sound environment was classified into nine categories differing in listening complexity based on the time-averaged SPL and SNR. Specifically, the time-averaged SPL and SNR were binned into "Low," "Medium," or "High" based on tertile splits within each participant's data from each hearing-aid model, and the nine categories were then formed from all combinations of the binned SPL and SNR. By binning the data based on individual splits, differences in the hearing-aid model's estimation of SPL and SNR, and lifestyle (and therefore in sound exposure) among the participants are controlled for.

Statistical Association Analysis • Associations between SSQ and EMA responses and independent variables were investigated using linear mixed-effects models fitted using the *lmerTest* packages in R (version 3.6.2, 2019 The R Foundation for Statistical Computing). Pairwise comparisons and effect size estimations were based on the estimated marginal means (EMMs) adjusted for multiple comparisons using Tukey's HSD computed using the *emmeans* package in R. The SSQ responses were modeled separately for each scale because each scale had a different number of Items. Furthermore, SSQ responses were predicted with a fixed effect for end-of-trial preference using a two-level categorical (preferred or nonpreferred) predictor in interaction with SSQ item. The random effects structure included participant ID to allow for a random intercept per participant. EMAs were predicted with the same end-of-trial preference as for the SSQ responses, and in addition included fixed effects for self-reported listening activity (eight levels), data-logged sound environment (nine levels), and EMA item (six levels) including all mutual two-way interactions. The random effects structure for the EMA data included participant ID, volume level, time of day (in hours), day of wear period, and wear period (first or second) in a crossed manner.

Prediction Models • We based a diagnostic prediction on mean differences in SSQ and EMA responses across wear periods and items. Here, "items" refer to the 49 questions on the SSQ and the 6 rating questions on the EMA. Recall that the aim of the diagnostic prediction was to classify if EMA responses (ratings) from unseen individuals (i.e., individuals not included in training) belonged to the preferred hearing aid.

Specifically, for the SSQ, for each participant, i, a mean difference across J items is:

$$\Delta M_i = \frac{1}{J}\sum_{j=1}^{J} r_{i,j,1} - \frac{1}{J}\sum_{j=1}^{J} r_{i,j,2} \qquad (1)$$

where r is the rating and the subindex "1" and "2" refers to wear periods 1 and 2, respectively. Mean differences for the EMA data were calculated similarly using Eq. (1), except that EMAs for each item were first averaged across time. A diagnostic prediction accuracy then equals the proportion of participants with mean differences favoring their preferred hearing-aid model (i.e., a positive value would indicate a preference toward the hearing-aid model worn in period one and vice versa).

Furthermore, the longitudinal and repeated-measures nature of the EMAs allowed for a prognostic prediction, which we applied using supervised random forest (RF) classifiers with five-fold cross-validation (Saeb et al. 2017; Bzdok & Ioannidis 2019). The aim of the prognostic prediction was to predict if ratings in unseen EMA responses (i.e., responses not included in training) from individuals belonged to a preferred hearing aid. The cross-validation was performed by training with 80% of all responses, randomly sampled without replacement and with equal proportions among participants to overcome class imbalance from inter-individual differences in number of responses, and testing was then performed on the remaining 20% of responses. The overall model accuracies were then calculated as the mean accuracies across folds, while the accuracies for individual participants were estimated by stratifying model predictions.

Four RF prediction models that varied by the number of contextual features included were trained. The simplest model (model M1) only included hour of day, day of wear period, participant ID, EMA item, and EMA rating magnitude associated with each EMA response. Three models additionally included either the self-reported listening activity (model M2) or the data-logged sound environment (model M3), or both the self-reported listening activity and the data-logged sound environment (model M4).

Hyperparameters of the models were optimized during training, with the number of trees set at 500 and a minimal terminal node size of 1. The models were trained using the *randomForest* package in R. For each model, the feature importance and partial dependency of common features were derived (using functions from the *randomForest* package in R). Feature importance estimates how much removing a feature would negatively impact the model performance. This is computed as a mean decrease in impurity (Gini index) and as a change in prediction accuracy in percentage (Hastie et al. 2001). The partial dependency characterizes relative relationships between individual feature levels and predicted probabilities of preference obtained from the RF models (Hastie et al. 2001; Cutler et al. 2007). Note that for both the diagnostic and the prognostic prediction of preference with EMAs, item Q1 ("How noisy was it?") was excluded as this relates to a perception of the sound environment rather than to a perception of the quality of the sound from the hearing-aid model.

RESULTS

Participants

Participants (18 females, 22 males) were aged between 26 and 79 years (mean [M] = 64.8; SD = 12). The audiograms of 10 participants were completed within the prior 2 years; the audiograms of the remaining 30 participants had been completed in the prior 2 to 3 years. This was considered acceptable because the UK NHS recommends reassessment of hearing every 3 years unless significant changes are reported.

Hearing-Aid Preference

Seventeen participants chose to keep HA1 and 23 chose HA2. Preference does not appear to be biased by recency effects in that only 21 of the 40 participants selected the most-recent hearing-aid model as their final preference.

Speech, Spatial, Qualities Questionnaire Responses

Of the 49 SSQ items, 17 (35.0%) were completed by all participants, 16 (32.5%) had N/A responses from 1 or 2 participants, and the remaining 16 items (32.5%) had N/A responses from between 3 and 16 participants. The items with the most N/A responses were items 5 ("Follow conversation without missing start of new talker") and 6 ("Having conversations in echoic environments") on the Speech scale with 12 and 16 N/A responses across the two SSQ's, respectively, and item 2 ("Judge distance of vehicle") on the Spatial scale with 11 N/A responses across the two SSQ's. Data from one participant (number 13) were removed from subsequent SSQ analysis due to outlier scores. Specifically, the participant only gave ratings of 10 and 0 for almost all items.

Figure 1A shows distributions and boxplots of the SSQ responses averaged across the items within each SSQ scale (panels) and separated by whether the data were for the preferred or nonpreferred hearing aid (blue versus dark gray). The grand averages were 7.24 (SD = 1.29), 7.53 (SD = 1.26), and 8.03 (SD = 1.10) for the SSQ Speech, Spatial, and Qualities scales respectively. Ratings for the preferred hearing-aid model were significantly higher than for the non-preferred hearing-aid (type III F-test, $F(1,3646) = 25.12$, $p < 0.001$). Pairwise comparisons of the EMMs showed that effect size was largest for the Speech scale, smaller for the Qualities scale, and lowest and nonsignificant for the Spatial scale (Speech scale: rating difference = 0.45, $Z = 2.86$, $p = 0.004$, Cohen's $d = 0.24$; Qualities scale: rating difference = 0.29, $Z = 2.86$, $p = 0.004$, Cohen's $d = 0.15$; spatial scale: rating difference = 0.20, $Z = 1.94$, $p = 0.053$, Cohen's $d = 0.11$).

Ratings on the SSQ responses also varied between items within each scale. This can be seen on Fig. 1B, which shows the EMM difference for each item and scale. For most of the 49 items (all but six), the preferred hearing-aid model received a higher rating than the non-preferred hearing-aid model. This difference in rating was significant for six items (three on the Speech scale, one on the Spatial scale, and two on the Qualities scale). If one assumes that SSQ responses reflect preference, one would conclude that these six items were the most influential in determining preference. Table 1 shows these six items ordered by the magnitude of their estimated rating difference. The Spatial and Qualities scales contain individual items that exhibited the largest differences.

EMAs and Data-Logging

Across the two wear periods, participants submitted a total of 3633 EMAs. As shown in Figures 2A, B, the number of EMAs completed, and the listening context in which they were completed, varied by time of day. The binning of sound environmental data into tertile splits resulted in comparable boundaries of the absolute levels across participants (see median and SD for each split in Table 2). The total number of EMAs completed also varied extensively across participants ($M = 90.83$, SD = 58.28, range = 12 to 299). This appears to be primarily driven by the time for which the hearing aid had a Bluetooth connection to the associated smartphone as measured by the daily hearing-aid logging time ($r = 0.714$, df = 78, $t = 9.02$, $p < 0.001$). While an alternative explanation is that some participants might simply have chosen not to troubleshoot connection failures. However, we consider this unlikely because participants were timely in their communications and showed flexibility around their online appointments.

On a group level, the completion of EMAs roughly followed the temporal pattern of hearing-aid logging time (Fig. 2C) with around 0.15 EMAs per hour across both wear periods. Furthermore, Christensen et al. (2023), using the same set of EMA data, documented that EMAs were completed in comparable sound environments as those experienced outside of performing EMAs (see Fig. 2 in Christensen et al. 2024). This suggests that EMAs represent the "typical" daily-life hearing-aid experience both in terms of how often and in which environments they were worn.

Contextual Effects • As with the SSQ outcomes, we assessed the utility of the EMAs by evaluating how well they discriminated between the preferred and nonpreferred hearing-aid model. The role of contextual data (i.e., self-reported listening activity and logged sound environment) was considered by comparing how well models of increasing complexity accounted for ratings on the EMAs. Discriminability of individual levels of the contextual predictors was then assessed by estimating marginal means separated by preference. The simplest model (LM1) used only the information about end-of-trial preference and EMA item as fixed effects to predict ratings, the next model (LM2) also included the reported listening activity as a categorical fixed effect, and finally, the most complex model (LM3) additionally included information about the sound environment using a categorical predictor with nine levels (the nine categories of the sound environment). The partial R^2 (explained variance by fixed effects alone) was 18.4% for LM1, 25.0% for LM2, and 27.6% for LM3. The higher partial R^2 for LM3 was not confounded by overfitting from the increase in degrees of freedom because the Akaike's information criterion (AIC) was lower than for both LM1 and LM2 (LM3 versus LM1 ΔAIC = −2416; LM3 versus LM2 ΔAIC = −741) and likelihood ratio tests comparing LM3 to LM1 and LM2 were significant (LM3 versus LM1 $\chi^2(105) = 2626.20$, $p < 0.001$; LM3 versus LM2 $\chi^2(21) = 782.74$, $p < 0.001$). Thus, from the model comparisons, the self-reported listening activity in combination with information about the sound environment produced the best fit to the listening experiences collected with EMA. In Table 3, all main effects, and interactions of LM3 are listed based on type III F-tests for fixed effects. There were significant main effects for all predictors and all interaction terms are significant. These interaction effects document that hearing-aid preference is evident in EMA ratings, but that the

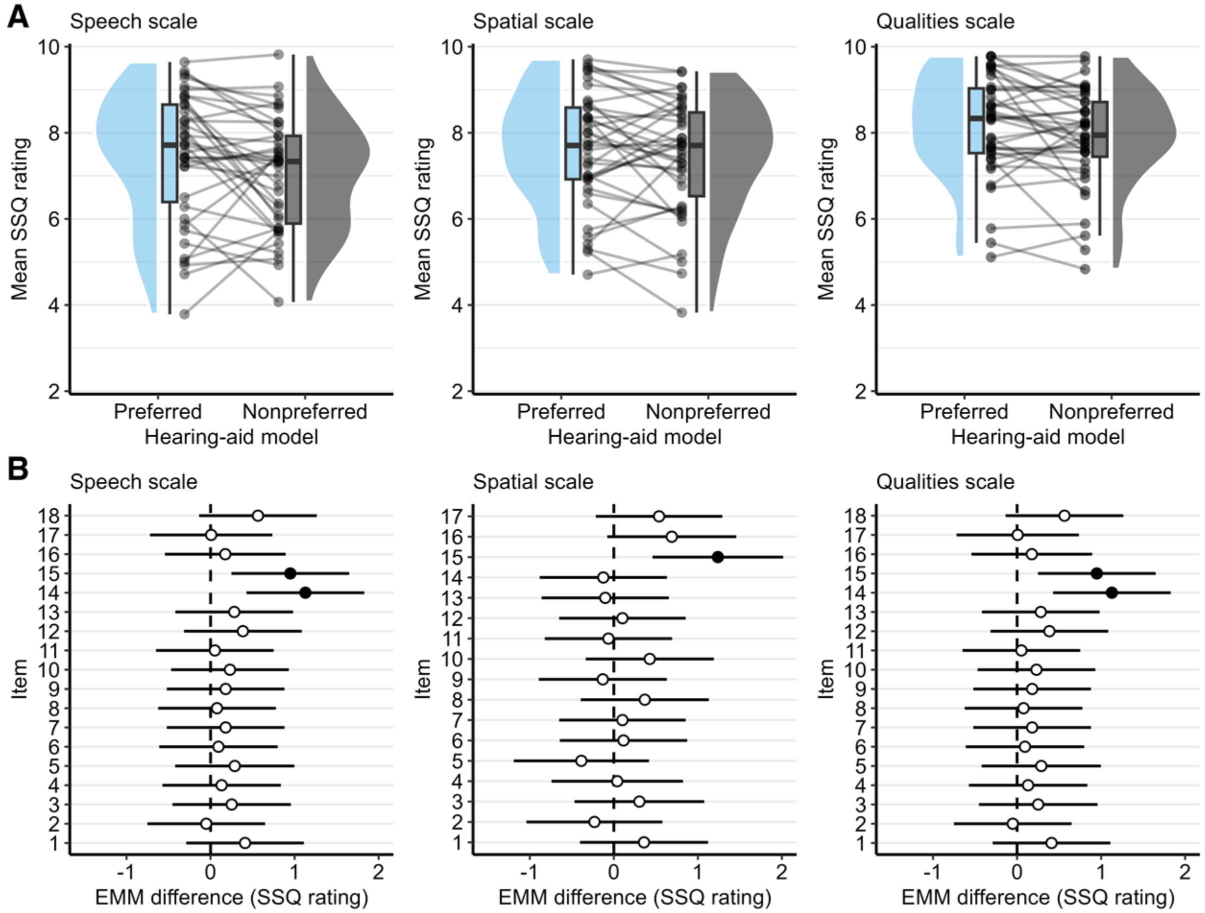

Fig. 1. SSQ ratings. A, Distributions and boxplots of mean SSQ ratings across items from each subscale of the SSQ. B, EMM differences between the preferred and non-preferred hearing aid. Error bars indicate the 95% confidence intervals and the point fill indicates statistical significance (black: $p < 0.05$; white: $p \geq 0.05$) from paired comparisons of the EMMs with Tukey's HSD. EMM indicates estimated marginal mean; SSQ, Speech, Spatial, and Qualities of Hearing Scale.

strength of this is moderated by EMA items and contexts. Note that during statistical modeling, we compared the outcome coefficients of the models applied to the full data with those from models applied to a subset of the dataset where datapoints associated with participant's indication of whether the listening situation was still happening at the time of EMA were removed (23% of the data points). However, because the model outcomes (magnitude and direction of coefficients) were close to identical, we preserved all observations for the subsequent and final analysis.

Next, we assessed discriminability of individual levels of the predictors by estimating marginal mean ratings separated by preference. In Figure 3, the marginalization is shown for self-reported listening activity (A, top), data-logged sound environment (B, middle), and EMA item (C, bottom). Listening activity "People talking," "Television," "Music (live)," "Sounds around me," and "Nothing in particular" were all significantly sensitive to hearing-aid preference with the largest effect observed for "Music (live)" (Cohen's d = 0.243). In addition, the listening

TABLE 1. EMM differences in ratings between the end-of-trial preferred and nonpreferred hearing-aid model with the SSQ questionnaire

Scale:Item	Question	EMM Diff. (SE)	p	Cohen's d
Spatial:15	Do the sounds of people or things you hear, but cannot see at first, turn out to be closer than expected when you do see them?	1.24 (0.40)	<0.001	0.78
Qualities:14	Do you have to concentrate very much when listening to someone or something?	1.13 (0.36)	<0.001	0.75
Qualities:15	Do you have to put in a lot of effort to hear what is being said in conversation with others?	0.95 (0.36)	0.002	0.63
Speech:8	Can you have a conversation with someone when another person is speaking whose voice is the same pitch as the person you're talking to?	0.85 (0.36)	0.020	0.548
Speech:1	You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the person you're talking to says?	0.82 (0.35)	0.020	0.528
Speech:14	You are listening to someone on the telephone and someone next to you starts talking. Can you follow what's being said by both speakers?	0.81 (0.35)	0.023	0.520

Only scales and items exhibiting significant differences are shown.
EMM, estimated marginal mean; SSQ, Speech, Spatial, and Qualities of Hearing Scale.

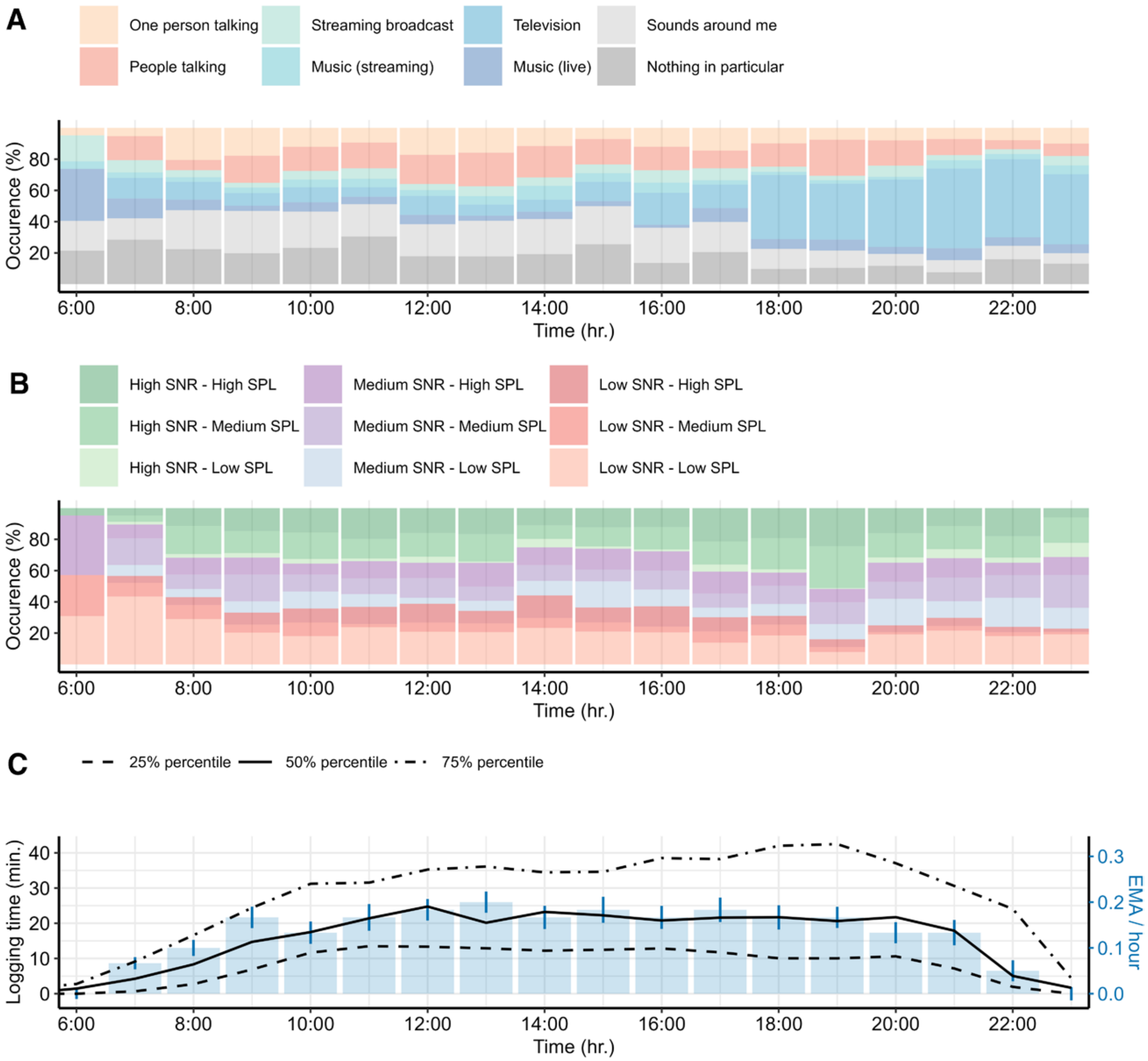

Fig. 2. Distribution of self-reported listening activity and logged sound environment information (averaged across participants) during EMA completion (A and B), and hearing-aid logging time and EMA completions per hour (C). A, Reddish colors represent speech listening, bluish colors represent focused listening, and grayish colors represent nonspecific listening according to the Common Sound Scenarios (Wolters et al. 2016). C, The hearing-aid logging time (percentiles) is shown on the left y axis and average EMA per hour across participants is shown on the right y axis. Error bars represent the standard errors. Note that the figure displays data from 6 A.M. to 11 P.M. for sake of illustration. There were 55 EMAs completed outside of this time, and these were also included in the analysis. EMA indicates Ecological Momentary Assessment.

conditions with the highest sensitivity to hearing-aid preference were those with predominantly low or medium SNR (largest for “Medium SNR – High SPL,” Cohen’s *d* = 0.283). Lastly, all EMA items but Q1 exhibited significant rating differences between the preferred and nonpreferred hearing aids. EMA item Q2 (“Sound satisfaction”) exhibited the highest sensitivity (Cohen’s *d* = 0.232) and the lowest was for Q5 (“Spatial localization,” Cohen’s *d* = 0.087).

Classification of Ratings by Preference

On a group level, SSQ and EMA responses favored the chosen end-of-trial hearing-aid model in that mean ratings were higher for the preferred than the nonpreferred model. When considering mean rating differences (shown in Fig. 1A) 26 of the 39 participants (66.7%) had higher SSQ ratings for the preferred than for the nonpreferred hearing aid. When looking only at the six SSQ items that showed a significant discriminability, the number of participants with higher mean SSQ for the preferred hearing-aid model increased to 28 (71.8%). For the EMA data, the number of participants with higher mean rating for the preferred hearing-aid model was 24 (60.0%) and 29 (72.5%) when considering either all ratings or only ratings performed in contexts with a significant association to preference as per Figure 3. Twenty participants had overlapping data, that is, both SSQ responses and EMAs reflected their hearing-aid preference, the preference of 8 participants was correctly predicted by the SSQ data only, and the preference of 9 participants was correctly predictedby the EMA data only (Fig. 4A). The preference in 3 participants could not be predictedwith neither SSQ nor

TABLE 2. Median (and SD) of the SPL and SNR for each binned tertile split

Split (Percentile Range)	SPL (dB)	SNR (dB)
Hearing-aid model 1		
Low (0–33%)	57.78 (6.03)	3.29 (3.44)
Medium (33–66%)	65.00 (4.75)	7.47 (2.72)
High (66–100%)	72.87 (5.45)	11.74 (2.50)
Hearing-aid model 2		
Low (0–33%)	52.00 (6.05)	4.74 (2.30)
Medium (33–66%)	61.07 (5.43)	9.07 (2.08)
High (66–100%)	68.91 (4.98)	13.58 (1.94)

Note that the tertile splits for SPL and SNR were computed independently of each other.
SNR, signal to noise ratio.

EMA data. These results yield a combined diagnostic prediction accuracy of 92.5%.

We further investigated how well the repeated EMAs could aid in a prognostic prediction of hearing-aid model preference by utilizing the contextual information available for each EMA. For example, EMAs performed at specific times of the day or in specific situations, might differ in their contribution to a final hearing-aid preference because they reflect changes in fatigue or situational importance, which is also indicated by the contextual effects identified by the linear mixed-effects modeling (Fig. 3). The overall prediction accuracy of the RF models across folds ranged from 79.6% (SD = 0.5%) for the EMA data alone (model M1), 88.0% (SD = 0.6%) with self-reported listening activity (model M2), 88.3% (SD = 0.7%) with data-logged sound environment information (model M3), and to 93.8% (SD = 0.3%) when both self-reported listening activity and data-logged sound environment information was included as features (model M4). These accuracies represent proportions of ratings pooled among all participants that could be correctly predicted.

Figure 4B shows the prediction accuracies by participant using EMA alone (i.e., without contextual information, model M1) on the x axis and with contextual information (y axis) about listening activity (black dots, model M2) and the sound environment (orange dots, model M3) and the combination of listening activity and sound environment (blue dots, model M4).

The contextual information provided by self-reported listening activity increased the prediction accuracy for 39 of the 40 participants (ΔM = 7.8%-point, SD = 4.5 %-point) while only adding sound environment information increased the prediction accuracy of all 40 participants (ΔM = 7.8%-point, SD = 4.7 %-point). The largest increase in prediction accuracy was when both self-reported listening activity and sound environment information were added (ΔM = 12.7%-point, SD = 6.3%-point). Notably, the largest benefit of including contextual information was for participants with lower prediction accuracies on the EMAs alone, suggesting that the contextual information for these individuals contributed relatively more to the preference prediction than for individuals with already high prediction accuracies. This is further corroborated by the feature importance shown on Figure 5A, which documents that inclusion of contextual data (models M2-M4) lowers the importance of the EMA rating magnitude itself in predicting preference. That is, for certain individuals, the EMA rating magnitude is relatively less important for predicting preference when the contextual information (i.e., the environments in which EMAs were completed) is provided. It is interesting that including EMA items yields a negative mean decrease in accuracy (Fig. 5A, bottom), suggesting that prediction accuracy would increase if this feature was left out of the RF models.

Figure 5B shows the partial dependency of the four features included in all RF models. The partial dependency represents how feature values differ in their relative importance for the prediction accuracy independently of other features (Cutler et al. 2007). Most notably, the dependency on EMA responses drops after 8 days of testing, while responses made in the morning are most important for the prediction accuracy. The partial dependency on EMA rating magnitude documents that ratings higher than six are important for shaping preference, while ratings between four and six add to a nonpreference. The partial dependency on EMA items exhibits large across-fold variations (error bars), and for models M2-M4 the items contributed equally to the prediction.

Lastly, there was a significant negative correlation between numbers of EMAs completed and the prognostic prediction accuracy for model M4 ($r = -0.54$, df = 38, $p = < 0.001$).

DISCUSSION

We compared everyday listening experiences with two hearing-aid models using retrospectively completed SSQ questionnaires and in-situ EMAs to document the utility of each approach for predicting a hearing-aid preference. On balance, EMA and SSQ ratings were similarly predictive of participants' hearing-aid preference. More specifically, EMA and SSQ ratings were equally effective at a diagnostic prediction of preference for about half the participants, EMA ratings were superior to SSQ ratings for about 20% of participants, and SSQ ratings were superior to EMA ratings for an additional 20% of participants. This illustrates the value of combining data from

TABLE 3. Type III F-tests with Satterthwaite's method for fixed effects (model LM3)

Coefficient	Df1	Df2	*F*	*p*
EMA item	5	21,724	657.47	<0.001
Preference	1	21,680	48.55	<0.001
Listening activity	7	11,074	105.40	<0.001
Sound environment	8	21,495	61.95	<0.001
EMA item: Preference	5	21,724	3.32	0.005
EMA item: Listening activity	35	21,724	19.79	<0.001
EMA item: Sound environment	40	21,724	8.14	<0.001
Preference: Listening activity	7	21,676	3.15	0.003
Preference: Sound environment	8	21,674	3.91	<0.001

":" indicate an interaction between coefficients.
Df1, numerator degrees of freedom; Df2, denominator degrees of freedom; EMA, Ecological Momentary Assessment.

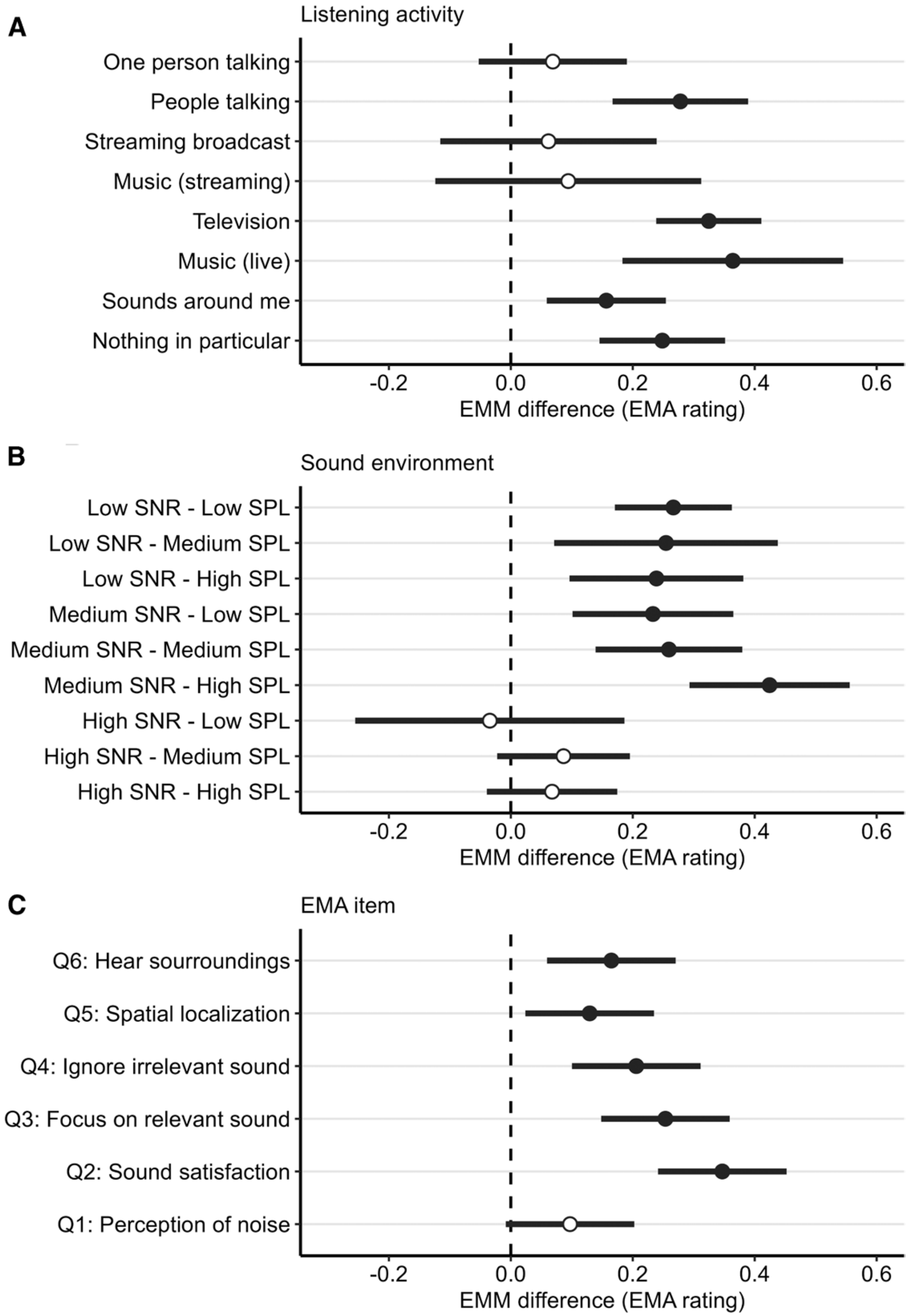

Fig. 3. EMM differences and 95% confidence intervals in EMA ratings of the preferred vs the nonpreferred hearing aid. The marginalization is shown for self-reported listening activity, sound environment information (logged by the hearing aids), and EMA items. Note that open symbols indicate that the EMM difference was not significantly larger than 0 ($\alpha = 0.05$) when tested with paired comparisons. EMA indicates Ecological Momentary Assessment; EMM, estimated marginal mean.

SSQ and EMA if one is interested in understanding hearing-aid preferences around noise reduction features. However, neither EMA nor SSQ data correctly predicted preference for some participants. This might be because those individuals did not have a strong preference for either hearing aid. Indeed, some participants said they did not think there was much difference

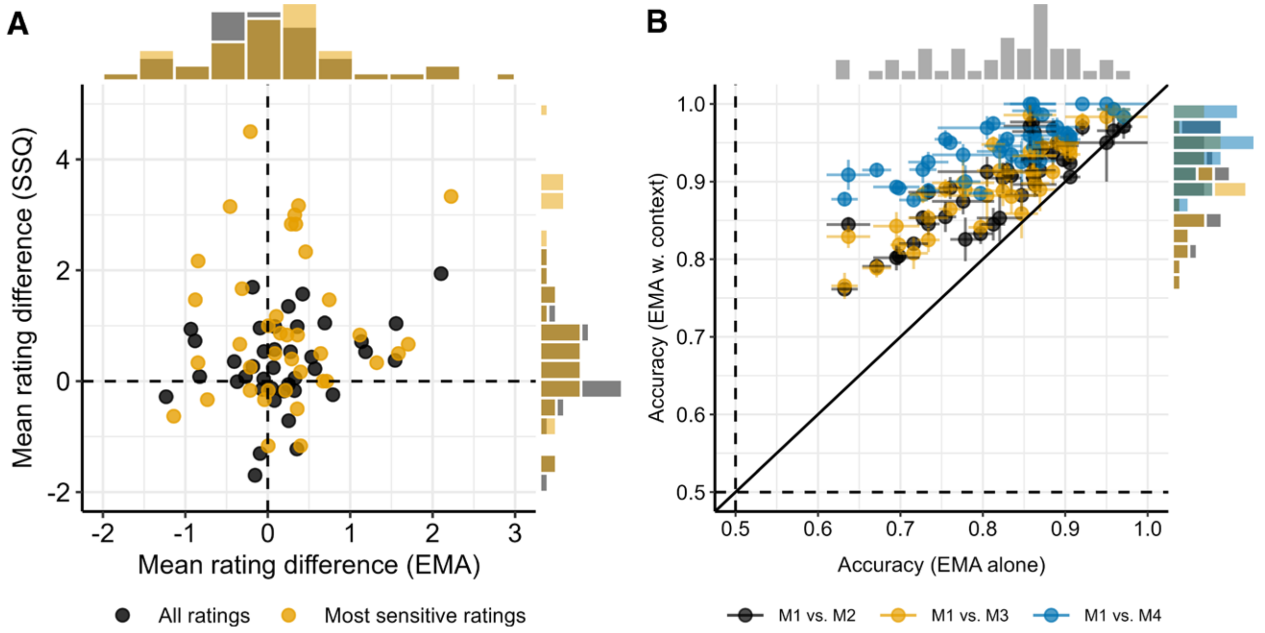

Fig. 4. Mean rating differences (A) and prediction accuracy per participant from Random Forest classifiers (B). A, The dashed lines represent the classification decision criterion, while the colors separate mean differences based on either all ratings (black) or selected subsets of ratings belonging to the most sensitive SSQ items or EMA contexts (orange, see Figs. 1, 3). B, The dashed lines represent the 50% accuracy. Predictions with the EMA data included features for EMA item, hour of day, day of wear period, and EMA rating magnitude (M1, x axis), self-reported listening activity (M2, y axis) or sound environment information from data-logging (M3, y axis), or both self-reported listening activity and sound environment information (M4, y axis). Error bars for each point reflect the across-fold SD. In both panels, marginal histograms of the data are show. EMA indicates Ecological Momentary Assessment; SSQ, Speech, Spatial, and Qualities of Hearing Scale.

between the two models of hearing aid, while others chose based on a nonauditory feature such as comfort. To examine this, it would be of interest to document strength of preference in addition to the ultimate preference.

The contexts that contributed most strongly to predicting the preferred hearing aid were complex listening situations. Specifically, this was hearing speech in noise, spatial localization of nonvisible sound sources and listening effort for the SSQ, and listening to music, television, and people talking in poor SNRs for EMA. This suggests that challenging and effortful situations mediate decisions about hearing-aid preference which is unsurprising because most people with mild-moderate hearing loss have little difficulty hearing in noncomplex situations, such as one-to-one conversations in high SNR environments (Schinkel-Bielefeld et al. 2023).

Only six SSQ items showed a significant difference between the preferred and nonpreferred hearing-aid model, which suggests that when predicting hearing-aid noise reduction preferences, the SSQ could be reduced to fewer items. Of these six items, one was related to spatial listening and two were related to listening effort. These aspects of hearing are not assessed by measures typically used in clinical practice, for example, the Client Oriented Scale of Improvement or the Hearing Handicap Inventory for Adults (Newman et al. 1990; Dillon et al. 1999). The fact that these items are similar to those identified as significant contextual predictors from the EMA data supports the notion that these listening situations are important for outcomes and should be routinely assessed in a clinical setting.

Our data also illustrate the value of supplementing EMAs with contextual information in the form of a self-reported listening activity and/or data-logged sound environment information, and that together they add complementary information. The evidence for this is the finding that when using EMA alone the RF model for a prognostic prediction reached almost 80% accuracy, but that this increased to almost 90% when either form of contextual information was added separately, and to over 93% when both were added. This strongly supports the use of EMA for individualized hearing outcome investigations (Timmer et al. 2018) and suggests that EMAs during clinical trials could be limited to relevant situations only—as long as those relevant situations for each individual are identified up front. Indeed, hearing-aid users typically use their hearing aids proactively to master situation-specific demands (Williger & Lang 2015) which would explain why we observe a strong effect of context on the sensitivity of EMAs. Moreover, the results suggest that contextual information lowers the relative reliance on EMA rating magnitudes for understanding hearing-aid preference (as shown in Fig. 5B). In turn, this seems to improve the prediction accuracy, particularly for those with lower accuracy when relying solely on the EMA (Fig. 4B). Thus, incorporating data-logging from hearing aids could increase the validity of outcomes from EMA studies. It is interesting that a negative correlation was observed between the number of EMAs and the accuracy of prognostic predictions. This indicates that individuals who complete fewer EMAs tend to provide highly meaningful responses that largely align with a preference. On the other hand, those who complete more EMAs may be less predictable because their EMAs are completed in diverse situations. For these individuals, the inclusion of contextual data appears to aid prediction.

Lastly, the Random Forest models applied to EMA ratings not only provide highly accurate predictions of hearing-aid

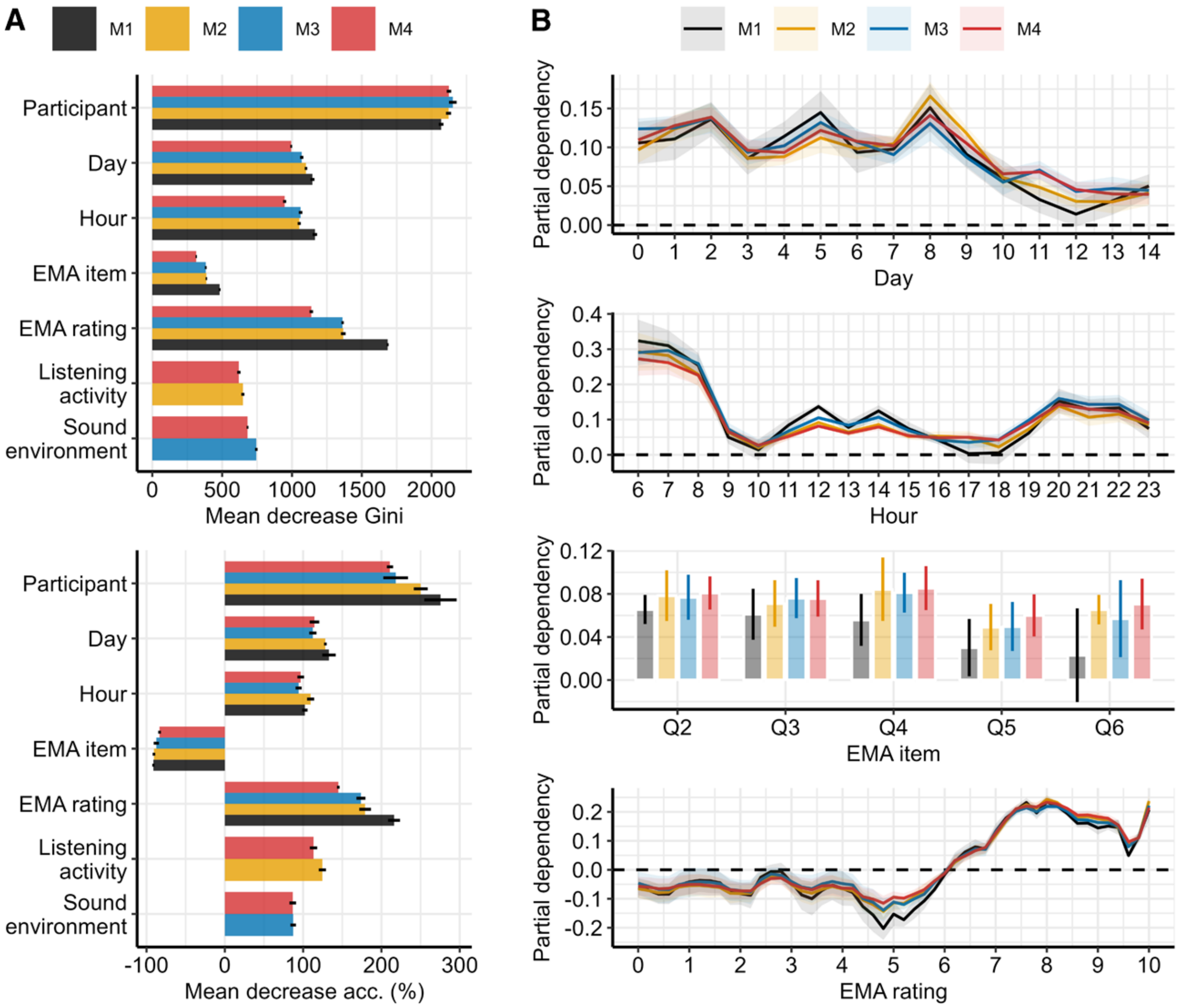

Fig. 5. Feature importance (A) and partial dependency (B) derived from the four RF models. The importance represents how much an RF model's performance would suffer if the associated feature was removed. It is defined in either a mean decrease in impurity (Gini index) or in mean decrease in prediction accuracy. PD ranks how independent feature levels contribute to the probability of predicting a preference. PD indicates partial dependency.

preference but also help uncover the factors influencing the choice. This underscores the method's utility as an alternative to black-box machine learning approaches. For example, the partial dependency (Fig. 5B) reveals that EMAs beyond day 10 seem to have less impact on the chosen hearing aid indicating that preference beyond 10 days is settled. It is interesting that EMAs completed in the early morning hours (between 6 A.M. and 8 A.M.) appear to have a greater influence on which hearing aid is chosen compared with other times of the day. We speculate that this could be because morning listening activities (e.g., listening to the radio) are conducted when there is relatively more time available than at other times of day and that performing EMAs therefore are less intrusive. Consequently, the EMAs completed during these hours might more accurately reflect the individual's true preference. Another possible explanation could be that the person is fresh and less fatigued at these times, thus, completing more consistent EMAs.

It is also noteworthy that the preference decision is most sensitive to EMA ratings above 7. In contrast, ratings between 4 and 6 are associated with negative partial dependencies, suggesting they are more likely to predict the nonpreferred hearing-aid model (see bottom panel, Fig. 5B).

Limitations

First, in the present study, two models of hearing aid were evaluated. The most significant difference between them was in the way noise reduction is implemented. Thus, the sensitivity of the SSQ responses and the EMAs possibly reflect differences specific to the hearing-aid models, and the contextual sensitivity identified with the self-reporting and data-logging might indicate situations where this specific difference is perceptually largest. Thus, whether this can be generalized to other hearing aids or hearing aid features needs to be determined. However, the methodology used for identifying the contextual sensitivity holds for comparisons of any technology.

Second, all participants here were experienced hearing-aid users, who might be differently sensitive to subtle differences in hearing-aid processing than new users. Again, this needs to be examined. The fact that audiometric thresholds measured within the prior 3 years were used to program the hearing aids and the fact that real ear verification was not used could have resulted in suboptimal hearing-aid fittings. While this is not ideal, it was unavoidable because the study was conducted during the COVID-19 pandemic when face-to-face research appointments were not permitted. However, all participants were experienced hearing-aid users who presumably knew how they wanted their hearing aids to sound, and fine-tuning of all fittings was provided, so the impact of these issues is expected to have been minimal. Third, while sound levels were measured differently by the two models of hearing aid (due to differences in frequency weighting of the SPL estimates), we do not expect this will have impacted the current results. This is because data were split based on tertiles for each individual and hearing-aid model and because SPL and SNR from the two hearing-aid models differed only by an offset and not by scaling (Christensen et al. 2023). Lastly, data collection took place between August 2021 and June 2022. While there were no formal lockdowns in the UK during this time period, COVID-19 was still rampant and face masks were mandatory in public spaces, thus participants likely limited their activities outside the home. This will undoubtedly have impacted the variety of listening situations participants experienced.

CONCLUSIONS

Real-world listening experiences from retrospective SSQ responses or momentarily sampled by EMAs can equally well reflect a hearing-aid preference with diagnostic prediction accuracies of 71.8% and 72.5%, respectively, and 92.5% when combined. Furthermore, when combined with self-reported listening activity and estimations of the sound environment, EMAs can provide a highly accurate prognostic prediction (over 93%). This highlights the importance of integrating EMA trials with objective data-logging for personalized and contextualized hearing outcomes. In clinical trials, the burden on participants can be reduced by directing their ratings toward the most relevant contexts. If this is not possible, it is important to consider the listening conditions under which self-reports are made. This will allow for a better understanding of how these conditions influence preferences or benefits.

ACKNOWLEDGMENTS

This project was supported by the NIHR Manchester Biomedical Research Centre (NIHR203308) and funded in part by Oticon A/S. The funding organizations had no role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; or in the decision to submit the article for publication; or in the preparation, review, or approval of the article. The views expressed are those of the author(s) and notnecessarily those of the NIHR or the Department ofHealth and Social Care.

J.H.C. and G.H.S. conceived original idea and design, M.L., H.W., and J.R. contributed to the final protocol. H.W. and M.L. collected the data. J.H.C. and J.C.G.-C. processed, analyzed, and visualized the data. J.H.C. wrote the initial draft of the manuscript, all authors contributed to revisions. All authors read and approved the final version of the manuscript.

Ethical approval for the study was provided by The University of Manchester—University Research Ethics Committee 3 (UREC3) 2020-10322-16539. All participants gave their written informed consent before enrollment in the study.

The authors have no conflicts of interest to disclose.

Address for correspondence: Jeppe H. Christensen, Eriksholm Research Centre, Oticon A/S, Rørtangvej 20, Snekkersten, 3070, Denmark. E-mail: jych@eriksholm.com

Received November 15, 2023; accepted April 07, 2024

REFERENCES

Andersen, A. H., Santurette, S., Pedersen, M. S., Alickovic, E., Fiedler, L., Jensen, J., Behrens, T. (2021). Creating clarity in noisy environments by using deep learning in hearing aids. *Semin Hear*, *42*, 260–281.

Andersson, K. E., Andersen, L. S., Christensen, J. H., Neher, T. (2021). Assessing real-life benefit from hearing-aid noise management: SSQ12 questionnaire versus ecological momentary assessment with acoustic data-logging. *Am J Audiol*, *30*, 93–104.

Andersson, K. E., Neher, T., Christensen, J. H. (2023). Ecological momentary assessments of real-world speech listening are associated with heart rate and acoustic condition. *Front Audiol Otol*, *1*, 1275210. https://doi.org/10.3389/fauot.2023.1275210.

Bosman, A. J., Christensen, J. H., Rosenbom, T., Patou, F., Janssen, A., Hol, M. K. S. (2021). Investigating real-world benefits of high-frequency gain in bone-anchored users with ecological momentary assessment and real-time data logging. *J Clin Med*, *10*, 3923.

Bradburn, N. M., Rips, L. J., Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, *236*, 157–161.

Bzdok, D., & Ioannidis, J. P. A. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci*, *42*, 251–262.

Christensen, J. H., Saunders, G. H., Havtorn, L., Pontoppidan, N. H. (2021). Real-world hearing aid usage patterns and smartphone connectivity. *Front Digit Health*, *3*, 722186. https://doi.org/10.3389/fdgth.2021.722186.

Christensen, J. H., Whiston, H., Lough, M., Gil-Carvajal, J. C., Rumley, J., Saunders, G. H. (2024). Evaluating real-world benefits of hearing aids with deep neural network-based noise reduction: An Ecological Momentary Assessment study. *Am J Audiol*, *18*, 1–12. doi: 10.1044/2023_AJA-23-00149. Epub ahead of print. Erratum in: Am J Audiol. 2024 Apr 18; 1. PMID: 38354098.

Collins, G. S., Reitsma, J. B., Altman, D. G., Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Circulation*, *131*, 211–219.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*, 2783–2792.

Dillon, H., Birtles, G., Lovegrove, R. (1999). Measuring the outcomes of a national rehabilitation program: Normative data for the Client Oriented Scale of Improvement (COSI) and the Hearing Aid User's Questionnaire (HAUQ). *J Am Acad Audiol*, *10*, 67–79.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation*. *Bell Syst Tech J*, *12*, 377–430.

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

Hart, A., Reis, D., Prestele, E., Jacobson, N. C. (2022). Using smartphone sensor paradata and personalized machine learning models to infer participants' well-being: Ecological momentary assessment. *J Med Internet Res*, *24*, e34015.

Hastie, T., Friedman, J., Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer. https://doi.org/10.1007/978-0-387-21606-5.

Holube, I., von Gablenz, P., Bitzer, J. (2020). Ecological momentary assessment in hearing research: Current state, challenges, and future directions. *Ear Hear*, *41*(Suppl 1), 79S–90S.

Kates, J. M. (2008). *Digital Hearing Aids*. Plural Publishing.

Kjems, U., & Jensen, J. (2012). Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement. 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 295–299.

Klyn, N. A. M., Robler, S. K., Bogle, J., Alfakir, R., Nielsen, D. W., Griffith, J. W., Carlson, D. L., Lundy, L., Dhar, S., Zapala, D. A. (2019). CEDRA—A tool to help consumers assess risk for ear disease. *Ear Hear*, *40*, 1261.

Lelic, D., Wolters, F., Schinkel-Bielefeld, N. (2024). Measuring hearing aid satisfaction in everyday listening situations: Retrospective and in-situ assessments complement each other. *J Am Acad Audiol*. https://doi.org/10.1055/a-2265-9418. Epub ahead of print. PMID: 38336116.

Ling, D. (1976). Hearing aids and the use of residual hearing. *Aust J Hum Commun Disord*, *4*, 9–14.

Newman, C. W., Weinstein, B. E., Jacobson, G. P., Hug, G. A. (1990). The hearing handicap inventory for adults. *Ear Hear*, *11*, 430–433.

Noble, W., Jensen, N. S., Naylor, G., Bhullar, N., Akeroyd, M. A. (2013). A short form of the Speech, Spatial and Qualities of Hearing scale suitable for clinical use: The SSQ12. *Int J Audiol*, *52*, 409–412.

Pasta, A., Petersen, M. K., Jensen, K. J., Pontoppidan, N. H., Larsen, J. E., Christensen, J. H. (2022). Measuring and modeling context-dependent preferences for hearing aid settings. *User Model User-Adapt Interact*, *32*, 977–998.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychol Bull*, *128*, 934–960.

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience*, *6*, gix019.. https://doi.org/10.1093/gigascience/gix019.

Schinkel-Bielefeld, N. (2020). Laboratory experiments versus ecological momentary assessment? The quest to evaluate real life hearing aid performance. *Forum Acusticum*, 91–98. https://doi.org/10.48465/fa.2020.0545.

Schinkel-Bielefeld, N., Kunz, P., Zutz, A., Buder, B. (2020). Evaluation of hearing aids in everyday life using ecological momentary assessment: What situations are we missing?. *Am J Audiol*, *29*, 591–609.

Schinkel-Bielefeld, N., Ritslev, J., Lelic, D. (2023). Reasons for ceiling ratings in real-life evaluations of hearing aids: The relationship between SNR and hearing aid ratings. *Front Digit Health*, *5*, 1134490.

Shiffman, S., Stone, A. A., Hufford, M. R. (2008). Ecological momentary assessment. *Annu Rev Clin Psychol*, *4*, 1–32.

Souden, M., Benesty, J., Affes, S. (2010). A study of the LCMV and MVDR noise reduction filters. *IEEE Trans Signal Process*, *58*, 4925–4935.

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Ann Behav Med*, *24*, 236–243.

Timmer, B. H. B., Hickson, L., David, M., Launer, S. (2018). Do hearing aids address real-world hearing difficulties for adults with mild hearing impairment? Results from a pilot study using ecological momentary assessment. *Trends Hear*, *22*, 233121651878360.

Uwe Simmer, K., Bitzer, J., Marro, C. (2001). Post-filtering techniques. In M. Brandstein & D. Ward (Eds.), *Microphone Arrays: Signal Processing Techniques and Applications* (pp. 39–60). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-04619-7_3

van Smeden, M., Reitsma, J. B., Riley, R. D., Collins, G. S., Moons, K. G. (2021). Clinical prediction models: Diagnosis versus prognosis. *J Clin Epidemiol*, *132*, 142–145.

Walden, B. E., Surr, R. K., Cord, M. T., Dyrlund, O. (2004). Predicting hearing aid microphone preference in everyday listening. *J Am Acad Audiol*, *15*, 365–396.

Williger, B., & Lang, F. R. (2015). Hearing aid use in everyday life: Managing contextual variability. *Gerontology*, *61*, 158–165.

Wolters, F., Smeds, K., Schmidt, E., Christensen, E. K., Norup, C. (2016). Common sound scenarios: A context-driven categorization of everyday sound environments for application in hearing-device research. *J Am Acad Audiol*, *27*, 527–540.

Wu, Y.-H., Stangl, E., Chipara, O., Gudjonsdottir, A., Oleson, J., Bentler, R. (2020). Comparison of in-situ and retrospective self-reports on assessing hearing aid outcomes. *J Am Acad Audiol*, *31*, 746–762.

Wu, Y.-H., Stangl, E., Zhang, X., Bentler, R. A. (2015). Construct validity of the ecological momentary assessment in audiology research. *J Am Acad Audiol*, *26*, 872–884.

Yellamsetty, A., Ozmeral, E. J., Budinsky, R. A., Eddins, D. A. (2021). A comparison of environment classification among premium hearing instruments. *Trends Hear*, *25*, 233121652098096.