BUZZ: An Auditory Interface User Experience Scale

Brianna J. Tomlinson
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
btomlin@gatech.edu

Brittany E. Noah
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332, USA
brittany.noah@gatech.edu

Bruce N. Walker
School of Psychology & School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
bruce.walker@psych.gatech.edu

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada

ACM ISBN 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188659

Abstract
Auditory user interfaces (AUIs) have been developed to support data exploration, increase engagement with arts and entertainment, and provide an alternative to visual interfaces. Standard measures of usability such as the SUS [4] and UMUX [8] can help with comparing baseline usability and user experience (UX), but the overly general nature of the questions can be confusing to users and can present problems in interpretation of the measures when evaluating an AUI. We present an efficient and effective alternative: an 11-item Auditory Interface UX Scale (BUZZ), designed to evaluate interpretation, meaning, and enjoyment of an AUI.

Author Keywords
User experience (UX); UMUX; Usability Evaluation

ACM Classification Keywords
H.5.2. Information interfaces and presentation: User Interfaces – Evaluation/Methodology.

Introduction
Understanding how best to measure the usability of a display is an ongoing discussion. Numerous scales have been developed as tools to quickly compute usability ratings for systems and applications, including the SUS [4], the UMUX [8], and the UMUX-Lite [13]. Even after collecting this information, it can be hard to identify

UX Measure	Peripheral Display Metrics
Helpful	Distraction, effects of breakdowns (helpfulness for task completion)
Interesting	Appeal (usefulness)
Pleasant	Appeal (aesthetics)
Understandable	Awareness (accuracy of info recall)
Relatable	Learnability (recall previous info)

Table 1: A table which relates each of the main UX measurement keywords with their metrics from the peripheral displays (and how they relate).

pain points and have a clear interpretation of the scales' reflection of the overall UX. Bangor et al. found the addition of an 11th question to the SUS, asking for a verbal description of the user-friendliness of the system, provided a reasonable subjective label for interpreting the SUS [2].

Further discussion between Bosley [3], Cairns [5], and Finstad [10] about the overall reliability, validity, and sensitivity of the UMUX scale present examples of continuing concerns about usability scale design. Sauro [17] discusses additional challenges with measuring the UX of a system, and suggests the best methods support utilizing multiple task benchmarks, collecting perceived UX measures, and evaluating a system based on a combination of multiple measures.

Current scales can provide generalizable comparisons of system usability; but, using them to evaluate the design of multimodal or AUIs can create additional complexities. People have relatively low experience interacting with auditory displays, and more general usability scales may result in individual differences for item interpretation. The most widely-used AUIs, such as navigation systems, are primarily text-to-speech based, instead of utilizing non-speech audio. Non-speech audio presents a unique usability problem since people may have associated meanings for the sounds used. Being able to understand intricacies in comprehension of auditory display mappings, ease of use, and appeal would provide better information for iterative design and evaluation of these displays.

UX Evaluations for Auditory UIs

Auditory components of multi-modal UIs and standalone auditory experiences have a long history of UX evaluation. Pendse et al.'s work on the Accessible Aquarium explored which features users perceived through video-taped interviews and think aloud sessions [15]. Antle et al. evaluated the UX of Sound Maker, an embodied music learning tool, through subscales from an intrinsic motivation inventory and individual items about learning difficulty [1]. While many of these evaluations include collecting feedback on the qualitative UX, there remains a need to be able to evaluate unique aspects of auditory displays such as data-to-sound mappings, comprehension, and aesthetic aspects of the displays.

Audio UX Scale Design

We present a scale composed of 11 questions that can efficiently and effectively evaluate the salient aspects of an AUI. The first five items (see Sidebar on pg. 3) were inspired by Matthews et al.'s work with evaluation of peripheral displays, specifically from the importance of perception and content interpretation (see Table 1 in Sidebar on this page) [14]. Similarly, many AUIs rely on temporary representations of the information, presented in a sequential or concurrent manner, all depending on the users' task [18].

The other six scale items (again, refer to Sidebar on pg. 3) were chosen to elicit feedback on meaning and interpretation (items 6 and 7), enjoyment (items 8 and 9), and comprehension of the auditory mappings (items 10 and 11). Both positive and negative questions were included, following the work of Simms [19].

Scale Response Categories

Following Finstad's approach for the UMUX [8], we used Diefenbach et al.'s suggestion of a seven-point scale, which had the best representation of a person's

Full set of Audio UX Statements:

1. The sounds were helpful.
2. The sounds were interesting.
3. The sounds were pleasant.
4. The sounds were easy to understand.
5. The sounds were relatable to their ideas.
6. It was easy to match these sounds to their meanings.
7. It was difficult to understand how the sounds changed from one variable to the next.
8. It was fun to listen to these sounds.
9. It was boring to listen to these sounds.
10. It was confusing to listen to these sounds.
11. It was easy to understand what each of the sounds represented.

opinions, the highest ease of use, and most accuracy compared other scales [7]. Dawes [6] and Preston and Coleman [16] found that both five- and seven-point scales provide high reliability and validity, while the seven-point scale was preferred for ease of use and discriminating power. Finstad also found that people are less likely to interpolate on a seven-point scale compared to five-point Likert items [9]. With these factors in mind, we chose to parallel the UMUX's statement structure and anchors for this scale.

Methods

Participants

A total of 52 participants (20 females) with an average age of 20.1 (SD=1.7) from a large research university in the southeastern United States participated.

Materials

Audio stimuli were presented using Sony MDR-7506 Studio headphones. Participant responses were collected in a computer lab, with each student working at their own pace. A previously-designed sonification model of the solar system [20] was used as the referent auditory display for all trials. To understand the effect of prior knowledge on perceived ease of use, participants reported their last astronomy class and any other relevant activities (e.g., planetarium visits).

Procedure

Participants listened through a two-part (audio-only) sonification presenting information about the *solar system* and *planets*. The *solar system* perspective conveyed details related to scaling and size (e.g., mass or distance from the sun) and the *planetary* perspective conveyed details specific to 'experiencing' each planet (e.g., mean temperature, rings, and composition). The sonification was carefully designed to scaffold non-speech audio-only comparisons, and used verbal descriptions to highlight details or concepts. Details were introduced in short chunks, and were grouped by topic; some pairs were played together to support easier comparison between features.

Participants answered task-specific questions for each section during or after listening to the auditory display. The *planetary* perspective contained more complex questions requiring some transfer of information learned, compared to the *solar system* questions which directly paralleled the content of the display. After each section, users answered the 11 questions (see Sidebar) based on the prompt, 'For the sounds in the previous section representing [set of concepts], please rate how much you agree or disagree with the following statements.' Then they completed the four UMUX questions following the prompt, 'Thinking about the sounds you listened to for the [section number], please rate how much you think the sounds could help you compare one planet to another.'

Results

Prior to analysis, negatively-worded items were converted to the same scale as positively-worded items (1-7). Usability scales for each perspective (*solar system* and *planetary*) were analyzed individually, to explore if they both contain the same components, or if they might have different components based on the characteristics of the listening task.

Solar System Perspective

PRINCIPAL FACTOR ANALYSIS (PFA)

PFA using Varimax rotation with Kaiser Normalization was used for dimension reduction, resulting in two

Factor	Alpha
Enjoyment and Appeal	0.88
Ease of Use	0.85
Overall	0.88

Table 2. Reliability summary table for the Solar System Perspective.

Factor	Alpha
Enjoyment and Appeal	0.91
Ease of Use	0.86
Understanding	0.69
Overall	0.83

Table 3. Reliability summary table for the Planetary Perspective.

factors for the solar system survey: one factor contains items related to enjoyment and appeal (items 1-3, 8, 9) and the second factor contains items related to ease of use (items 4-7, 10, 11).

RELIABILITY

Cronbach's alpha served as a measure of internal consistency or reliability for the items within those two factors. Reliability of 0.88 was found for the items in factor one (enjoyment and appeal) and 0.85 for items in component two (ease of use), with an overall Cronbach's alpha of 0.88 for all items. Table 2 summarizes the reliability levels. A value of 0.7 is generally considered acceptable [11,12] (note: the UMUX and SUS both have reliability over 0.9 [8]).

CORRELATION WITH UMUX

Correlating this audio UX survey with the UMUX (an already validated and highly used scale) can provide some evidence for validity. The correlation between the UX questions and the UMUX was $r=0.68, p < .001$. Therefore, the data from the UMUX and BUZZ are highly positively related.

Planetary Perspective

PRINCIPAL FACTOR ANALYSIS (PFA)

The PFA, from the planetary perspective sonifications, using Varimax rotation with Kaiser Normalization resulted in three factors: one factor compiled items relating to enjoyment and appeal (2, 3, 8, 9); the second factor combined items relating to ease of use (1, 4, 7, 10, 11); and, the third factor included two items relating to understanding (5 and 6). The differences between the number of factors may be a result of task difficulty or may result from variation in the types of data (i.e., the details) presented between the *solar system* and *planetary* perspectives. One possibility may be that the simplicity of the interpretation tasks for the *solar system* view did not lead to a factor impact between ease of use and relatability of the display mappings. Further exploration is necessary to better understand the differences between the *solar system* and *planetary* displays.

RELIABILITY

Cronbach's alpha (reliability) for the items in the first factor (enjoyment and appeal) was 0.91. Factors 2 and 3 had Cronbach's alpha values of 0.86 and 0.69, respectively. The overall reliability statistic for the entire set of items was 0.83; all values are summarized in Table 3.

CORRELATION WITH UMUX

The correlation between the UX survey on the second task with the UMUX was $r = .74, p < .001$, again presenting a strong, positive relationship between BUZZ and UMUX.

Discussion

This work introduces BUZZ, an auditory-specific usability scale, that has been shown to be highly correlated with the existing, widely used UMUX scale. These scales expand the ability to measure usability specifically for AUIs, a display type which may benefit from its own measurement tools. BUZZ may lead to insights for design of auditory displays that may not be captured with traditional usability measures like UMUX and SUS. For example, if there were low ratings on one of the subscales, that could indicate that the user is not understanding or does not like the experience of listening to the auditory displays. If understanding

Using the BUZZ Scale
This scale should be utilized for assessing the usability of auditory displays. It should be used in its entirety for now, as it is uncertain whether the factors, and therefore the sub scales, are stable with differing listening tasks. We recommend that a PFA be conducted when using this scale prior to calculating subscale scores, as it has not been widely validated.

To calculate the total score, recode the negative items to the same scale as positive ones by subtracting the current number from eight. Then sum the total score for all items. The maximum score is therefore 77. The total score can be used as a measure of usability and for comparison between systems in usability testing.

When taken together, the whole set is intended to provide an interpretation about the overall effectiveness and efficiency of an auditory display.

items rate lower, that could indicate a need to re-evaluate the mapping of sound to data in the display.

Limitations
This is a small-scale validation study, which used complex AUIs, and further studies should be conducted to ensure broader generalizability. These studies should include varied context for the displays and different display types (e.g., auditory icons, earcons, or other shorter AUIs [21]). Randomization of scale items was not done during this study, and further exploration of possible order effects could be completed.

Future Work
Further evaluation and validation of this scale for measuring auditory UX is needed before using it to complete formal evaluations. Replicating this study may help provide internal validity (to its ability to correctly measure the enjoyment, ease of use, and relatability for an auditory display). Using these same scales in other comprehension and interpretation tasks could help provide additional external validity.

Additional use of the SUS as a secondary validation measure could help provide a more thorough understanding of the external validity of these scales. Both SUS and UMUX are highly-used for UX evaluation; but, due to their general nature, using them for evaluation of AUIs provides complications in their interpretation and raises questions about a person's ability to accurately understand these UIs. Developing a usability scale for AUIs may provide a more-direct and insightful way to understand these experiences.

Acknowledgements
We thank Emily McDonald and Fernbank Science Center for their help and support on this research. Portions of this work were supported by funding from the National Science Foundation (NSF) and from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR).

References
1. Alissa N. Antle, Milena Droumeva, and Greg Corness. 2008. Playing with the sound maker: do embodied metaphors help children learn? *Proceedings of the 7th international conference on Interaction design and children - IDC '08*: 178–185.

2. Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3: 114–123. https://doi.org/66.39.39.113

3. John J. Bosley. 2013. Creating a short usability metric for user experience (UMUX) Scale. *Interacting with Computers* 25, 4: 317–319.

4. John Brooke. 1996. SUS-A quick and dirty usability scale. In *Usability Evaluation in Industry*, Ian L. Jordan, Patrick W.; Thomas, Bruce; Weerdmeester, Bernard A.; McClelland (ed.). Taylor & Francis, 189–194.

5. Paul Cairns. 2013. A commentary on short questionnaires for assessing usability. *Interacting with Computers* 25, 4: 312–316.

6. John Dawes. 2008. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research* 50, 1: 61–77.

7. Michael A. Diefenbach, Neil D. Weinstein, and Joseph O'reilly. 1993. Scales for assessing perceptions of health hazard susceptibility. *Health Education Research* 8, 2: 181–192. https://doi.org/10.1093/her/8.2.181

8. Kraig Finstad. 2010. The usability metric for user experience. *Interacting with Computers* 22, 5: 323-327.

9. Kraig Finstad. 2010. Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies* 5, 3: 104–110.

10. Kraig Finstad. 2013. Response to commentaries on "The usability metric for user experience." *Interacting with Computers* 25, 4: 327–330.

11. Thomas K. Landauer. 1997. Chapter 9 - Behavioral Research Methods in Human-Computer Interaction. *Handbook of Human-Computer Interaction (Second Edition)*: 203–227.

12. James R. Lewis. 2013. Critical review of "the usability metric for user experience." *Interacting with Computers* 25, 4: 320–324.

13. James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2099.

14. Tara Matthews, Tye Rattenbury, and Scott Carter. 2007. Defining, designing, and evaluating peripheral displays: An analysis using activity theory. *Human-Computer Interaction* 22, 1: 221–261.

15. Anandi Pendse, Michael Pate, and Bruce N. Walker. 2008. The accessible aquarium: identifying and evaluating salient creature features for sonification. *Proceedings of the 10th International ACM Conference on Computers and Accessibility (ASSETS '08)*: 297–298. https://doi.org/10.1002/acp.1291.

16. Carolyn C. Preston and Andrew M. Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104, 1: 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

17. Jeff Sauro. 2016. The challenges and opportunities of measuring the user experience. *Journal of Usability Studies* 12, 1: 1–7.

18. Jonathan H. Schuett, Riley J. Winton, Jared M. Batterman, and Bruce N. Walker. 2014. Auditory weather reports: demonstrating listener comprehension of five concurrent variables. 1–7.

19. Leonard J. Simms. 2008. Classical and Modern Methods of Psychological Scale Construction. *Social and Personality Psychology Compass* 2, 1: 414–433.

20. Brianna J. Tomlinson, R. Michael Winters, Christopher Latina, Smruthi Bhat, Milap Rane, and Bruce N. Walker. 2017. Solar System Sonification: Exploring Earth and its Neighbors Through Sound. In *Proceedings of the 23rd International Conference on Auditory Display (ICAD 2017)*.

21. Bruce N. Walker and Michael A. Nees. 2011. Theory of Sonification. In *The Sonification Handbook*, Thomas Hermann, Andy Hunt and John G Neuhoff (eds.). Logos Verlag, Berlin, 9–39.