Evaluating the Effectiveness of Sonification-Enhanced Auditory-Based Learning

Yeana Lee Bond, Gabrielle Ferro, Utkarsha Mohan, Marc Nassar

ABSTRACT

The impact of auditory sonification in academic audiobooks on reading comprehension is largely unexplored. By employing pretest-posttest measures and NASA-TLX surveys, the effectiveness of serial and parallel audio highlights across diverse academic topics can be investigated. While topic-specific tests revealed evidence of learning, sonification exhibited no significant effect on short- or long-term information retention. However, serial sonification showcased potential in improving secondary task performance in comparison to parallel. Understanding participant engagement and distractions with auditory highlights proved intricate due to varied responses in exit surveys. Furthermore, potential benefits of user-enabled earcon insertion were highlighted, fostering social interactions among audiobook users. Findings suggest potential demand for sonification in audiobooks, signaling the need for further exploration to enhance user experience and augment learning outcomes. This research offers valuable insights into sonification of audiobooks on learning, advocating for deeper investigation into its application and effectiveness.

INTRODUCTION

Journal articles can be read out loud through a variety of text-to-speech software currently available (Adobe Acrobat, NaturalReader, Microsoft Word, etc.). However, these options often have problems within their software and user interface that could negatively affect reading comprehension. Audiobooks, a modern tool to aid reading comprehension, have become increasingly popular, and over half of adults reported that they are using them in 2022 [1]. This increase in audiobook popularity has shed light on the multitude of benefits associated with auditory-based literature, including the ability to multitask, accommodations for visually impaired and neurodivergent individuals, promoting imagination, improving attitudes toward reading, and more [2]. Some literature suggests that audiobook comprehension is comparable to that of e-text, but there is a paucity of studies looking at consuming scientific literature in this modality [3]. Perhaps, the shortcomings associated with text-to-speech software reduce reading comprehension. However, sonification (the use of auditory cues), which has been used to aid comprehension in a variety of scenarios and educational materials, may be an answer to this problem [5].

The purpose of this study is to evaluate the effectiveness of an audiobook application designed for scholarly articles, with the capacity to cue listeners and alert them of important information. Given that sonification in the background of an audiobook narration facilitates improved comprehension of the text compared to speech-only audiobooks, it is possible that sonification could help students consume scholarly literature more effectively [4]. One issue is that audiobook listeners are often performing other tasks simultaneously. Therefore, the effect of a dual-task on the comprehension of an audiobook with sonification will be investigated as well. Lastly, there are several ways to implement sonification: serial and parallel audio. For Serial Sonification, a given audiobook is interrupted by an audio cue, usually on a natural pause in the literature, while in Parallel Sonification, audio cues play underneath a given audiobook to highlight an important section (see Figure 1). The primary hypothesis is that listeners will have better comprehension when sonification is present.

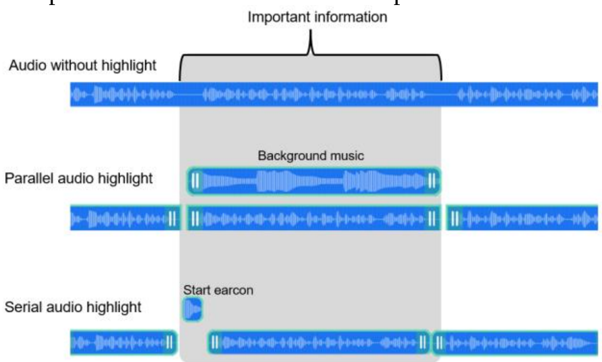


Figure 1: Examples of the audio tracks present in each interface: unedited speech (top), parallel audio highlighting (middle), and serial audio highlighting (bottom). The gray area shows the highlighted speech.

METHODS

Participants

Fourteen participants, 6 female and 8 male adults, with a mean (SD) age of 30 (11.5) years, comprised a convenience sample from the university population. Inclusion criteria consisted of a self-reported secondary education and minimal familiarity with the topics presented in the study. To ensure that all participants match these criteria, a brief screening quiz was used to gauge their familiarity with the following subjects: computer science, disease ecology, and Indian cuisine. All participants reported minimal prior knowledge of the topics. Four participants had an undergraduate education and the remaining ten reported a graduate education. Exclusion criteria included any self-reported hearing impairment and color blindness. No participants were excluded. Lastly, regarding our experiment environment, we conducted seven sessions virtually and seven sessions in person.

Apparatus and Equipment

Participants engaged in the “Lipuzz Water Sort” online game to simulate concurrent tasks using either personal computers or mobile devices [6]. Audiobook content was experienced through noise-canceling headphones or speakers. Text samples for each topic were generated by ChatGPT 3.5 [7], fact-checked, corrected, and transformed into audiobooks via Microsoft Azure AI Speech Studio [8]. The AI voices used were female and had a youthful tone. Three distinct natural speech models were employed: Microsoft Sonia (English UK) for Topic A, Microsoft Jane (English US) for Topic B, and Microsoft Jenny (English US) for Topic C. Audio highlights for the parallel mode were extracted in varying lengths from tracks such as ‘Haggstrom,’ ‘Subwoofer Lullaby,’ and ‘Sweden’ by C418 [9, 10, 11], ‘Indian Summer’ by Jai Wolf [12], and ‘Clarity (Orchestral Version)’ by Zedd [13]. For serial audio highlight, a 2-second-long sample from ‘Haggstrom’ by C418 was used [9].

Experimental Design

Repeated pretest-posttest measures under between-subject design within a single 45-minute session were employed. The independent variable was the sonification condition, which had three levels: no sonification, serial highlights, and parallel highlights. The dependent variable was the performance score on a multiple-choice test that assessed the participants’ comprehension of three topics: Topic A, Food and Culture of Varanasi; Topic B, Computer Vision; and Topic C, Rocky Mountain Spotted Fever. Utilizing a Paired Latin square design, the presentation order of the three sonification conditions and the three topics was randomized across fourteen participants and between topics. This method ensured a balanced distribution and minimized order effects, such as fatigue, learning, or primacy/recency, as well as the effects of individual differences and carryover.

Experimental Measures

Objective Measures

The first objective measure was a pretest assessment of initial knowledge Before the interventions, a pretest was conducted to evaluate participants' baseline understanding of the topics. This served to ensure their initial knowledge and confirmed the unfamiliarity with the subjects. This contained three multiple-choice questions and two fill-in-the-blank questions per topic, for a total of 15 questions. The second objective measure was a topic specific test, where participants underwent tests at the conclusion of each intervention. These tests comprised three multiple-choice questions and two fill-in-the-blank questions per topic, aimed at evaluating both their recognition and recall of each subject.

The third objective measure was an overall comprehension test post-interventions. Upon completion of all interventions and a 5-minute rest, participants took an overall comprehension test. This test had the same questions as the pretest. The order of appearance of the questions and the order of options in multiple-choice questions were randomized. All tests were conducted using Google Forms, and the order of appearance of questions were randomized.

Subjective Measures

The first subjective measure was the NASA-Task Load Index (TLX) Survey [14]. Participants were administered NASA-TLX [14] after being subjected to each condition. The second subjective measure was a preference survey after completion of the post-test. Participants were provided a comprehensive explanation of the experimental setup, detailing the auditory highlights and addressing any inquiries before administering the survey. The questionnaire aimed to gather subjective insights into participants' preferences concerning types of highlighting, evaluation of audio features, perceived effectiveness of auditory highlighting, emotional responses to earcons [15], and specific preferences regarding auditory highlighters. Additionally, participants were encouraged to provide feedback on their experience with the auditory highlights within the exit survey.

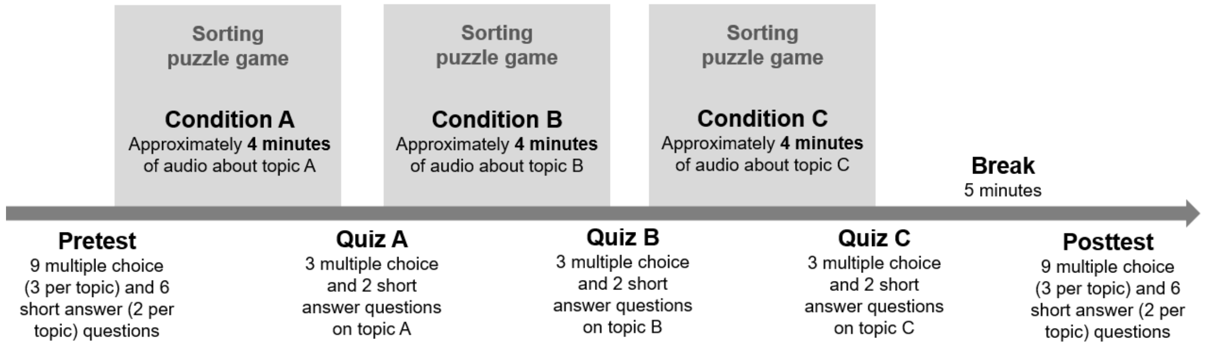

Sorting puzzle game
Condition A
Approximately 4 minutes of audio about topic A
Sorting puzzle game
Condition B
Approximately 4 minutes of audio about topic B
Sorting puzzle game
Condition C
Approximately 4 minutes of audio about topic C
Break
5 minutes
Pretest
9 multiple choice (3 per topic) and 6 short answer (2 per topic) questions
Quiz A
3 multiple choice and 2 short answer questions on topic A
Quiz B
3 multiple choice and 2 short answer questions on topic B
Quiz C
3 multiple choice and 2 short answer questions on topic C
Posttest
9 multiple choice (3 per topic) and 6 short answer (2 per topic) questions

Figure 2: Diagram of experiment procedure

Procedure

A pretest was used to eliminate participants who were too familiar with the material, as well as establish a baseline familiarity level of the topics for those who passed screening. Each recording covered a different topic and had a different sonification condition. After each recording, participants took a short quiz on the most recent topic presented and a mental workload survey using NASA-TLX for each audio they experienced [14]. After completing all recordings, participants were asked to take a 5-minute break and then complete a post-test on all topics presented. Following the test, participants completed a survey on the subjective effectiveness of the interventions. Likert scales [16] were used to gather opinions from participants regarding their experiences.

Statistical analysis

Several statistical analyses were performed. Test scores were used in a three-way, repeated measures ANOVA to investigate the effects of audio condition (none, parallel, or serial), test (pre, mid, or post tests), topic (A, B, or C), and all interactions. No three-way interactions were significant. Level reached in the game and NASA-TLX ratings were used in separate two-way, repeated measures ANOVA to investigate the effects of audio condition and topic. All the necessary pairwise comparisons were done using Student's T-test. All statistical analyses were performed using JMP Pro 16 with a significance level of 0.05. All summary statistics are reported as means (SD).

RESULTS

The mean (SD) values of all the reading comprehension tests are displayed in Table 1. For test scores, main effects were found for topics ($p < 0.001$) and different tests ($p = 0.006$), but no main effect was found for audio conditions ($p = 0.41$). No interaction effects were found between topics, audio conditions, or tests. Across all conditions and topics, the topic specific test, or mid-test, scores were significantly higher than pretest scores ($p = 0.0016$). No statistical differences were found between pretest and post-test scores or mid-test and post-test scores. However, the mean post-test score of 36.7% +/- 30.3% was higher than the mean pretest score of 27.1% +/- 20.2% and lower than the mean mid-test score of 43.8% +/- 23.9%. There were no differences in test scores found between different audio highlighting conditions. However, the mean test scores for the mid and post tests were higher in our control, no audio highlight conditions than in the parallel or serial conditions.

	Control	Parallel	Serial
Comprehension measures			
Pretest Score (%)	27.1 (20.2)	27.1 (18.6)	27.1 (23.0)
Midtest Score (%)	50.0 (21.8)	40.0 (26.0)	41.4 (24.1)
Posttest Score (%)	42.9 (32.2)	28.6 (26.8)	38.6 (31.8)
Secondary task performance			
Max Level Reached	8.4 (1.4)	7.8 (1.5)	9.0 (1.4)
NASA TLX			
Mental	4.1 (2.0)	5.1 (1.5)	4.9 (1.7)
Physical	3.1 (1.9)	3.7 (2.4)	3.4 (2.0)
Temporal	3.8 (1.0)	4.1 (1.2)	4.1 (1.4)
Performance	2.9 (1.5)	2.5 (1.1)	2.9 (1.2)
Effort	4.1 (1.8)	4.9 (1.4)	4.6 (1.8)
Frustration	3.6 (2.0)	4.1 (1.5)	4.0 (2.1)
	Topic A	**Topic B**	**Topic C**
Comprehension measures			
Pretest Score (%)	31.4 (23.2)	20.0 (13.6)	30.0 (21.8)
Midtest Score (%)	41.4 (18.3)	30.0 (21.8)	60.0 (22.2)
Posttest Score (%)	41.4 (25.4)	14.3 (16.5)	54.3 (32.7)
Secondary task performance			
Max Level Reached	8.9 (1.3)	8.3 (1.5)	7.9 (1.6)
NASA TLX			
Mental	4.2 (1.6)	6.0 (1.4)	3.9 (1.5)
Physical	3.2 (2.3)	3.8 (2.3)	3.2 (1.8)
Temporal	4.0 (1.1)	4.6 (1.3)	3.5 (0.9)
Performance	3.0 (0.9)	1.8 (0.7)	3.6 (1.4)
Effort	4.4 (1.7)	5.1 (1.4)	4.0 (1.7)
Frustration	3.2 (2.0)	5.1 (1.8)	3.5 (1.4)

Table 1: Mean (SD) performance in primary and secondary tasks and mental workload during auditory conditions and topics, A, B, and C

There was significant effect of the topic on participant's test performance. Topic B test scores were parallel conditions significantly lower than topic A scores (p = 0.012) and topic C scores (p < 0.001) when compared across all tests and between tests. No differences were found between topic A and C. The distribution in number of correct responses was unequal across the topics, as shown in Figure 4.

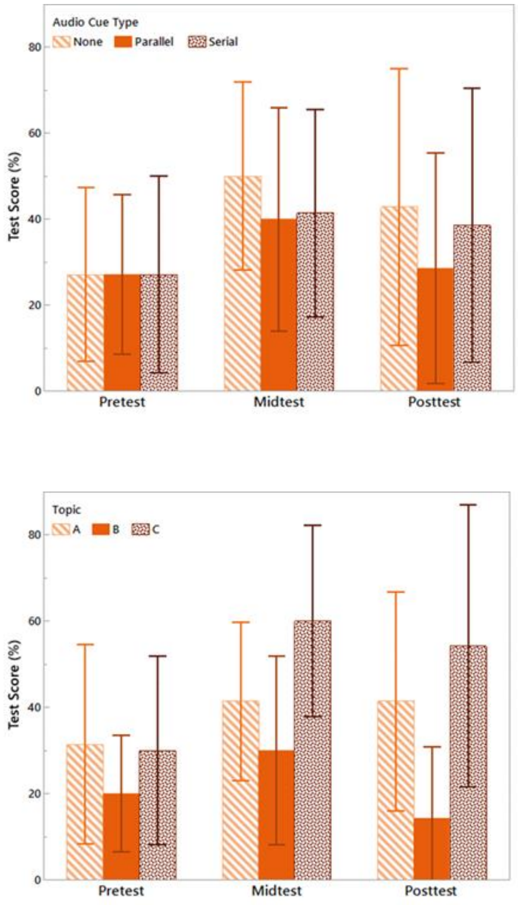


Figure 3: Test scores for the pretest, topic specific test (mid-test), and post-test. The graph on the top shows the test scores by which audio highlighting condition the information was presented in. The graph on the bottom shows the test scores by topic.

Three questions related to topic B in the pre- and post- tests were not answered correctly by a single participant. However, all topic A and C questions were answered correctly at least once on the post-test. Participant performance in the secondary task was assessed by the highest level reached in the sorting puzzle game. There was a main effect of auditory condition (p = 0.032), but no main effect of topic (p = 0.069) on the highest level reached. Across all topics, participants performed significantly better during serial auditory highlights than parallel ones (p = 0.009). The mean values of secondary task performance are shown in Table 1.

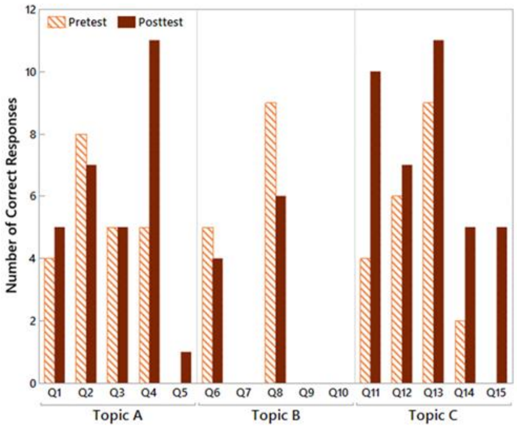


Figure 4: Total number of correct responses to each question on the pre and post tests.

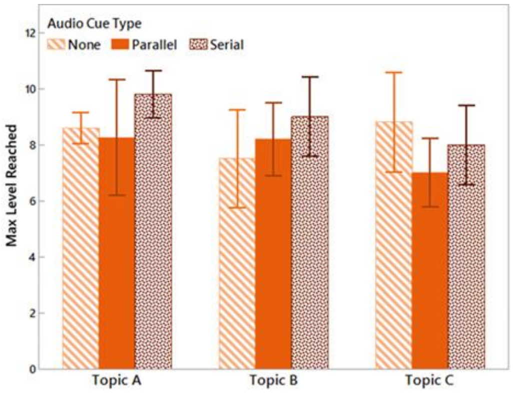


Figure 5: The highest level reached during each audio condition.

Participants' self-reported mental workload was collected with a NASA-TLX survey after each condition [14]. Main effects were found by topic but not audio highlight conditions for the following components: mental (p = 0.001), performance (p < 0.001), and frustration (p = 0.021). No main effects were found for the other components (physical (p = 0.74), temporal (p = 0.078), and effort (p = 0.23)). Participants found topic B was significantly more mentally demanding than topics A (p =0.004) and C (p < 0.001). Participants reported a significantly lower perceived performance after topic B than topic A (p = 0.003) and C (p < 0.001), as well. Additionally, participants reported having significantly more frustration during topic B conditions than topic A (p = 0.029) and topic C (p = 0.009).

Based on the exit survey, serial audio highlight was preferred by half of the participants. The second most preferred type was the parallel type, and the three participants said that they liked neither of them (See Appendix A.6). There were six participants that expressed indifference or gave negative responses on at least one of the audio highlights. Half of them said that they felt that at least one of the audio highlights was distracting or disrupting. There were eight positive responses, of which three expressed that the serial audio highlights helped them retain key information. Six out of Seven participants reported feeling distracted or disrupted, and 5 participants found at least one audio highlight helpful. Remaining participants voiced concerns regarding technical language and volume issues.

Participants were asked what emotion(s) they felt while listening to cues. Five participants mentioned feeling 'Neutral' or having 'No Feelings,' while another five expressed a sense of 'urgency.' Four participants reported experiencing 'stress,' while the emotions 'surprise' and 'confusion' received three and two mentions respectively. Additionally, there was one mention each for 'boredom,' 'irritation,' 'depression,' 'annoyance,' 'anger,' 'fear,' 'violence,' and 'relief.' (See Appendix A.5). Lastly, eleven participants reported that they would recommend the comprehension tools they experienced to others.

DISCUSSION

The potential effectiveness of an audiobook application system equipped with auditory sonification using earcons to highlight key information was explored. While the topic specific test scores showed evidence of learning, sonification tools had no effect on retention of presented material. However, the secondary task performance was improved in serial sonification compared to parallel. One possible confounding issue is that the IVs did not consider the reduction in cognitive resources caused by stimuli sharing sensory channels. According to Wickens' Multiple Resource Theory (MRT), it is plausible that serial sonification was preferred because it did not conflict with the information being highlighted, whereas parallel sonification may have reduced listeners' ability to retain new information [17].

Topic B yielded the poorest mean scores across all pre, mid, and post-tests. Additionally, significant differences in mental workload, performance, and frustration were identified using the NASA-TLX Likert 7-Point Scale [14]. This indicates that Topic B was the most challenging.

Half of the participants did not feel engaged between auditory highlights and spoken text. Interestingly, while participants didn't strongly perceive the auditory highlights as distracting, the qualitative feedback revealed that seven participants found at least one type of auditory highlight distracting. This pilot study encountered challenges in distinguishing whether these auditory cues were ultimately disruptive or merely perceived as such due to mixed responses in the exit survey.

Temporal pressure, notably the prevalent sense of urgency, as reported by five participants, might have negatively impacted participants' ability to retain information. Other studies have shown that in the presence of time constraints, performance impairments are often caused by media multitasking [18]. In reality, users rarely suffer from time constraints, and rather consume audiobooks leisurely. Historically, audiobooks have been regarded merely as "talking books", as opposed to viable substitutes for visually reading texts [19]. Listening to audiobooks has typically been seen as a secondary activity while engaging in other tasks, as users seldom focus solely on audiobooks as a primary task.

Some participants highlighted the potential benefits of enabling users to insert earcons at preferred points in audiobooks. This feature not only allows users to annotate information audibly but also creates auditory bookmarks at desired locations. Additionally, the application could accumulate and share these recorded bookmarks among users, fostering social interactions akin to traditional book clubs. Our proposal's novelty lies in initially inserting earcons for enhanced listening-based learning, with potential expansion into social computing aspects.

Despite seven negative responses regarding the effectiveness of audio highlights in retaining one-time heard information, the exit survey yielded eleven positive responses about the potential use of sonification in audiobooks. This dichotomy underscores the potential demand for sonification among audiobook users and suggests that auditory highlights could significantly enhance user experience with audiobooks, warranting further research and exploration.

LIMITATIONS

The participant pool lacked representation from frequent audiobook users, notably excluding visually impaired learners who heavily rely on audio-based learning. This omission might have affected the understanding of how auditory enhancements and narration-rhythm interact, particularly for non-audiobook users. Moreover, the interpretation of auditory cues and spoken content by participants who speak English as a second language may have been impaired in comparison to native English speakers. Accounting for individual participant preferences regarding audiobook format, pace, and style of narration might have enriched the study's insights into engagement and perception of auditory cues. The inconsistency in volume balance between audio tracks—narration and highlights—could have affected participants' experiences [20], [21].

Variability in study settings was notable, spanning virtual and in-person environments, with participants using different audio devices and employing various methods for secondary tasks. This variability may have influenced outcomes. Additionally, fixed variables restricted customization, limiting participant choices in topic selection, AI voice type, and placement of highlights. The single playthrough of topics might not have provided a fair assessment of learning, impacting perceived difficulty levels. The limited exposure to each topic might not align with real-world audiobook engagement, potentially overlooking prolonged impacts on user engagement, comprehension, and retention. Longer studies with a larger participant cohort and topic pool may uncover more extensive learning outcomes.

FUTURE WORK

Future research aims to enhance audiobook applications for optimized learning experiences. This includes incorporating repetition features and ensuring consistent volume levels across various audio outputs for equitable sonification. Evaluating the impact of native language audiobooks with sonifications compared to English versions on comprehension levels also stands as a promising avenue for exploration. Standardizing assessment methods, particularly in audio highlight placements and language considerations, can enrich user experiences. Offering customization options like topic selection, narration speed adjustments, and highlight positioning can boost user engagement. Exploring user-driven variables like preferred narrator models or pace of playback as independent factors will unveil preferences and learning outcomes. These efforts seek to revolutionize audiobook applications, making them more engaging and effective for diverse users, catering to individual learning needs.

CONCLUSION

According to the exit survey, 11 participants reported that they believed future users would enjoy learning from sonification-enhanced audiobooks. An important lesson from user feedback is that the user experience of these tools would be improved if they are granted the ability to customize audio cues and narration voices. This indicates that the demand for sonification learning tools may increase in the future. The field of sonification remains largely unexplored with regards to enhancing scholarly literature. These ideas can be applied to future research or to audiobook system design guidelines for better user experience(s).

REFERENCES

[1] APA, “Research Surveys Press Release,” APA Sales and Consumer Data, https://www.audiopub.org/surveys (accessed Sep. 21, 2023).

[2] A. Miranda-Cueva and M. Cabanillas-Carbonell, “Benefits of using an audiobook application as an educational entertainment tool for children: A review of the scientific literature in the years 2006-2019,” *2020 IEEE Congreso Bienal de Argentina (ARGENCON)*, 2020. doi:10.1109/argencon49523.2020.9505385

[3] B. A. Rogowsky, B. M. Calhoun, and P. Tallal, “Does modality matter? the effects of reading, listening, and dual modality on comprehension,” *SAGE Open*, vol. 6, no. 3, p. 215824401666955, 2016. doi:10.1177/2158244016669550

[4] G. Kramer et al., “Sonification Report: Status of the field and Research Agenda,” DigitalCommons@University of Nebraska - Lincoln, https://digitalcommons.unl.edu/psychfacpub/444/ (accessed Sep. 21, 2023).

[5] Walker, B. N., & Nees, M. A. (2011). Theory of sonification. The sonification handbook, 1, 9-39.

[6] Coolmath Games. (n.d.). Lipuzz: Water sort [Video game]. Retrieved from Coolmathgames.

[7] OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]

[8] Microsoft. (n.d.). Speech Studio [Website]. Retrieved from Azure AI

[9] C418. (2013, April 6). Minecraft Volume Alpha - 7 - Haggstrom [Video]. YouTube. https://www.youtube.com/watch?v=3-922W32n4k

[10] C418. (2011). Subwoofer lullaby [Song]. On Minecraft - Volume Alpha [Album]. C418.YouTube. https://www.youtube.com/watch?v=o3_InDEtpLA

[11] C418. (2011). Sweden [Song]. On Minecraft - Volume Alpha [Album]. C418.

[12] Wolf, J. (2015). “Indian summer” [Recorded by Jai Wolf]. On Indian summer [YouTube video]. Retrieved from Jai Wolf - Indian Summer (Official Music Video)

[13] Zedd. (2012). “Clarity (Orchestral Version)” [Recorded by Zedd]. On Clarity [CD]. Interscope Records.

[14] Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, 904-908. Santa Monica: HFES. https://humansystems.arc.nasa.gov/groups/tlx/

[15] Meera M Blattner, Denise A Sumikawa, and Robert M Greenberg. 1989. Earcons and icons: Their

structure and common design principles. Human–Computer Interaction 4, 1 (1989), 11–44.

[16] Bertram, D. (2007). Likert scales. Retrieved November, 2(10), 1-10.

[17] Wickens, C. D. (2002). Multiple resources and performance prediction. Theoretical issues in ergonomics science, 3(2), 159-177.

[18] Aagaard, J. (2019). Multitasking as distraction: A conceptual analysis of media multitasking research. Theory & Psychology, 29(1), 87–99.

[19] Singh, A., & Alexander, P. A. (2022). Audiobooks, print, and comprehension: What we know and what we need to know. Educational Psychology Review, 34(2), 677-715.

[20] Audiobooks with Sound Effects; how to Create an Immersive Listening Experience. https://canaritaudiobooks.com/audiobooks-with-sound-effects/

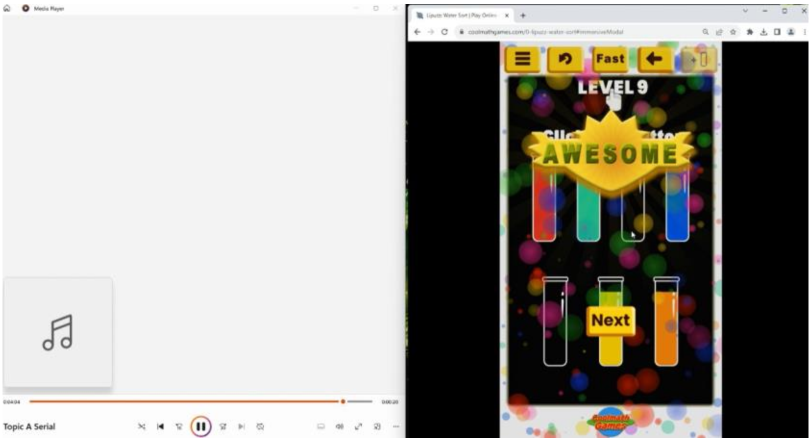
[21] 5 Tips for Creating High-Quality Audiobook Recordings. https://canaritaudiobooks.com/audiobook-recording/

APPENDIX

A.1 . Hierarchical task analysis for the task of playing an audiobook as part of the planned multi-tasking.

Task step	Error consequence	Error probability	Criticality or error
1. Plan to listen to an audiobook while doing a chore such as folding dry clothes or driving	Cancellation of the plan	Medium	Low
2. Start the planned chore	Cancellation of the plan or an unexpected interruption	Medium	Low
3. Locate a mobile device	Unable to locate a mobile device	Low	High
4. Open an audiobook application on the device	Unable to open the application	Low	High
5. Select a paper	The selected paper is not loaded	Low	High
6. Select a voice type, a desired pace, and earcons	Unable to load and/or save the preferences	Low	High
7. Decide the delivery type 7.1 Speaker of the device 7.2 Wireless earphones or a headset	Speaker of the device does not work Unable to connect the wireless earphones or a headset	Low	Low
8. Play the paper	Unable to play the paper	Low	High
9. Pause and mark a location to place an earcon	Unable to pause or the mobile app unexpectedly stops	Low	High
10. Go back to 8 after closing/restarting the application or the mobile device	Unable to play or the mobile app unexpectedly stops	Low	High

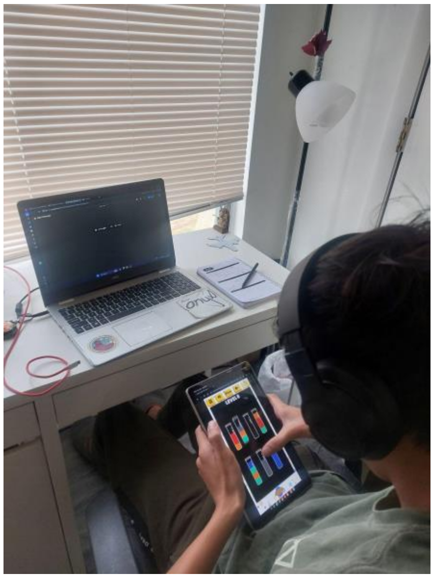
A.2 Sample of a screenshot of virtual session



A.3 Link to Sample Audios with Sonification:

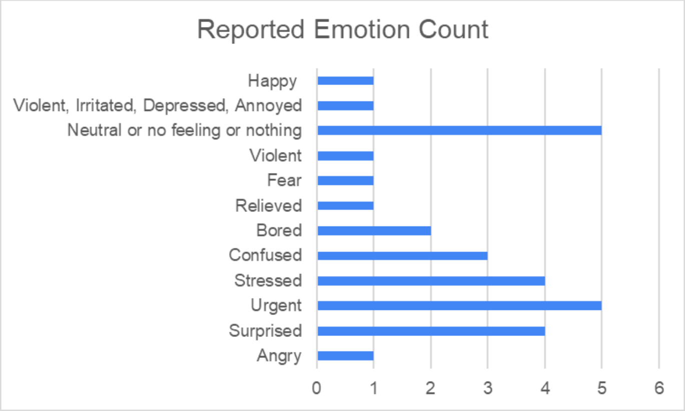
Sample Audio Serial Sonification
Sample Audio Parallel Sonification

A.4 Photograph of a in-person Session:

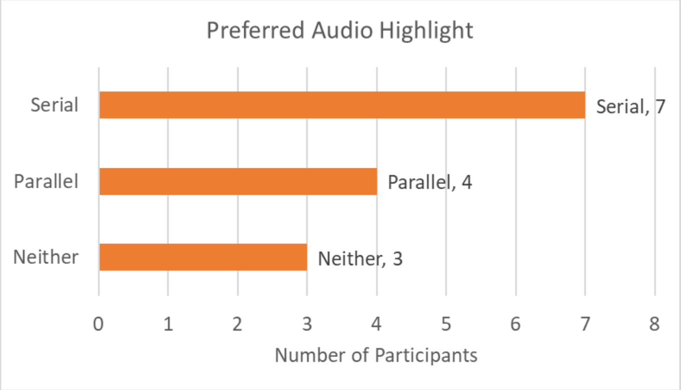


The participants attempt the secondary task on the iPad, while listening to the audiobook on headphones. The computer in front of them is used to play the audiobooks and complete the tests and surveys for the study.

A.5 Graph : Results from the likert scale questions from the exit survey



A.6 Graph : Results from the Highlighting Preference


Preferred Audio Highlight
Serial
Parallel
Neither
Serial, 7
Parallel, 4
Neither, 3
Number of Participants