Evaluating user-operated tests of audible contrast sensitivity towards efficient hearing healthcare services

Raul Sanchez-Lopez[1,2,3] Johannes Zaar[2,4,] Søren Laugesen[1]

[1]*Interacoustics Research Unit, Kongens Lyngby, Denmark.*

[2]*Hearing Systems Section, Dept. of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark.*

[3]*Institute of Globally Distributed Open Research and Education (IGDORE).*

[4]*Eriksholm Research Centre, Snekkersten, Denmark.*

Abstract

The user-operated audiometry (UAud) project aims at introducing an automated system for user-operated audiometric testing into everyday clinical practice. Here, we focus on the Audible Contrast Threshold (ACT™) test, which has been proposed as a language-independent alternative to aided speech-in-noise tests. Five distinct user-operated ACT (U-ACT) test candidates were evaluated in terms of performance, reliability, and usability against two established test benchmarks. The study involved 28 participants with diverse hearing and cognitive abilities. The results demonstrated that the primary factor influencing the reliability of test results was the type of *task*. In terms of usability, the participants reported positive experience with all test candidates, participants reporting the Yes/No task as having the lowest perceived difficulty and best understanding. Overall, a test candidate with a custom adaptive procedure inspired by audiologists' experiences with ACT was chosen for the UAud protocol as a user-operated test of hearing-in-noise perception.

Introduction

With the perspective of an increasing global burden of hearing loss, there exists an urgent need to improve efficiency and reduce costs in hearing healthcare to ensure access to quality services[1,2]. The currently used audiometric tests, which are time-consuming and require active participation of a hearing care professional, represent a bottleneck that effectively limits access to hearing healthcare. To meet the demand for hearing assessments, the availability of well-trained audiologists and clinical facilities is crucial[3]. However, audiologists are responsible for various important tasks beyond hearing diagnostic tests such as hearing-aid fitting and verification as well as counselling. The User-operated Audiometry (UAud) project seeks to introduce an automated system that enables user-operated audiometric testing within everyday clinical practice, aiming to support audiologists in prioritizing hearing healthcare services and thereby allowing them to treat more patients and provide better hearing rehabilitation[4].

Currently, audiological assessments are mainly based on pure-tone audiometry, which consists of the detection of pure tones played at different frequencies and sound pressure levels to obtain the patient's hearing thresholds. However, there has been a growing focus on additionally evaluating supra-threshold abilities, such as the patient's ability to discriminate speech in the presence of noise when the speech is fully audible. The clinical evaluation of realistic, or ecologically valid, aided speech-in-noise perception presents several challenges, such as the need for multi-loudspeaker setups and speech material in multiple languages. This has motivated the search for alternative tests capable of capturing the fundamental abilities employed in challenging 'cocktail-party' scenarios, which are characterized by multiple interfering sound sources including speech. Recently, the Audible Contrast Threshold (ACT™) test[5] was introduced as a quick, simple, and

clinically viable manual version of a spectro-temporal modulation (STM) sensitivity test[6–8]. The results of ACT and its precursor, the STM test[6], showed high correlations with speech reception thresholds measured in hearing-in-noise tests. The speech reception threshold is defined as the speech sound level, relative to the interfering noise, at which an individual can repeat 50% of the speech material (sentences in this case) in an adaptive test varying the noise level. Notably, this correlation was stronger when using realistic speech-in-noise conditions with spatially distributed speech interferers[7]. Although the addition of ACT in current clinical practice is less resource-demanding than other time-consuming audiometric tests, an automated implementation of ACT offers significant benefits for busy clinical services, making it a practical choice alongside the automatic method for testing auditory sensitivity (AMTAS)[9,10].

In the present study, we therefore evaluated distinct candidate user-operated tests of audible contrast (U-ACT), which assess the supra-threshold hearing-in-noise abilities of the patient with language-independent stimuli and can thus complement the pure-tone audiometric testing. Both the ACT and the STM tests were considered here as test *benchmarks,* to which the different U-ACT test candidates were compared. While ACT is a test administered by an audiologist, STM is a user-operated test conducted using a research platform for psychophysical experiments. For the U-ACT tests we considered different *tasks* and *procedures*. In this context, a *task* is the type of patient action that produces a response in the test (e.g., pressing a button upon detecting an event), while a *procedure* is the method used to seek the threshold, for example, decreasing the contrast level by a certain value every time there is a response until no further responses are obtained.

The main goal of the present study was to identify an optimal candidate for a user-operated ACT test that is clinically feasible and adheres to the stated objectives. The objectives for a successful user-operated test were: 1) The test should be easy to perform such that the patient immediately understands what needs to be done. 2) The test should not be unnecessarily long; it must allow patients to maintain attention and performance effectively. 3) The test must be reliable and repeatable, demonstrating good agreement with standard or well-accepted alternatives (benchmarks). 4) Overall, the test should be accessible to a broad group of patients without requiring any special modifications for subpopulations, including those with low cognitive abilities. 5) The test should have an optimal quality of interaction so the patients can perform the test seamlessly. In the present study, we proposed a user-centered multi-dimensional evaluation which aims at providing evidence of the feasibility and adequacy of the new user-operated test in a realistic case scenario and with a diverse group of participants.

Results

A diverse group of participants

Since one of the study's goals was to obtain an evaluation from a diverse group of participants with varying hearing and cognitive abilities, care was taken during recruitment to obtain said diversity in the tested cohort (see Methods section). Thus, a total of 28 hearing-impaired (HI) participants with diverse hearing and cognitive abilities, all of them above 50 years old, were recruited along with 6 young normal-hearing (NH) listeners aged below 30 years. Figure 1 shows the results from pure-tone audiometric assessments, Reverse Digit Span (RDS) scores, and participants' ages. The HI participants were divided into two groups based on their RDS scores: a group of 11 people with nominally higher

cognitive abilities (HI_{hi}) and a group of 11 with lower cognitive abilities (HI_{lo}). This categorization is visually represented by a dashed line in the corresponding panel (third panel from the left) of Figure 1.

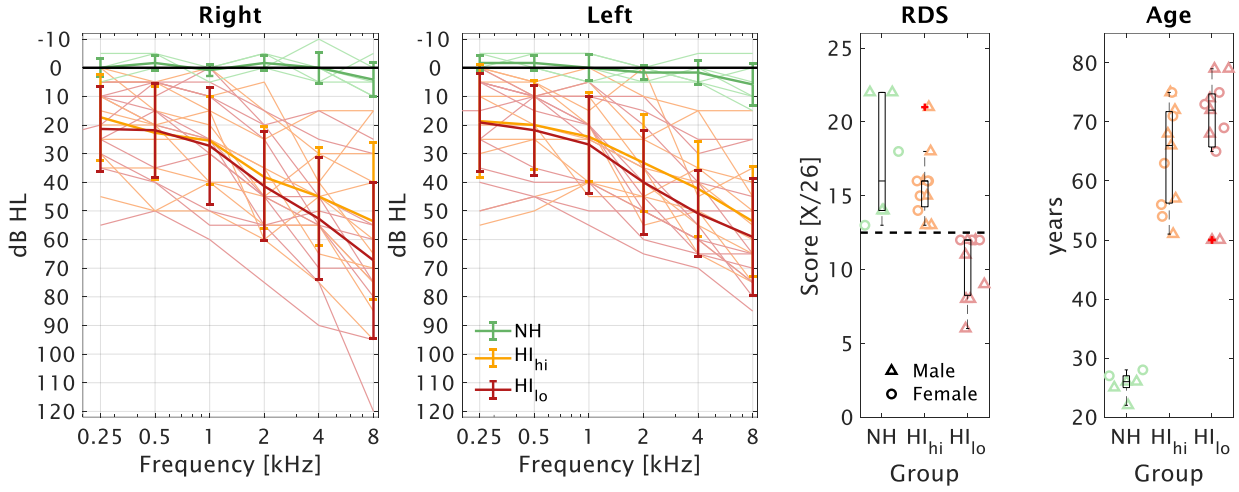

Figure 1: Summary of the participant characteristics. First and second panel from the left represent the average pure-tone thresholds (bold lines) with maximum and minimums (whiskers) and raw data (thin lines) corresponding to each of the three participant groups (NH: normal hearing in green, HI_{hi}: Hearing-impaired with higher cognitive abilities in orange, HI_{lo}: Hearing-impaired with lower cognitive abilities in red). Third panel from the left: Reverse Digit Span (RDS) scores, which were the basis for splitting the hearing-impaired group into HI_{hi} and HI_{lo} (separation indicated by black dashed line). Rightmost panel: Age, for the three participant groups.

The average hearing thresholds among participants with hearing loss showed no noteworthy differences between the two groups, with average thresholds below 1 kHz of about 20 dB hearing level (HL) and a gradual decline in hearing at higher frequencies (between 40 and 60 dB HL). On average, the hearing thresholds were slightly lower in the left ear than in the right ear in HI listeners. This was an incidental finding, as there were no specific symmetry requirements for audiograms during recruitment. The NH group scored between 14 and 22 on the Reverse Digit Span test (maximum score of 28), averaging approximately 17.8 points. The participants with hearing loss and higher cognitive abilities scored between 14 and 21 whereas the ones with lower cognitive abilities scored between 6 and 12. In terms of age, the HI_{lo} group was slightly older (median = 72 years) than the HI_{hi} group (median = 66 years).

Audible Contrast Threshold test benchmarks

Two benchmark tests of the participants' ability to detect spectro-temporal modulations were used, namely (i) a research version (STM) using psychoacoustic methods[6] and (ii) the clinically viable ACT test[5] which is administered by an audiologist. While STM is user-operated and uses a transformed up-down adaptive procedure[11], ACT is carried out by an audiologist, who presents the target stimulus and follows a Hughson-Westlake procedure (similar to the pure-tone audiometry)[12] to seek the threshold of the patient based on their responses. The results from the two benchmarks across all participants are displayed and compared in Figure 2.

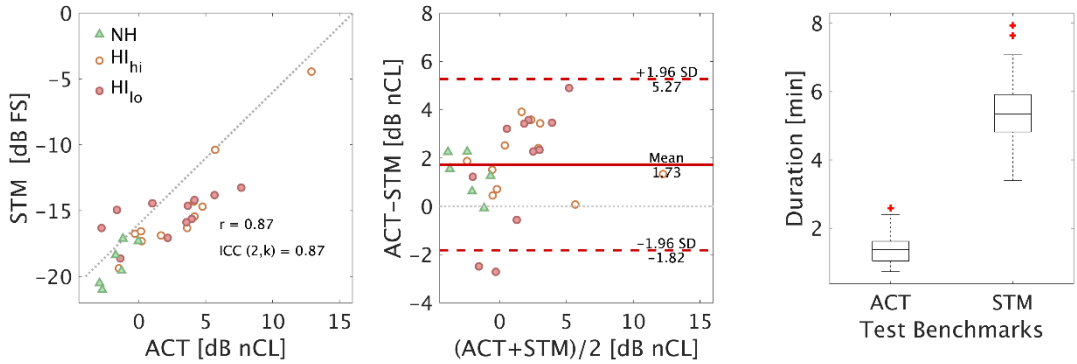

Figure 2: Comparison between the spectro-temporal modulation detection (STM) test and the Audible Contrast Threshold (ACT) test. Left panel: Scatter plot of agreement between tests. Middle panel: Bland-Altman plot of differences between tests. Right panel: Differences in duration, boxes represent the interquartile difference and whiskers the maximum and minimum values not considering outliers. Red crosses indicate outliers.

The left panel shows the STM results as a function of the ACT results. ACT (and the result of the U-ACT variants) is measured in normalized contrast level (nCL), a quantity where the physical magnitude, i.e., modulation level, is mapped to perceived contrast. Thus, an ACT score of 0 dB nCL represents the performance typically observed for listeners with normal hearing[5] (i.e., an elevated threshold indicates a contrast loss). In contrast, the STM test result is expressed in modulation depth in dB full scale (FS), where 0 dB FS corresponds to fully modulated and it is equivalent to 16 dB nCL[5]. For the Bland-Altman

analysis, the dB-FS results from STM were transformed to dB nCL by adding 16 dB. The intraclass correlation coefficient between the STM and ACT test results was very high (ICC(2,k) = 0.87) indicating excellent agreement. The participants' thresholds spanned a range of -2.95 to 12.9 dB nCL, with a mean value of 1.8 dB nCL. The differences, depicted in the middle panel, were smaller than $\pm$2 dB nCL with a bias of 1.73 dB showing that ACT values were slightly higher than STM results, which is consistent with Zaar/Simonsen et al.[5]. This result demonstrates the equivalence and agreement of both benchmark tests despite the substantial difference in testing time (Figure 2, right panel), with ACT requiring much shorter testing time in line with Zaar/Simonsen et al.[5]

User-operated Audible Contrast Threshold test candidates

We designed and implemented a total of five U-ACT test candidates which differed in terms of *task* and *procedure*. In a non-automatic test, the *task* is related to the instructions given to the participant while the *procedure* consists of step-by-step actions followed by the examiner. In user-operated tests, the *procedure* is automated. Each U-ACT candidate was numbered from 0 (U-ACT-0) through 4 (U-ACT-4). U-ACT-0 mirrors the standard ACT test by using the same *task* and *procedure* (but in an automatic manner), while U-ACT-4 aligns closely with the STM test's *task* and *procedure*.

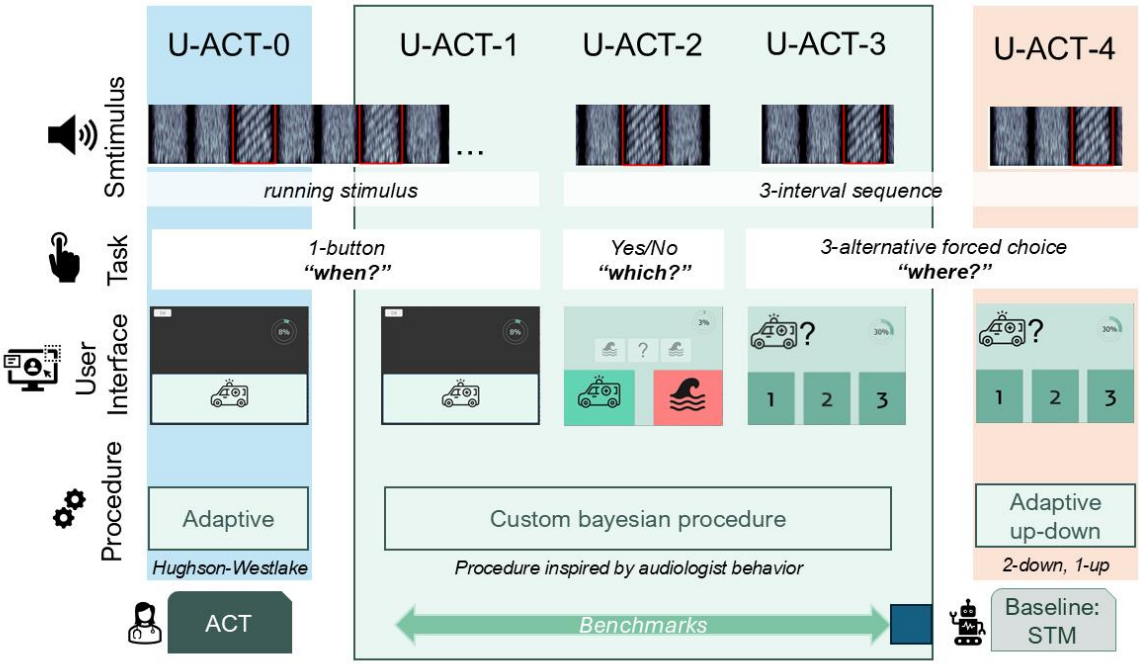

Figure 3: Illustration of the essential aspects of the user-operated Audible Contrast Threshold (U-ACT) test candidates, in terms of task, procedure, and similarity with the test benchmarks. In the user-operated domain, we introduced a visual communication system in which the target stimulus is referred to as an 'ambulance', and the reference sound is described as a 'wave'. From left to right, the test candidates are designated with an increasing number related to their resemblance with the clinical Audible Contrast Threshold (ACT) test. U-ACT-0 mirrors the clinically viable ACT benchmark (but automated) while U-ACT-4 resembles the spectro-temporal modulation detection (STM) test. The differences are described in terms of stimulus presentation (first row), where the target is inside the red square, task (second row), user interface (third row), procedure (fourth row) and resemblance with the benchmarks (bottom row).

Figure 3 depicts the five U-ACT candidates showing their progression from quick-and-simple ACT-like (U-ACT-0) to the research-oriented STM-like (U-ACT-4) designs. In terms of the stimuli presentation, U-ACT-1, and the previously mentioned U-ACT-0, used a "running stimuli" paradigm requiring responses 'when' the target was presented. As depicted in Figure 3, the user interface consisted of a single response button. The test candidates U-ACT-2, U-ACT-3, and U-ACT-4 employed a sequential presentation of three stimuli with brief pauses between intervals. The sequential presentation tests differed in terms of *task*: in U-ACT-2 the first and third intervals were always references, while the second interval contained either a target or reference (a 'which' *task*) and the user interface accordingly consisted of two buttons, a target (left button) or reference (right

174 button). U-ACT-3 and U-ACT-4 applied a 3-Alternative Forced Choice (3-AFC) *task* (a
175 'where' task) where the target is in one of the three intervals and the user interface
176 consists of three buttons corresponding to the three intervals. In terms of *procedure*, U-
177 ACT-1, U-ACT-2 and U-ACT-3 share a Bayesian-based custom tracking *procedure* (see
178 Methods section) instead of typical up-down adaptive procedures. The *procedure* was
179 inspired by the experience of audiologists with the manual ACT and it utilizes a Bayesian
180 approach to decrease the uncertainty of the measure (the participant's threshold) with
181 each new stimulus presentation. The present multidimensional evaluation aims to compare
182 three tests sharing a Bayesian-based procedure but differing in tasks with U-ACT-0 and U-
183 ACT-4, which resemble the benchmarks more closely.

184 The study protocol involved two *visits*, each comprising two *blocks*. Within each *block*,
185 participants completed one run of each test candidate, presented in random order. Data
186 from all four *blocks* were utilized for various investigations into the feasibility of the test
187 candidates.

188 Efficacy of non-verbal instructions

189 In the present study, we specifically focused on exploring the utility of language-
190 independent instructions in user-operated tests. To achieve this, we implemented a set of
191 non-verbal instructions consisting of four main steps: 1) putting the headphones on 2)
192 familiarization, 3) training, and 4) demonstration. These instructions were refined through
193 multiple iterations with naïve users interacting with the system[13]. During this investigation,
194 we identified key graphical signifiers (a set of icons), that facilitated correct instruction
195 usage while allowing for user mistakes. We defined efficacy based on a success rate.
196 Thus, a test run in *block* 1 was considered successful if the ACT value obtained was within
197 4 dB of the average results obtained in *block*s 2-4. The decision to use 4 dB as a criterion

was taken because it is two steps apart in the clinical procedure and is considered within the limits of agreement of the test[5]. *Block* 1 served as an experimental *block* where some participants were expected to perform less well due to the absence of training or previous explanations and lack of familiarity with the manual ACT test at this point in time.

Table 1: Success rates obtained with the different candidate paradigms in the three participant groups. Definition of success: U-ACT-x in block 1 was less than 4 dB apart from the U-ACT-x mean from the three subsequent blocks.

	U-ACT-0	**U-ACT-1**	**U-ACT-2**	**U-ACT-3**	**U-ACT-4**
NH	100%	100%	100%	83%	100%
HI_{hi}	72%	63%	54%	81%	100%
HI_{lo}	54%	45%	54%	63%	63%

Table 1 shows the success rate of each U-ACT test candidate across the three participant groups. The NH group demonstrated high success rates for most U-ACT candidates, while HI sub-groups showed lower performance, particularly those with lower cognitive abilities HI_{lo} (45-63%). Notably, the 3-AFC task (U-ACT-3 and -4) achieved the highest success rates across both HI groups. For those paradigms, 7/11 participants with lower cognitive abilities and 9/11 with higher cognitive abilities performed reliably without prior knowledge and only with the help of the non-verbal instructions. The test with the lowest success rate in the two HI sub-groups was U-ACT-2 while the U-ACT-1, with running stimulus, showed low success rate only in the group with lower cognitive abilities. Note that U-ACT-2 is the only candidate that has catch trials.

Effect of cognitive abilities, task, and procedure on the measured thresholds

To explore the effect of *task*, *procedure*, and the likely influence of cognitive abilities, we analyzed the differences between the thresholds obtained from all five U-ACT test candidates and those from the manual ACT, with which the participants were also tested. Figure 4 shows the threshold differences for each of the three groups of participants.

The 1-button tasks (U-ACT-0 and -1) with running stimuli showed minimal differences compared to the manual ACT values. In contrast, tasks involving a 3-interval sequence (U-ACT-2-4) yielded significantly lower thresholds than those from the manual test in agreement with Zaar/Simonsen et al.[5]. In the analysis of variance (ANOVA) with a mixed effects model, no significant differences were observed across the participant *groups* suggesting that any differences in mean performance among ACT and the various U-ACT tests were largely independent of participants' cognitive abilities. However, we can't estimate with the current dataset to what extent cognitive abilities affect the ACT results based on the present findings.

Examining the overall data*, task*-specific differences were evident: there were significant mean threshold differences between the 1-button tasks (U-ACT-0 and U-ACT-1) and the 3-interval yes/no choice of U-ACT-2 (t.ratio = 8.96, $p < 0.001$) and between the 1-button tasks and the 3-AFC tasks, i.e., U-ACT-3 and -4, (t.ratio= 15.17, $p < 0.001$). No significant differences were observed among the tests with sequential presentation (U-ACT-2, U-ACT-3 and U-ACT-4). Notably, the self-paced sequential presentation allows the participant to reflect on the response after the presentation of the three intervals, while the 1-button task is done while there is a running stimulus and it has a fixed response window (0.6-1.4 seconds after the onset of the target stimulus).

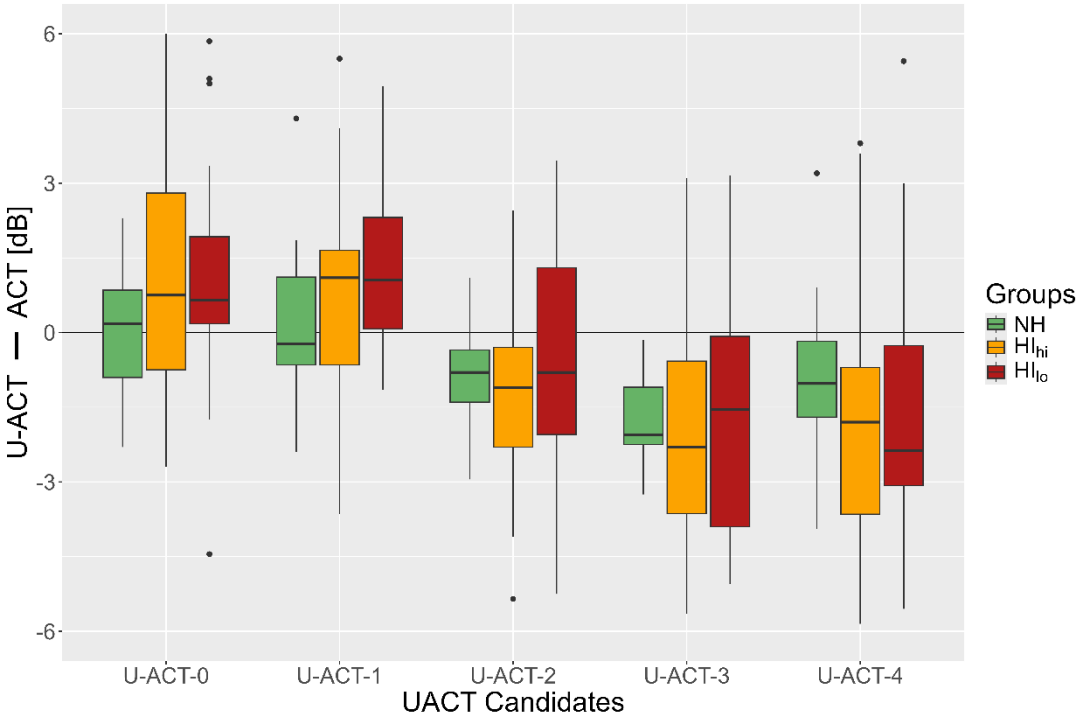

Figure 4: Differences between user-operated Audible Contrast Thresholds (U-ACT) and manually obtained Audible Contrast Thresholds (ACT). In the boxplots, the horizontal lines indicate the median, boxes indicate interquartile difference, and vertical lines indicate the range between the max and minimum values within ±1.5 times the interquartile range (whiskers). Individual data points outside of the whiskers are also indicated.

Quality of measurements: Reliability and agreement with benchmarks

The test-retest reliability of the U-ACT test candidates was evaluated using the thresholds obtained in the second *block* of each visit (that is *block* 2 in visit 1 and *block* 4 in visit 2). For the analysis of the reliability and repeatability, three different metrics were employed: the inter-class correlation (ICC)[14], the coefficient of repeatability (CoR)[15] from the Bland-Altman plots[16,17] and a metric based on information retrieval named Fractional Rank Precision (FRP)[18]. While ICC and FRP are quantified on a scale from 0 to 1, where 1 indicates ideal reliability, the CoR represents the range within which repeated measurements are likely to fall. The CoR is thus expressed in dB, and a smaller range indicates a more reproducible measure. These metrics provided a comprehensive assessment of the tests' consistency.

Table 2 shows the reliability results of the U-ACT test candidates and the two benchmarks. The STM test demonstrated the highest reliability. U-ACT-2 emerged as the second-best test, and thus the best U-ACT candidate in terms of reliability, across all considered metrics. The manual 1-button ACT was substantially more reliable than its user-operated counterparts (U-ACT-0 and U-ACT-1).

Table 2: Test-retest reliability of the user-operated Audible Contrast Threshold (U-ACT) test candidates and test benchmarks analyzed using data from blocks 2 and 4. The three metrics are the inter-class correlation (ICC), coefficient of repeatability (CoR), and fractional rank precision (FRP). The rightmost column summarizes the average test duration. Bold letters indicate the best values among the U-ACT test candidates.

Test paradigm	ICC (2,1)	CoR (dB nCL)	FRP (ratio 0-1)	Average Duration (s)
U-ACT-0	0.71	3.72	0.75	**90**
U-ACT-1	0.69	4.11	0.70	97
U-ACT-2	**0.84**	**3.27**	**0.78**	*202
U-ACT-3	0.80	3.59	0.73	*145
U-ACT-4	0.78	3.62	0.74	*292
ACT	0.83	3.51	0.75	82
STM	0.87	3.04	0.75	324

* In hindsight, U-ACT-2,3 and 4 included an unnecessary 1-second pause, which added about 25 seconds to the reported test duration.

To assess the agreement between U-ACT candidates and benchmarks, the same metrics were applied to the average thresholds obtained across *blocks* 2, 3, and 4. Table 3 highlights that the U-ACT-2 test candidate showed remarkable agreement with both benchmarks ACT and STM achieving an ICC > 0.9, a low coefficient of repeatability (lower

than the one corresponding to ACT, i.e., CoR < 3.5) and the best fractional rank precision (FRP = 0.8) compared to the other candidates. Although the other U-ACT test candidates showed an acceptable agreement with both benchmarks, it is remarkable that U-ACT-4 exhibited an excellent agreement with the STM test (ICC(2,k) = 0.97) despite being shorter and using larger step sizes (see Methods).

Table 3: Agreement between the U-ACT test candidates and the test benchmarks analyzed using the mean across blocks 2, 3, and 4. The three metrics were the inter-class correlation (ICC), coefficient of repeatability (CoR) and fractional rank precision (FRP).The best value for each metric is highlighted in bold font.

	ICC (2,k)		CoR (dB nCL)		FRP (ratio 0-1)	
Agreement	ACT	STM	ACT	STM	ACT	STM
U-ACT-0	0.86	0.69	3.46	4.18	0.77	0.69
U-ACT-1	0.88	0.76	3.70	4.37	**0.78**	0.66
U-ACT-2	**0.93**	0.94	**3.16**	2.90	0.77	0.82
U-ACT-3	0.80	0.95	4.15	2.81	0.66	0.81
U-ACT-4	0.86	**0.97**	3.77	**2.43**	0.69	**0.83**

Quality of interaction

Participants completed a subjective evaluation immediately after each test (except the initial *block*). The evaluation consisted of a series of statements using a continuous visual analog scale (VAS)[19] from "completely disagree" to "completely agree". The statements were divided into two categories: test attributes and self-confidence with the test. Test attributes included difficulty, perceived length, demand, and boredom. Self-confidence

questions aimed at determining if participants understood the instructions, found the test useful, thought that the results were good, and what their overall opinion of the test was. The evaluation was conducted twice for each attribute, once with a positive statement and once with a negative statement, referred to as *wording*. Figure 5 and Figure 6 display the standardized data (see Methods) of the usability evaluation for the five U-ACT candidate tests.

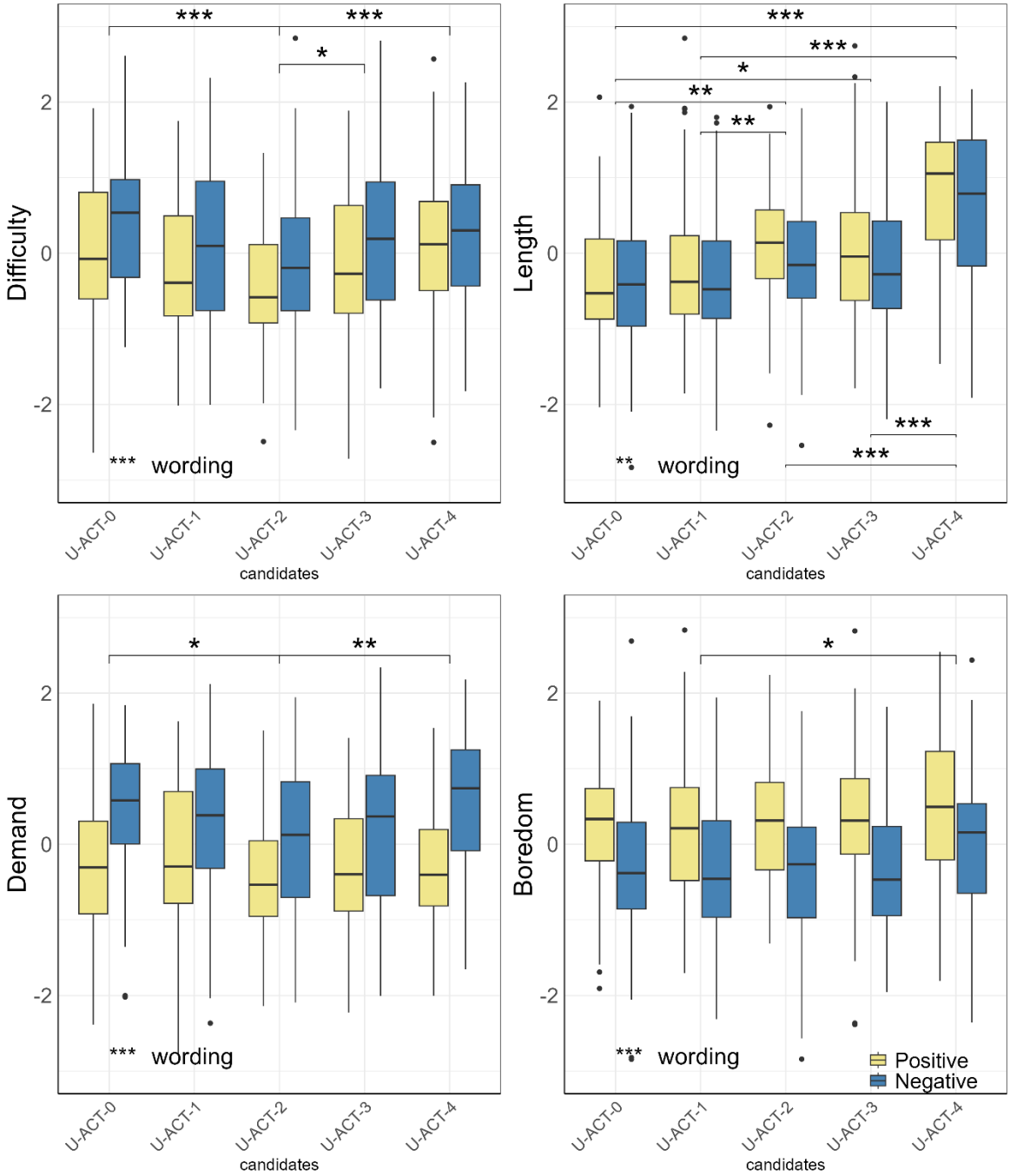

Figure 5: Subjective evaluation of the usability of the user-operated Audible Contrast Threshold (U-ACT) candidate tests (U-ACT-0 to U-ACT-4). Standardized subjective ratings of the attributes related to the test: Difficulty, Length, Demand, and Boredom. Boxplots depict the median, interquartile ranges, and whiskers as in Figure 4. The colors represent Positive (yellow) or Negative (blue) wording. Asterisks indicate statistically significant differences between groups ($p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).*

.

Figure 5 compares subjective ratings of attributes related to the test experience. Lower ratings indicate better results, meaning less difficult, shorter, less demanding, or less boring. The *wording* had a significant impact on all four attributes. This suggests that the phrasing of statements can affect ratings. For example, participants found U-ACT-2 to be less “difficult" than U-ACT-4 but still equally “easy". The most significant results showed that U-ACT-4 was perceived as significantly longer than the other tests and U-ACT-2 was less demanding than U-ACT-0 and U-ACT-4.

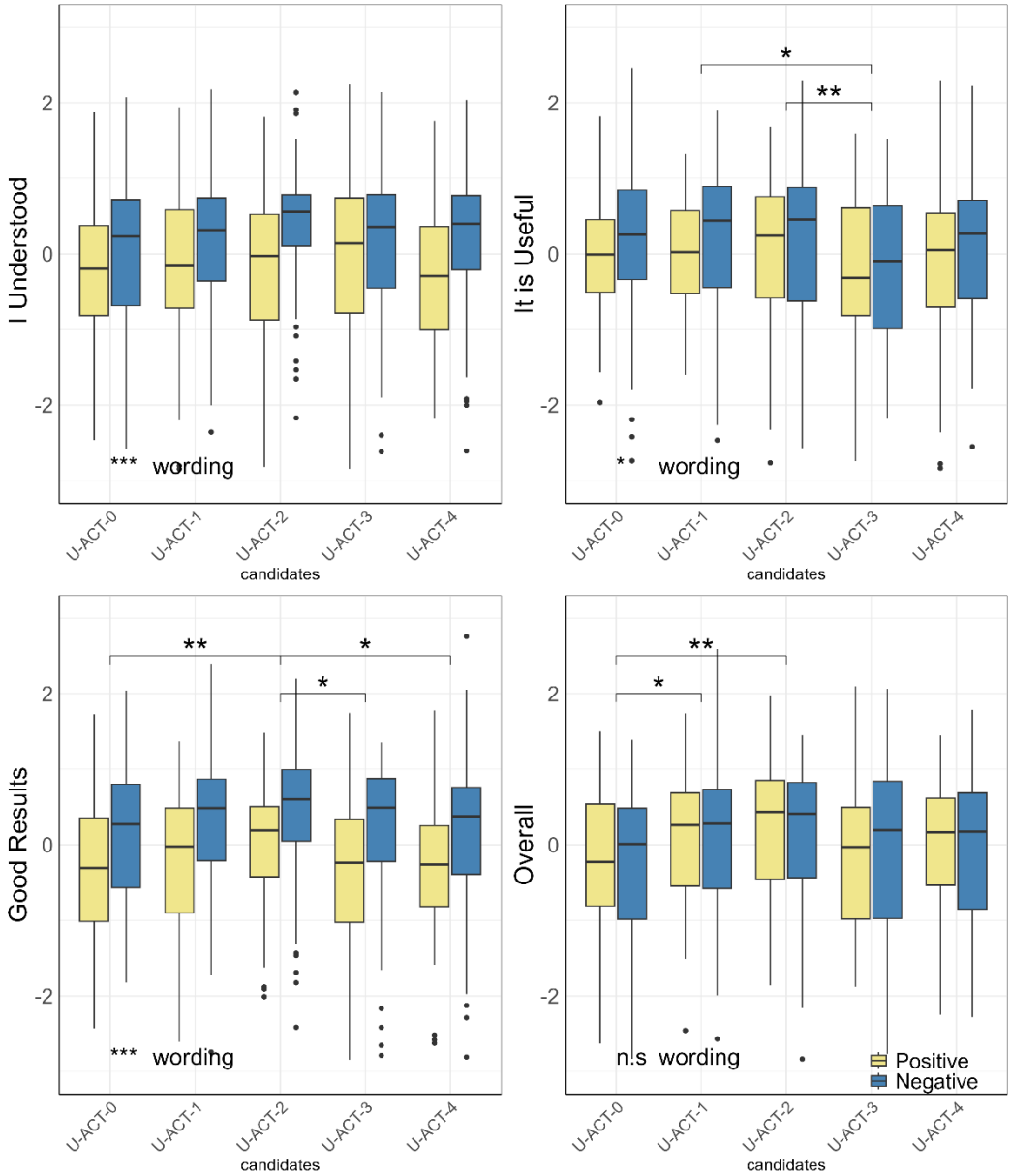

*Figure 6: Evaluation of the usability of the user-operated Audible Contrast Threshold (U-ACT) candidate tests. Average subjective ratings of the attributes related to self-confidence with the tests: Understanding of Task, Usefulness, Goodness of Results, and Overall opinion. Boxplots depict the median, interquartile ranges, and whiskers as in Figure 4. The colors represent Positive (yellow) or Negative (blue) wording. Asterisks indicate statistically significant differences between groups (n.s. p>0.05 , * p < 0.05, ** p < 0.01, *** p < 0.001).*

Figure 6 shows subjective ratings reflecting attributes related to participants' confidence with the tests. Higher ratings indicate better results, meaning that the participant understood the task, found the test useful, believed they did well, or had a positive overall opinion of the test. From these attributes, only the overall opinion did not differ significantly

in terms of *wording*. U-ACT-3 was perceived as significantly less useful than U-ACT-1 and U-ACT-2. Additionally, participants reported that the results obtained using U-ACT-2 were significantly better than the ones from U-ACT-0, U-ACT-3 or U-ACT-0.

Discussion

The present study consisted of a multi-dimensional evaluation of various test candidates for a user-operated auditory test. The study aimed to generate quantitative data on test reliability, observations from tracking user interactions with the system, and users' self-reported opinions about usability. Instructions are critical for obtaining meaningful results in user-operated tests. Initial results from *block* 1, conducted without additional instructions, indicated that a substantial proportion of hearing-impaired listeners may struggle understanding their task, particularly those with lower cognitive abilities. All test candidates showed good agreement with benchmarks and good test-retest reliability. While the procedures and tasks differed from classical psychoacoustic tests[7,11], results were comparable to benchmarks using perceptual evaluation. Some insights emerged, such as perceived differences in test difficulty and length among candidates, and a difference in results based on attribute wording which had a strong and clear significance in user evaluations. However, it remains inconclusive whether negative formulations of the same evaluated attribute consistently elicited stronger opinions, as might be expected by the concept of negativity bias[20].

The chosen test candidate for the UAud clinical trial

The study evaluated user interaction in a realistic scenario and aimed to identify the "best candidate" for user-operated ACT. The selected candidate was U-ACT-2, which exhibited high reliability and repeatability, with excellent agreement compared to ACT and STM

benchmarks. Overall, the test was appropriate in terms of length and acceptable, in the sense the user's evaluation of U-ACT-2 was average or better than average in all dimensions considered. While generally easy to perform due to its simplified target/reference discrimination task, U-ACT-2 initially presented comprehension challenges for some participants during the first block. However, addressing this through improved instructions and combined verbal/non-verbal cues might be sufficient[21].

Relation to other studies

Previous studies have proposed quick and reliable automated tests of spectro-temporal modulation perception[22–24]. In most cases, the process of developing and validating the test has been meticulously done in various steps, adding complexity in each new study towards a realistic application. In contrast, we have directly tested U-ACT in a realistic environment, with a diverse sample in terms of auditory and cognitive abilities. Furthermore, we have evaluated the use of instructions and the quality of interaction in automated auditory tests. This is the first time, to our knowledge, that an auditory test has been selected and evaluated in a multimodal fashion, including with the user experience in the center and the RAMES framework[25] as a guidance in the field of audiology.

Strengths and limitations of the study

The UAud approach aims to implement user-operated hearing assessments in healthcare services. We proposed a protocol for multi-dimensional evaluation involving reliability, efficiency, efficacy, and subjective evaluations. Key strengths of the study include a diverse participant selection regarding hearing losses and cognitive abilities, and a comprehensive evaluation using a simple two-visit protocol. *Block* 1 evaluated instruction efficacy, blocks 2 & 4 assessed measurement reliability and quality, while block 3 provided a fresh run after one week. Inclusion of benchmark test measurements, subjective

evaluations, and interviews further strengthened the evidence supporting our novel user-operated test design.

While the multi-dimensional evaluation was valuable, there are limitations to consider. Firstly, participants were sourced from the university's database, potentially biasing results due to prior hearing research experience affecting task interactions. Secondly, repeated assessments of the same listening ability within each block may have positively influenced results through training effects. Additionally, the protocol lacked speech intelligibility in noise tests as a reference measure, as included in Zaar et al.[6]

Outlook

The auditory tests used for audiological diagnosis should provide a reliable estimate of a patient's hearing status; however, manual and user-operated tests need not be based on identical tasks or procedures[9]. While the reliability of test results is influenced by user behavior, it is the responsibility of both the audiologist (human) and the system (represented by the system designer) to present the test in a manner that minimizes influencing factors unrelated to hearing status, such as attention, understanding of the task, and fatigue[26,27]. In user-operated tests, it is reasonable to think that this difference in the allowed response time is the most influential factor. The results here support using sequential presentation in user-operated tests (as also demonstrated in AMTAS studies[10]), while emphasizing the importance of the examiner's role in manually operated ACT. In research, the current version of U-ACT opens the opportunity to test other populations (pediatric, cognitively impaired) using a single and similar framework for audibility and contrast tests, and to explore the use of U-ACT as a remote test (as in the case of the audiometry[28]). Overall, U-ACT can help to investigate further the role of contrast loss in auditory perception and communication handicap.

Methods

The study design was inspired by the Roles Activities Materials Environment System (RAMES) framework which supports user-centered evaluations in research studies[25]. The objective of the framework is to check all the factors affecting the outcomes of the evaluation, knowing the interaction among the different elements. Besides, it was guided by the ISO/IEC 25022:2016[29], standard considering the quality in use with focus on the context coverage. The context coverage is related to efficacy, efficiency and satisfaction and it is defined as the extent to which the system can be used by people without specific knowledge, skills or experience. The definition of best candidate was based on the criteria of Table 4.

Table 4: Criteria of the definition of best candidate.

Psychoacoustics	User interaction	Audiometric groups	Cognition groups
High Reliability Low training effect	High success rate Low time in task Low Self-reported Difficulty	All groups can perform the test	All groups can perform the test

The following sections explain how these results were assessed in the study.

Study design and participants

Participant recruitment

The potential participants of the study were selected from the database of the Hearing Systems Section of the Technical University of Denmark (DTU). In the recruitment process, care was taken to ensure a balanced number of male and female participants,

and a reasonable variety of hearing losses and cognitive abilities as reflected in their audiometry and RDS scores. The potential participants were divided into three groups already during the recruitment based on the information contained in the database. We aimed at a total of 30 participants where there would be a control group formed by 6 young NH listeners (<30 years old), a group of people with hearing loss and with high cognitive abilities (i.e., higher scores in the RDS test, $HI_{hi.}$), and another group of participants with hearing loss with lower cognitive abilities (i.e., lower scores in the RDS test HI_{lo}). Each HI group contained 12 participants according to the screening data and consisted of older individuals with a minimum of 50 years of age. However, the RDS used for the stratification was the one performed during the study. The only exclusion criteria were a hearing loss above the range where the ACT can be performed (i.e., exceeding 95 dB HL) or an inability to operate a tablet due to dexterity or vision problems. Thus, a total of 30 participants with diverse hearing and cognitive abilities were recruited for the study. Two participants abandoned the study after the first visit and 28 participants completed the protocol. This study was carried out in accordance with the ethical approval of the Danish Science-Ethics Committee granted by the Capital Region Committee (H-16036391). All participants gave written informed consent and were offered an economic compensation for their participation.

Equipment and testing environment

The study was conducted in the laboratory of the Hearing Systems Section at the DTU, specifically in the audiology clinic (which resembles a typical audiology clinic). This facility includes a sound-isolated booth equipped with a clinical audiometer, videoscope, middle-ear analyzer, and computer for conducting hearing-in-noise tests and cognitive assessments. For the present study, we employed a clinical research prototype of Audible

Contrast Threshold testing implemented in MATLAB® (Version 9.8, The MathWorks Inc.). The prototype consists of an RME UC soundcard (RME Audio, Haimhausen, Germany) connected to a Lake People G103-P MKII amplifier (Lake People, Überlingen, Germany). The prototype included a pair of headphones Radioear DD450 (Middelfart, Denmark), and a push button. For the user-operated tests, we used an iPad (Apple Inc., Cupertino, CA, USA) connected to the computer functioning as an external screen using the software Splashtop Wired Xdisplay (Splashtop Inc., San Jose, CA, USA).

Protocol

The participants were invited for two visits that occurred at least one week and no more than four weeks apart. In the first visit, they were introduced to the study and were informed that the test involved a usability test and that we wanted the participants to interact with the system. After this briefing, participants underwent a standard pure-tone audiometry examination administered by a trained audiologist before they conducted the main experiment. The second visit began with the administration of the RDS test followed by the main experiment. The main experiments consisted of a total of four testing *block*s (two *block*s in each visit), where participants evaluated five alternative U-ACT candidates, with candidates presented in balanced order across groups and *block*s. In each visit, both test benchmarks (that is, ACT and STM) were administered in between the two testing *block*s once the examiner had explained the test to the participant. After *block* 4, semi-structured interviews were conducted with each participant. The interview included questions about the opinions of having user-operated tests in the clinic and the utility of the U-ACT to complement the audiometry. The qualitative results of these interviews are not within the scope of the present investigation.

Auditory tests of audible contrast sensitivity

Stimuli

The stimuli used in all the tests performed in the study were identical to those used in the ACT as described by Zaar/Simonsen et al.[5], which has been thoroughly explained in previous studies using STM detection[6]. In general terms, the stimuli consist of 1-second-long broadband noise carriers (354-2000 Hz) – serving as reference stimuli -, onto which spectral modulations of 2 cycles per octave and temporal modulations of 4 Hz are imposed to create the target stimuli. The experimental variable was the modulation depth (m), which was transformed into dB FS by $20\ log(m)$ for the STM test, and into dB nCL for all the ACT tests (both manual and user operated).

Test benchmarks

Clinical assessment of contrast sensitivity has emerged as an important diagnostic tool for evaluating a patient's ability to extract subtle acoustic information from complex sounds, and its use in hearing rehabilitation is supported by evidence[5,6]. In this study, we compare the performance of U-ACT candidates with two benchmarks that have been previously used in relevant studies[5,6]. The first benchmark, referred to as the STM detection test[6,7], is a psychoacoustic measure that utilizes a three-alternative forced choice (3-AFC) task and a 2-down 1-up procedure that approximates the 70.7%-point on the psychometric function[11]. The psychometric function represents the probability of detecting the target sound and it is often modelled as a logistic regression function with a midpoint (threshold) and a slope that define the performance for any level of the tracking variable. The second benchmark is the clinical version of the Audible Contrast Threshold (ACT™)[5] test, which has been widely used in previous studies.

Test tasks

Three *task*s were considered for the U-ACT test candidates. These were:

1. 1-button task: The stimuli are presented in 1-second-long successive waves of noise, and occasionally the wave contains the target stimulus (i.e. it is modulated), thus differing from the other (unmodulated) reference waves. We use the term ‘wave’ because the 1-second-long noise has a fade-in and fade-out and it is followed immediately by another 1-second-long noise so it sounds like waves in the ocean. A large virtual button, referred to as the '1-button', was placed on the touch screen occupying 40% of the screen's area at the bottom. After the presentation of the target stimulus, the participant had a 600-ms response window after the stimulus had stopped (1.6 s after the stimulus start) to touch the button. If the participant did not touch the screen, the trial was marked as MISS, whereas the trial was marked as a HIT if the participant touched the screen during the response window. The screen touches outside of the response window were registered as false positives. This task was used in U-ACT-0 and U-ACT-1.
2. Yes/No task: The stimuli were presented in triplets, the first and last intervals were always references, while the second interval could be either a reference or a target stimulus. The user’s task was to press either the button depicting an ambulance when detecting a target, or the button depicting a wave indicating the detection of a reference, i.e., no detection of a target. This task resembles the one used in AMTAS[10] and includes catch trials with feedback when the participant gets caught pressing the button when there was no target stimulus presented. This was done to ensure that the performance and attention is maintained during the test. The catch trials consisted of trials containing a reference stimulus in the second interval. There were 5 catch trials which were assigned at the beginning of the test to be presented before specific trials. The first catch trial is always between the first and second trials, the remaining were assigned randomly. The feedback was presented only when the participant was caught in a catch trial which is expected to minimize the bias and have less influence on the results of the actual test [30,31]. There were no restrictions on response time. This task was used in U-ACT-2.

3. 3-AFC task: The stimuli were presented in triplets. The screen showed three virtual buttons with the numbers 1-3, which were highlighted sequentially in synchrony with the playback of the three intervals. The target stimulus was randomly placed in one of the three intervals. The task of the participant was to indicate which of the three intervals was different from the other two by pressing the button numbered with that interval. If no differences were heard, guessing was compulsory to advance the test. This task was used in U-ACT-3 and U-ACT-4.

Test procedures

Three *procedure*s were considered for the U-ACT test candidates:

- Modified Hughson-Westlake procedure[12]: This procedure was identical to that used in the ACT test, but automatic. After each response (either HIT or MISS), the contrast level of the next target stimulus was adjusted by, respectively, decreasing or increasing the contrast level. The step size was 2 dB nCL. After a HIT, the contrast level was decreased by 2 steps (i.e., 4 dB nCL). After a MISS, the contrast level was increased by 1 step (i.e., 2 dB nCL). The testing stopped when three out of five positive responses occurring after a non-response happened at the same contrast level. This is referred to as ‘the ascending method’ in pure-tone audiometry[32]. This procedure was used in U-ACT-0.
- Custom Bayesian procedure: The presented procedure is an adaptation of the BAYES-FIG[33]. In their procedure, Remus and Collins combined Bayesian inference with the Fisher information matrix and Theory of Optimal Experiments. Overall, the Bayesian inference aims to reduce the uncertainty by obtaining new evidence (a response at a certain level) and updating the probability of a given combination of midpoint and slope of the psychometric function to be the true result. The Fisher information matrix is obtained using the derivative of the psychometric function with respect to the midpoint and the slope and estimating which value of the tracking variable is the one that should be tested next to maximize the information provided by the new evidence. However, our method incorporates distinct variations:
    - Bayesian updates: Unlike BAYES-FIG, our method utilizes the last two trials (n-1 and n) to update probabilities, rather than relying solely on the last trial

537 as in other psychoacoustic procedures based on Bayes theorem. This was
538 done under the assumption that using two trials makes the estimation more
539 robust to momentary lapses in participant attention. This assumption was
540 tested and confirmed using Monte Carlo simulations.

541 - Informational gain function: Instead of using Fisher's informational gain, our
542 method employs a different information gain function that is derived solely
543 from the slope's derivative.

544 - Optimal next CL algorithm: In the BAYES-FIG the selection of the level for
545 the next trial is based on the maximum of the information gain function. In
546 contrast, our method alternates between the maximum and minimum values
547 corresponding to the 82% and 18% points of the psychometric function,
548 respectively, on even and odd trials. The exact points of the psychometric
549 function are defined by the partial derivative of the psychometric function with
550 respect to the slope.

551 - Exception inclusion: To mirror the manual test experiences reported by
552 audiologists, our procedure incorporates additional catch trials (used in the
553 Yes/No task) or presentations above threshold to maintain patient attention
554 and ensure consistent performance. These presentations occurred based on
555 erratic patient responses that might suggest that the patient has lost the
556 internal reference of the sound that they have to detect.

557 - Catch Trials: In the Yes/No task, catch trials were included in the procedure.
558 By default, the system considers 4-5 catch trials assigned randomly to
559 specific trial numbers with 4-8 trials in between.

560 Overall, the Bayesian procedure consisted of 25 trials that started with 8 pre-test
561 trials for a first approximation of the threshold using the modified Hughson-Westlake
562 method and then continued alternating easier and more difficult presentations using
563 the Optimal next CL algorithm. This procedure is employed in U-ACT-1,2 and 3.

564 • Modified 2-down 1-up procedure: This procedure was based on the 2-down 1-up
565 described by Levitt[11], with a slight modification to make it shorter. Thus, the contrast

level of the stimulus was decreased only if there were two consecutive correct responses and increased otherwise. The entire procedure consisted of two iterations (sub-runs), each of them consisting of 4 upper reversals with decreasing step size. An upper reversal is a change of direction that takes place in ascending direction after two consecutive positive responses. At the start, the step size was 8 dB after the first upper reversal it decreased to 4 dB, and after the second to 2 dB. The procedure then continued until there was a correct response after a negative response. The second iteration started 4 dB nCL above the last presentation level and the procedure was repeated identically. This procedure was employed in U-ACT-4

Post-processing and threshold estimates

For each test run, the levels of the tracking variable and the participant responses constituted a trace. In the context of adaptive testing procedures, a trace is a step-by-step history of the tracking variable from the start and until it converges to the threshold. The trace was fitted to a psychometric function[34]:

$$\psi = \frac{100}{1+e^{4s_{50}\cdot(x_{50}-x)}} \ ,$$

where the slope (s_{50}) and midpoint (x_{50}) are the parameters, x is the contrast level and ψ is the % correct. In the fitting process, the slope was kept fixed ($s_{50} = 0.2518\ \mathrm{dB}^{-1}$) and x was defined in the closed interval [-8, 16] with a step size of 0.1. The variable x was thus allowed to go below the minimum value used with ACT (-4 dB nCL) to ensure accurate representation of the entire psychometric function, particularly for cases where threshold levels were very low. For all the ACT tests (manual and user-operated) the ACT value was taken as the $\psi = 70.216\ \%$-point on the psychometric function, following the recommendations from Zaar/Simonsen et al.[5]. The STM benchmark followed the same procedure as in the Zaar et al.[6,7] studies (i.e. a transformed 2-down 1-up procedure[11] with

a final step size of 1 dB). However, the our test procedure of the STM was extended by 8 more reversals compared to the studies of Zaar et al.[6,7] which allows an investigation of the likely effects of fatigue (not considered here). The threshold estimate (STM_{th}) was the resulting mean across the values of the first 8 reversals (4 upper and 4 lower) at a step size of 1 dB In this case, the threshold corresponds to the $\psi = 70.7\%$-point[11].

Non-verbal instructions

The instructions consisted in a succession of assignments and were designed so they only contained non-verbal audiovisual information[35]. The participants had to go through four instruction phases, each of which comprised a specific assignment and a goal that they had to achieve by themselves. However, the participants could make use of different types of support during the instruction phases. 1) Help: If the user pressed a question mark in the bottom left corner of the touch screen, minimalistic written information was added to the screen. 2) Written instructions: If the user pressed the blue “information” button, thorough written instruction appeared on the screen. 3) Call the examiner: The user always had the option to call the examiner for task explanation. The participants could go backwards and forwards and were not obliged to complete the assignment that was presented in each step of the instruction phase. A cartoon of the instructions and main visual elements is depicted in Figure 7.

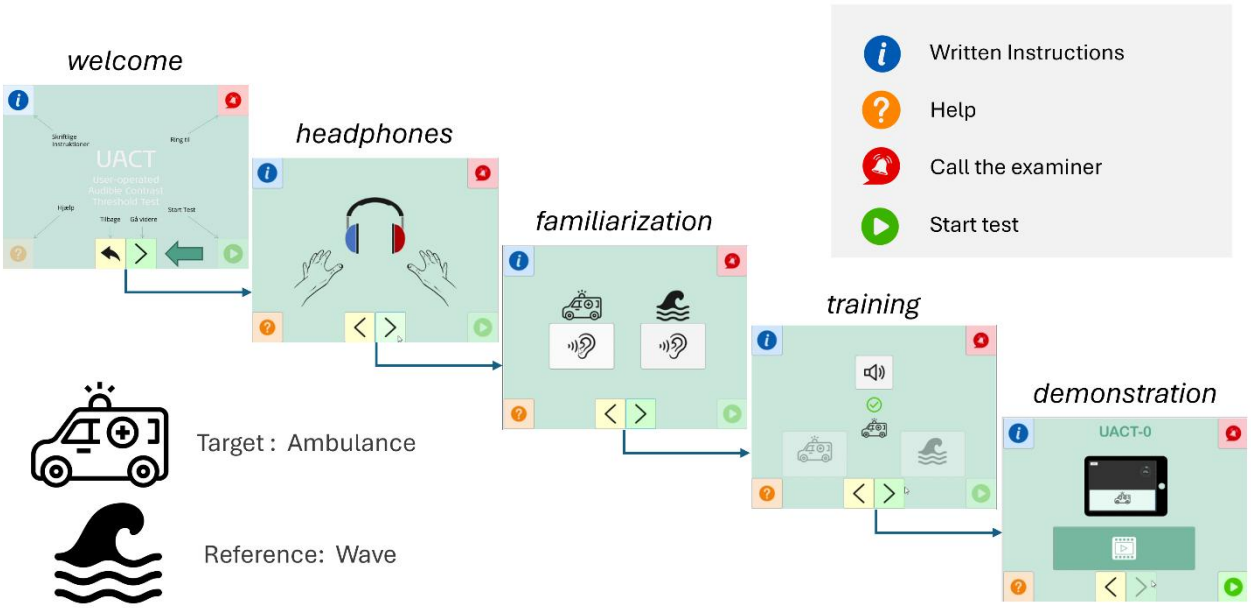

Figure 7: User workflow for the U-ACT test. The participant progresses from a welcome screen through headphone setup, sound familiarization, and training, before starting the test demonstration. The interface uses an "Ambulance" as the target-sound icon and a "Wave" as the reference icon, with on-screen help always available. The participant can make use of 3 types of help: extra information, written instructions, or call the examiner to obtain verbal instructions.

The instructions aimed at achieving four goals for the participant:

1. Correct headphone placement. The screen displayed an image of the headphones indicating which side corresponds to which ear.
2. Become familiar with the stimuli: The screen showed an ambulance and a wave, each of them with a button with an icon representing a loudspeaker.
3. Try out a detection test. The screen showed three buttons. The button with a speaker was enabled and played a sound when pressed. The participant was then required to press one of the other two buttons (ambulance or wave). After the button was pressed by the participant, a sign appeared on the screen indicating whether their choice was correct or incorrect.
4. Watch a demonstration. The participants should watch a short video showing how to perform the *task.*

The use of the instructions was investigated in *block* 1. All the events (i.e., next phase starts, go forward, go back, etc.), button presses, and response times were captured by the system.

Subjective evaluation

The evaluation was carried out using the touch screen. The data collection process involved the use of a visual analogue continuous scale (VAS), following each run in *block*s 2, 3, and 4. On the screen, a statement was written starting always with “The test was …” and below it, a continuous slider served to collect the judgments. To mimic a Likert scale, the slide had the three labels: “strongly disagree”, “neutral” and “strongly agree” and the result was a number between -50 and 50. The collected attributes included difficulty, self-perceived duration, demanding, boredom, as well as evaluations on the test's outcome such as understanding the instructions, usefulness, goodness of results, and overall impression. These attributes were assessed twice: once using positive statements like "The test was easy", and another time employing negative statements such as "The test was difficult".

Analyses

Statistical analysis of threshold differences

All statistical analyses were carried out using linear mixed-effects models and based on an analysis of variance (ANOVA) with subsequent post-hoc analyses of pairwise differences. The analyses were run in RStudio (Version 12; Posit team, 2024) using the packages `lmerTest` and `emmeans`. More details about the investigations can be found in the supplementary material.

For the investigation of the training effects and the effects of *task* and *procedure* on the thresholds, we made use of the difference between the U-ACT test candidate results and the benchmark ACT value of each individual participant (denoted ΔACT). This allowed a fairer comparison of the training effects. For this purpose, *block* 1 was disregarded since the purpose of that *block* was to evaluate the instructions and the participants did not have any previous experience with the test candidates at this point. The effect of the remaining repetitions was included in the factor *block.*

Statistical models included all the possible fixed effects and their interaction with the participant `group` while participant `ID` was included as a random effect. The models had the following form in R notation:

```
ΔACT ~ group*(task + procedure + block) + (1 | ID).
```

Estimated marginal means (EMMs) for the fixed factors were calculated with Satterthwaite's degrees-of-freedom approximation. Pairwise contrasts between task conditions were evaluated using Bonferroni-corrected t-tests to account for multiple comparisons. In case of significant interactions, an analysis of the subset of the data related to specific conditions or groups was performed.

Test-retest reliability metrics

To evaluate the test-retest performance of the test candidates and the agreement with the test benchmarks, the following metrics were considered:

- Intra-Class Correlation (ICC)[14,36]: This metric is commonly used in test-retest reliability studies. It measures consistency and agreement across different occasions, conditions or procedures, on the same subjects and it makes use of mixed models. For the measure of absolute test-retest reliability, we employed a 2-way random effect model (2 different measurements) with a single rater measurement (the measurement is not the average of various repetitions). In the manuscript we refer to this as ICC(2,1). For the comparisons between tests, we employed a 2-way random effects model, to quantify the absolute agreement, among multiple repetitions. In the manuscript we refer to this as ICC(2,k)[14,36].
- Mean difference and Coefficient of Repeatability (CoR)[15,17]: These are obtained from Bland & Altman plots used, for example, in the middle panel in Figure 2, where the mean difference represents the bias, and the CoR defines the limits within which 95% of the differences between paired measurements lie. Mathematically, $CoR = 1.96\sqrt{2 \ * \ \text{within subject standard deviation}}$.
- Fractional Rank Precision (FRP)[37]: This is a metric used to assess the precision of a method in ranking objects or samples and thus related to the mean average precision, which is a standard tool for information retrieval. The higher the FRP value, the better the precision in rankings.

Postprocessing and data modelling of subjective evaluation data

Prior to the analysis, data from the subjective evaluations underwent preprocessing to account for variations in the use of the VAS scale among participants and to focus on the differences between U-ACT test *candidate.* To standardize the ratings, each raw judgement $(rating_{i,b})$ was transformed using a participant-*block*-specific Z-score:

$$Zrating_{i,b} = \frac{rating_{i,b} - \mu_{i,b}}{\sigma_{i,b}},$$

where $\mu_{i,b}$ and $\sigma_{i,b}$ represent the mean and standard deviation, respectively, calculated across all rated attributes but separately for each participant i within each experimental *block* b..

Additionally, both types of wording for each attribute (positive and negative wording) were transformed so that they shared the same sign and represented by the factor *wording*. For example, difficulty remained unchanged while easiness was multiplied by -1. Subsequently, linear models were created for each attribute in the following format:

```
Attribute = candidate + wording
```

followed by pairwise contrasts between UACT test candidates using Bonferroni-corrected t-tests for comparison. Note that we use a linear model here with no random effects because of the standardization.

References

1. Haile, L. M. *et al.* Hearing loss prevalence and years lived with disability, 1990–2019: findings from the Global Burden of Disease Study 2019. *The Lancet* **397**, 996-1009, (2021).
2. Clark, J. L. & Swanepoel, D. W. The World Report on Hearing – a new era for global hearing care. *Int. J. Audiol.* **60**, 161–161 (2021).
3. Windmill, I. M. & Freeman, B. A. Demand for Audiology Services: 30-Yr Projections and Impact on Academic Programs. *J. Am. Acad. Audiol.* **24**, 407–416 (2013).
4. Sidiras, C. User-Operated Audiometry Project (UAud) – Introducing an Automated User-Operated System for Audiometric Testing Into Everyday Clinic Practice. *Front. Digit. Health* **3**, 7 (2021).

4. Zaar, J. / Simonsen, L. B., Sanchez-Lopez, R. & Laugesen, S. The Audible Contrast Threshold (ACT) test: A clinical spectro-temporal modulation detection test. *Hear. Res.* **453**, 109103 (2024).
6. Zaar, J., Simonsen, L. B. & Laugesen, S. A spectro-temporal modulation test for predicting speech reception in hearing-impaired listeners with hearing aids. *Hear. Res.* **443**, 108949 (2024).
7. Zaar, J., Simonsen, L. B., Dau, T. & Laugesen, S. Toward a clinically viable spectro-temporal modulation test for predicting supra-threshold speech reception in hearing-impaired listeners. *Hear. Res.* **427**, 108650 (2023).
8. Bernstein, J. G. W. *et al.* Spectrotemporal Modulation Sensitivity as a Predictor of Speech-Reception Performance in Noise With Hearing Aids. *Trends Hear.* **20**, (2016).
9. Margolis, R. H. & Morgan, D. E. Automated Pure-Tone Audiometry: An Analysis of Capacity, Need, and Benefit. *Am. J. Audiol.* **17**, 109–113 (2008).
10. Margolis, R. H., Glasberg, B. R., Creeke, S. & Moore, B. C. J. AMTAS® : Automated method for testing auditory sensitivity: Validation studies. *Int. J. Audiol.* **49**, 185–194 (2010).
11. Levitt, H. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* **49**, Suppl 2:467+ (1971).
12. Carhart, R. & Jerger, J. F. Preferred Method For Clinical Determination Of Pure-Tone Thresholds. *J. Speech Hear. Disord.* **24**, 330–345 (1959).
13. Ericsson, K. A. & Simon, H. A. How to Study Thinking in Everyday Life: Contrasting Think-Aloud Protocols With Descriptions and Explanations of Thinking. *Mind Cult. Act.* **5**, 178–186 (1998).
14. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **15**, 155–163 (2016).

15. Bartlett, J. W. & Frost, C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet. Gynecol.* **31**, 466–475 (2008).

16. Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Int. J. Nurs. Stud.* 6 (2010).

17. Giavarina, D. Understanding Bland Altman analysis. *Biochem. Medica* **25**, 141–151 (2015).

18. Dorr, M. *et al.* New Precision Metrics for Contrast Sensitivity Testing. *IEEE J. Biomed. Health Inform.* **22**, 919–925 (2018).

19. Jebb, A. T., Ng, V. & Tay, L. A Review of Key Likert Scale Development Advances: 1995–2019. *Front. Psychol.* **12**, (2021).

20. Rozin, P. & Royzman, E. B. Negativity Bias, Negativity Dominance, and Contagion. *Personal. Soc. Psychol. Rev.* **5**, 296–320 (2001).

21. Sørensen, C. B., Pedersen, C., Pedersen, E. R., Schmidt, J. H. & Nielsen, J. Observed issues and their prevalence when patients conduct user-operated audiometry: International Symposium on Auditory and Audiological Research. in (2023).

22. Aronoff, J. M. & Landsberger, D. M. The development of a modified spectral ripple test. *J. Acoust. Soc. Am.* **134**, EL217–EL222 (2013).

23. Landsberger, D. M., Dwyer, R. T., Stupak, N. & Gifford, R. H. Validating a Quick Spectral Modulation Detection Task. *Ear Hear.* **40**, 1478–1480 (2019).

24. Lelo De Larrea-Mancera, E. S. *et al.* Portable Automated Rapid Testing (PART) for auditory assessment: Validation in a young adult normal-hearing population. *J. Acoust. Soc. Am.* **148**, 1831–1851 (2020).

25. Larusdottir, M. K., Gulliksen, J. & Hallberg, N. RAMES – Framework supporting user centred evaluation in research and practice. *Behav. Inf. Technol.* **38**, 132–149 (2019).

26. Poling, G. L., Kunnel, T. J. & Dhar, S. Comparing the Accuracy and Speed of Manual and Tracking Methods of Measuring Hearing Thresholds: *Ear Hear.* **37**, e336–e340 (2016).

28. Mahomed, F., Swanepoel, D. W., Eikelboom, R. H. & Soer, M. Validity of Automated Threshold Audiometry: A Systematic Review and Meta-Analysis. *Ear Hear.* **34**, 745–752 (2013).

29. ISO/IEC 25022: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of quality in use (2016)

30. MIkhalevskaya, M. B. & FInkel', N. V. The Influence of Feedback on the Detectability of a Signal. *Sov. Psychol.* **25**, 50–60 (1986).

31. Hoglund, E. M. & Feth, L. L. Bayesian analysis of the impact of incomplete feedback on response bias. in 050006 (Sydney, Australia, 2023). doi:10.1121/2.0001851.

32. Suh, M. J. *et al.* Improving Accuracy and Reliability of Hearing Tests: An Exploration of International Standards. *J. Audiol. Otol.* **27**, 169–180 (2023).

33. Remus, J. J. & Collins, L. M. Comparison of adaptive psychometric procedures motivated by the Theory of Optimal Experiments: Simulated and experimental results. *J. Acoust. Soc. Am.* **123**, 315–326 (2008).

34. Brand, T. & Kollmeier, B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J. Acoust. Soc. Am.* **111**, 2801–2810 (2002).

35. Zabalbeascoa, P. The nature of the audiovisual text and its parameters. in *The Didactics of Audiovisual Translation* vol. 77 21–37 (John Benjamins Publishing, 2008).

36. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1**, 30–46 (1996).

37. Dorr, M. *et al.* New Precision Metrics for Contrast Sensitivity Testing. *IEEE J. Biomed. Health Inform.* **22**, 919–925 (2018).

38. Margolis, R. H., Saly, G. L., Le, C. & Laurence, J. Qualind: A method for assessing the accuracy of automated tests. *J. Am. Acad. Audiol.* **18**, 78–89 (2007).

Acknowledgements

This work was supported by Innovation Fund Denmark Grand Solutions 9090-00089B and the William Demant Foundation and carried out at the Technical University of Denmark. We thank T. Dau and R.S. Sørensen for their support, the participants for their insights, time and effort, and L.B. Simonsen for her contributions to the custom procedure. We want to acknowledge the discussion among the UAud project partners and several colleagues from Demant A/S and DTU who tried out earlier versions of the tests.

Author Contributions

RS-L: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization.

JZ: Conceptualization, Writing– Review & Editing, Investigation.

SL: Conceptualization, Writing – Review & Editing, Supervision, Project administration, Funding acquisition, Investigation, Methodology.

Artificial Intelligence statement

During the preparation of this work, the authors used Gemini (version 2.5 Pro) and Mistral (version Nemo Instruct) to enhance readability and correct grammatical and typographical errors. The authors reviewed and edited all AI-generated suggestions and take full responsibility for the final content of the manuscript.

Data availability

The data and code used in the analyses of this manuscript is available in the Zenodo repository https://doi.org/10.5281/zenodo.15730189. Most additional data can be made available on request.

Competing interests

The three authors are listed as co-inventors of the preliminary patent application EP4371480A1, and the patents EP4094685, EP4005474 which have been granted to Interacoustics A/S. Both patents are partially related to the test described in these studies. SL own stocks in Demant A/S, which owns Interacoustics A/S and Oticon A/S. The Audible Contrast Threshold test as described above is commercially available in instrumentation from Interacoustics A/S and other Demant A/S brands. The authors declare that the research was conducted in the absence of any other commercial or financial relationships that could be construed as a potential conflict of interest.