

# Development and Application of an Annotation Procedure to Assess the Impact of Hearing Aid Amplification on Interpersonal Communication Behavior

Trends in Hearing  
Volume 22: 1–17  
© The Author(s) 2018  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/2331216518816201  
journals.sagepub.com/home/tia



Markus Meis<sup>1,2</sup>, Melanie Krueger<sup>1,2</sup>, Petra v. Gablenz<sup>2,3</sup>,  
Inga Holube<sup>2,3</sup>, Maria Gebhard<sup>1,2,3</sup>, Matthias Latzel<sup>4</sup>, and  
Richard Paluch<sup>1,2,5</sup>

## Abstract

Hearing impairment is associated with a decrease in speech intelligibility and health-related quality of life, such as social isolation and participation restriction. However, little is known about the extent to which hearing impairment and hearing aid fittings change behavior in acute communication situations as well as interrelated behavior patterns. Based on a pilot study, in which the basis for annotating communication behavior was laid, group discussions in noise were initiated with 10 participants using three different hearing-aid brands. The proposed offline annotation scheme revealed that different hearing aids were associated with changes in behavior patterns. These behavioral changes were congruent with speech recognition threshold results and also with subjective assessments. Some of the results were interpreted in terms of participation restriction and activity limitation following the framework of the International Classification of Functioning, Disability and Health. In addition to the offline annotation scheme, a procedure for instantaneous coding of eight behavior patterns was iteratively developed and used for the quick examination of lab studies with good to excellent interrater reliability values.

## Keywords

hearing loss, hearing aid, international classification of functioning, disability and health, quality-of-life, interpersonal communication behavior

Date received: 1 February 2018; revised: 16 October 2018; accepted: 6 November 2018

## Introduction

The benefit of hearing aids (HAs) is typically shown by means of clinically oriented test procedures in the lab, such as speech-in-noise tests estimating the speech recognition thresholds (SRTs). In addition to those efficacy procedures, hearing-specific questionnaires (e.g., Chisolm et al., 2007; Cox & Alexander, 2002), generic questionnaires (e.g., Hunt, McEwen, & McKenna, 1985; Robinson, Gatehouse, & Browning, 1996), or assessments with open-ended questions (Dillon, James, & Ginis, 1997) are used to determine the hearing-aid benefit in everyday life. To date, measures for behavioral patterns of communication abilities of hearing-impaired people are seldom used. The disease-oriented view in clinical standards of diagnostics might explain why the development of communication performance measures

was neglected for a long time, so that sensorineural hearing loss is addressed but not the social and behavioral aftereffects. Since the International Classification of Functioning, Disability and Health (ICF) was approved in 2001, the social and behavioral perspectives on the

<sup>1</sup>Hörzentrum Oldenburg GmbH, Germany

<sup>2</sup>Cluster of Excellence Hearing4all, Oldenburg, Germany

<sup>3</sup>Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany

<sup>4</sup>Phonak AG, Stäfa, Switzerland

<sup>5</sup>Carl von Ossietzky University, Oldenburg, Germany

## Corresponding author:

Markus Meis, Hörzentrum Oldenburg GmbH, Marie-Curie-Strße 2, D-26129 Oldenburg, Germany.

Email: m.meis@hoerzentrum-oldenburg.de



individual who is living with hearing impairment have changed.

The ICF (Deutsches Institut für Medizinische Dokumentation und Information, 2005; World Health Organization, 2001) describes and organizes information on functioning and disability. The biopsychosocial model considers the dynamic interaction between the components: body functions or structure, activities, participation, and also contextual factors, that is, environmental and person-centered factors. In recent publications from Granberg, Möller, Skagerstrand, Möller, and Danermark (2014) and Granberg, de Swanepoel, Englund, Möller, and Danermark (2014), the ICF core set was linked to the domain of audiology. Therefore, it might meet the requirements for external evaluation and observation when using it as a basis for the development of a code or annotation system for communication behavior in real life. Because of its holistic view, the ICF framework is a candidate for reflecting auditory ecology. Gatehouse, Elberling, and Naylor (1999) proposed an *auditory ecology* approach, which takes the objective physical characteristics of everyday listening environments and the individual listener's demands in those environments into account. The term *auditory ecology* is related to *ecological validity*, following Brunswik (1956) or Bronfenbrenner (1977). They referred to the potential utility of various cues for organisms in their ecology (or natural habitat). In this sense, the evaluation of hearing aid amplification in the respective habitat of the hearing-impaired person forms an important goal.

Health-related quality of life (HrQoL) questionnaires, as outcome tools, are focusing on self-administered questionnaires over a past, longer period of time. Therefore, they can be regarded as a measure of long-term HrQoL. Prominent tools of long-term HrQoL are outcome inventories such as the hearing handicap inventory for adults and elderly, the HHI-A (Newman, Weinstein, Jacobson, & Hug, 1990) and the HHI-E (Ventry & Weinstein, 1982), the Glasgow health status inventory (Robinson et al., 1996), and the Nottingham health profile (Hunt et al., 1985), with the subscale *social isolation* as generic tools. In all of the listed disease and generic questionnaires, long-term behavioral aspects are included in different degrees. In the HHI-E or HHI-A variants, two factors are identified as perceived and reported handicaps induced by hearing impairment: the emotional and the social scale. In the social scale, many items are directly related to concrete behavioral patterns, for example, "Does a hearing problem cause you difficulty when visiting friends, relatives, or neighbors?" "Does a hearing problem cause you to talk to family members less often than you would like?" (HHI-A items; Newman et al., 1990). Another prominent questionnaire for assessing disease specific abilities is the Speech,

Spatial and Qualities of Hearing Scale (SSQ), based on Gatehouse and Noble (2004). The reported questionnaires aim toward measuring cumulative effects, usually over a period of 4 weeks, and thus summarize experiences retrospectively. However, the responses are possibly biased by interlocked effects of long-term memory and inference (Bradburn, Rips, & Shevell, 1987; Shiffman, Stone, & Hufford, 2008). To bridge the temporal gap between perception and assessment of the perception, ecological momentary assessment methods have recently been applied in audiological research (e.g., Bitzer, Kissner, & Holube, 2016; Kowalk, Kissner, von Gablenz, Holube, & Bitzer, 2018; Wu, Stangl, Zhang, & Bentler, 2015) to avoid distortions by averaging auditory experiences over a long period of time. The development of measures to assess behaviors was established in this realm.

To our knowledge, the first study conducted to observe communication behavior in virtual life scenarios was a study from Paluch, Latzel, and Meis (2015). They showed that communication behavior changes depending on different HA modes. Ten elderly hearing-aid users were fitted with different beamformers and different modes. The participants were encouraged to discuss four topics of general interest, both in a rather quiet and in a loud scenario while the conversation was recorded on video. In addition to questionnaires, external raters watched the video and rated the participant's communication behavior. Most important, Paluch et al. (2015) qualitatively confirmed two core dimensions of communication behavior: *forms of interaction* (face-to-face [F-t-F] vs. group communication) and *interdependence* (symbolic gestures vs. spoken words). However, the pilot study by Paluch et al. (2015) showed some limitations. The *n* of test participants was small, and four codes as well as the frequency of annotations were possibly not sufficient.

Based on the theoretical concept and the empirical data from this exploratory study, the present contribution outlines the development of an annotation system for application: First, the definition of a meaningful communication scenario and its validation, second, an offline behavior-code system of interpersonal communication for laboratory use, and third, the development of an instantaneous code system in two iterations for a quick examination of the data, possibly suitable for the usage in real-life settings.

It is important to note that we use the terms interaction, communication, and conversation. Interaction refers to reciprocal actions of the communication partners, which occur during the experimental settings. We analyzed only the interpersonal communication behavior both in listening and talking for each person. We defined verbal communication as conversation.

## Description and Creation of the Communication Scenario

We developed a communication situation simulating a typical group conversation known to be challenging for hearing-impaired individuals. We oriented ourselves toward a scenario, described in the SSQ with Item 4 of the SSQ Speech scale: “You are in a group of about five people in a busy restaurant. You can see everyone else in the group. Can you follow the conversation?” (Speech, spatial and qualities of hearing scale [SSQ], Gatehouse & Noble, 2004). In contrast to questionnaire tools, we were interested not only in the subjective evaluation of such a scene but also in the behavior patterns as a measurable outcome tool.

We implemented this scenario as a group discussion within the communication acoustic simulator, a room with virtual acoustics in the House of Hearing in Oldenburg, Germany. The reverberation time of the communication acoustic simulator was set to  $T_{60} = 0.38$  s for the frequencies 250 Hz to 4 kHz. The illuminance was  $>1.000$  lux to enable lip reading. Four loudspeakers were used to create a diffuse sound field simulating an open restaurant inside a shopping mall. An average level (time period of 15 min for an entire group discussion) of  $L_{Aeq} = 67$  dB SPL was measured in the center of a table arrangement covering an area of 1.5 m to 2.1 m.

The participants, all hearing-impaired and experienced hearing-aid wearers, were seated around the table with near and distant communication partners for three group sessions with a duration of 15 min each (see Figure 1 for an example). Within each group discussion, the participants were motivated by a trained moderator to discuss four topics related to hearing impairment and HAs. To stimulate participant involvement, it was made

clear that the content of the discussions would later be analyzed. The discussion sessions were video-recorded with three cameras. The recorded material allowed for replays and ratings by different individuals. All the data reported in this article refer to this basic scenario displayed in Figure 1.

The first outcome measure was named Analyses of Interpersonal Communication in Realistic Acoustical Experimental Settings (AICRAS<sup>©</sup>) and consisted of a questionnaire with seven items (see chapter subjective assessment). The participants completed the AICRAS questionnaire directly after the group discussions. The participants were encouraged to discuss the several topics, assuming a general interest of all participants, so that all would be both active (talking) and passive (listening) participants in the discussions. In addition, a second outcome measure was used and was named Video-based Analyses of Interpersonal Communication in Realistic Acoustical Experimental Settings (VIB-AICRAS<sup>©</sup>). External raters watched video recordings of the group discussions and rated the interpersonal communication behavior.

For all the methods referring to the video-based ratings, the assessors were blinded to the conditions they were analyzing. In addition, they were naïve to the hearing-aid programs which were contrasted.

## Validation of the Communication Scenario With Different In-The-Ear Brands

### Methods

**Subjects.** In total, six males and four females, experienced in hearing-aid usage, participated in the experimental



**Figure 1.** Picture of the moderated communication situation.

study. The mean age was 72.6 years ( $SD$ : 7.6 years). The participants' average pure-tone threshold (0.5, 1, 2, and 4 kHz) of the better ear was 49.7 dB HL ( $SD$ : 6.7 dB HL) and 55.1 dB HL ( $SD$ : 6.9 dB HL) of the worse ear. The participants were divided into two groups of five subjects, which were invited successively.

**Hearing aids.** Three different brands of custom-made, in-the-ear (ITE) devices per ear were used for bilateral hearing-aid provision. They were built from identical ear impressions of each individual ear. The vents of the ITEs were individually chosen on the basis of the pure-tone audiogram and the HA characteristics. The power levels of the devices were specified in order to ensure the same power levels across all three test devices. The ITEs were fit to the individual hearing losses of the participants using the default first-fit settings recommended by the manufacturers. The fittings were based on the brand-specific speech-in-noise program and represented the typical variation between devices on the market. For Device 1, the speech-in-noise program with a narrow beamformer that adjusts to the position of the speaker was chosen. The fitting of Device 2 was based on a comparable program with an adaptive beamformer. The program with speech-in-noise functionality was used for Device 3. Beamformers which adjust adaptively have a wider directional characteristic than Device 1. The differences between the devices are manufacturer specific; thus, it was ensured that similar microphone settings—typical and suitable for the current group communication situation—were chosen in the programs. All participants used the same ITE brands within each session. The test order of the three brands turned out to be challenging, as no complete randomization for three different brands could be realized in two group discussion sessions. Therefore, based on the SRT, it was determined that the order of the tested hearing-aid brands with the highest and lowest SRT results would be randomized. A clear contrast in the SRT suggests that differences in behavior may also occur. The order of the HAs was selected in Group 1 with Brands 1, 2, and 3 and for Group 2 with Brands 3, 2, and 1.

**Speech-in-noise test.** Speech intelligibility measurements, as used here, ensure a clinically relevant, external evaluation of possible differences between the devices. This is an additional performance indicator and is needed for the interpretation of the behavioral results. Speech intelligibility in noise for the three ITEs was determined for each participant using the Oldenburg sentence test (Wagner, Kühnel, & Kollmeier, 1999). Using an adaptive procedure for speech-level changes, the SRT for a speech recognition score of 50% was estimated. The level of an ICRA5-250 noise (Wagner, Brand, & Kollmeier, 2006) was kept constant at 65 dB SPL. While the speech

was presented from a loudspeaker in front of the participants ( $0^\circ$ ), the noise was presented from 11 different directions around the participant's seat ( $30^\circ$ ,  $60^\circ$ ,  $90^\circ$ ,  $120^\circ$ ,  $150^\circ$ ,  $180^\circ$ ,  $210^\circ$ ,  $240^\circ$ ,  $270^\circ$ ,  $300^\circ$ , and  $330^\circ$ ) using uncorrelated segments of the noise. Prior to the measurements, one test list was presented for training. The speech test was conducted in a sound-proofed, free-field cabin.

**Subjective assessment.** Between and after the group discussion sessions, participants rated their subjective impressions of the sessions on different scales: subjective speech intelligibility from 1 (*nothing*) to 7 (*all*), listening effort from 1 (*extremely effortful*) to 7 (*effortless*), and overall satisfaction with amplification from 1 (*very dissatisfied*) to 6 (*very satisfied*). Four other Likert-type scales were overall loudness perception from 1 (*much too soft*) to 5 (*much too loud*), loudness perception of the noise (1–5), spatial awareness from 1 (*very dissatisfied*) to 6 (*very satisfied*), and sound quality from 1 (*very unpleasant*) to 6 (*very pleasant*). For the measures, overall satisfaction and sound quality, we chose 6-point scales to avoid the possibility of a medium assessment. For the loudness scales, we chose 5-point scales with a possible medium assessment *adequate*; (Paluch et al., 2015).

**Statistical analysis.** The data were analyzed with SPSS version 22. Nonparametrical procedures were used for repeated measurement and explorative analyses using boxplots. The Friedman test was used to compare all three conditions, and the Wilcoxon ranked sign test was used for pairwise comparisons. For the Friedman test, the significance level was set to  $\alpha = .05$ . For pairwise comparisons, the level was set to  $\alpha = .017$ , due to Bonferroni corrections.

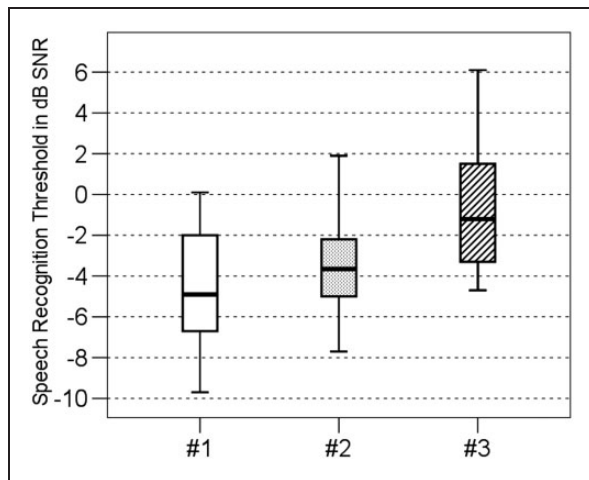
## Results

**Speech test.** Figure 2 shows the results of the speech test for the three different ITEs. The SRT was lowest for ITE 1 (median:  $-4.9$  dB signal-to-noise ratio [SNR]), increased by 1.3 dB for ITE 2 (median:  $-3.6$  dB SNR), and increased by 3.7 dB for ITE 3 (median:  $-1.2$  dB SNR). Friedman test results revealed significant differences ( $p < .001$ ) between the three ITEs. The post hoc Wilcoxon test supported significant differences between all three devices: ITE 1 versus ITE 2 ( $p = .012$ ), ITE 1 versus ITE 3 ( $p = .005$ ), and ITE 2 versus ITE 3 ( $p = .005$ ). These results substantiated clinically measurable differences between the three brands.

**Subjective assessment.** The results of the subjective assessments of speech intelligibility, overall satisfaction with

amplification, and listening effort are shown in Figure 3. The simulated shopping mall scenario was very strenuous for most of the participants, with ratings for subjective speech intelligibility from *little* to *medium*, and subjective listening effort from *very effortful* to *clearly effortful* even for the ITE with the best SRT results (box area of ITE 1).

Friedman tests revealed significant differences between the three ITEs for subjective speech intelligibility ( $p = .007$ ), overall satisfaction ( $p = .005$ ), and listening effort ( $p = .061$ , trend only). For subjective speech intelligibility, the post hoc Wilcoxon tests resulted in differences between ITE 1 and ITE 2 ( $p = .035$ , trend only) as well as between ITE 1 and ITE 3 as a significant result ( $p = .016$ ) but not between ITE 2 and ITE 3 ( $p = .102$ ). For overall satisfaction, significant differences were



**Figure 2.** Speech recognition thresholds (SRT) in dB SNR using the OLSA for the three ITEs 1 to 3 ( $N = 10$ ). SNR = signal-to-noise ratio; OLSA = Oldenburg sentence test.

found between ITE 1 and ITE 2 ( $p = .014$ ) as well as between ITE 1 and ITE 3 ( $p = .016$ ), and a statistical trend was observed between ITE 2 and ITE 3 ( $p = .075$ ). The rating of listening effort showed a statistical trend between ITE 1 and ITE 3 ( $p = .045$ ) but not between the other ITE combinations.

For the other scales, we found no significant differences between the three devices. The overall loudness perception was rated for ITE 1 as adequate in contrast to ITE 2 and ITE 3 (median values *little too soft*), whereas the loudness of noise was rated *somewhat too loud* for ITE 1, and *much too loud* for ITE 2 and ITE 3. The sound quality was rated for all three ITEs as *a little bit unpleasant*. The participants were a little dissatisfied with the spatial hearing impression by ITE 1, and ITE 2, and dissatisfied by ITE 3.

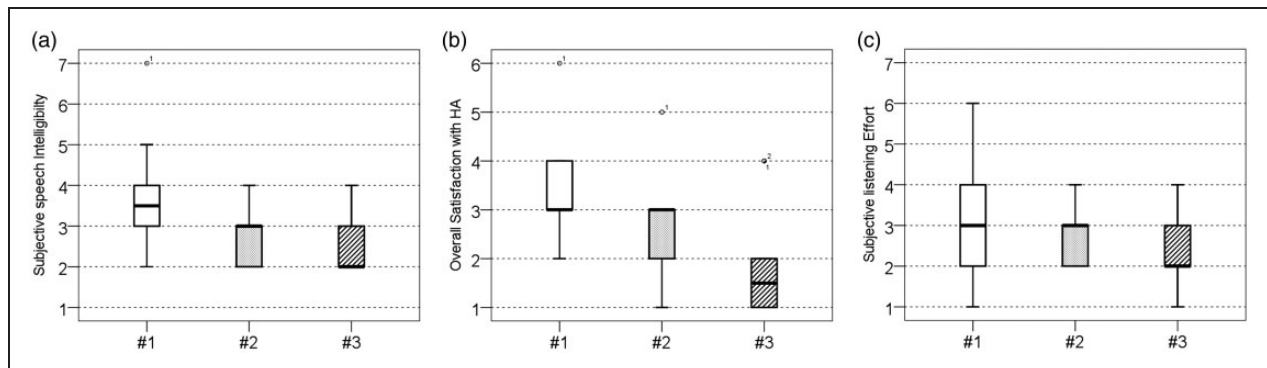
### Summary

Significant contrasts were found for subjectively perceived and objectively measured speech intelligibility between the three different ITE brands. Based on this finding, our assumption was that these differences will correspond to a change of interpersonal communication behavior. This assumption was tested in the first iteration.

## First Iteration: Development of an Offline Behavior Code System and Evaluation With ITEs

### Methods

One goal of the experimental study was to replicate the findings of Paluch et al. (2015), as described in the introduction, for a larger number of participants and other hearing devices. Paluch et al. (2015), the basis of the



**Figure 3.** (a) Subjective speech intelligibility from “1 = nothing” to “7 = all,” (b) overall satisfaction with amplification from “1 = very dissatisfied” to “6 = very satisfied,” and (c) subjective listening effort from “1 = extreme” to “7 = effortless” for ITEs 1, 2, and 3 in the group discussions ( $N = 10$ ). HA = hearing aid.

following iterations, identified a  $2 \times 2$  scheme of core dimensions for interaction: symbolic gestures versus verbal communication and F-t-F communication versus group communication. Preliminary analysis of the video recordings revealed that ITE 3 featured worst speech intelligibility but no significant change in the behavior codes according to the scheme used. It turned out that qualification of the results was difficult and more dimensions, as well as a higher resolution, were required. In the next step, we analyzed the whole data set with Mangold INTERACT<sup>®</sup>, a platform for synchronized viewing and analysis of video footage and audio files in observational research. It allows for content coding and event logging. This platform allows for setting clear time stamps according to the two core dimensions as reported in Paluch et al. (2015).

We assumed that shorter communication episodes may indicate possibly *unsuccessful* communication attempts, in particular, meta-communication induced by the challenging hearing scenario, such as “I can’t understand you” and “Please repeat the last sentence.” We divided the core dimension “forms of interaction” into episodes  $\leq 4$ s and  $>4$ s, derived by distribution functions, but no differences of the three devices were found. Based on the grounded theory (GT) approach (Glaser & Strauss, 1967), we inspected the entire video material over the course of several sessions. The GT approach allows the iterative evaluation of qualitative data beginning with neutral descriptions up to the interpretation of the data (Strauss, 1987). The neutral description of the video material was a chronological report and sequenced representation of phenomena that occurred (Przyborski & Wohlrab-Sahr, 2009, p. 64). For the interpretation, phenomena were classified as indicators with which assumptions can be justified (e.g., leaning forward, looking at a group, or talking face to face). By comparing video recordings, indicators were highlighted that showed similarities, differences, and interrelations of interactions. This was in line with the GT coding paradigm (Strauss, 1987, pp. 27–28). It was worked out qualitatively, which conditions possibly led to which forms of interaction and what consequences this had for the subjects.

It seemed that the higher ratio of F-t-F communication of Device 3 might be associated with more problems for the participants to communicate with the more distantly placed partners. We concluded that a higher rate of F-t-F is once again an important dimension but not necessarily an indicator of a successful communication episode. Furthermore, we inspected the data to assess whether we could observe systematic differences regarding moving, shaking, and symbolic gestures like nodding the head, or blocking the ears for nonunderstanding. Qualitatively, we found no systematic differences, but we found indications that using Device 3 was associated

with leaning more forward as a nonverbal gesture to the respective communication partner.

We developed a new annotation procedure with a higher resolution (frequency of annotated behavior units) to annotate each behavior occurring, even those not considered most striking, as reported in the Paluch et al. (2015) study. We also included more qualitative aspects of communication behavior in groups: Different proxemics regarding near versus distant torso movements (forward–backward) to the dialogue partner (DP) and communication with the distant versus near DP. The categories of the revised offline behavior code system, the results of the first iteration step, and the corresponding rater instructions are shown in Table 1.

This new annotation scheme was used by three assessors to reanalyze all video recordings with Mangold INTERACT<sup>®</sup>, applying clear time stamps for annotating each behavior. The behavior units contained each listener or speaker movement and each talker’s conversational turn for the 15-min discussions.

The outcome scheme included a total of 18 codes organized in a hierarchic scheme of interdependent codes in four dimensions. Each of the 18 codes contains the specific information of the four dimensions with the following coding order I to IV. For F-t-F communications, 12 codes were possible, and for group communication, 6 codes, because the exclusive F-t-F dimension *Distance to the dialogue partner* was missing. In total, 2,299 communication units, with a frequency of  $\approx 13$ s per behavior unit and test person, were assessed for the three ITEs.

## Results

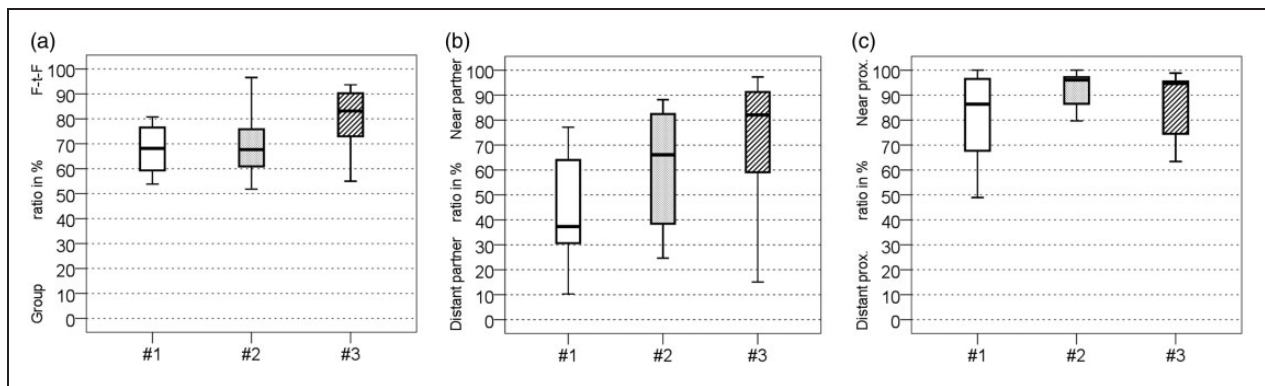
Using the revised annotation scheme given in Table 1, the core dimension *forms of interaction* revealed a statistical trend according to the Friedman test ( $p = .067$ ). Figure 4(a) shows the ratio of F-t-F communications to the sum of F-t-F and group communications. This ratio was highest for ITE 3, indicating approximately 15% more F-t-F communications when compared with the other two devices. The differences showed a statistical trend for the paired comparisons of ITE 1 and ITE 3 ( $p = .047$ ), and for ITE 2 versus ITE 3 ( $p = .059$ ); no effect was observed for the ITE 1 versus ITE 2 ( $p = .709$ ).

No differences were found for the core dimension *interdependence*, the comparison of the amount of verbal versus nonverbal, that is, gesture communication. An additional analysis of the data included the number of verbal communications (F-t-F and group communication), while neglecting purely nonverbal behavior. In total, 498 verbal communication episodes were counted for ITE 1, 522 for ITE 2, and 558 for ITE 3. The

**Table 1.** Dimensions of the Revised Laboratory Behavior Code System.

I—General forms of interaction	II—General forms of interdependence	III—Forms of interaction: Distance to the dialogue partner	IV—Interdependence: proxemics
<p><b>F-t-F vs. group communication</b> to distinguish between those two general communication situations.</p> <p>[F-t-F: The conversation takes place in direct contact with only one person, so a total of two people are involved in the interaction. Group: verbal communication in a group is when the speaker turns to several people and listens to more than one person]</p>	<p><b>Speech vs. gestures vs. combined gestures and speech communications.</b></p> <p>To distinguish between these communication patterns.</p> <p>[Speech contributions and gestures: all nonverbal gestures, such as moving, shaking, and nodding head, blocking ears, moving arms and torso, but classifying in each to the dichotomy near vs. distant proxemics]</p>	<p><b>Near vs. distant dialogue partner (only for F-t-F communications)</b></p> <p>[Near dialogue partners are those who sit to the right and left of a person as direct neighbors; distant dialogue partners are those who sit diagonally opposite or directly opposite the person to be observed, see screenshot of the setting]</p>	<p><b>Near vs. distant torso movements</b></p> <p>[Near: Sitting position of the upper body leaned forward (<math>&lt;90^\circ</math>) to the conversation partner; Distant: Sitting position of the upper body in neutral upright position or leaning back (<math>\geq 90^\circ</math>) on the chair]</p>

Note. Instructions for the rater are given in square brackets.



**Figure 4.** Communication behavior of (a) F-t-F/(F-t-F + group communication) in % based on all 2.299 communication episodes, (b) near dialogue partner/(near dialogue partner + distant dialogue partner) in % based on 1.698 F-t-F communication episodes, and (c) proxemics near/(proxemics near + proxemics distant) in % based on all 2.299 communication episodes of 10 participants.

differences between the three ITEs were not significant (Friedman test,  $p = .900$ ).

The analysis of the F-t-F category “Distance to dialogue partner” (data basis: 1.698 communication episodes) resulted, for the ratio of near dialogues to the sum of near and distant dialogues, in a significant group effect (Friedman test,  $p = .003$ ). As shown in Figure 4(b), the participants interacted in more of the assessed communication units with their near DP when using ITE 3 compared with when using ITE 1 (medians: 82% vs. 38%; Wilcoxon test,  $p = .009$ ). The ratio difference showed a statistical trend between ITEs 1 and 2 ( $p = .022$ ) and between the ITEs 2 and 3 ( $p = .059$ ). As a subanalysis of the F-t-F category “Distance to dialogue partner,” the data were inspected while ruling out purely nonverbal behavior, to obtain data related only to conversation (data basis: 1.318 communication

episodes). Based on communication episodes with verbal contributions, the participants conversed in 36% of the episodes with the respective near partner when using ITE 1, in 67% when using ITE 2, and in 80% when using ITE 3 (all medians). This indicates that the communication pattern observed in Figure 4(b) is also related to verbal communication episodes (conversation). The Friedman test was significant ( $p = .014$ ), with a significant paired comparison of ITE 1 versus ITE 3 ( $p = .009$ ) and statistical trends for the comparison ITE 1 versus ITE 2 ( $p = .047$ ), and ITE 2 versus ITE 3 ( $p = .093$ ).

The analysis of proxemics (see Figure 4(c)) when including all behavior units revealed a significant difference between the ITEs for the percentage of leaning forward (data basis: 2.299 communication episodes, Friedman test,  $p = .025$ ). Participants tended to lean

forward more when using ITE 2 (median: 98%) and ITE 3 (median: 96%), compared with using ITE 1 (median: 88%). For the comparison of ITEs 1 and 2, a statistical trend was observed ( $p = .043$ ), but for the comparisons of ITEs 1 versus 3 ( $p = .203$ ) and ITEs 2 versus 3 ( $p = .285$ ), the differences were not statistically significant.

### Summary

We found some indications that the offline annotated communication behavior showed differences between the three brands. These results were in line with the subjective and more objective data regarding speech intelligibility. However, this annotation procedure is possibly suitable for the described lab situation. Three main points were missing: the theoretical widening of the approach, the question of the reliability of the annotation procedure, and the need to establish an instantaneous or on-the-spot procedure which is suitable in real-life settings, where an offline annotation is not possible, because of data protection aspects. Possible solutions were elaborated in the next two iteration steps.

## Second Iteration: ICF Expansion, Scaling, and On-The-Spot-Coding

### Methods and Theoretical Background

Based on the ICF core set for hearing loss (Granberg, de Swanepoel, et al., 2014; Granberg, Möller, et al., 2014), ICF codes possibly relevant for behavior observations were identified by means of expert interviews. A revised behavior codes system, extended and linked to ICF codes, was built and tested in ethnographical walks in real-life settings in a cafeteria of the University of Oldenburg (Paluch et al., 2018). In the course of this iteratively structured process, a preliminary behavior code system was derived with experts that included relevant and observable behavioral categories linked to ICF codes, possibly suitable for the current scenario. For details of the process and also for other real-settings, see Meis et al. (2018) and Paluch et al. (2018).

The second-level ICF category attention functions [b140] that refers to aspects such as sustaining, shifting, dividing, and sharing attention, as well as to concentration and distractibility, is moderately suitable for observations (Meis et al., 2018; Paluch et al., 2018). For addressing spatial awareness for example, *attention functions* were included in the second iteration of the behavior code system. The experts stated that the most prominent ICF component levels for observational studies belong to *activities and participation* [d]. The ICF third-level categories basically differentiate between *conversing with one person* [d3503] and *conversing with many people* [d3504], that is, the number of conversation

partners. The corresponding abilities are described as initiating, maintaining, shaping, and terminating a dialogue or interchange with one person or many people, respectively. Particularly in group conversations, the spatial distance between potential conversation partners needs to be considered in the observational assessment of communicative behavior. Accordingly, the distinction between near versus distant communication partners was assumed to relate to the ICF third-level category [d3504] in the first iteration of the offline behavior code system. From the ethnographical walks, we found two additional categories as candidates for behavior observations not included in the offline annotation scheme: non-understanding gestures and speech supporting gestures (symbolic gestures), as well as the total amount of verbal communication.

**Qualifiers for behavior assessment.** The use of qualifiers is a crucial component in ICF. In general, a 5-point scale is used, such that 0 = *no impairment* to 4 = *complete impairment* for the domain *body functions or structures* and 0 = *no problem* to 4 = *complete problem* for *activities and participation*. Qualifiers for *environmental factors* range from 0 = *no barrier* to 4 = *complete barrier*. This scheme of qualifying was roughly adopted for the behavior code system, but without qualifying interpretation from the participants' perspective.

To reduce too frequent annotation activities for the rater that would be not manageable in a field situation, 5- or 7-point rating scales were used, depending on the code. For F-t-F communications versus group communications, a 7-point scale with three qualifiers of the direction toward F-t-F (3 = *solely*, 2 = *predominantly*, 1 = *slightly*, and 0 = *balanced*) was used. For *sustained attention* (face of conversation partner), the behavior was qualified from 1 = *very little concentrated* to 5 = *extremely concentrated*, for *frequency of verbal communication* from 1 = *never* to 5 = *very frequent*, and so on.

A manual with the detailed descriptions of the extended ICF categories and rating scales was established to provide a clear reference for the evaluation of the different characteristics. In the end, we chose eight codes for instantaneous coding, see Table 2. For the short manual including the rater sheet and the qualifiers, see the Appendix A.

### Results

The second iteration code system was reviewed for interrater reliability (IRR) using the video material derived in the offline lab experiment described earlier. Three different raters, two of them already involved in the first iteration step, jointly and repeatedly watched the videos and assessed the behavior of each conversation partner separately in a staggered sequence. Ratings referred to



**Table 2.** IRR for Extended ICF Categories.

Rater	ICF (sub-) categories or scale	A–B		B–C		A–C	
		$\kappa$	$r_{Sp}$	$\kappa$	$r_{Sp}$	$\kappa$	$r_{Sp}$
<b>b140_1</b>	Sustained attention face partner (low-medium-high)	.39	.58	.32	.56	<b>.44</b>	.65
<b>d3504_1</b>	Communication (F-t-F-balanced-group)	<b>.47</b>	.58	.36	.38	<b>.57</b>	.70
<b>d3504_2</b>	Frequency verbal comm. (seldom-sometimes-frequent)	<b>.51</b>	.72	<b>.52</b>	.68	<b>.43</b>	.70
<b>d3504_3</b>	Communication partner (near-balanced-distant)	<b>.59</b>	.73	<b>.62</b>	.70	<b>.72</b>	.79
<b>d3504_4</b>	Proxemics (forward-balanced-backward)	<b>.57</b>	.68	.38	.52	<b>.50</b>	.59
<b>d3504_5</b>	Change torso position (seldom-sometimes-frequent)	.13	.26	.33	.56	.39	.57
<b>d3504_6</b>	Nonunderstanding gestures (seldom-sometimes-frequent)	.07	.29	.35	.40	.16	.32
<b>d3504_7</b>	Speech supporting gestures (seldom-sometimes-frequent)	.24	.51	.26	.39	<b>.46</b>	.57

Note. ICF = international classification of functioning, disability and health.

Cohen's kappa, indicating moderate or substantial agreement, are in bold. Qualifiers in short for the ratings in brackets (b140\_1 to d3504\_7).

Legend: A–C = 3 raters;  $\kappa$  = Cohen's kappa,  $r_{Sp}$  = Spearman's rho.

Cohen's kappa: Agreement:  $\kappa < 0$  = poor,  $\kappa 0-0.20$  = slight,  $0.21-0.40$  = fair,  $0.41-0.60$  = moderate,  $0.61-0.80$  = substantial, and  $0.81-1.00$  = almost perfect; see Landis and Koch (1977).

3-min sections, in total five ratings, in a 15-min conversation. For IRR calculation, Cohen's Kappa ( $\kappa$ ) (Cohen, 1960) and correlations (Spearman's rho  $r_{Sp}$ ) were used (SPSS v. 22). IRR were calculated for pairs of raters. For this purpose, the rating scales were condensed into 3-point ordinal scales. The results are displayed in Table 2.

IRR statistical values for the categories b140\_1 and d3504\_5 to d3504\_7 were predominantly poor to slight. Moderate IRR values were found for the *communication* categories F-t-F versus group [d3504\_1], the *frequency of verbal communication* [d3504\_2], and the *proxemics* categories, torso position leaning forward versus backward vertical axis [d3504\_4]. Substantial IRR values were gathered for the communication episodes with the distant versus near communication partner [d3504\_3].

### Summary

All in all, the annotation of the behavior by means of scaling showed IRR to be too weak to be suitable for subsequent use for on-the-spot annotation in the field. The three raters reported difficulties when *averaging* the behavior over a period of 3 min, so that scaling methods are possibly a dead-end method for rating one's behavior. Moreover, the ICF subcategories [b140\_1], [d3504\_5], [d3504\_6], and [d3504\_7] had too much latitude for interpretation which led to poor IRR values. We decided to drop these possible subcategories. The idea for the next iteration step was to use the most sensitive categories from this offline annotation scheme, which showed moderate to substantial IRR values from the scaling method, but with a simplified scheme for instantaneous, binary annotation in order to avoid overtaxing the raters.

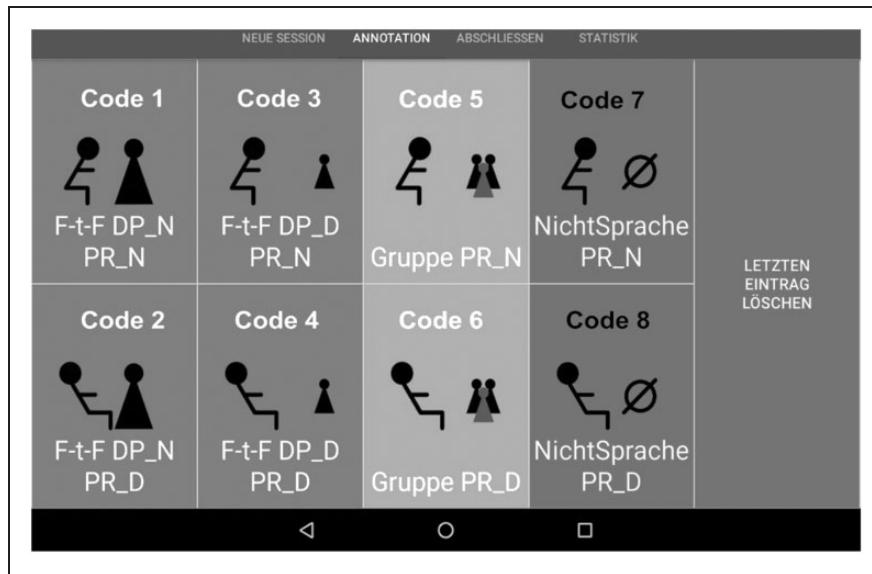
### Third Iteration: App Development, Instantaneous Coding, and Reliability

The entire video material, but also the evaluations of the offline annotation, was viewed again. A revised annotation scheme was developed that simplified the above offline scheme (see chapter first iteration), but contained all relevant dimensions that were significant, when comparing the different HA brands. The main simplification was that the number of nonverbal behaviors to be annotated was reduced. For nonverbal behavior, only the torso movements near versus far were annotated. Initial explorative studies have shown that it is possible to annotate a total of eight behavior codes simultaneously.

### Methods

As a consequence of the above, we developed a running application for instantaneous coding of the behavior in the lab and the field (see Figure 5), with a direct compute to gather frequencies of the codes and an export to excel, based on the main codes of the offline annotation. The goal was to achieve higher frequencies of annotation within the 15-min sessions, a higher test–retest reliability, and substantial values of IRR. The instantaneous coding was simplified by means of pictograms, together with the abbreviation of the codes for German raters.

As shown in Figure 5, we used the binary codes F-t-F communication versus group, DP: near (N) versus distant (D) and proxemics torso: near (N) versus distant (D). The frequency of verbal communication is not a subcategory but can be computed by the app. Codes 1 to 6 also contained the information of proxemics. For the short manual of the instantaneous coding, including



**Figure 5.** Annotation app for behavior coding in conversation episodes (screenshot of the German version). F-t-F: Face-to-Face communication versus group (German: *Gruppe*), DP: dialogue partner: near (N) versus distant (D), and nonverbal Proxemics PR (torso): near (N) versus distant (D); code numbers 1 to 8 were included for the figure, but do not appear in the test mode. The right-hand column means “delete last entry.” Codes 7 to 8 referred to nonverbal behavior (“NichtSprache”).

the hierarchical structure and the qualitative description of the codes, see Appendix B.

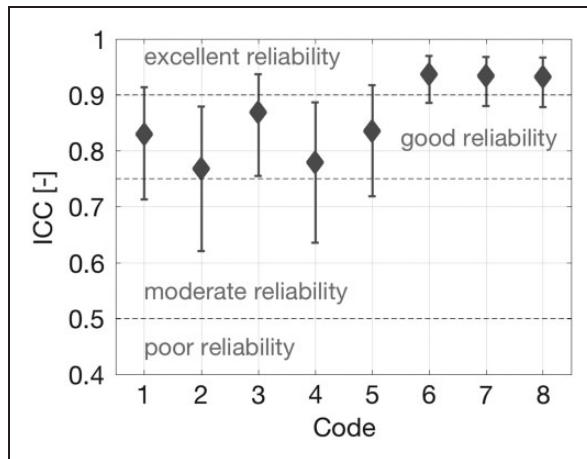
**Training sessions, exploratory ICC, and test-retest analyses.** A training concept, including an animated manual with embedded sketches, sample videos, and short descriptions, for a total of four raters was elaborated and included an introduction, training with existing video material, exercises in the field, and a final annotation in the lab with users carrying out a communication episode behind a one-way window. The in-house training lasted 2 days.

**Reliability analyses by intraclass correlation coefficients.** To test IRR for the eight codes for four raters, we used intraclass correlation coefficients (ICCs). ICC assesses the reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and subjects. This method is used for estimating the IRR for continuous variables, that is, ordinal, interval, and ratio variables. In this study, the ICC values were computed by absolute agreement in the ratings, and two-way random effects model,  $k$  raters, the most strict testing according to Koo and Li (2016). The ICC values range from 0 to 1. ICC values of less than 0.5 suggest poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 suggest good reliability, and values greater than 0.9 indicate excellent reliability (Koo & Li, 2016).

## Results

**Exploratory ICC and test-retest analyses.** Within the training sessions, we conducted exploratory test-retest analyses. We analyzed first the ICC scores via SPSS v 22.0 for the six categories (Codes 1–6) with verbal content, see Figure 5. We used a video tape with a communication session (video with the coffee table situation from the offline annotation) with five participants testing directionality of HAs. The data for all of the codes in the test session were significant, with moderate to excellent ICC values, ranging from 0.60 (Code 4; 95% CI from 0.19 to 0.71) to 0.94 (Code 6; 95% CI from 0.81 to 0.99). This pattern of results indicated that the training sessions have to be extended to obtain better ICC values and less scatter. After the same training session, but using different video material, the retest ICC values showed an improvement compared with the first test session: The ICC values ranged from 0.81 (Code 4; 95% CI from 0.46 to 0.97) to 0.95 (Code 1; 95% CI from 0.82 to 0.99). All in all, good to excellent ICC values were obtained, but with some large scatter, perhaps caused by the low  $n$  of cases.

**Final ICC analyses.** We conducted two sessions in order to establish a broad database for IRR analyses. The four raters analyzed five video-recorded group communication sessions with five participants sampled from three different projects testing different hearing-aid brands, settings, and background noise in the lab (coffee table



**Figure 6.** ICC values from four raters of 23 participants. Code 1: F-t-F DP\_N PR\_N, Code 2: F-t-F DP\_N PR\_D, Code 3: F-t-F, DP\_D PR\_N, Code 4: F-t-F DP\_D PR\_D, Code 5: Group PR\_N, Code 6: Group PR\_D, Code 7: nonverbal proxemics near, and Code 8: nonverbal proxemics distant. PR = proxemics: Near versus Distance, F-t-F = Face-to-Face communication, and Group = Group communication; DP = dialogue partner: Near versus Distance. For an overview of the codes, see also Table 1 and Figure 5. The Koo and Li (2016) benchmarks or boundaries for IRR values were included as dashed lines. ICC = intraclass correlation coefficients.

scenario). The raters jointly and repeatedly watched these videos and assessed the behavior of each participant and each session. The sequence of the observed communication partners was not staggered in these trials. The experimenter-in-charge, however, was present during the rating trials and monitored their proper conduct. From a total of 25 different participants, we were able to use the observations from 23, because material from two participants in the training sessions was not used. All material for the subsequent ICC analyses had never previously been annotated by the four raters. The annotation time for each rater was 5.75 hr. The ICC values estimated from these sessions are shown in Figure 6. ICC values ranged from 0.76 to 0.95, thus confirming a good to excellent reliability.

## Discussion

One essential aim of this study was the development of new outcome tools, based on behavioral data of communication and conversation, for use in the lab with virtual acoustics, the offline behavior code system, and in real-life settings for instantaneous coding.

The study of Paluch et al. (2015) showed that communication behaviors changed as a result of different modes of hearing aid amplification using a pilot offline annotation scheme, the basis for the three iterations.

Based on the results of this pilot study, a more sophisticated annotation scheme with a higher

frequency of behavior using 18 codes was iteratively developed, here referred to as the first iteration. It was shown that speech intelligibility scores, measured by standard efficacy methods in the lab, were in line with the results of subjectively perceived speech intelligibility and changes of behavioral patterns. Those patterns included more group communications (in comparison to F-t-F communications), more distant proxemics, and more dialogues with the distant communication partners for the beamformer with the best speech intelligibility. In this respect, a higher number of F-t-F communications did not reflect a successful communication setting. From the data obtained here, we have no data and knowledge of normal hearing participants, a possible important reference to qualify in the direction of successful communication. In further studies, this point is relevant to address.

In addition, participants were possibly forced to change proxemics behavior by leaning further forward when the conversation was more strenuous for them. This behavior pattern was in line with results from Brimijoin, Hadley, Naylor, and Whitmer (2017). They found an impact of increasing noise level on measured behaviors: Participants spoke louder, moved closer together, and pointed their head and eyes toward their partner, with less variability. To clarify and objectify this point, it would be valuable to additionally monitor head movements using ambient and not-too-obvious head- and eye-tracking technology (Brimijoin et al., 2017) to improve the offline behavior code system. Nevertheless, any additional equipment to be worn by the participants might influence their behavior and might have an impact on the observation results. Moreover, in subsequent studies, the annotation of the angle leaning forward or backwards could be improved in different categories, because we obtained medians of 98%, 96%, and 88% for leaning forward.

The data regarding “Distance to the dialogue partner” can be interpreted in terms of activity limitation and participation restriction, that is, as a limitation of conversation activities induced by the microphone mode of the HAs. Subjects with suboptimal hearing-aid fittings were still able to communicate and converse with the near DP, but not as often with the respective distant DP. The resulting pattern suggested the interpretation of data along theoretical and practical models of HrQoL, for example, with questions from HHI-E or HHI-A variants. Especially items from the social scale related to concrete behavior, such as Item S-21 “Does a hearing problem cause you difficulty when in a restaurant with relatives or friends?” (Newman et al., 1990), are related to conversations with DPs as well as Item 4 from the SSQ speech scale. In summary, a first link to the ICF framework, namely the code [d3504], was observed.

Ringdahl and Grimby (2000) and Meis et al. (2007) concluded that people with severe hearing losses and insufficiently supplied hearing devices are more likely impacted by social isolation, a scale from the generic HrQoL questionnaire Nottingham Profile (Hunt et al., 1985). In the validation study, we did not find evidence for a lower rate of verbal phrases, which would have been a possible sign of withdrawal. We assume that the underlying reason was the induced lab situation: The participants were instructed to discuss different topics. The announced summary of the results after the group discussion might have fostered performance motivation, which is not typical in everyday life, and small-talk communication. Hence, effects of withdrawal were not measured with the current research design. In general, it seems to be difficult to realize such an approach in the realm of a laboratory setting. An alternative is the observation of behavior in real-life communication settings without provoking verbal *performance*, for example, in challenging restaurant settings or family celebrations using a field behavior code system.

Unfortunately, from the results of the forms of interaction, we were not able to classify the F-t-F communication into *successful* or *unsuccessful* communication episodes. A deeper conversation analysis using verbal transcripts of the dialogues (see Egbert & Deppermann, 2011) might be necessary to qualify conversations as successful or unsuccessful. Egbert and Deppermann (2011) proposed using content-driven analyses of conversations, for example, whether people with hearing loss are more likely to use requests for clarification when the conversation partner is familiar rather than unfamiliar. Moreover, Egbert, Golato, and Robinson (2009) analyzed conversations regarding *repair mechanisms*. One example for a repair mechanism is the experience of hearing difficulties when listening to a phrase and subsequent repair initiations by the listener. Egbert and Deppermann (2011) listed numerous approaches to conversation analyses in audiology, which could be the next step for the offline approach presented here. Nevertheless, this extension could be applied in a lab study, but not in the field, due to ethical and privacy issues.

The interpretation of the offline study, the first iteration, is additionally limited with respect to some further aspects. A weakness of our approach is that we did not consequently distinguish between passive listening and active speaking communication behavior as a possible category. Moreover, the IRR was not systematically analyzed. Three annotators were instructed after intensive training and control sessions with clear definitions of the 18 code combinations. Unfortunately, a complete data set for all three independent raters is no longer available. But we developed an app for instantaneous annotation of the communication behavior (third iteration). This annotation scheme contained the most

relevant eight codes of the offline code scheme, used in the experimental study. The ICC values also included the material from the experimental study with the three brands. We observed good to excellent IRR values from four independent raters, so that it is very supposable that the results of the 18 code system were based on reliable data.

Another aspect affects auditory ecology. The noise level was set to  $L_{Aeq} = 67$  dB, a situation that hearing-impaired people typically try to avoid (see e.g., Paluch et al., 2015). The noise level was chosen because most of the beamformers were activated by the HA's classification algorithms at this noise level or above. Taking this aspect into account, the situation tested here is not typical for everyday life. The participants reported high values of subjective listening effort and were annoyed by the loud noise scenario. Therefore, the data obtained here are limited to this especially challenging conversation scenario.

To sum up, the 18 code system might be a promising basis for analyzing behavior in the lab. Possible additions and modifications could be the validation of the behavior with head- and eye-tracker technology, conversation analysis, IRR, as well as the variation of scenes and virtual acoustics.

The expert review and the ethnographical walks showed that the ICF framework is, in general, useful to widen the approach of an annotation system for behavior analyses in the domain of audiology. The main codes were the second-level categories *attention functions* [b140] and *conversation* [d350]. Owing to the hierarchical structure of the ICF classification, the subordinate categories are of preferred use for behavior. This is the case for the subcategories *conversation with one person* and [d3503] and *conversation with many people* [d3504] that specify and, thus, replace the superordinate category [d350] in the field behavior system.

Overall, only few categories out of the ICF core set for hearing loss proved to be applicable for an observational approach aiming toward *auditory ecology*. Most categories of the core set refer to *body functions* [b] and *body structures* [s] and are thus addressed by audiological tests. Nevertheless, behavioral methods that were aligned to the ICF framework could complement knowledge about how hearing impairment, as well as hearing rehabilitation, impact real-life behavior.

The data showed that the on-the-spot behavior coding (second iteration) using scales is not a reliable way to annotate the communication behavior. The 3-min time interval is too long as raters appear to be inconsistent in how they perceive average behavior across the interval. Moreover, the additional subcategories *Sustained attention face partner*, *Change horizontal torso position*, *Non-understanding gestures*, and *Speech supporting gestures* might be less reliable categories for further research.

The revised code system for instantaneous behavioral observation by means of an app (third iteration) represents a method for reliably recording observable behavior. In addition to field use, this method is also suitable for laboratory studies, as the behavioral dimensions examined so far are valid and can also be evaluated quickly. However, reliability tests are still required for field experiments, as many factors that are difficult to control or spontaneously occur in the field need to be taken into account. The app should involve a reduced set of codes in combination with M-HrQoL items and acoustical readings of the situations as a holistic EMA research plan. Contextual and environmental factors, as

postulated by the ICF framework, that possibly influence the behavior and behavioral change were integrated into the app as a description of the situation to consider the environmental conditions with regard to acoustics and light and, for example, the use of hearing devices.

All in all, the results presented here referred to a special scenario of interpersonal communication and could be used for hearing-aid evaluation. In future, the instantaneous annotation scheme needs validation in the field with possibly *confounding* variables in the real-life listening environments and a widening toward other significant situations in the field. The items of the SSQ might be candidates to define such additional scenarios.

## Appendix A: Interrater Reliability for the Extended ICF Categories

### a) Rater sheet

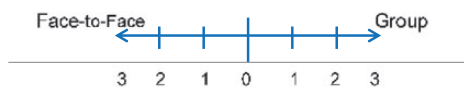
ID: \_\_\_\_\_ Time: 3 6 9 12 15

Name Rater: \_\_\_\_\_ Condition: \_\_\_\_\_

#### b140\_1 Sustained attention face of communication partner

- 1  very little focused
- 2  slightly focused
- 3  moderately focused
- 4  highly focused
- 5  extremely focused

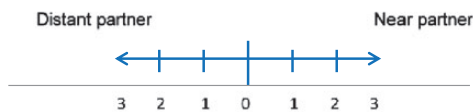
#### d3504\_1 Communication group vs. F-t-F



#### d3504\_2 Frequency verbal communication units

- 1  never
- 2  seldom
- 3  sometimes
- 4  often
- 5  very often

#### d3504\_3 Distance communication partner



#### d3504\_4 Seating position/leaning (proxemics)



#### d3504\_5 Frequency change sitting position

- 1  never
- 2  seldom
- 3  sometimes
- 4  often
- 5  very often

#### d3504\_6 Non-understanding gestures/behavior

- 1  never
- 2  seldom
- 3  sometimes
- 4  often
- 5  very often

#### d3504\_7 Speech supporting gestures

- 1  never
- 2  seldom
- 3  sometimes
- 4  often
- 5  very often

### b) Qualifiers

b140_1	To what extent (sustained attention) the person to be observed concentrates on the face and in particular the mouth image of the person(s) to be able to follow and participate in the conversation. Very little concentration' means that hardly any concentration is to be recognized on the mouth picture of the speaking person, however with 'extremely focused' the observed person "hangs" on the lips of the speaking person in order to be able to understand him.
d3504_1	Specifies whether communication with the group is active during the call or conversations take place in direct contact with only one person. The conversations are divided into face-to-face and group conversations, the predominantly taking place form of communication is evaluated. Characteristic for communication in the group is that the speaker turns to several people and listens to them accordingly > 1 person. In contrast, an interaction between a total of 2 persons involved can take place. The following two example images are used to illustrate this.

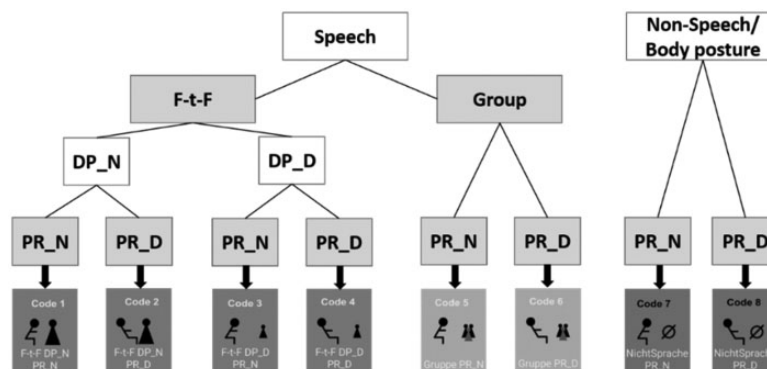
(continued)

Continued









- d3504\_2 The frequency of participation in the conversation is evaluated in the form of active, spoken contributions, which include initiating, maintaining, shaping and ending communication contributions. "Never/seldom" means that a conversation partner withdraws almost completely, the person is "out of the action" and observes rather than shapes the scene. If "very often" is indicated on the rating scale, a person is almost always in the middle of the action and communicates a lot verbally. The "sometimes" rating alternates sections of rare and frequent episodes.
- d3504\_3 Possible preference or active selection of the discussion partners of the group, based on the distance. Differentiation between close and distant interlocutors. Near interlocutors are those who sit to the right and left as direct neighbors; distant interlocutors are those who sit diagonally or opposite. The following example images serve as a visual representation. The respective selected numerical value on the scale describes the specification even more precisely.  
 0 = balanced, communication with near and distant interlocutors takes place to a similar degree  
 1 = slight tendency to identify near or distant interlocutors  
 2 = mainly, is communicated with near or distant interlocutor  
 3 = almost exclusively, is communicated with near or distant interlocutors  
 It must be taken into account when assessing that in the case of a very low participation of a test person, it is not generally evaluated with 0, but rather the few word amounts are evaluated according to the tendency to the interlocutor.
- d3504\_4 The sitting position is changed by shortening the physical distance to the person speaking. The predominant sitting position is evaluated during the discussion with strength of the characteristics forwards and backwards. The following picture serves as a visual representation. If a respondent is static and moves accordingly little, this static position is also evaluated. The dynamics of the movements are evaluated with the following code (see below). The respective selected numerical value on the scale describes the specification even more precisely.  
 0 = predominantly neutral, upright posture or balanced forward or reclined position  
 1 = mainly slightly forward or backward leaning position  
 2 = mainly middle forward or backward position  
 3 = pronounced forward or backward position
- d3504\_5 The frequency of changes in the overall sitting position in the evaluation sequence is evaluated.
- d3504\_6 The frequency of nonunderstanding gestures that are used spontaneously or consciously to indicate to interlocutors that what is said cannot be followed because it is not heard in an understandable way is evaluated. Gestures of incomprehension and behavior are judged as shrugging of the shoulders, horizontal shaking of the head, placing hands and hands behind the ear, aligning the ear toward the source, averting one's gaze (from the conversation scene).
- d3504\_7 The frequency of nonverbal communication is evaluated in the form of conversation-supporting gestures. Gestures are spontaneous or consciously used movements of the body, especially of the hands and the head, for example agreeing nodding of the head. Gestures can accompany someone's words or replace actions or messages that are intended to express something indirectly.

## Appendix B: Codes and Qualitative Description of the Codes (Instantaneous Coding)

a) Hierarchical structure of the coding process



## b) Codes and qualitative description

Codes	Qualitative Description
<p>Code 1</p>  <p>F-t-F DP_N PR_N</p>	<p>Annotations, when the person to be observed is actively speaking. Any change in sitting position while speaking is evaluated. The conversation takes place in direct contact with only one person, so a total of two people are involved in the interaction.</p> <p>Close discussion or dialogue partners are those who sit to the right and left of a person as direct neighbors. Sitting position of the upper body leaned forward (<math>&lt;90^\circ</math>) to the interlocutor.</p>
<p>Code 2</p>  <p>F-t-F DP_N PR_D</p>	<p>Annotations, when the person to be observed is actively speaking. Any change in sitting position while speaking is evaluated. The conversation takes place in direct contact with only one person, so a total of two people are involved in the interaction.</p> <p>Close discussion or dialogue partners are those who sit to the right and left of a person as direct neighbors. Sitting position of the upper body in neutral upright position or leaning back (<math>\geq 90^\circ</math>) to the back of the chair.</p>
<p>Code 3</p>  <p>F-t-F DP_D PR_N</p>	<p>Annotations, when the person to be observed is actively speaking. Any change in sitting position while speaking is evaluated. The conversation takes place in direct contact with only one person, so a total of two people are involved in the interaction. Distant conversation or dialogue partners are those who sit diagonally opposite or straight opposite the person to be observed. Sitting position of the upper body leaned forward (<math>&lt;90^\circ</math>) to the interlocutor.</p>
<p>Code 4</p>  <p>F-t-F DP_D PR_D</p>	<p>Annotations, when the person to be observed is actively speaking. Any change in sitting position while speaking is evaluated. The conversation takes place in direct contact with only one person, so a total of two people are involved in the interaction. Distant conversation or dialogue partners are those who sit diagonally opposite or straight opposite the person to be observed. Sitting position of the upper body in neutral upright position or leaning back (<math>\geq 90^\circ</math>) to the back of the chair.</p>
<p>Code 5</p>  <p>Gruppe PR_N</p>	<p>Annotations, when the person to be observed is actively speaking. Any change in sitting position while speaking is evaluated. Characteristic of verbal communication in the group is that the speaker turns to several people and listens to more than (<math>&gt;</math>) 1 person.</p> <p>Sitting position of the upper body leaned forward (<math>&lt;90^\circ</math>) to the interlocutor.</p>
<p>Code 6</p>  <p>Gruppe PR_D</p>	<p>Annotations, when the person to be observed is actively speaking. Any change in sitting position while speaking is evaluated. Characteristic of verbal communication in the group is that the speaker turns to several people and listens to more than (<math>&gt;</math>) 1 person.</p> <p>Sitting position of the upper body in neutral upright position or leaning back (<math>\geq 90^\circ</math>) to the back of the chair.</p>
<p>Code 7</p>  <p>NichtSprache PR_N</p>	<p>Annotations, when the person to be observed is not actively speaking, but listens. Every change in the sitting position is evaluated. Sitting position of the upper body leaned forward (<math>&lt;90^\circ</math>) in the direction of possible interlocutors.</p>
<p>Code 8</p>  <p>NichtSprache PR_D</p>	<p>Annotations, when the person to be observed is not actively speaking, but listens. Every change in the sitting position is evaluated. Sitting position of the upper body in neutral upright position or leaning back (<math>\geq 90^\circ</math>) to the back of the chair.</p>

## Authors' Note

Part of the results were presented as a poster at the International Hearing Aid Conference (IHCON), Lake Tahoe, CA, in 2016 and Meis et al. (2018) at the International Symposium on Auditory and Audiological Research 2016 (ISAAR) in Nyborg. The experimental study was approved by the ethics committee of Carl von Ossietzky University Oldenburg, No. 58/2015 for video observation and Drs.-Nr. 36/2015 for audiological measurements in general.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Cluster of Excellence EXC 1077/1 "Hearing4all" and SFB1330 "Hearing acoustics" (HAPPAA), both funded by the German Research Council (DFG), the Hearing Industry Research Consortium (IRC), 2016 Grant IHAB-RL, and from the federal resources of Niedersächsisches Vorab within the research focus "Hearing in everyday life (HALLO)."

## References

- Bitzer, J., Kissner, S., & Holube, I. (2016). Privacy-aware acoustic assessments of everyday life. *Journal of the Audio Engineering Society*, *64*(7), 395–404. doi:10.17743/jaes.2016.0020
- Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, *236*(4798), 157–161. doi:10.1126/science.3563494
- Brimijoin, W. O., Hadley, L. F., Naylor, G. M., & Whitmer, W. M. (2017). Parametric measurements of natural conversation behavior reveal effects of background noise level on speech, movement, and gaze. Abstract/Talk S1.1. *Proceedings of the 6th International Symposium on Auditory and Audiological Research (ISAAR)*, Nyborg. Retrieved from [http://www.isaar.eu/programme/isaar2017\\_programme.pdf](http://www.isaar.eu/programme/isaar2017_programme.pdf)
- Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513–531. doi:10.1037/0003-066X.32.7.513
- Brunswik, E. (1956). Historical and thematic relations of psychology to other sciences. *The Scientific Monthly*, *83*(3), 151–161.
- Chisolm, T. H., Johnson, C. E., Danhauer, J. L., Portz, L. J. P., Abrams, H. B., Lesner, S., . . . Newman, C. W. (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. *Journal of the American Academy of Audiology*, *18*(2), 151–183.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104
- Cox, R. M., & Alexander, G. C. (2002). The International Outcome Inventory for Hearing Aids (IOI-HA): Psychometric properties of the English version. *International Journal of Audiology*, *41*(1), 30–35.
- Deutsches Institut für Medizinische Dokumentation und Information (DIMDI) (Hrsg.). (2005). *Internationale Klassifikation der Funktionsfähigkeit, Behinderung und Gesundheit (ICF)*. [German Institute of Medical Documentation and Information (DIMDI) (2005). International Classification of Functioning, Disability and Health (ICF)]. Geneva, Switzerland: WHO.
- Dillon, H., James, A., & Ginis, J. (1997). Client Oriented Scale of Improvement (COSI) and its relationship to several other measures of benefit and satisfaction provided by hearing aids. *Journal-American Academy of Audiology*, *8*(1), 27–43.
- Egbert, M., & Deppermann, A. (2011). Introduction to conversation analysis with examples from Audiology. In M. Egbert, & A. Deppermann (Eds.), *Hearing aids communication integrating social interaction, audiology and user centered design to improve communication with hearing loss and hearing technologies* (pp. 40–47). Mannheim, Germany: Verlag für Gesprächsforschung.
- Egbert, M., Golato, A., & Robinson, J. D. (2009). Repairing reference. In J. Sidnell (Ed.), *Comparative studies in conversation analysis* (pp. 104–132). Cambridge, England: Cambridge University Press.
- Gatehouse, S., Elberling, C., & Naylor, G. (1999). Aspects of auditory ecology and psychoacoustic functions as determinants of benefits from and candidature for nonlinear processing in hearing aids. In A. N. Rasmussen, P. A. Osterhammel, T. Anderson, & T. Poulsen (Eds.), *Proceedings of the 18th Danavox Symposium on auditory models and nonlinear hearing instruments* (pp. 221–233). Copenhagen, Denmark: Holmens Trykkeri
- Gatehouse, S., & Noble, W. (2004). The Speech, Spatial and Qualities of Hearing Scale (SSQ). *International Journal of Audiology*, *43*(2), 85–99. doi:10.1080/14992020400050014
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research (Observations)*. Chicago, IL: Aldine.
- Granberg, S., Möller, K., Skagerstrand, Å., Möller, C., & Danermark, B. (2014). The ICF core sets for hearing loss: Researcher perspective, Part II: Linking outcome measures to the International Classification of Functioning, Disability and Health (ICF). *International Journal of Audiology*, *53*(2), 77–87. doi:10.3109/14992027.2013.858279
- Granberg, S., Swanepoel de, W., Englund, U., Möller, C., & Danermark, B. (2014). The ICF core sets for hearing loss project: International expert survey on functioning and disability of adults with hearing loss using the international classification of functioning, disability, and health (ICF). *International Journal of Audiology*, *53*(8), 497–506. doi:10.3109/14992027.2014.900196
- Hunt, S. M., McEwen, J., & McKenna, S. P. (1985). Measuring health status: A new tool for clinicians and epidemiologists. *The Journal of the Royal College of General Practitioners*, *35*(273), 185–188.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability



- research. *Journal of Chiropractic Medicine*, 15(2), 155–163. doi:10.1016/j.jcm.2016.02.012
- Kowalk, U., Kissner, S., von Gablenz, P., Holube, I., & Bitzer, J. (2018). An improved privacy-aware system for objective and subjective ecological momentary assessment. *Proceedings of the International Symposium on Auditory and Audiological Research*, 6, 25–30B.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Meis, M., Krueger, M., Gebhard, M. V., Gablenz, P., Holube, I., Grimm, G., & Paluch, R. (2018). Development and application of a code system to analyse behaviour in real life listening environments. *Proceedings of the International Symposium on Auditory and Audiological Research*, 6, 31–38.
- Meis, M., Lesinski-Schiedat, A., Plotz, K., Dillier, N., Walger, M., Wechtenbruch, J., & Hessel, H. (2007). A prospective longitudinal quality of life study before and after cochlear implantation in post-lingually deafened adults. *Proceedings of the 8th EFAS Congress/ Joint meeting with the 10th Congress of the German Society of Audiology*, Heidelberg, Germany
- Newman, C. W., Weinstein, B. E., Jacobson, G. P., & Hug, G. A. (1990). The hearing handicap inventory for adults: Psychometric adequacy and audiometric correlates. *Ear and Hearing*, 11(6), 430–433. doi:10.1097/00003446-199012000-00004
- Paluch, R., Krueger, M., Hendrikse, M. M., Grimm, G., Hohmann, V., & Meis, M. (2018). Ethnographic research: The interrelation of spatial awareness, everyday life, laboratory environments, and effects of hearing aids. *Proceedings of the International Symposium on Auditory and Audiological Research*, 6, 39–46.
- Paluch, R., Latzel, M., & Meis, M. (2015). A new tool for subjective assessment of hearing aid performance: Analyses of interpersonal communication. *Proceedings of the International Symposium on Auditory and Audiological Research*, 5, 453–460.
- Przyborski, A., & Wohlrab-Sahr, M. (2009). *Qualitative Sozialforschung. Ein Arbeitsbuch. 2. korrigierte Auflage*. München, Germany: Oldenbourg Verlag.
- Ringdahl, A., & Grimby, A. (2000). Severe-profound hearing impairment and health-related quality of life among post-lingual deafened Swedish adults. *Scandinavian Audiology*, 29(4), 266–275.
- Robinson, K., Gatehouse, S., & Browning, G. G. (1996). Measuring patient benefit from otorhinolaryngological surgery and therapy. *Annals of Otolaryngology & Laryngology*, 105(6), 415–422. doi:10.1177/000348949610500601
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4(1), 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415
- Strauss, A. L. (1987). *Qualitative analysis for social scientists*. New York, NY: Cambridge University Press.
- Ventry, I. M., & Weinstein, B. E. (1982). The hearing handicap inventory for the elderly: A new tool. *Ear and Hearing*, 3(3), 128–134.
- Wagener, K. C., Brand, T., & Kollmeier, B. (2006). The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners. *International Journal of Audiology*, 45(1), 26–33. doi:10.1080/14992020500243851
- Wagener, K. C., Kühnel, V., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests. *Zeitschrift für Audiologie (Audiological Acoustics)*, 38(1), 4–15.
- World Health Organization. (2001). *International Classification of Functioning, Disability and Health (ICF)*. Geneva, Switzerland: Author.
- Wu, Y. H., Stangl, E., Zhang, X., & Bentler, R. A. (2015). Construct validity of the ecological momentary assessment in audiology research. *Journal of the American Academy of Audiology*, 26(10), 872–884. doi:10.3766/jaaa.15034