Journal on Multimodal User Interfaces (2021) 15:323–334
https://doi.org/10.1007/s12193-021-00365-z

ORIGINAL PAPER

Identifying and evaluating conceptual representations for auditory-enhanced interactive physics simulations

Brianna J. Tomlinson[1] · Bruce N. Walker[1] · Emily B. Moore[2]

Received: 17 January 2020 / Accepted: 27 January 2021 / Published online: 15 March 2021

Abstract
Interactive simulations are tools that can help students understand and learn about complex relationships. While most simulations are primarily visual due to mostly historical reasons, sounds can be used to add to the experience. In this work, we evaluated sets of audio designs for two different, but contextually- and visually-similar simulations. We identified key aspects of the audio representations and the simulation content which needed to be evaluated, and compared designs across two simulations to understand which auditory designs could generalize to other simulations. To compare the designs and explore how audio affected a user's experience, we measured preference (through usability, user experience, and open-ended questions) and interpretation accuracy for different aspects of the simulation (including the main relationships and control feedback). We suggest important characteristics to represent through audio for future simulations, provide sound design suggestions, and address how overlap between visual and audio representations can support learning opportunities.

Keywords Interactive simulations · Usability · Audio user experience · Multimodal interactive tools · Surveys · formative evaluations

1 Introduction

Multiple modalities have long proven their positive influence on educational tools [27] (typically through speech, text description, and images). The auditory display community has clearly demonstrated the ability of sounds to support exploration and comprehension of data [10]. The present work investigates the design and impact of auditory displays specifically in the context of interactive simulations. It compares a pair of physics simulations, and explores how audio can impact a learning experience. Simulations are typically visual, and the application of an extended multimedia principle, using auditory displays, has not been evaluated systematically with a pair of simulations to understand generalizable design concepts.

2 Related work

Interactive simulations are educational tools that allow students to explore real-world phenomena in a controlled context [12]. Simulations support learner inquiry in a simplified version of a complex system and encourage exploration of hypothetical changes which may be impossible to conduct in real life [21]. They can also allow students to isolate specific scientific concepts for comprehension in an authentic context [14]. Simulations enhance traditional learning experiences, particularly for lab activities [32].

The quality of an interactive simulation typically hinges on the design of its visual representations, though recent work has begun investigating the enhancement of simulations with auditory displays [25,29,35]. Evaluations of visual displays mostly rely on interpretation of user-driven actions to observe patterns and changes which reflect rules of the dynamic system [11]. Often, evaluations emphasize students' conceptual understanding of the concepts represented within the simulation environment [13]. Multimedia (visual and speech)

This material is based upon work supported by the National Science Foundation under Grant No. DRL-1621363.

✉ Brianna J. Tomlinson
btomlin@gatech.edu

Bruce N. Walker
bruce.walker@psych.gatech.edu

Emily B. Moore
emily.moore@colorado.edu

[1] Georgia Institute of Technology, Atlanta, GA, USA

[2] University of Colorado Boulder, Boulder, CO, USA

Table 1 Overview of study details

Evaluation	Design versions	Total participants
Ohm's Law Evaluation I	4	79
Ohm's Law Evaluation II	8	29
Resistance in a Wire Evaluation I	5	30
Resistance in a Wire Evaluation I	5	27

learning tools have some history of evaluation [26]. Evaluation approaches for auditory displays within simulations present unique challenges distinct from the evaluation of their visual displays [40]. Evaluations of most sound-enhanced learning experiences have focused on learners with vision impairment [24], but have not systematically evaluated auditory or multimodal representations for a general audience (for one that has, see [38]).

2.1 Design and evaluation of sounds

Auditory displays can encompass a broad category of speech and non-speech audio representations. In this work, we focus specifically on evaluating how learners interpret, experience, and use visual representations with associated non-speech audio mappings. The audio mappings include sonification (non-speech sounds whose mappings are driven by data or an underlying model) [22], auditory icons which use a metaphor for interpretation [19], and earcons which use a rapidly acquired association for interpretation [8]. For brevity, we will refer to these categories collectively using the colloquial terms "sounds" or "sound designs" throughout the paper, with the exception of Tables 2, 3, and 4, where auditory subcategories are specified.

Within the auditory display community, designers evaluate audio mappings to ensure they properly present their embedded information in an understandable and comprehensible way. Typical evaluation methodologies include identification tasks (measuring reaction time or accuracy), discrimination tasks (e.g., magnitude estimation), sound card sorting (i.e., matching), among others [7]. This methodological variety supports flexibility in the evaluation's focus, as each of these methods can support qualitative feedback (the user's preferences and enjoyment) and quantitative feedback (accuracy of interpretation) during any stage of the design.

In qualitative evaluations, such as a sound card sorting, users listen to an entire series of sounds and match them with researcher-selected labels or categories. This is a commonly-used technique when in exploratory stages of sound design [34]. Late-stage qualitative evaluations typically take a task-oriented approach, where the sound design is integrated into a system, and designers may rely on task accuracy, video and written observation, questionnaires to understand the user's interpretation, and the success of the display for supporting a task [5].

Quantitative evaluations can provide other means for comparing sound designs: accuracy [6,43], reaction or task time [36], situation awareness [33], and workload [37]. For complex sound designs (those which present multiple streams of information), random sampling for stimuli evaluation can be used to probe comprehension and accuracy for individual variable mappings [33]. Additional measures such as learnability, intuitiveness (recognition or interpretation time), and memorability (retention of association) are frequently used to characterize a successful sound design [18,20].

One difficulty in audio evaluation emerges from general lack of critical examination of sounds and sound designs. People have lots of experience interacting, observing, and critiquing visual designs. The ubiquity of visual displays means that people can use a large body of common language to critique visual displays; however, as audio-only or multimodal displays are less widespread, people lack experience in similar analysis or critiques. When evaluating sound designs, people may imagine or expect changing visual representations, causing them to consider the audio changes within the context of visual ones. To actually evaluate this combined multimodal approach, researchers and designers should ask context-specific and generalizable questions.

Notably, even with the large body of work for evaluating sound designs, there is no consistent way to evaluate one sound design against another. Researchers and designers rely on a variety of methods, making it difficult to compare across designs in a more general sense. The lack of a standardized scale for evaluation of sound designs makes general comparisons difficult, and researchers typically rely on ad hoc Likert questions to elicit feedback on other factors (e.g., aesthetics).

Table 2 Details about each of the four sound designs included as different sim versions in Ohm's Law Evaluation I

OL designs	Current	Sliders	Circuit diagram features
OL 1.1	Pitched loop	Same timbre, diff. pitches	Earcons
OL 1.2	Pitched loop (after interaction)	Same timbre, diff. pitches	Earcons
OL 1.3	Pitched loop	Same timbre, diff. pitches	Earcons
OL 1.4	Synthesized tempo	Same pitch range, diff. timbres	Earcons (after interaction)

Table 3 Details about each of the eight sound designs included as different sim versions in Ohm's Law evaluation II

OL Designs	Current	Sliders	Circuit diagram features
OL 2.1	Pitch change	N/A	N/A
OL 2.2	Tempo change	N/A	N/A
OL 2.3	Fading tempo/pitch	N/A	N/A
OL 2.4	N/A	Same pitch range, diff. timbres	N/A
OL 2.5	Tempo change	Same pitch range, diff. timbres	N/A
OL 2.6	Synthesized tempo	N/A	Earcons
OL 2.7	Pitch change	Modulation rate, diff. timbres	N/A
OL 2.8	Synthesized tempo	Same pitch range, diff. timbre	Earcons

Additionally, these methods are used in a summative manner, and do not explicitly inform iteration for the sound designs. (As a comparison, within the field of Human-Computer Interaction, standardized usability scales are common: see the System Usability Scale [9].)

2.2 What should be evaluated

Visual simulation evaluation is heavily explored due to their prevalence in learning materials. For example, PhET Interactive simulations [30] has a long history of focused visual design for their simulations. As part of that work, Lancaster et al. [23] highlight five different characteristics important to their visual simulation design goals. This work focuses on two of those goals: (1) the need to engage learners in scientific exploration; and (2) the need to develop conceptual understanding across multiple representations. Expanding this set of concepts to encompass design goals for sound in multimodal simulations is a novel area, and has not been systematically explored through structured evaluations.

To meet Goal 1 (supporting scientific exploration), the sound design needs to be interpretable by the learner, and should help a learner make predictions and compare against prior knowledge. To meet Goal 2 (develop conceptual understanding through multiple representations), sound designs should work in tandem with the visuals to provide additional methods of emphasizing the relationships embedded within the sim. Each visual design goal can be adapted to analyze the sound design goals, though covering each of them in detail is beyond the scope of this paper.

Intersecting methods and strategies from the visual simulation design community and the auditory display community could create a cohesive approach for the evaluation of multimodal simulations. Prior work has explored three important concepts that should be evaluated in sound designs, including mapping interpretation [15], user experience [39], and usability [7]. Each of these aspects could impact how a learner is building knowledge with a multimodal learning tool (for both the sound and visual designs). Throughout this work, we will use "interpretation" to mean how a learner understands audio/visual changes within an interactive simulation's context. User experience and usability will be evaluated cohesively. Poor mapping design could lead to difficulty in interpretation of the content, an unenjoyable experience, and frustration, which would reduce the effectiveness of the careful visual and sound design. We will use the term "preferences" to mean the overall user experience and usability of the designs. In these evaluations, we focused on general questions about layering sounds and visuals, and specific questions on individual mappings for conceptual or relationship representations.

3 Methods

In this work, we tested whether or not similar sound designs work for multimodal content with a similar visual designs, layouts, and topics. Four evaluations were completed, two for each of the simulations (see Table 1). We used two surveys per simulation to iterate the sound designs and discover which had the best interpretation and the highest learner preferences. Within-simulation comparisons (between sound designs in Ohm's Law Evaluation I and II) often changed individual mappings or properties, to directly compare the effect of that multimodal version on participant interpretations and preferences.

Across the four evaluations we measured both specific metrics (e.g., accuracy for interpretation of relationship based on sound mapping) and generalizable metrics (e.g., usability), to compare within the designs for one simulation or across designs for both simulations. Most differences between simulation versions across the pairs of studies (i.e., between Evaluation I and Evaluation II for both sims) were simplifications or changes to the total number of concepts presented through sound.

In this work, our goal was to:

- Identify important aspects of audio representations within simulations that should be evaluated.

- Compare audio designs for a pair of simulations to understand what aspects of audio design can generalize across simulations.
- Understand the relationship between conceptual visual and audio representations for simulations.

3.1 Motivation for simulation choices

We selected two of the more than 140 free math and science simulations from the PhET Interactive Simulations Project [30]. PhET's simulations (sims) are available in 87 languages and widely-used (more than 100 million times a year). Their sims follow a set of specific design goals, including supporting interactivity, giving dynamic feedback, and supporting multiple representations (e.g., [23]). Their iterative design process supports evaluation to ensure these key design goals, and includes numerous interviews with learners, rounds of heuristic evaluation with expert designers and teachers, and thorough formal evaluations [3,23]. PhET sims are widely-used and thoroughly-evaluated from a visual design standpoint and their historical success in supporting education made them excellent candidates for exploring auditory and multimodal design within this work.

Their sims rely on a principle of implicit scaffolding to encourage student exploration and use without needing explicit, guided instructions; this is similar to Quintana et al.'s [31] concept of instructional scaffolding. They include three main scaffolding strategies: providing visual representations that build on initial understanding; allowing direct control and observation of phenomena; and enabling learners to explore the content through multiple views and representations, especially for complex concepts. Standardized evaluation methods have already been thoroughly explored within their visual sims, and their ability to successfully support learning for the visually-included concepts provided a good baseline to compare against.

Ohm's Law and Resistance in a Wire were chosen for their similar visual representation and interaction, but differing levels of complexity in user-controlled variables. They are widely-used sims from 6th grade onward (until university), and their relatively simple controls made them prime candidates for use in this preliminary study.

In Ohm's Law (OL), users can adjust sliders to explore relationships between current (I), voltage (V), and resistance (R) in a circuit. In this simulation (Fig. 1), the Ohm's Law equation is displayed (V = I R). Below the equation is a circuit containing a resistor and multiple 1.5 Volt batteries in series. A readout indicates the circuit's current in milliamps.

By adjusting the voltage or resistance slider, users change the value of the voltage or resistance in the equation and in the circuit. As the voltage or resistance is changed, the equation letter sizes change (e.g., increasing voltage results in increase to the size of letter V and letter I, indicating a proportional relationship). Corresponding changes occur to the circuit, including a change in number of batteries shown as voltage is changed (full or partial 1.5 V batteries appear or disappear) or in the amount of impurities in the resistor as resistance is changed.

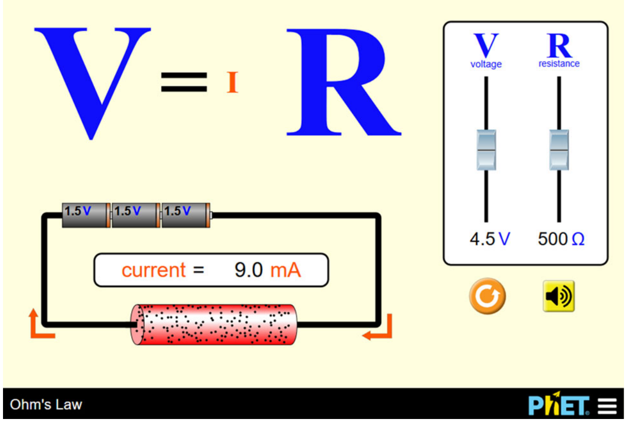

Fig. 1 Screenshot of PhET simulation Ohm's Law

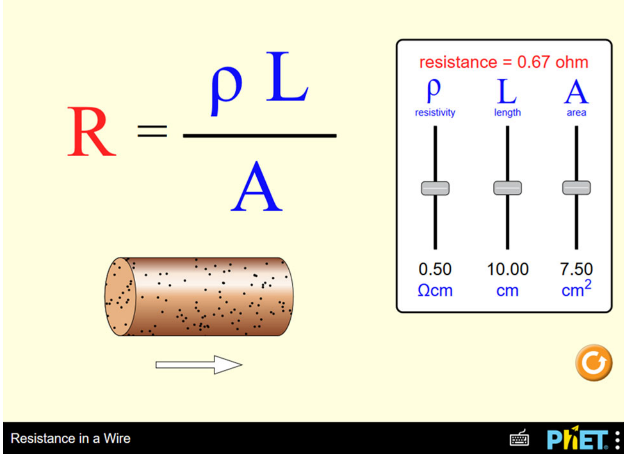

Fig. 2 Screenshot of PhET simulation Resistance In A Wire

In Resistance in a Wire (RIAW), users can adjust sliders to explore the relationships between resistance in a wire (R), and the resistivity (ρ), length (L), and area (A). In the simulation (Fig. 2), the equation is displayed (R = ρ L / A). Below the equation is a section of wire. By adjusting the resistivity, length, and area sliders, users change the value of these variables in the equation and in the wire. As the variables are changed, the equation letter sizes change (e.g., increasing area decreases resistance, indicating an inverse relationship). Corresponding changes occur to the wire, including a change in the amount of impurities in the wire as resistivity is changed, and changes to the length or area of the wire as length or area are changed.

3.2 Audio design process

Multiple sound designs were developed for each sim, implemented through Web Audio [41], after initial prototyping using the audio design software SuperCollider [28], Max/MSP [1], GarageBand [4], and Ableton [2]. Weekly meetings of the interdisciplinary design team were used to brainstorm the auditory display designs and to decide the different conditions or representations which would be compared in each of the surveys. Previous research in parameter mapping was leveraged to inform portions of the design (e.g., slider values relied on pitch mappings, where low values were lower pitched, and high values were higher pitched, following [42]. In part, the lack of leverageable research forced our evaluation of multiple sound designs [17]. Prototypes presented different sound designs for representing changes in: directly interactive variables (i.e., those represented through sliders); the indirectly interactive variable (Current in Ohm's Law and Resistance in Resistance in a Wire); and physical properties (batteries and resistor in Ohm's Law circuit, wire in Resistance in a Wire). See the supplemental videos for demonstrations of each prototype design.

3.3 Audio-enhanced simulation designs

Each evaluation had research questions specific to the sim design (detailed below). However, all of these questions sought to understand if the sounds helped encourage learner exploration, and if the sounds effected conceptual understanding, our two focuses for considering the success of potential sound-enhanced simulation designs [23].

3.3.1 Ohm's Law (OL) evaluation I

In Ohm's Law, it is important for a learner to explore the presented relationships through direct changes to the two variables (voltage and resistance) and to understand their effects on current. The design team was primarily interested in: (1) Should the main relationship or concept (current) be represented through sound?; (2) Should the sliders be represented through sound?; and (3) Should the physical circuit (i.e., voltage or resistance properties, like the number of batteries) be represented through sound? To investigate these questions, OL Evaluation I included four sound-enhanced sim versions with different sound designs (OL 1.1–1.4, listed in Table 2). Each design contained three categories of sounds: representations for current (mapping: pitch vs. tempo; timing: during vs. after); slider mappings (pitch vs. timbre); and physical circuit properties (earcons vs. no earcons).

3.3.2 Ohm's Law (OL) evaluation II

From OL Evaluation I we gained insight into learners' interpretations and preferences for the sound mappings. In Ohm's Law Evaluation II, we aimed (1) to understand the effect of sound design complexity on interpretation and preference and (2) to identify potential opportunities to simplify the sound designs. OL Evaluation II utilized eight sound designs (OL 2.1–2.8, see Table 3), most with less complex mappings than versions in OL Evaluation I (OL 1.1–1.4). For example, most versions included in OL Evaluation II used sound to represent either changes in current, or interaction with sliders, but not both. A single complex design (OL 2.8) was included to serve as a contrasting design comparison.

3.3.3 Resistance in a wire (RIAW) evaluation I

Resistance in a Wire Evaluation I included five different sound designs (RW 1-5) shown in Table 4. This sim has the same set of interaction controls (sliders) and similarly-designed visual displays as Ohm's Law. However, Resistance in a Wire was a contrasting case to the Ohm's Law sound designs, since it presented a more complex set of relationships (three variables instead of two). The two main research questions were: (1) Should the main relationship or concept (resistance) be represented through sound?; and (2) Can the resulting sound mappings from OL Evaluation I & II for the sliders and physical properties (e.g., length of the wire) generalize to inform the design of other sims? The five sim versions contained categorical comparisons: resistance (mapping: pitch loop vs. tempo loop vs. pluck; timing: during vs. after); slider mappings (pitch vs. timbre); and wire properties (earcons vs. no earcons). Inspired by the wire in the visual display, some versions incorporated a wire 'pluck' sound instead of a general pitch or tempo-mapped sound clip.

3.3.4 Resistance in a Wire (RIAW) Evaluation II

Though it will be discussed in more detail later, much of the RIAW Evaluation I feedback centered around the difficulty in interpreting what mapping represented resistance. The main goal of RIAW Evaluation II was to investigate whether or not differences in sound design emphasis through volume could affect interpretation of the concepts within the sim. The same main sound design mappings were used in RIAW Evaluation II (see Table 4).

3.4 Participants

Participants were recruited from the undergraduate population at a technical university in the United States, all with self-reported normal or corrected-to-normal hearing. Participants were recruited through the Psychology Department,

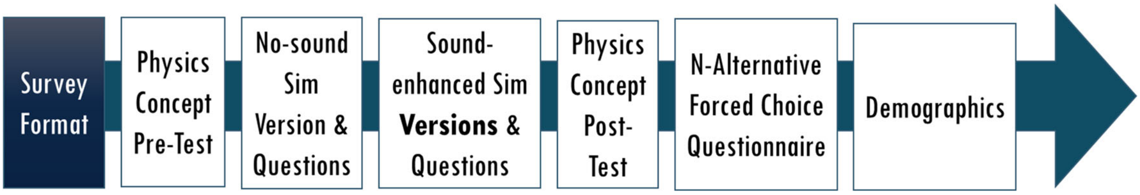
Survey Format
Physics Concept Pre-Test
No-sound Sim Version & Questions
Sound-enhanced Sim **Versions** & Questions
Physics Concept Post-Test
N-Alternative Forced Choice Questionnaire
Demographics

Fig. 3 General format of all surveys

and had a range of knowledge about the sim topics, representing a diverse group of novice to expert learners. Importantly, we recruited this participant group as they were likely to have some exposure to the concepts, and could discuss a nuanced perspective of relating the auditory representations to the visual ones without having a potential negative impact of their initial learning (which, though unlikely, could have happened with novice learners, and thus was something we wanted to avoid). More students were recruited for the initial survey (OL Evaluation I), to capture a broad range of feedback. In the additional surveys, we found that a smaller number of participants still provided a broad range of responses, and so reduced the total number of participants. OL Evaluation I and II had 79 and 29 student respondents, respectively (108 participants total, ages 18–24, mean = 19.6, 55 females). RIAW Evaluation I and II had 30 and 27 responses, respectively (57 participants total, ages 18–37, mean = 19.6, 29 females). Each survey took 60–90 min to complete; total duration was dependent on the number of simulation versions included within each survey.

3.5 Evaluation materials and procedure

Each evaluation (OL Evaluation I, OL Evaluation II, RIAW Evaluation I, and RIAW Evaluation II) was given through a comprehensive, single survey. First, participants were prompted to review the ethics committee approved consent form, and indicated their consent before proceeding to the survey (see Fig. 3 for overview).

3.5.1 Physics concept pre-test

Second, the next section was a short two-question (OL Evaluation I and II) or three-question (RIAW Evaluation I & II) pre-test on physics concepts relevant to the sim. Questions were provided with a graphic of a simple physics circuit diagram (OL Evaluation I and II) or a piece of wire (RIAW Evaluation I & II), labeled with relevant variables (e.g., V, I, and R). No other supporting materials (i.e., equations or definitions) were provided.

3.5.2 No-sound sim questions

Third, each participant had a minimum of 30 s (enforced, as the survey would not proceed sooner) using a no-sound (visual-only) version of the sim. After 30 s, three comprehension questions were displayed, probing for specific quantitative values (e.g., “What is the value of the current (I) in the circuit when you move the voltage (V) slider to 2.4 V and the resistance (R) slider to approximately 750 Ω?”). A couple of questions about interpretation of the visual sim components (e.g., “What do you think the black dots in the wire represent?”) were then asked.

3.5.3 Audio-enhanced sim questions

After using the visual-only version of the sim, interaction and exploration with the multimodal sim versions followed a similar pattern. Each sim had a minimum 30 s interaction before the participant could move to the interpretation and preference questions. Three value-finding questions were asked for each sim, with similar goals of encouraging sim use to answer. Then participants mapped sound descriptions to interactions and concepts, while following a prompt like this one from Ohm’s Law: “Adjust the Voltage and Resistance sliders by moving them up or down. What sounds did you hear while you were adjusting the two sliders?” A list of high-level descriptions for each sound mapping was provided. They would assign each description to a concept or control within the sim.

Then, they were asked different sets of questions on ease of interpretation (for the sound mappings), preference about sound length, and how clear they felt the slider mappings were. Next, each participant ranked the sounds by preference. Afterward, they provided open-ended feedback about how each of the sounds impacted their understanding of the concepts.

Additional Likert questions were asked on the following topics: enjoyment; whether they would recommend the sim to a friend learning the main concept (e.g., Ohm’s Law); and their confidence in interpretation. Then, they answered the four user experience questions (standardized UMUX) [16] and a subset of audio user experience questions (BUZZ) [39] focused on their ease of interpretation for the sound mappings, as well as enjoyment and aesthetic preferences about the sounds.

3.5.4 Physics concept post-test

After completing all of the individual questions for each sound-enhanced sim version, they answered a post-test comprised of the exact same questions from the pre-test.

3.5.5 N-alternative forced-choice questionnaire

The final sim used was a version that allowed participants to select between different sound options to create a customized auditory display in real-time. The included sound options came from the different designs they had experienced throughout each of the sim versions. They were asked to report their choices for the directly interactive variables (i.e., the sliders), the indirectly interactive variables (i.e., current or resistance), and physical representations (e.g., the batteries or resistance/resistivity). When reporting their sound design choices, participants explained their reasoning for the choices. Finally, they ranked the available sound choices from most to least favorite and answered the exact same set of Likert questions used in each of the audio-enhanced sim question sets.

3.5.6 Demographics

Lastly, participants self-reported demographics including gender, age, major, related classes, prior use of PhET simulations, and primary language.

3.6 Analyses

Descriptive statistics were calculated for responses to each survey, as this work was largely exploratory and did not require more complex statistical analyses to discover patterns in responses. Open-ended responses were categorized by two members of the study team as either “positive” or “negative”. In this work, we present data related to user comprehension interpretation and user preference (including user experience and usability). The first set of data (interpretation) related to the educational nature of these simulations; without clear evidence for successful interpretation of the multimodal representations, it would be difficult to argue for more learning resources to include the time, effort, and evaluation for auditory display development for future tools. The second half of the data (user preference) evaluated overall usability and enjoyment of the displays; without support for an enhanced, enjoyable experience, learners may mute or turn-off the sound, resulting in a lower chance of engaging with the additional representations.

3.6.1 Interpretation

Interpretation was measured three ways. First, differences in pre-post scores for the physics concepts provided a baseline about the concepts. Second, accuracy in the interpretation of the visual sim components provided insight to their understanding of the visual-only sim. Third, accuracy in matching sounds to their concepts and interpretation of their mappings provided insight to their understanding of the audio and visual layers used within the multimodal sims.

3.6.2 User preferences

We used the UMUX scale, an overall metric for user experience that includes sub-components of effectiveness, satisfaction, overall ease of use, and efficiency. The BUZZ audio user experience questions focused on enjoyment and aesthetics of the sound designs. Scores were calculated according to their scale designs, following Finstad [16] and Tomlinson, Noah, & Walker [39]. UMUX was scored out of 24, and the BUZZ Aesthetics Subscale was scored out of 21. Descriptive statistics were calculated for the other Likert and ranking questions.

4 Results

4.1 OL evaluation I & II

OL Evaluation I had three main questions: (1) Should the main relationship or concept (i.e., current) be represented through sound?; (2) Should the sliders be represented through sound?; and (3) Should the physical circuit (i.e., voltage or resistance properties, like the number of batteries) be represented through sound? After the first survey, the sound designs were simplified for OL Evaluation II. OL Evaluation II’s focus was on comparing these different simplified sound designs to (1) understand the effect of sound design complexity on interpretation and preference, and (2) to identify opportunities to simplify the final sound designs.

4.1.1 Interpretation

Pre-post scores for physics concepts were consistently high across both surveys (over 90%). In Ohm’s Law, interpretation accuracy of the visual-only sim components was highest for the circuit diagram (OL Evaluation I: 95%, OL Evaluation II: 93%) and lowest for the dots representing resistance (OL Evaluation I: 49%, OL Evaluation II: 48%).

Sound mappings used in all multimodal sim versions were scored for accuracy. Across all designs in OL Evaluation I, participants had low accuracy (40–50%) interpreting the sound mappings for current, the most pedagogically relevant

Table 4 Details about each of the eight sound designs included as different sim versions in Resistance in Wire Evaluation I & II

RIAW designs	Resistance	Sliders	Wire features
RIAW 1	Pitched loop	Same timbre, diff. pitches	Earcons
RIAW 2	Pitched loop (after use)	Same timbre, diff. pitches	Earcons
RIAW 3	Tempo change	Same timbre, diff. pitches	Earcons
RIAW 4	Pitch change "pluck"	Same timbre, diff. pitches	N/A
RIAW 5	Pitch change "pluck" (after use)	Same timbre, diff. pitches	N/A

sound. In contrast, they were highly accurate (92–100%) interpreting the slider sounds across the different sim versions. The earcons for the circuit diagram had 65% accuracy across the four multimodal versions.

Based on results from OL Evaluation I, representations for current were simplified for OL Evaluation II. In OL Evaluation II, sound designs with current alone (OL 2.1–2.3, and 2.6) had higher interpretation accuracy for current (80%, 67%, 81%, and 70%, respectively) than those with both current and slider sounds (OL 2.7: 35% and OL 2.8: 40%). OL 2.1 and 2.3 had the overall highest accuracy for current interpretation (81%). Slider interpretation accuracy was high (around 80% for OL 2.4 and 2.5) when it was the only sound or when it was paired with a tempo and pitch mapping for current. Sliders used in sound designs with more layers had the lowest accuracy for interpretation (OL 2.7: 14 people were accurate for both; OL 2.8: 13 people were accurate for both). Earcon interpretation was highest in OL 2.6 (Voltage: 81%), and lowest in the most complex simulation version (OL 2.8, about 50% for both earcon mappings).

4.1.2 User preferences

Overall, designs OL 1.1 and 1.4 had the highest UMUX scores (1.1: 16.96; 1.4: 17.03). While OL 1.4 had the highest overall score for aesthetics (BUZZ: 10). Table 5 contains the average usability and audio user experience scores for each sim version. In Ohm's Law Evaluation I, there were no major differences in scores for the subset of standard audio scale questions. OL 1.4 had the highest average score for being fun (4.07 out of 7) and OL 1.2 had the highest average score for being recommended to help someone learn about Ohm's Law (4.85 out of 7). Participants all reported the sound length for current as being too long. Through the open-ended questions, 60–93% of participants reported that the earcons were distracting or annoying, and many found them to be unnecessary double-mappings when compared to the sliders or general current representations

In OL Evaluation II, designs OL 2.4 and 2.3 had the highest UMUX scores (2.4: 18.25, 2.3: 18.24). OL 2.4 also had the highest aesthetics score (13.18). Table 6 contains the average usability and audio user experience scores for each sim version. Participants rated their confidence in interpretation of sound designs highest for the two which used a simple tempo-only mapping for Current (OL 2.2: 5.07 out of 7; OL 2.5: 4.92 out of 7). OL 2.4 and 2.5 had the two highest ratings for being fun (4.46 and 4.35 respectively), and had the highest ratings for ease of interpretation for the sound mappings (2.4: 4.77 out of 7; 2.5: 4.5 out of 7).

Table 5 Mean UMUX scores (out of 24) and BUZZ aesthetics subscale (out of 21) in Ohm's Law evaluation I

OL designs	UMUX scores	BUZZ scores
OL 1.1	16.96	9.19
OL 1.2	16.61	9.65
OL 1.3	15.96	9.08
OL 1.4	17.03	10

OL Evaluation II included shorter sounds (across all representations), participants reported the length for all sounds was fine (reported as "neither too long nor too short"). In the forced-choice questions, the highest preferred Current representation was through tempo change (OL 2.2 and 2.5). Participants chose pitch-based slider representations (OL 2.4, 2.5) over to the non-pitched based ones (OL 2.7, 2.8).

4.2 RIAW evaluations I & II

RIAW Evaluation I had two main questions: (1) Should the main relationship or concept (resistance) be represented through sound?; and (2) Can the resulting sound mappings from OL Evaluation I & II for the sliders and physical properties (e.g., length of the wire) generalize to inform the design of other sims? After the first survey, the sound designs were adapted to discover whether or not differences in sound design emphasis through volume could affect the interpretation of the concepts within the sim.

4.2.1 Interpretation

Pre-post scores for physics concepts were high, except for two notable relationships for Resistance in a Wire: length & resistance (overall accuracy in the RIAW Evaluation II: 13 to 27), and area & resistance (overall accuracy in the RIAW Evaluation I: 16 to 30; RIAW Evaluation II: 11 to 27). In

Table 6 Mean UMUX scores (out of 24) and the BUZZ Aesthetics Subscale (out of 21) in Ohm's Law Evaluation II

OL Designs	UMUX scores	BUZZ scores
OL 2.1	18.17	13.21
OL 2.2	18.1	11.93
OL 2.3	18.24	12.04
OL 2.4	18.25	13.18
OL 2.5	17.75	12.38
OL 2.6	15.32	11.79
OL 2.7	14.71	10.97
OL 2.8	13.42	9.93

Table 7 Mean UMUX scores (out of 24) and the BUZZ aesthetics subscale (out of 21) in Resistance in a Wire Evaluation I

RIAW designs	UMUX scores	BUZZ scores
RIAW 1.1	16.63	9.5
RIAW 1.2	16.6	9.77
RIAW 1.3	15.93	10.43
RIAW 1.4	16.59	8.55
RIAW 1.5	16.8	9.03

Table 8 Mean UMUX scores (out of 24) and the BUZZ Aesthetics Subscale (out of 21) in Resistance in a Wire Evaluation II

RIAW designs	UMUX scores	BUZZ Scores
RIAW 2.1	14.31	9.77
RIAW 2.2	14.62	10.46
RIAW 2.3	13	10.15
RIAW 2.4	13.42	9.38
RIAW 2.5	15.27	11.5

Resistance in a Wire, interpretation accuracy of the visual sim component (resistivity) was lowest (RIAW Evaluation I: 43%, RIAW Evaluation II: 59%).

In RIAW Evaluation I, the slider mapping accuracy score was highest (RIAW 1.4 and 1.5: 75–85% accuracy, across the three sliders) when the resistance sound was a single plucked note. Accuracy mappings for resistance was 35–45% for all five simulation designs. After the re-balanced sound levels for RIAW Evaluation II, accuracy in resistance mapping increased to 75% for the plucked sound representations.

4.2.2 User preferences

When asked to report their choices for the final, Alternative Forced Choice version of the sim, half (55%) preferred the plucking sound to a looping one. Mean UMUX scores were calculated for each sim, and RIAW 5 had the highest rating (16.8 out of 24).

Mean UMUX scores were calculated for each sim, and there were no major differences (Table 7), though RIAW 1.5 had the highest rating. RIAW 1.3 had the highest score calculated from the BUZZ Aesthetics Subscale; participants reported that there were too many different sounds (making them difficult to differentiate). When asked to report their choices for the forced-choice version, over half (55%) preferred the plucking sound to represent resistance.

Participants reported having an easier time interpreting the concept for the plucked sounds (RIAW 2.5 was selected the preferred representation for the resistance sound). In the open-ended responses, 16 participants specifically listed positive feedback about the plucked sounds. In most cases they liked it because it was a concise representation and was the 'least distracting and easiest to identify.'

Almost all of the BUZZ Aesthetics Subscale scores were higher for every sim design in RIAW Evaluation II (Table 8), though many had lower general usability scores compared to the RIAW Evaluation I. This may imply that there are differences between what factors audio user experience scales measure compared to general usability.

5 Discussion

Through our studies of these four sound-enhanced, multimodal sims, we:

- Identified key audio representations that should be represented within these two physics simulations.
- Compared audio designs between the two simulations to understand what aspects of audio design can generalize across different physics simulations.
- Explored the relationship between visual and audio representations for physics simulations.

5.1 Important aspects to represent through audio

The surveys conducted in OL Evaluations I & II addressed three specific questions related to which concepts should be represented through an audio layer. (1) Should the main relationship or concept be represented through sound? Yes: participants could successfully interpret sounds for the concept of current, particularly when it was mapped to a pitch or pitch and tempo change.

(2) Should the sliders be represented through sound? Yes: participants could easily interpret the slider mappings; they can be used as long as they do not overlap with the pitches for current.

(3) Should the physical circuit be represented through sound? Maybe: earcons could successfully increase preference scores (e.g., OL 1.4). Using diverse tempo, pitch, and earcon mappings made the sounds easier to differentiate; participants could interpret all of them. However, some felt the earcons were too repetitive compared to the visual changes, and did not want those sounds as well as slider sounds.

OL Evaluation II further tested whether simplifying sound designs would lead to differences in interpretation and preference, particularly since participants were less accurate and did not rate the experience highly for the complex designs. Simpler designs without earcons had higher interpretation and preference scores. These simulations had better alignment between understandability and aesthetics for the sound mappings.

RIAW Evaluations I & II sought to discover similar constructs, but within a different set of simulations. (1) Should the main concept be represented through sound? Yes. Using a single, plucked representation for resistance, with balanced volume levels (to reduce emphasis of the sliders), and without earcons, provided the best interpretation and preference scores. Timing for the sound changes also affected these scores; participants preferred just-in-time feedback as opposed to feedback after they completed their interaction.

(2) Could similar sound design mappings be used across sims (i.e., between Ohm's Law and Resistance in a Wire)? Yes. For the sound designs which paralleled Ohm's Law, the sliders had high interpretation accuracy. Instead of using a parallel sound design between resistance and current (with a pitched loop), we found that participants could interpret and also preferred a shorter sound representing the overall resistance value. This highlights how some factors between sim sound designs could transfer, but others may need to be adapted, depending on the concepts and interactions.

5.2 Sound design suggestions

In general, participants preferred focused, simpler sound designs for both sims. Across all versions, the slider mappings were well understood, but could be confused when used in conjunction with other pitch-based representations (e.g., current in Ohm's Law). Since sliders could distract from the main concept, we suggest using a plain wood-block timbre (or other similarly plain sound) to convey this type of feedback. Using a complex tone for the main concept can be successful, as long as it is clearly balanced with the volume for any other sounds. Both current (Ohm's Law) and resistance (Resistance in a Wire) had high interpretation and preference scores when it was emphasized as the most important sound through volume, pitch, and tempo/timing. Lower scores for the earcons led us to forgo them in the final audio designs for both sims.

Designs with the highest interpretation accuracy also had higher preference scores (i.e., higher audio user experience and usability). Integrating these types of questions could help narrow a larger set of designs to a finalized subset during formative evaluations through the use of standardized questions, instead of needing to ask a significant amount of interpretation and specialized preference questions. A subset of preference and interpretation questions may be enough for formative studies.

5.3 Overlapping visual and audio design

At the outset, we provided two important needs that audio-enhanced sims would have to meet in order to support successful learning. First, the sound design needed to be interpretable by the learner, and should help them make predictions as well as compare against prior knowledge. The sound-enhanced sims with the highest accuracy and preference scores did accomplish this. Through the sim's visual and sound representations, participants could identify changes and interpret how it represented the main concept. Many discussed their interpretation of the sound designs, and why they chose an option for the N-Alternative Forced-Choice section, "It was clear that it was an effect of the change in the slider by only playing afterwards. Furthermore, it was succinct and clear by changing tones and playing very quickly, which maximized its effectiveness while keeping it short enough to not become annoying." Participants often discussed how the sounds emphasized the changes in the main concept, without distracting them. Many cited the sound brevity as a positive in this case: it supported making many comparisons within their brief exploration time.

Second, the sound designs should work in tandem with the visuals to provide additional methods of emphasizing the relationships embedded within the sim. Participants often discussed both the visuals and the sounds in their interpretation and preference responses: "Both voltage and resistance changed the sound so I assumed it's representing current." The open-ended responses allowed us to integrate participants' reasoning with their interpretations and preferences to understand how the overlapping visual and sound designs supported their experiences. In general, participants gave sims higher preference scores when their interpretations for the sounds and visuals were more accurate.

6 Future research and limitations

This work sought to address the evaluation of multimodal simulations, particularly those using non-speech auditory representations as an extension of the multimedia principle. We found that a sample of about 30 participants provided enough feedback to evaluate the interpretation and preference, and enough guidance to iterate on formative designs. Future research should explore a variety of simulations, to discover whether or not these findings hold across a larger set of simulations. In addition, structured evaluation of the ability for sound-enhanced simulations to support all five of Lancaster et al.'s goals [23] is imperative to comparing the success of these multimodal simulations to visual-only versions. One limitation is that these evaluations were done with a sample of college students, who are one potential group of simulation users. Therefore future research should seek to evaluate formative designs of multimodal simulations with other groups of learners (e.g., primary and secondary school students).

7 Conclusion

Sound-enhanced multimodal simulations can provide a better learning experience for students, since simulations can present relationships and concepts through multiple means. This study presents straightforward design recommendations for the design of sound-enhanced simulations through comparisons between four different rounds of sound design evaluation for two similarly-structured physics simulations. This evaluation highlights key aspects to consider when designing sounds to layer with visual representations; it provides specific, actionable design suggestions which can be implemented for other simulations. Additionally, it provides a foundational methodological approach that can be used in future evaluations, and suggests a subset of methods to use for facilitation of faster iteration on multimodal simulation designs.

Acknowledgements We thank John Blanco and Taliesin L. Smith (PhET Interactive Simulations), and Mike Winters (Georgia Tech) for their significant contributions to the design and implementation of auditory display in PhET sims. We thank Jared Batterman, Prakriti Kaini, and Siyan Zhou (Georgia Tech) for their help with data collection and survey design. This material is based on work supported by the National Science Foundation under DRL-1621363.

References

1. '74, C.: Max/msp 7. https://cycling74.com/ (2014)
2. Ableton: Ableton live. https://www.ableton.com/en/live/ (2013)
3. Adams WK, Reid S, Lemaster R, McKagan SB, Perkins KK, Dubson M, Wieman CE (2008) A study of educational simulations Part 1: engagement and learning. J Interact Learn Res 19(3):397–419
4. Apple: Garageband. https://apple.com/mac/garageband/ (2004)
5. Basballe DA, Breinbjerg M, Fritsch J (2012) Ekkomaten: an auditory interface to the 18 th century city of aarhus. Nordic conference on human-computer interaction, pp 742–745. https://doi.org/10.1145/2399016.2399130
6. Bonebright TL (1996) An investigation of data collection methods for auditory stimuli: paired comparisons versus a computer sorting task. Behav Res Methods Instrum Comput 28(2):275–278. https://doi.org/10.3758/BF03204780
7. Bonebright TL, Flowers JH (2011) Evaluation of auditory display. In: T. Hermann, A. Hunt, J.G. Neuhoff (eds.) The sonification handbook, pp 111–144. http://sonification.de/handbook
8. Brewster SA, Wright PC, Edwards ADN (1993) An evaluation of earcons for use in auditory human-computer interfaces. In: Proceedings of the SIGCHI conference on Human factors in computing systems–CHI '93 (April), pp 222–227. https://doi.org/10.1145/169059.169179
9. Brooke J (1996) SUS: a quick and dirty usability scale. Usability Evaluat Ind 189(194):4–7
10. Brown LM, Brewster SA (2003) Drawing by ear: interpreting sonified line graphs. In: Proceedings of the 2003 international conference on auditory display, pp 152–156
11. Colella V (2014) Participatory simulations: building collaborative understanding through immersive dynamic modeling. J Learn Sci 9(4):437–469
12. D'Angelo C, Rutstein D, Harris C, Haertel G, Bernard R, Borokhovski E (2014) Simulations for STEM learning: systematic review and meta-analysis report overview. Tech. rep
13. Dorić B, Lambić D, Jovanović Ž (2019) The use of different simulations and different types of feedback and students 'Academic performance in physics'. Research in Science Education pp 1–21. https://doi.org/10.1007/s11165-019-9858-4
14. Edelson DC, Gordin DN, Pea RD (1999) Addressing the challenges of inquiry-based learning through technology and curriculum design. J Learn Sci 8(3–4):391–450. https://doi.org/10.1080/10508406.1999.9672075
15. Ferati M, Pfaff MS, Mannheimer S, Bolchini D (2012) Audemes at work: investigating features of non-speech sounds to maximize content recognition. Int J Human Comput Stud 70(12):936–966. https://doi.org/10.1016/j.ijhcs.2012.09.003
16. Finstad K (2010) The usability metric for user experience. Interact Comput 22(5):323–327
17. Frauenberger C, Stockman T, Bourguet ML (2007) A survey on common practice in designing audio in the user interface. In: Proceedings of HCI 2007 The 21st British HCI group annual conference University of Lancaster, UK 21, pp 1–9
18. Garzonis S, Jones S, Tim J, O'Neill E (2009) Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1513–1522. ACM
19. Gaver WW (1986) Auditory icons: using sound in computer interfaces. Hum Comput Interact 2:167–177
20. Giannakis K (2006) A comparative evaluation of auditory-visual mappings for sound visualisation. Organised Sound 11(3):297–307. https://doi.org/10.1017/S1355771806001531
21. van Berkum JA, de Jong T (1991) Instructional environments for simulations. Educat Comput 6(3–4):305–358. https://doi.org/10.1016/0167-9287(91)80006-J
22. Kramer G (1994) An introduction to auditory displays. In: G. Kramer (ed.) Auditory display: sonification, audification and auditory interfaces, chap. 1. Addison-Wesley Publishing Company
23. Lancaster K, Moore EB, Parson R, Perkins KK (2013) Insights from using PhET's design principles for interactive chemistry simulations. ACS Symp Ser 1142:97–126. https://doi.org/10.1021/bk-2013-1142.ch005
24. Levy ST, Lahav O (2012) Enabling people who are blind to experience science inquiry learning through sound-based mediation. J Comput Assist Learn 28(6):499–513. https://doi.org/10.1111/j.1365-2729.2011.00457.x
25. Levy ST, Peleg R, Lahav O, Chagav N, Talis V (2016) Listening versus looking: learning about dynamic complex systems in science among blind and sighted students. In: Annual international national

association for research in science teaching (NARST). Baltimore, MD

26. Mayer RC, Clark Richard E (2008) Learning by viewing versus learning by doing: evidence-based guidelines for principled learning environments. Perform Improv 47(9):9–16. https://doi.org/10.1002/pfi
27. Mayer RE (2002) Multimedia learning. In: The annual report of educational psychology in Japan vol 41, pp 27–29. https://doi.org/10.1016/S0079-7421(02)80005-6
28. McCartney J et al. (1996) Supercollider. https://supercollider.github.io/
29. PhET Interactive Sims: Phetsims/Tambo (2018). https://github.com/phetsims/tambo
30. PhET Interactive Simulations: PhET Interactive Simulations. http://phet.colorado.edu/
31. Quintana C, Reiser BJ, Davis EA, Krajcik JS, Fretz E, Duncan RG, Kyza E, Edelson DC, Soloway E (2014) A scaffolding design framework for software to support science inquiry. J Learn Sci 13:37–41. https://doi.org/10.1207/s15327809jls1303
32. Rutten N, Van Joolingen WR, Van Der Veen JT (2012) The learning effects of computer simulations in science education. Comput Educ 58(1):136–153. https://doi.org/10.1016/j.compedu.2011.07.017
33. Schuett JH, Walker BN (2013) Measuring comprehension in sonification tasks that have multiple data streams. In: Proceedings of the 8th audio mostly conference. ACM
34. Shortridge W, Gable TM, Noah BE, Walker BN (2017) Auditory and head-up displays for eco-driving interfaces. In: The 23rd international conference on auditory display, pp 107–114
35. Smith TL, Lewis C, Moore EB (2017) Description strategies to make an interactive science simulation accessible. JTPD **5**(22), 225–238. http://scholarworks.csun.edu/handle/10211.3/190214
36. Sodnik J, Jakus G, Tomažič S (2011) Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users. Int J Human Comput Stud 69(1–2):100–112. https://doi.org/10.1016/j.ijhcs.2010.10.004
37. Stevens RD (1996) Principles for the design of auditory interfaces to present complex information to blind people. Ph.D. thesis, The University of York. http://www.cs.manchester.ac.uk/~stevensr/papers/thesis-rds.pdf
38. Tomlinson BJ, Kaini P, Harden EL, Walker BN, Moore EB (2019) A multimodal physics simulation: design and evaluation with diverse learners. J Technol Persons Disabil
39. Tomlinson BJ, Noah BE, Walker BN (2018) BUZZ: an auditory interface user experience scale. In: CHI'18 Extended Abstracts. Montreal, QC, Canada. https://doi.org/10.1145/3170427.3188659
40. Tomlinson BJ, WalkerBN, MooreEB(2020) Auditory display in interactive science simulations: description and sonification support interaction and enhance opportunities for learning. In: Proceedings of the 2020 CHI conference on human factors in computing system. ACM, https://doi.org/10.1145/3313831.3376886
41. W3C: Web audio api. https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API (2011)
42. Walker BN, Kramer G (2005) Mappings and metaphors in auditory displays: an experimental assessment. ACM Trans Appl Percept 2(4):407–412
43. Zhao H, Plaisant C, Shneiderman B (2005) "I hear the pattern": interactive sonification of geographical data patterns. In: Proceedings of ACM CHI 2005 conference on human factors in computing systems, vol 2, pp 1905–1908. https://doi.org/10.1145/1056808.1057052

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.