IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 16 October 2024, accepted 27 November 2024, date of publication 9 December 2024, date of current version 23 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3513712

RESEARCH ARTICLE

Non-Visual Interfaces for Visual Learners: Multisensory Learning of Graphic Primitives

STACY A. DOORE[1], (Member, IEEE), JUSTIN R. BROWN[2], SAKI IMAI[1,3], JUSTIN K. DIMMEL[4], AND NICHOLAS A. GIUDICE[2,5]

[1]INSITE Laboratory, Department of Computer Science, Colby College, Waterville, ME 04901, USA
[2]VEMI Laboratory, The University of Maine, Orono, ME 04469, USA
[3]Khoury College of Computer Sciences, Northeastern University, Boston, ME 02115, USA
[4]IMRE Laboratory, School of Learning and Teaching, The University of Maine, Orono, ME 04469, USA
[5]Spatial Computing Program, School of Computing and Information Science, The University of Maine, Orono, ME 04469, USA

Corresponding author: Stacy A. Doore (sadoore@colby.edu)

This work was supported by the National Science Foundation under Grant DRL 2048394 and Grant IIS 1822800.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by The University of Maine IRB under Application No. 2018-06-16.

ABSTRACT Multimodal learning systems have been found to be effective in studies investigating cognitive theory of multimedia learning. Yet this research is rarely put into practice in Science, Technology, Engineering, and Math (STEM) learning environments, which are dominated by visual graphics. Introducing multimodal learning systems into STEM settings and allowing students to access dual channel cues beyond visual perception may help more students process information in their preferred modality. The purpose of this study was to investigate the usability, effectiveness, and design of multimodal interfaces for enhancing access to graphical representations. We used existing theories of multisensory information processing to study how sighted participants could learn and interpret spatial primitives and graphical concepts presented via three non-visual conditions: natural language (NL) descriptions, haptic renderings, and a NL-Haptic combination. The results showed that access to haptic-only renderings produced the least accurate responses, whereas NL descriptions with and without haptics led to similar performance by participants when learning graphical content without vision. Performance was also impacted by the complexity of the graphical content, with the highest level of accuracy observed for closed forms, compared to paired line segments and line/polygon intersections. We argue that universally designed, multimodal learning environments can transcend traditional, visual diagrams by utilizing non-visual channels and commercial hardware to support learners with different sensory abilities, preferences, and processing needs. Findings contribute to extending theoretical insights of non-visual information processing to better understand multisensory learning in sighted individuals.

INDEX TERMS Education and learning interfaces, multimodal interfaces, interaction design, multisensory information, non-visual information access.

I. INTRODUCTION

Advanced information interpretation skills will be needed for the successful implementation of AI-based decision support systems. One result of the AI information revolution is the automated aggregation of raw data into many different types of graphical representation, i.e. graphs, charts, diagrams, etc. New techniques are needed to accurately convey AI-generated data visualizations for humans in the loop to evaluate the system's recommended actions. As human-AI communication will increasingly occur through multiple channels (i.e., multimodal environments), all learners will need new skills and strategies to interpret and draw conclusions about graphical data. A significant limitation of current data aggregation and communication (manual or AI-based), is the bias towards focusing on visual representations

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

of information. There are many inclusive settings where multimodal user interfaces (MUIs) may present additional support for data interpretation beyond just blind and low vision learners; i.e., the demographic that is conventionally discussed and studied when designing and evaluating multisensory access to visually-based graphical content. There is a gap in the research literature about the optimal presentation methods for the comprehension of graphical information in multimodal systems as the majority of extant studies are focused on individual sensory channels in isolation from other sensory channels [1], [2]. Likewise, few studies have investigated the interaction between modalities in the effective representation of graphical information with sighted participants, however, there are notable examples [3], [4]. This study is specifically interested in identifying the optimal combination of several non-visual modalities for communicating graphical representations to sighted individuals. To isolate the potential complexities introduced with traditional charts, diagrams, and graphs, we have focused our comparisons here on stimuli of spatial primitives (points, straight/curved lines, and open/closed regions) using several combinations of non-visual sensory information.

In this paper, we review several non-visual components of interactive multimodal interfaces and discuss how they may be used to promote Science, Technology, Engineering, and Math (STEM) learning, maximize attention, and facilitate information synthesis. Next, we report the results of our study that investigates the learning performance of blindfolded, sighted participants' interpreting graphical primitive stimuli under non-visual conditions: 1) two different unimodal conditions (vibrotactile and natural language descriptions) for conveying simple images, and 2) a synchronized bimodal condition (abbreviated overview description along with a vibrotactile representation). The motivation for the study was to provide a basic comparison of multisensory learning of non-visual stimuli in sighted individuals. Previous studies suggest that sighted individuals can learn graphical content non-visually, however, to our knowledge, research directly comparing the different non-visual conditions evaluated in this study has not been conducted. Likewise, evaluating the functional equivalence between these sensory inputs has not previously been studied with sighted participants. Removing vision from the equation was a necessary, intentional, and incremental study design decision to increase our understanding of non-visual learning modes. This is also the first step in evaluating the effectiveness of the unimodal/bimodal representations to be applied in interface designs for multimodal learning systems. The primary research questions addressed by this research are the following:

- RQ1: Can sighted individuals accurately learn and mentally represent basic STEM-based non-visual graphical content?
- RQ2: Which non-visual presentation methods (haptics alone, natural language (NL) descriptions alone, or bimodal combination of haptics and short NL overview) are most effective for building up a mental model in multimodal learning environments?
- RQ3: Does the type/complexity of graphical content impact performance as a function of presentation modality?

.

We hypothesized that learning performance would not differ as a function of information presentation condition or as a function of image complexity, as all presentations were designed to ensure that each conveyed the requisite information to support accurate learning. This information matching between presentation conditions is a key baseline criterion to meet when directly comparing learning outcomes between different sensory/information modalities. That is, it is important to ensure that the same information is available from all comparison channels to allow fair multimodal evaluations, as has been described in previous research on sensory substitution and multimodal learning [5], [6], [7], [8]. This study contributes new theoretical insights into how a multimodal learning system can augment the interpretation of STEM visual instructional graphics by leveraging non-visual channels.

II. RELATED WORK

A. MULTIMODAL LEARNING PRINCIPLES

The cognitive theory of multimedia learning (CTML) [9] provides guidance in the development of multimedia learning interfaces. Based on CTML dual channel theory [10], learning can take place over two information processing channels - e.g., visual / pictural channel, auditory / verbal channel, or haptics / tactile channel. However, this learning process is subject to capacity limitations, as learners can process only a limited amount of material in each channel at any given time, which means that reducing extraneous material is critical. Likewise, CTML's active processing assumption requires that learners must be engaged by handling key words to be processed in the verbal channel and identifying key graphics features to be processed in either of the other channels in working memory. The brain is optimized for the integration of simultaneous redundant input from separate modalities, allowing for parallel and modality-independent processing. This improves people's ability to develop and access concepts and mental models in a flexible way [11], [12].

Theories on the coordination of dual processing, or the role and timing of complementary and redundant sensory signals, have been applied to the design of multimodal interfaces [13]. This research shows that organizing words, visuals, and haptic/tactile sensations with one another, along with long-term memory-activated prior knowledge, helps learners to create connections between the information using carefully coordinated dual-modality cues. Compared to unimodal approaches, coordinated dual-modality signals support information processing that aligns spatially matched, temporally matched, or semantically matched visual or haptic/tactile information [14]. If dual channel cues are spatially or temporally mismatched, they present an extraneous overload that can hinder cognitive processing in working memory [15].

The functional equivalence hypothesis of spatial information [16], emphasizes the underlying similarity and perceptual salience of information that can be specified between multimodal channels. Since spatial information is common to multiple senses, learning the same spatial stimuli across different perceptual modalities (e.g., audition, touch, or vision)is possible. In addition, cognitively mediated input, such as spatial language, leads to the development of a unitary, amodal (sensory-independent) representation in the brain, called the spatial image. The spatial image supports functionally equivalent behavior -i.e., statistically indistinguishable performance, irrespective of the input source [8]. The functional equivalence hypothesis has been demonstrated with both sighted and blind participants in a variety of tasks and modalities. With respect to haptic comparisons, a similar study with blind and low vision participants showed equivalent performance when learning simple diagrams between natural language descriptions and haptically rendered vibrotactile images with a brief natural language overview [17]. Functional equivalence was also found with sighted participants when learning haptic and visual spatial arrays [18] and with both sighted and blind participants for haptic and visual learning of simple vibrotactile graphs [19] and maps [20]. Functional equivalence has been shown with both sighted and blind participants for spatial updating of target locations learned from spatialized audio and spatial NL conditions [16], and with sighted participants for vision and spatial NL conditions [21] and between visual, spatialized audio, and spatial NL conditions [22].

B. MULTIMODAL INTERFACES

The multimodal learning system evaluated in this study is designed to provide a variety of sensory channels for learners to perceive, interpret, and interact with stimuli. These types of systems often consist of a combination of interaction modalities, including vibrotactile / haptic, verbal descriptions, audio and sonification cues, and high-contrast visuals [23]. We designed our system in order to compare learning outcomes using non-visual channels (verbal natural language descriptions and vibrotactile/haptic, plus a combination of both modalities) because: 1) they can be employed using standard commercial smartphone hardware without expensive or specialized technologies [24], 2) they have already proven effective in learning environments in a variety of applications [25], and 3) they are not often studied together as non-visual learning support for sighted participants [26].

Multimodal information is rarely implemented as a primary UI in inclusive STEM learning environments [27]. This lack of application is in conflict with clear evidence-based support for their efficacy and empirically driven theories such as the cognitive theory of multimedia learning [28], [29]. Although multimodal user interfaces (MUIs) can lead to some improvement in task efficiency when compared to unimodal interfaces, their greatest advantage over using traditional unimodal approaches are that they lead to more reliable performance, greater precision (especially for spatial tasks), and superior ability to support individual learning preference through inclusive design (see [13], [30], [31]). While MUIs can incorporate both complementary and redundant channels of information and be used for supporting input and output interactions, the focus of most systems is on multimodal output, as is studied here. As such, we next provide an overview of vibrotactile representations of graphics and then a summary of verbal descriptions using natural language.

C. HAPTIC AND VIBROTACTILE INTERFACES

Traditional haptic perception of graphical content is done using embossed stimuli based on pressure-based mechanoreceptors. These receptors are innervated by movement and deformation of the skin as the user feels the stimuli, (e.g., feeling a graph, map, figure, etc.) that is produced using a tactile embosser, thermoform machine, or via heat sensitive swell paper [32], [33]. Force-feedback devices such as the PHANToM or a force-feedback Joystick are also used in many haptics studies [34]. In this study, we employ a recent form of dynamic haptic vibrotactile stimulation from a touchscreen vs. traditional pressure-based tactile or force feedback approaches. Advances in control of embedded vibration motors in commercial touchscreen-based devices (e.g., phones and tablets) provide a broad range of textures and vibration patterns that can be easily implemented and broadly deployed. These haptic engines are available on commercial hardware and do not require expensive purpose-built components that are the foundation of other force-feedback or pin-based haptic solutions. Designing a system for mobile devices that are already used by most learners and implemented in many STEM educational contexts allows for a more inclusive instructional setting. In addition, this approach has already garnered significant research, relating to both the psychophysical and usability parameters needed to design perceptually salient and usable vibration-based MUIs via the touchscreens of smart devices [35], [36]. The haptic interface has been shown to support accurate learning of similar graphic stimuli as are used here (e.g., oriented lines and shapes) by sighted and blind participants [19], [37].

Studies employing touchscreen-based vibrotactile interfaces have demonstrated that these non-visual signals can communicate many types of spatial information in different real-world contexts, including navigation and wayfinding [38], conveying and interpreting on-screen 2D images [39], and supporting mobile interface interactions [40], [41]. There is also substantial evidence to suggest that well-designed haptic representations are effective in helping individuals accurately perceive and follow line graphs, graph shapes, and graph patterns [19], [42]. Gorlewicz and colleagues [35] provides a set of recommendations for graphical images using haptic vibrations which include: 1) graphical elements with object line widths of at least 2mm for object edge/boundary detection and object line widths of at least 4mm to allow for object edge/boundary finger tracing, 2) any angled object lines should have a width of at least 4mm to increase detection accuracy, 3) there should be different haptic feedback signals for any points of significance, (e.g., endpoints, vertices, and inflection points), and 4) gap widths between lines and objects should be at least 4mm [35], [43]. In this study, we have used this collection of design principles as well as other recent related studies on the design of systems to increase access and interpretation of graphical information [17], [44].

D. NATURAL LANGUAGE INTERFACES

Another non-visual technique for representing graphical information is through natural language descriptions. Natural language (NL) is a term used in psychology, linguistics, and computer science for the communication and representation of any language that has evolved naturally in humans through use and diffusion [45]. Natural language text descriptions are often used by blind and low vision individuals using screen readers to access alt text descriptions for digital figures in documents, webpages, and e-books. However, sighted individuals also use NL information, even when they think they are relying on visual graphical access. Sighted learners often do not notice when they are provided additional support for learning graphical material in a multisensory environment. The theory of multimedia design emphasizes the importance of providing natural language (or narration) to align signaling cues and the temporal contiguity of information presentation [15].

While there is significant evidence that both longer natural language descriptions and shorter overviews help to support learning, the challenge is how to best structure and standardize these descriptions to be most useful. Similarly, the poor quality of many descriptions perpetuates unnecessary barriers for anyone who might benefit from well-formed and presented natural language descriptions. Often, if an alt text or auditory description is provided, it is incomplete, inaccurate, and/or does not adequately convey the complexity of graphical content to answer questions, make inferences, or draw conclusions [46]. This is problematic not only for people who rely on them (e.g. blind and low vision learners), but also for anyone else who is using NL as the primary interaction style for learning, as is the focus of the participants in this study. Significant effort was made in the protocol development to standardize the longer descriptions as well as the shorter overviews that served as the verbal stimuli. Specifically, these NL descriptions were based on terms used by STEM instructional experts to help with the construction of a mental image [47]. The descriptions provided comparable spatial information that was available in the verbal and vibrotactile representations of the images.

III. MULTIMODAL LEARNING SYSTEM

As described above, accessible technologies and UIs can greatly benefit from adopting multimodal design approaches. As such, the multimodal learning system described in this paper is an adaptation of the multimodal touchscreen-based system implemented by Giudice et al. [19]. The original system was developed using an Android touchscreen tablet and later rebuilt to iOS for use in studies with sighted and blind/low vision participants [17], [44].

The multimodal learning system used here leverages the built-in speakers and vibration motors found in an Apple iPhone XS. Experiment materials (see Experiment Stimuli) are designed with high-contrast, colored lines and points that are programmatically identified in the system and tagged with researcher-provided NL descriptions. Haptic feedback and natural language audio, customized within the iOS Swift program scripts, are activated through touch interactions on the touchscreen with a single finger. For example, placing a fingertip on a line segment (4mm width) yielded a constant vibration (230 Hz, intensity: 1.0, sharpness: 0.5) emitted from the vibration motor in the iPhone. Endpoints, vertices, or points of intersection were rendered as pulsing vibration patterns (230 Hz, intensity: 1.0, sharpness 0.5, 0.1 second intervals) that represented a discrete point. NL descriptions were activated by a double-tap gesture and read by the built-in iOS text-to-speech engine in their entirety or as brief overviews depending on the experimental condition. Touching the boundary zones produced auditory clicks, rendered as 'bars' placed at the top and bottom of the screen. These signals were used to identify the outer frame of the active window and also prevented the participant from accidentally closing out of the app or activating the notification bar. Line widths, gap widths, and vibration frequencies of the elements were chosen according to previously discussed design guidelines and have been well studied with blind and sighted participants to determine their perceptual salience and efficacy for rendering such basic elements [35], [36], [48], [49]. Figure 1 highlights the interactions described above with one of the high-contrast designs used in the study.

By leveraging these responsive touchscreen gestures, we can streamline user interactions with the multimodal learning system. This seamless activation of haptic feedback and NL descriptions allows us to employ all three presentation conditions–Haptic Only, Natural Language Only, and Haptic + Overview–within a single, unified system. This integration not only enhances the versatility of the system, but also provides a robust framework for examining the effects of different sensory inputs on information access and comprehension.

IV. STUDY DESIGN AND METHODS

The study consisted of a within-subjects design with two independent variables, information presentation (with three conditions) and image stimuli (with three experimental sets). Information presentation was manipulated in three different conditions (see Section IV-D): 1) Haptic Only, 2) Natural Language Only, and 3) Haptic + Overview. The image stimuli were grouped by similarity into four image sets, one set was used to practice the different presentation methods,

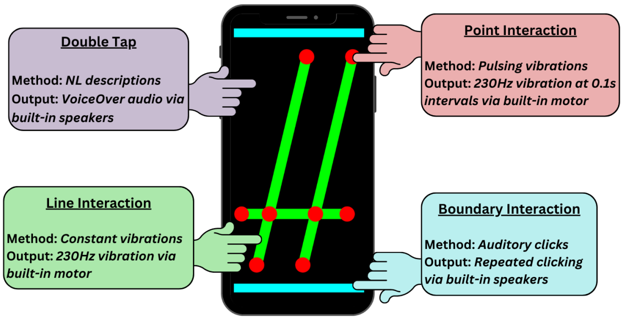
Double Tap
Method: NL descriptions
Output: VoiceOver audio via built-in speakers
Point Interaction
Method: Pulsing vibrations
Output: 230Hz vibration at 0.1s intervals via built-in motor
Line Interaction
Method: Constant vibrations
Output: 230Hz vibration via built-in motor
Boundary Interaction
Method: Auditory clicks
Output: Repeated clicking via built-in speakers

FIGURE 1. A diagram of the interactions used by participants within the study.

while the remaining three were used in the experimental analysis. Each image set contained three different image stimuli where one stimulus was used in each condition of the study. This means that each condition incorporates one stimulus from each set of images, totaling four stimuli per condition. The order in which stimuli and conditions were experienced was counterbalanced between participants and all participants completed the experimental trials while blindfolded.

The dependent variables within the study were response accuracy (captured using multiple-choice questions) and quality of open response descriptions. Accompanying each image stimulus were four multiple-choice questions. These questions probed different aspects of information extraction and interpretation, relating to information readability, spatial properties, and global relations. For example, when presented with two curved line segments (shown in Figure 3 (b)), the four questions included were:

- Q1) "What is the total number of endpoints, vertices, and points of intersection in this image?" [A. 2, B. 4, C. 6],
- Q2) "Which option best describes this image?" [A. Two straight line segments, B. Two curved line segments, C. One straight line segment and one curved line segment],
- Q3) "Which line segment is shortest?" [A. Left line segment, B. Right line segment, C. The line segments are of equal length], and
- Q4) "Where would the line segments intersect if they were extended beyond the screen?" [A. Beyond the top right edge, B. Beyond the bottom left edge, C. They would never intersect].

After the multiple-choice questions, participants were tasked with describing the image stimulus in their own words. This was meant to qualify the level of participant understanding and characterize the mental representation of the stimulus they learned. In summary, every participant experienced four image stimuli (one practice, three experimental) in each condition of the study, totaling 12 stimuli. They all answered a total of 36 multiple-choice questions and 9 open response questions from the experimental trials, which were used in the subsequent analyses.

A. PARTICIPANTS

Twenty-four sighted participants were recruited for this study through campus-wide email and flyers (self-identified F = 9 and M = 15, ages 18–35). The experiment took between 1–1.5 hours to complete. Participants were blindfolded for the duration of the study, excluding the initial training period. This study was approved by the University IRB; all participants signed an informed consent form and were compensated with a $30 Amazon gift card for their participation.

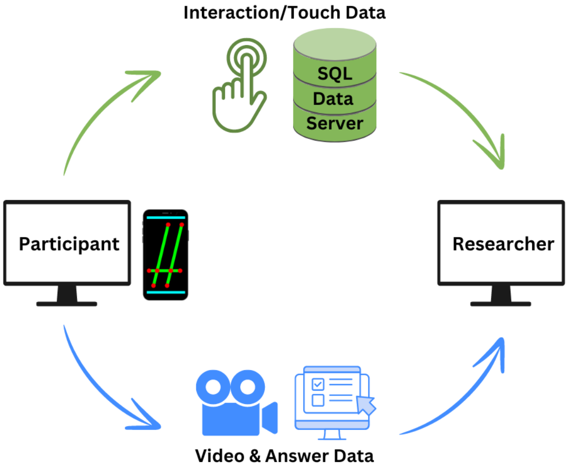

FIGURE 2. A diagram of the experimental data flow from the study.

B. EQUIPMENT

Two computer stations were required for the experimental setup in the lab to facilitate the study according to the pandemic safety protocols implemented at the time. At the participants' station, individuals sat at a table in front of an iMac (21.5-inch, Core i5, 2.7 model). An iPhone XS, with the experimental app pre-loaded and open, was plugged into the computer. This wired connection allowed the iPhone screen to be shared in real-time via Zoom with the researcher's computer while minimizing any end-to-end system lag. The iMac front-facing camera was turned on to monitor that participants' blindfolds stayed in place during the study. The microphone was also on to capture any audio from the experiment (e.g., participant responses to questions). The researcher's station was set up with an iMac (same model) that was logged in to Zoom. One side of the researcher's screen showed the iPhone XS screen-share and participant's camera feed in Zoom and the other side displayed a Qualtrics survey protocol used to record participant responses. This split-screen view allowed the researcher to proctor participants as they were using the app in real-time while simultaneously asking questions from the protocol for the experiment. The researcher recorded all experimental sessions via Zoom. Touch interactions were also recorded (timestamps, color, and x-y coordinates) on the iPhone and later uploaded to a SQL data server to be downloaded by the researcher conducting the experiment. Figure 2 depicts the flow of collected data from the study.

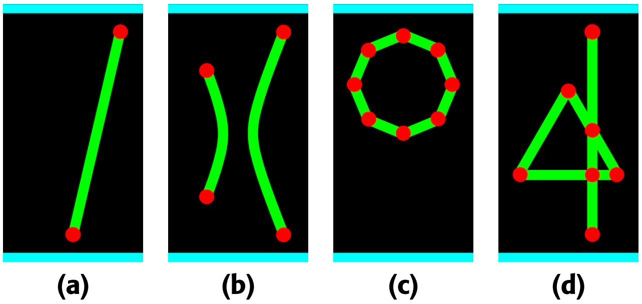

FIGURE 3. Example stimuli from each image set. (a) A single line segment from Practice Set (b) Paired curved lines from Image Set 1 (c) An octagon from Image Set 2 (d) A polygon/line intersection from Image Set 3.

C. EXPERIMENT STIMULI

The stimuli for this study were images representing spatial primitives, such as line segments with vertices, points of intersection, simple closed regions, and shapes. There were four image sets, each containing three unique stimuli that varied by the type and number of primitives in the stimuli (see Figure 3 for example images). The image sets increased in complexity as more spatial primitives were added to the stimuli, with the combined elements representing spatial relations such as intersections and parallels. The Practice Set contained single line segment images, Image Set 1 included closed form images, Image Set 2 consisted of images with pairs of line segments that do not intersect on screen, and Image Set 3 was the most complex and had images with lines and polygons that intersect on screen. The stimuli were designed by the experimenters to represent some of the most common primitive spatial relationships among points and curves that are realized in geometry diagrams [47]. The stimuli were not designed to be representative of all possible primitive spatial relationships that are realized in STEM diagrams, but rather to capture a range of the diagrammatic objects that are encountered in high school geometry classes [50].

D. EXPERIMENT CONDITIONS

1) HAPTIC ONLY CONDITION

Participants were presented with the stimuli on the iPhone screen (see Figure 4). By moving their hand around the screen and feeling the elements via vibrations on one finger, participants learned the image and constructed a mental model of the represented information. As described previously, line segments produced constant vibrations (230Hz) and points produced pulsating vibrations (230Hz, 0.1s intervals) that could be traced with search strategies introduced to the participant by the researcher during the practice session. These exploratory procedures/strategies build on work by Klatzky and Lederman using traditional pressure-based haptics [51], [52] but are optimized for use with vibrotactile stimuli on touchscreens based on participant observation data from prior work. For example, using zigzag finger movements for line tracing, employing finger circling and four-directional scanning at intersections, and using the device boundary as a reference when following global contours [36], [44], [53].

2) NATURAL LANGUAGE CONDITION

When participants double-tapped the iPhone screen in this condition, the app read aloud a description of the image using the built-in iPhone VoiceOver feature. There was no haptic interaction other than tapping the screen to initiate the NL description. The carefully constructed descriptions consisted of information such as the location of endpoints, the curvature of line segments, and in some cases the specific shapes and/or global placement of the figure relative to the screen. To compare measures of functional equivalence between modal presentation, the information provided was matched between conditions, meaning that all information in one modality was also conveyed by the others, which is a critical controlling factor for inter-modal comparisons [8].

3) HAPTIC + OVERVIEW CONDITION

This condition employed a combination of the other two methods. Participants could feel the images (as described in the Haptic Only condition) and could also listen to a brief Natural Language overview description. The NL overview descriptions were always the first sentence of the full descriptions used in the NL condition for each image. The inclusion of the overview description is based on growing support for the critical nature of summary information to guide user attention and haptic search behavior [44], [54].

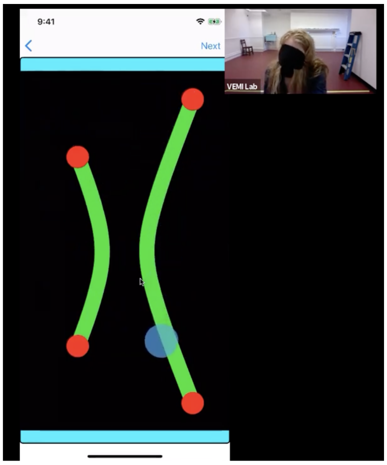

FIGURE 4. View of stimulus and single finger exploration in haptic only condition.

E. PROCEDURE

The first trial was used as a practice for each participant and involved a randomly selected image from the practice set, experienced under one of the three presentation conditions. These images were used as practice because they were the least complex. Conversely, the fourth trial was always a randomly selected image from Image Set 3 since they were the most complex in terms of number of elements rendered on the screen. We did not have a priori rationale for Image Sets 1 and 2 being more difficult than one another. As such, half of the participants had the image set order P, 1, 2, 3 and the other half had P, 2, 1, 3. None of the images in the stimuli set were repeated and each participant was exposed to all twelve images over the course of the study.

The study began with a sighted practice phase where participants were instructed on how to navigate the sequence menu, select buttons to begin trials, and access images in all three presentation conditions. For the Haptic Only condition, the researcher explained the different vibration patterns for image elements and gave some examples of best practice strategies for exploring the stimuli. In the Natural Language condition, the researcher trained the participant on how to start and stop the audio descriptions. There was no limit on how many times the description could be repeated. The participants also practiced the Haptic + Overview condition, which involved actions previously introduced in the other two conditions. After they finished the sighted practice, they were blindfolded and completed the practice phase again with the three conditions to ensure that they had a baseline understanding of how each condition functioned. The participants remained blindfolded for the rest of the experiment. Once the experimental sequence started, the participant was told which presentation condition to expect and was given a reminder of how to access the images in that condition.

The participants were informed that they would be answering five questions related to each image stimulus. The first four questions were multiple-choice and the last question was open response. The multiple-choice questions had three possible answers. This design choice was based on evidence that the three-option versions of multiple choice questions are usually as robust, if not optimal, over the four- or five-option versions [55], [56]. The questions and possible answers were read aloud, and the participants vocalized their answers. The first question was asked at the beginning of each trial/learning of an image. Subsequent questions were asked immediately following participants' responses to the prior question. As this was meant to be a perceptual task, rather than memory-based, they could simultaneously access the stimuli as they answered the questions. There was no time limit, but they were encouraged to answer as quickly and accurately as possible. They could also ask for questions and answer options to be repeated. The researcher recorded the answers to the four multiple-choice questions in Qualtrics, the device recorded the participants' iPhone interactions, and a video of the session recorded participant description of the perceived graphic using Zoom's record and transcript mode. This last question asked participants to describe their mental model of the image as a re-creation task. It is not unusual in perception studies for researchers to ask sighted participants to draw a representation of their mental model of an image, diagram, or map to evaluate learning of stimuli. However, given the length of time needed for this experiment and wanting to avoid the removal of the blindfold to record the participants' drawings for each image, we opted for a more streamlined process of recording their verbal description of the perceived image. This approach also allowed us to determine if they would merely repeat the terms used in the NL description or generate their own output representation of their mental model for the description.

F. ANALYSIS METHODS

The independent variables were Information Presentation Condition (3 levels: Natural Language (NL), Haptic Only, and Haptic + Overview) and Image Set (3 levels: Set 1 = closed forms, Set 2 = paired line segments, and Set 3 = line/polygon intersections). The dependent variable was response accuracy, measured by the multiple-choice questions, as percent of overall correct and incorrect responses. Response accuracy was analyzed using a two-way repeated measures ANOVA at a (95% confidence level ($\alpha = 0.05$)) to identify effects from manipulating the two independent variables. A post hoc power analysis with 2 independent variables, $\alpha = 0.05$, $N = 24$, and effect size = 0.303 for the ANOVA achieved a power of 0.87. A power analysis result over 0.8 supports the likelihood that the test is correctly rejecting the null hypothesis for a sample of this size [57]. Accuracy data were not analyzed for the Practice Set because the researcher gave feedback based on the participant's initial performance.

Analysis of the open response image descriptions was conducted using thematic analysis methodology, which offered a framework and set of practices for examining qualitative data and for describing potential themes, patterns, and trends within diverse phenomena situated in the context of the domain [58]. We developed first order coding processes from emergent, open coding methods [59] but the data categorization followed a deductive process for second order coding. This yielded 288 descriptions collected from 24 participants. We then collapsed the categories to Description Accuracy (correct/sufficient details, partially correct/insufficient detail, incorrect/no or wrong detail) and Misconception Types (insufficient detail, wrong spatial relation, wrong understanding of image). Two researchers labeled each description independently, and then resolved any disagreement through discussion [60]. This iterative process helped us to draw comparisons between description sentences containing explicit spatial and/or mathematical concept information that would be useful for forming a mental model of the image - e.g.,*"There is a line that goes from the bottom to the top of the screen"* to descriptions conveying a reference to a visually similar object to the image - e.g., *"It looks like a nine towards the top"*, *"It's like a hand* *fan facing left", "It has almost a butterfly shape that doesn't touch in the center".*

V. RESULTS

A. ACCURACY BY CONDITION

The overall grand mean accuracy, based on the correctness of 864 multiple-choice questions analyzed across all participants, conditions, and image sets was 72.45% (SE = 2.12%). By condition, participants were least successful with the Haptic Only condition with an average accuracy of 52.08% (SE = 3.37%). Performance significantly increased for the Haptic + Overview and Natural Language conditions with an average accuracy of 80.56% (SE = 3.12%) and 84.72% (SE = 2.73%) respectively (see Figure 5). The ANOVA test revealed that information presentation condition had a significant main effect on accuracy [F(2, 22) = 42.97, $p < 0.001$, $\eta_p^2 = 0.796$]. With a Bonferroni post hoc t-test pairwise comparison, the Haptic Only condition yielded a statistically lower performance from the Haptic + Overview (MD = -28.47%, SE = 4.10%, $p < 0.001$) and the Natural Language (MD = −32.64%, SE = 3.61%, $p < 0.001$) conditions, while the Haptic + Overview and Natural Language conditions did not statistically differ (MD = −4.17%, SE = 3.92%, $p = 0.896$). These results indicate that participants had significantly more difficulty with non-visual learning of the stimuli through the Haptic Only condition as compared to the Haptic + Overview and Natural Language conditions.

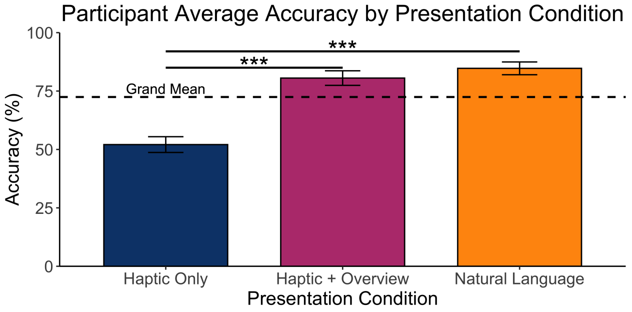

FIGURE 5. A bar graph of participant average accuracy for selected response questions by presentation condition (Haptic Only, Haptic + Overview, and Natural Language).

B. ACCURACY BY IMAGE STIMULI SET

By image set, collapsing across presentation condition, average accuracy for closed forms (Set 1) was 80.56% (SE = 1.85%), paired line segments (Set 2) was 67.01% (SE = 2.99%), and line/polygon intersections (Set 3) was 69.79% (SE = 3.36%). The ANOVA test revealed that image set had a significant main effect on accuracy [F(2, 22) = 14.61, $p < 0.001$, $\eta_p^2 = 0.570$]. Post hoc t-tests with Bonferroni correction revealed that closed forms were significantly better than both paired lines (MD = 13.54%, SE = 2.92%, $p < 0.001$) and line/polygon intersections (MD = 10.76%, SE = 2.86%, $p = 0.003$). While paired lines and line/polygon intersections were not statistically different (MD = -2.78%, SE = 3.71%, $p = 1.0$).

C. ACCURACY BY CONDITION AND IMAGE STIMULI SET

Looking at the interaction between condition and image set, no significant interaction effect was detected [F(4, 20) = 2.17, $p = 0.11$, $\eta_p^2 = 0.303$]. However, a significant simple effect of condition on image set was observed; information presentation condition had a significant effect in each image set - all with significance values less than 0.05 ($p < 0.001$). For closed forms, participants had an average accuracy of 59.38% (SE = 4.71%) with Haptic Only, 86.46% (SE = 3.68%) with Haptic + Overview, and 95.83% (SE = 1.94%) with Natural Language. For paired lines, participants had an average accuracy of 43.75% (SE = 5.27%) with Haptic Only, 78.13% (SE = 4.59%) with Haptic + Overview, and 79.17% (SE = 4.43%) with Natural Language. For line/polygon intersections, participants had an average accuracy of 53.13% (SE = 4.83%) with Haptic Only, 77.08% (SE = 4.23%) with Haptic + Overview, and 79.17% (SE = 4.68%) with Natural Language. These data, along with significance levels between the conditions within each image set, are visually represented in the clustered bar graph in Figure 6.

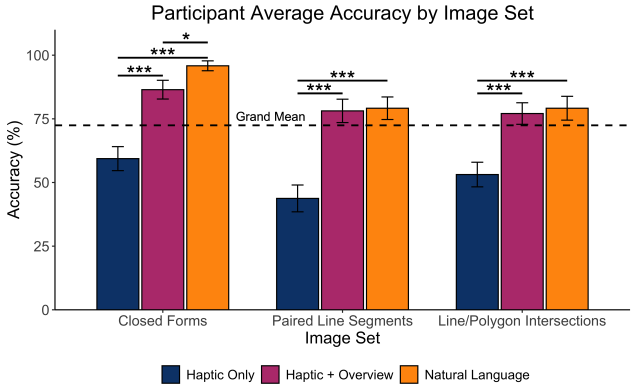

FIGURE 6. A bar graph of participant average accuracy for selected response questions by presentation condition (Haptic Only, Haptic + Overview, and Natural Language) and image stimuli set (Closed Forms, Paired Line Segments, and Line/Polygon Intersections).

D. POTENTIAL OF ORDER EFFECTS

The study was counterbalanced by alternating the order of the presentation conditions and the image sets experienced by the participants to minimize learning effects. When accuracy was grouped by block order (first, second, and third), the three blocks had averages of 71.53% (SE = 2.68%), 73.26% (SE = 3.19%), and 72.57% (SE = 2.68%) respectively with no significant effects of block order on accuracy [F(2,36) = 0.097, $p = 0.908$, $\eta_p^2 = 0.005$].

Given the consistent differences between the two haptic conditions, we were curious if the order in which the conditions were experienced influenced the accuracy of the subsequent haptic condition. That is, did participants who experienced the Haptic Only condition prior to the Haptic + Overview condition perform similarly to participants who had the reverse order? This between-subjects variable had 12 participants who experienced Haptic Only first and 12 participants who experienced Haptic + Overview first. The average accuracy for participants that had Haptic Only first was 65.28% (SE = 3.63%) and the average accuracy for participants that had Haptic + Overview first was 67.36% (SE = 3.63%). With similar accuracy, this suggests that there was not a significant difference between the two conditions based on the order the condition was experienced [F(1,22) = 0.165, $p = 0.689$, $\eta_p^2 = 0.007$]. With no significant difference in presentation order and a significant difference in average accuracy, we interpret that the simple NL overview, which is the element manipulated between the two haptic conditions, impacted participant performance rather than the condition order.

E. OPEN RESPONSE DESCRIPTIONS

We collected 24 open response descriptions for each of the image sets across each condition. In total, this resulted in 288 individual descriptions consisting of 524 natural language sentences or sentence fragments.

1) DESCRIPTION ACCURACY CLASSIFICATION

Looking at the description results, the Natural Language condition achieved the highest proportion of recreation descriptions coded as accurate, with almost 2/3rd of descriptions being coded as accurate for Image Sets 1 and 3. The Haptic Only condition achieved the lowest proportion of descriptions coded as accurate across all image sets. In the Haptic + Overview condition, no more than half of the descriptions were coded as accurate (see Table 1). A set of chi-square tests for goodness of fit were conducted to see if an observed frequency distribution of an accuracy rating matched an expected frequency distribution. In this case, there were 96 descriptions collected for each of the three presentation conditions across all image sets, totaling 288 descriptions. Under the null hypothesis for a chi-square goodness of fit test, it is assumed that the descriptions are equally distributed across all accuracy ranks and conditions. Thus a 3 × 3 table (Table 1) with 9 cells will have an expected value of 32 descriptions per accuracy rank.

TABLE 1. Observed and Expected Accuracy Classification per Condition.

	Haptic Only	Haptic + Overview	Natural Language	Expected
Accurate (A)	20	38	56	32
Insufficient (I)	29	48	36	32
Wrong (W)	47	10	4	32

A chi-square test of goodness of fit was performed to examine the relation between each individual presentation condition and the description accuracy rank (A, I, W). The relation between these variables was significant at a $p < 0.05$. [Haptic Only X2 (1, $N = 96$) = 11.812, p =.00272. Haptic + Overview X2 (1, $N = 96$) = 24.250, p =.00001. NL X2 (1, $N = 96$) = 43.0, p =.00001]. A chi-square test of independence was also performed to examine the relation between all three presentation conditions and the level of description accuracy. The relation between these variables was significant with a large effect size, [X2 (2, $N = 288$) = 75.2995, p =.00001, V = 0.3615].

In summary, the Natural Language image descriptions were significantly more likely than Haptic + Overview or Haptic Only to produce accurate descriptions. Haptic + Overview were significantly more likely to produce insufficient descriptions and Haptic Only was significantly more likely to produce inaccurate descriptions. The participant generated recreation descriptions (vs. the researcher generated unimodal NL stimuli in the protocol) are used here as a form of reconstruction of the stimuli. It served as a way to evaluate the amount and type of information conveyed by the unimodal/bimodal representations as well as the participant's underlying cognitive representation of the learned stimuli. Image Set 3 descriptions consistently had a greater proportion of accurate codes across modality conditions. This was an unexpected finding given that as image complexity increased, we predicted that the number of descriptions coded as accurate would decrease across all three conditions. Overall, Image Set 2 descriptions had fewer accurate codes than the other image set descriptions. This finding was also surprising as we expected that these images would be easier to interpret because they were almost parallel lines as opposed to different lines or shapes.

It should be noted that although the researchers had developed a set of 12 extended NL descriptions designed to follow existing linguistics-based guidelines and research for describing visual images for non-visual access, participants did not 'parrot' back the provided descriptions when asked to give their own description. Based on the video observations, these data reflected the participants' genuine efforts to produce output reflecting their mental model vs a simple recall of the exact wording or terms used in the NL condition or the Haptic + Overview conditions. This finding suggests that participants were attempting to provide novel descriptions to verbally convey their mental representations built up from learning, as was our intent.

2) INCORRECT DESCRIPTION CLASSIFICATION

All of the description responses were coded as accurate-detailed (A), accurate but insufficient detail (ID), wrong-spatial relations (WSD), or wrong-image understanding (WU). The Natural Language condition descriptions consistently produced a significant number of accurate responses, however there were many that fell into the insufficient detail category (>50%). This may be due to not receiving additional spatial information from the visual or tactile channels and therefore needing to rely on working memory to form a mental model of the global image.

The Haptic + Overview descriptions were mixed in terms of insufficient/incorrect response codes based on the image set. For example, the frequency of incorrect responses in the category of wrong understanding of the image fluctuates between a high of 50% of incorrect responses (Image Set 1) to a low 5% of incorrect responses (Image Set 3). This last observation is surprising because we predicted that Image Set 3 would be the most difficult to describe due to the combination of lines, shapes and intersections, hence the highest level of complexity. However, based on the results of the analysis, Image Set 3 resulted in the fewest number of codes that were evaluated to be wrong due to a complete misunderstanding of the image.

The Haptic Only descriptions consistently produced the greatest percentage of incorrect responses based on a wrong understanding of image across all four image sets. The earlier comparison of responses for Haptic Only and Haptic + Overview would suggest there was not a learning effect. There are several other alternative explanations for this pattern. One explanation could be the use of sighted blindfolded participants who are not familiar with processing haptic information into mental models. We find this explanation unlikely given success of sighted people on other haptic learning tasks/stimuli [18] and with both sighted and blind participants for haptic and visual learning of simple vibrotactile graphs [19], [20]. Another explanation could be the misalignment of asking participants to take information processed through a haptic input and provide a natural language format output as a replication task. In the next section, we further explore these results.

VI. DISCUSSION

Well-designed multimodal learning systems have been demonstrated to benefit students with a variety of sensory needs and preferences [61], [62]. Yet, there are few studies that directly compare the ways in which non-visual presentation methods might help to improve universally designed multimodal learning systems with sighted participants. This paper is part of a long-term research program to investigate non-visual information processing abilities in sighted individuals as an important theoretical contribution for its own sake and not just serving as a control for research into non-visual representations for blind and low vision learners. Our contribution to the body of literature on multimodal learning interfaces is the comparison of the effectiveness of three types of non-visual information presentation modes for sighted participants learning basic graphical primitives. In the next sections, we discuss our findings based on the research questions and how they can be used to better understand the ways in which the multimodal interfaces providing input through non-visual channels may substantively contribute to learning graphical STEM content.

A. LEARNING ACCURACY

The first research question focused on the ability of blindfolded sighted participants to learn spatial primitives using the three non-visual interface conditions as measured by response accuracy. Response accuracy was measured by the number of correct selected responses to the multiple choice questions and by the accuracy of the participant verbal recreation descriptions of their mental model of each image stimuli.

RQ1: Can sighted individuals accurately learn and mentally represent basic STEM-based non-visual graphical content?

The selected response results suggest that sighted participants were able to effectively learn the graphical primitives using two of the three conditions (Haptic + Overview and Natural Language). The Natural Language condition had the numerically highest overall level in response accuracy of the three conditions. In the NL condition, participants also produced more accurate and detailed descriptions in each of the three image sets. The accuracy of the NL condition descriptions were consistent for both the least complex (single polygon (Image Set 1)) and the most complex images (line/polygon intersections (Image Set 3)). Of interest, the difference in response accuracy for the NL condition was not statistically different from the response accuracy in the Haptic + Overview condition. This outcome provides evidence in support of the functional similarity of behavioral performance when learning with both non-visual graphical presentation modes.

The Haptic Only condition produced the least accurate descriptions of the three conditions across all image sets and had the lowest overall accuracy compared to the other two presentation conditions. Over half of the descriptions that were generated based on the Haptic Only condition resulted in inaccurate descriptions of the image stimuli. There are several possible interpretations of these results. It is possible that participants did not accurately learn the graphical information from the Haptic only condition. Some of the descriptions of this condition point to such a conclusion. For instance, below are examples of inaccurate descriptions for the image of the two parallel curved segments (Image Set 2) in the Haptic only condition:

> "Like a capital D" (P5)
> "It's sort of like an oval-ish." (P14)
> "I guess C? The letter?" (P11)
> "There is a horizontal (line) on the bottom, and there is a diagonal (line) going this way…starting at the bottom and going to the top right." (P24)

These were coded as inaccurate descriptions of the image because they do not faithfully represent the learned stimuli (wrong-image understanding/wrong spatial relations). The first two examples add information to the image that was not there (e.g., mentally adding a vertical line to the curved lines because it fits a prior knowledge schema of a capital D or oval). The next two are examples of inaccurate spatial relations (reversed from the letter C, only 1 line vs. 2 lines, and thinking one line was horizontal and the other diagonal vs two parallel curved lines).

Another possible interpretation is that the Haptic Only condition did not lead to fully elaborated mental representations.

In other words, the mental model may have only been partially formed based on the information provided in the Haptic Only condition. Examining the Haptic Only descriptions for the same image stimuli that were coded as accurate but insufficient detail and comparing them to the above inaccurate description provides some support for this interpretation.

> *"Curved line segments from the bottom left up towards the top right, but not, like, all of the way, like top middle I guess. And then from right to the right bottom as well."* (P9)

> *"So there's a line that starts at the bottom right corner. and it curves in towards the center of the screen. and then goes upwards and kind of just to the top of the screen. and then there's a second line that's on the left that curves, sharply into the center, sharply out towards the top left corner of the screen."* (P19)

In these examples of descriptions with accurate but missing details, the participants did not mention the parallelism of the two curved lines which was coded as one of the defining features or geometric properties of the image. One might ask if perhaps the Haptic Only condition was disadvantaged in some way in the recreation task that asked for the mental model to be represented as a NL response? However, although there was a lower rate of learning accuracy for the Haptic Only condition, there were instances of both accurate and complete detailed mental models. The following description in the Haptic Only condition represents an example that was coded as accurate-detailed based on that it provided the key attributes (2 lines, curves, parallelism, vertical orientation, centered image).

> *"Two parallel lines going from the center… close to the center bottom of the screen towards center right of the screen."*(P1)

The above description suggests that even in the Haptic Only condition, it was possible for sighted participants to interpret the image accurately, however, that this was not the norm for our sample as compared to the other two non-visual presentation conditions. Our results provide new evidence that non-visual information modes can be equally learned, accurately represented in memory, and acted upon in highly similar, statistically identical ways by sighted participants, i.e., the functional equivalence hypothesis of spatial information [16]. Given the reliably worse performance of the Haptic Only condition, our findings do not provide across the board support for functional equivalence or development of a common amodal spatial image [8]. These theories emphasize the importance of the information being compared, e.g., that all input modalities must convey the same information and that sufficient learning is allocated with each input for the potential of functional equivalence. While more research is needed in this domain, it is possible that the haptic-only condition would lead to equivalent performance given additional learning time or more information, e.g., as is provided in the Haptic+Overview condition that showed markedly improved performance.

This study contributes additional support for the impact of layering of multiple sensory channels. Participants learning in the Haptics Only condition produced the most incorrect answers to the multiple-choice questions and the recreation descriptions. In contrast, learning in the Haptic + Overview condition performed at a functionally equivalent level to the Natural Language condition. The first possible explanation is the participants' lack of familiarity with interpreting graph information through vibrotactile channels. Although most people use haptics in their phones on a daily basis, these are often single cues, or attention focusing signals (i.e., incoming calls in silent mode, or mobile game interactions). Sighted users rarely use haptic information for data extraction, tracing, or learning as a primary interaction style, especially using touchscreens as we employed here. Without an organizing schema for the haptic representations, participants were left to interpret the information based on inconsistent or inadequate search patterns. The Haptic + Overview condition's use of a NL image summary likely provided just enough organizational information to activate a previously learned schema to construct spatial images - e.g., *"This is a circle" or "Two parallel line segments"*. This mental schema inevitably assisted in more effective search behavior to extract the relevant data, were used to help reason through the selected response questions and provided some globally-coherent spatial information to assist in completing the recreation descriptions task. It is also possible that the learning accuracy and description recreation results in the Natural Language condition were the strongest of the three conditions because the combinations of the simple points, lines, and regions were within the amount of spatial information that could be held in short term memory to form a spatial image from which to reason in and answer the multiple-choice questions and recreate a description in the same modality as it was originally given.

B. METHOD EFFECTIVENESS

The second research question focused on the effectiveness of each non-visual interface in building sufficient mental models that allowed the participants to answer questions about the representations.

RQ2: Which non-visual presentation methods are most effective for building up mental models in multimodal learning environments?

Although functional equivalence has not been studied after learning the same stimuli used here, or evaluated when directly comparing NL and Haptic conditions, the theory would suggest that we should observe functionally equivalent performance given our emphasis on information-matching between conditions. Findings showing a lack of statistically reliable differences (i.e., null results) between the Natural Language and Haptic + Overview presentation modes would corroborate this hypothesis and suggest that these two non-visual inputs are equally sufficient for supporting the spatial reasoning tasks required in the selected response questions. Yet, while the Haptic + Overview condition produced functionally equivalent results, there was a significant difference in the accuracy and level of detail in the recreation descriptions between the Natural Language and Haptic + Overview conditions. This finding may be due to participants having just enough information with the brief overview and the haptic signals to construct a mental image, however, still be missing information to complete the full mental image to describe accurately or with sufficient detail. There are several ways to potentially test this interpretation in a future study, such as asking participants to draw the mental image or to perform a matching task with a number of similar images to select from after they have learned the image in each condition, e.g., an alternative forced choice task.

In this study, we do not include a full description of the image stimuli during the Haptic + Overview condition. Instead, we provide a brief description of the type of image stimuli and the context that serves more as an organizing caption than a description. This overview acts as an auditory cue to provide a schema prompt for the haptics information the learner will be experiencing. The results of this study suggest that this 'image summary' does provide enough information for the participant to answer the multiple choice questions about the spatial primitives in a functionally equivalent manner to the extended NL condition through the organization of effective search strategies that assist in extraction and interpretation of the spatial information and the construction of a spatial image in memory. There is significant research support for the efficacy of auditory summaries from theories of multimedia learning [15] and that are demonstrated in real world settings such as providing auditory overviews for navigation routes [54], [63] as well as our own experience of having participant feedback informally request these types of overviews when we have conducted previous studies using systems with touchscreen graphical access. There is a distinct advantage for the layering of spatial information within multimodal systems so that perhaps missing information in a single modality can be filled in using additional channels. Since we are comparing novel stimulus-modality pairings without clear precedent, the contributions of this study provide important new theoretical insights into how non-visual modalities can be used to support spatial learning. This will help guide future implementation of multimodal learning environments, maximizing inclusive graphical information access and the best sensory canvas available to both learners and educators.

C. IMAGE SET IMPACT

The final research question focused on the impact of graphical content complexity on participant performance in each of the three non-visual conditions.

RQ3: Does the type/complexity of the graphical content impact performance as a function of presentation modality?

This study focused on stimuli consisting of spatial primitives (points, lines, regions/polygons) based on the rationale that they are the building blocks of STEM graphical information (charts, graphs, maps). We were careful to introduce all the presentation conditions in a practice image set (single line segments) to ensure participants understood how the experimental stimuli would be presented in each condition, the actions needed to review the stimuli and answer questions, as well as how to revisit the stimuli to complete the question tasks before starting the experimental trials. We also provided sufficient and matched information between our three presentation conditions to ensure some level of baseline learning between conditions and counterbalanced the image sets to prevent any ordering effect. Our results suggest that the complexity of the image sets had a simple effect on the accuracy of the multiple-choice questions.

Participants were most successful learning closed forms (Set 1) with the Haptic + Overview and Natural Language conditions. The Haptic Only condition significantly under-performed across the three image sets. We conclude that participant performance was not affected by the order in which the condition was experienced and the Haptic + Overview condition did not influence accuracy for the Haptic Only condition and vice versa. These results further support the benefits of layering different types of spatial information to help construct spatial images from the simplest forms to more complex combined element representations. It is possible that although the Natural Language condition produced similar accuracy (functionally equivalent to Haptic + Overview) across all image sets, the image set complexity may not have been enough to reach the limits of cognitive load. Additional studies on the two functionally equivalent conditions with images of increased complexity (i.e., simple graphs and charts) would help determine the threshold ability for reasoning on spatial information through natural language and the impact of additional information layered through other intrinsically spatial modalities such as haptics and vibrotactile interfaces.

D. MULTIMODAL LEARNING SYSTEM FUNCTIONALITY

While the primary research questions focused on how well blindfolded sighted participants were able to receive, interpret, and respond to primitive graphical representations using only vibrotactile and natural language input, there are a few issues to discuss about functionality of the multimodal learning system itself. The system described in this paper was an earlier version of the system evaluated and reported in [17] and as such was revised based on research observations of participant touch screen search strategies and qualitative feedback collected from participants during the experiment sessions. For example, there were times that participants could accidentally exit the system by pushing the wrong buttons on the side of the phone while holding it to explore the screen. This was problematic because the researcher would have to leave their observation station, restart the system, and navigate the correct part of the protocol. In other cases, participants reported that the lack of a boundary about the representation on the screen made it difficult to understand where the target images extended or were oriented. These observations and interface design feedback were then incorporated into the update of the system for subsequent experiments. Further evaluation of the system functionality for sighted participants will be reported in a future paper on using the system for representing chart graphics with vibrotactile and natural language non-visual interfaces.

VII. LIMITATIONS AND FUTURE WORK

Although precautions were taken in the study design to: 1) provide the same information in all information presentation conditions in order to answer the same questions for each individual image stimuli, and 2) counterbalance the presentation condition and image set order, the participants were allowed to review the stimuli as often as they wanted during the experiment. It is possible that the Natural Language condition was somehow advantaged, thus producing the highest levels of accuracy. We would need to further investigate if there was a significant benefit for participants to revisit and learn the stimuli in the NL condition vs the Haptic Only condition. This may be true for the description accuracy results as there was an alignment of stimuli and recreation tasks (NL stimuli to NL description). In a future study, we could take an opposite approach and ask participants to draw the spatial image with their finger on a touchscreen and capture the finger trace. We could then compare the image recreation in a closer format to the haptics modality. In this experiment, there was no way to do both of these recreation tasks without exceeding the reasonable amount of time (60-90 minutes) for the participant to be engaged in the study trials. If the alignment of the recreation format truly had an impact, the haptic recreation would benefit and produce more accurate recreations in that modality and a deficit would be seen in the NL learning mode. Another interesting question would be to look at the recreation task from the perspective of spatial information organization, analyzing recreation descriptions for part/whole schema, frame of reference, spatial relationships, and spatial information ordering. It is possible that an intrinsically spatial analyses would be easier and more accurate when done after learning from a Haptic Only input modality in a layered multimodal system. Future studies will investigate these two effective non-visual learning conditions with graphical images of increased complexity, such as simple STEM graphs, charts, and maps, to study whether functionally equivalent results persist as the amount of spatial information increases.

VIII. CONCLUSION

This paper investigated the ability of individuals to learn simple spatial information through non-visual channels in three different presentation conditions: haptic only renderings, natural language descriptions, and haptic with a short natural language overview. The results of this study provide evidence for the ability of sighted individuals to effectively use non-visual methods in learning graphical spatial information, with at least two of our three non-visual presentation channels supporting accurate and similar learning outcomes. Results of the study also provides some support for the functional equivalence theory, as two of the three presentation conditions (Natural Language and Haptic + Overview) produced similar accuracy results. However, there was no evidence of functional equivalence between the Haptic + Overview condition and the Haptic Only condition. This would suggest that the presentation of even a small amount of natural language provides enough additional spatial information to benefit the building of mental spatial images to aid in reasoning and learning. It is unclear from this study if the effectiveness comes from the presentation of any amount of natural language or if it is the combination of the presentation modalities that allows for the filling in of missing pieces of spatial information to produce similar results of multisensory learning by sighted users in multimodal learning environments. These findings further support the need for research and development on layered information formats, multimodal interaction methods, and multimodal UIs to advance future universally designed STEM learning systems.

ACKNOWLEDGMENT

The authors thank the members of the INSITE Laboratory, the VEMI Laboratory, and the IMRE Laboratory, who helped to conduct the experiments during a challenging COVID-19 research time, especially Kaitlyn Haase. They thank their funding agencies for their support and the participants who helped to make this study possible.

REFERENCES

[1] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang, "Eviza: A natural language interface for visual analysis," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, Oct. 2016, pp. 365–377, doi: 10.1145/2984511.2984588.

[2] A. Srinivasan and J. Stasko, "Natural language interfaces for data analysis with visualization: Considering what has and could be asked," in *Proc. EuroVis*, B. Kozlikova, T. Schreck, and T. Wischgoll, Eds., Jun. 2017, pp. 55–59.

[3] A. Srinivasan and J. Stasko, "Orko: Facilitating multimodal interaction for visual exploration and analysis of networks," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 511–521, Jan. 2018, doi: 10.1109/TVCG.2017.2745219.

[4] A. Saktheeswaran, A. Srinivasan, and J. Stasko, "Touch? Speech? or touch and speech? Investigating multimodal interaction for visual network exploration and analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 6, pp. 2168–2179, Jun. 2020, doi: 10.1109/TVCG.2020.2970512.

[5] J.-K. Kim and R. J. Zatorre, "Generalized learning of visual-to-auditory substitution in sighted individuals," *Brain Res.*, vol. 1242, pp. 263–275, Nov. 2008, doi: 10.1016/j.brainres.2008.06.038.

[6] J.-K. Kim and R. J. Zatorre, "Can you hear shapes you touch?" *Exp. Brain Res.*, vol. 202, no. 4, pp. 747–754, Feb. 2010, doi: 10.1007/s00221-010-2178-6.

[7] J.-K. Kim and R. J. Zatorre, "Tactile–auditory shape learning engages the lateral occipital complex," *J. Neurosci.*, vol. 31, no. 21, pp. 7848–7856, May 2011, doi: 10.1523/jneurosci.3399-10.2011.

[8] J. M. Loomis, R. L. Klatzky, and N. A. Giudice, "Representing 3D space in working memory: Spatial images from vision, hearing, touch, and language," in *Multisensory Imagery*. New York, NY, USA: Springer, 2013, pp. 131–155, doi: 10.1007/978-1-4614-5879-1_8.

[9] R. E. Mayer, "Introduction to multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, 2014, doi: 10.1017/cbo9781139547369.002.

[10] R. E. Mayer, "Cognitive theory of multimedia learning," in *The Cambridge Handbook of Multimedia Learning*. Cambridge, U.K.: Cambridge Univ. Press, Jul. 2014, pp. 43–71, doi: 10.1017/cbo9781139547369.005.

[11] R. Moreno and R. Mayer, "Interactive multimodal learning environments: Special issue on interactive learning environments: Contemporary issues and trends," *Educ. Psychol. Rev.*, vol. 19, no. 3, pp. 309–326, Sep. 2007.

[12] P. Owens and J. Sweller, "Cognitive load theory and music instruction," *Educ. Psychol.*, vol. 28, no. 1, pp. 29–45, Jan. 2008, doi: 10.1080/01443410701369146.

[13] M. Turk, "Multimodal interaction: A review," *Pattern Recognit. Lett.*, vol. 36, pp. 189–195, Jan. 2014, doi: 10.1016/j.patrec.2013.07.003.

[14] Y.-H. Su, "Content congruency and its interplay with temporal synchrony modulate integration between rhythmic audiovisual streams," *Frontiers Integrative Neurosci.*, vol. 8, pp. 1–13, Dec. 2014, doi: 10.3389/fnint.2014.00092.

[15] R. E. Mayer and L. Fiorella, "Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles," in *The Cambridge Handbook of Multimedia Learning*. Cambridge, U.K.: Cambridge Univ. Press, Jul. 2014, pp. 279–315, doi: 10.1017/cbo9781139547369.015.

[16] J. M. Loomis, Y. Lippa, R. L. Klatzky, and R. G. Golledge, "Spatial updating of locations specified by 3-D sound and spatial language," *J. Exp. Psychol., Learn., Memory, Cognition*, vol. 28, no. 2, pp. 335–345, 2002, doi: 10.1037//0278-7393.28.2.335.

[17] J. R. Brown, S. A. Doore, J. K. Dimmel, N. Giudice, and N. A. Giudice, "Comparing natural language and vibro-audio modalities for inclusive STEM learning with blind and low vision users," in *Proc. 25th Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS)*, New York, NY, USA, Oct. 2023, pp. 1–17.

[18] N. A. Giudice, R. L. Klatzky, and J. M. Loomis, "Evidence for amodal representations after bimodal learning: Integration of haptic-visual layouts into a common spatial image," *Spatial Cognition Comput.*, vol. 9, no. 4, pp. 287–304, Nov. 2009, doi: 10.1080/13875860903305664.

[19] N. A. Giudice, H. P. Palani, E. Brenner, and K. M. Kramer, "Learning non-visual graphical information using a touch-based vibro-audio interface," in *Proc. 14th Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS)*, Boulder, CO, USA, Oct. 2012, p. 103. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2384916.2384935

[20] H. P. Palani, P. D. S. Fink, and N. A. Giudice, "Comparing map learning between touchscreen-based visual and haptic displays: A behavioral evaluation with blind and sighted users," *Multimodal Technol. Interact.*, vol. 6, no. 1, p. 1, Dec. 2021.

[21] M. N. Avraamides, J. M. Loomis, R. L. Klatzky, and R. G. Golledge, "Functional equivalence of spatial representations derived from vision and language: Evidence from allocentric judgments," *J. Exp. Psychol., Learn., Memory, Cognition*, vol. 30, no. 4, pp. 801–814, 2004, doi: 10.1037/0278-7393.30.4.804.

[22] R. L. Klatzky, Y. Lippa, J. M. Loomis, and R. G. Golledge, "Encoding, learning, and spatial updating of multiple object locations specified by 3-D sound, spatial language, and vision," *Exp. Brain Res.*, vol. 149, no. 1, pp. 48–61, Mar. 2003.

[23] Y.-J. Thoo, M. J. Medina, J. E. Froehlich, N. Ruffieux, and D. Lalanne, "A large-scale mixed-methods analysis of blind and low-vision research in ACM and IEEE," in *Proc. 25th Int. ACM SIGACCESS Conf. Comput. Accessibility*, New York, NY, USA, Oct. 2023, pp. 1–20. [Online]. Available: https://dl.acm.org/doi/10.1145/3597638.3608412

[24] N. A. Giudice, "Navigating without vision: Principles of blind spatial cognition," in *Handbook of Behavioral and Cognitive Geography*. Cheltenham, U.K.: Edward Elgar Publishing, 2018, p. 260, doi: 10.4337/9781784717544.00024.

[25] S. Oviatt, "Multimodal interfaces," in *The Human-Computer Interaction Handbook*. Boca Raton, FL, USA: CRC Press, 2007, pp. 439–458.

[26] M. J. Proulx, D. J. Brown, A. Pasqualotto, and P. Meijer, "Multisensory perceptual learning and sensory substitution," *Neurosci. Biobehavioral Rev.*, vol. 41, pp. 16–25, Apr. 2014, doi: 10.1016/j.neubiorev.2012.11.017.

[27] G. Volpe and M. Gori, "Multisensory interactive technologies for primary education: From science to technology," *Frontiers Psychol.*, vol. 10, pp. 1076–1083, Jun. 2019, doi: 10.3389/fpsyg.2019.01076.

[28] R. E. Mayer, J. Heiser, and S. Lonn, "Cognitive constraints on multimedia learning: When presenting more material results in less understanding," *J. Educ. Psychol.*, vol. 93, no. 1, pp. 187–198, 2001.

[29] R. E. Mayer and R. Moreno, "Nine ways to reduce cognitive load in multimedia learning," *Educ. Psychologist*, vol. 38, no. 1, pp. 43–52, Jan. 2003, doi: 10.1207/s15326985ep3801_6.

[30] B. Dumas, D. Lalanne, and S. Oviatt, "Multimodal interfaces: A survey of principles, models and frameworks," in *Human Machine Interaction: Research Results of the MMI Program*. Cham, Switzerland: Springer, 2009, pp. 3–26.

[31] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 116–134, 2007.

[32] J. Rowell and S. Ungar, "The world of touch: An international survey of tactile maps. Part 1: Production," *Brit. J. Vis. Impairment*, vol. 21, no. 3, pp. 98–104, Sep. 2003, doi: 10.1177/026461960302100303.

[33] J. Rowell and S. Ongar, "The world of touch: An international survey of tactile maps. Part 2: Design," *Brit. J. Vis. Impairment*, vol. 21, no. 3, pp. 105–110, Sep. 2003, doi: 10.1177/026461960302100304.

[34] I. Han and J. B. Black, "Incorporating haptic feedback in simulation for learning physics," *Comput. Educ.*, vol. 57, no. 4, pp. 2281–2290, Dec. 2011, doi: 10.1016/j.compedu.2011.06.012.

[35] J. L. Gorlewicz, J. L. Tennison, P. M. Uesbeck, M. E. Richard, H. P. Palani, A. Stefik, D. W. Smith, and N. A. Giudice, "Design guidelines and recommendations for multimodal, touchscreen-based graphics," *ACM Trans. Accessible Comput.*, vol. 13, no. 3, pp. 1–30, Aug. 2020, doi: 10.1145/3403933.

[36] H. P. Palani, P. D. S. Fink, and N. A. Giudice, "Design guidelines for schematizing and rendering haptically perceivable graphical elements on touchscreen devices," *Int. J. Hum.-Comput. Interact.*, vol. 36, no. 15, pp. 1393–1414, Apr. 2020, doi: 10.1080/10447318.2020.1752464.

[37] H. P. Palani, G. B. Giudice, and N. A. Giudice, "Haptic information access using touchscreen devices: Design guidelines for accurate perception of angular magnitude and line orientation," in *Proc. 12th Int. Conf. Universal Access Hum.-Comput. Interact. Methods, Technol., Users*, in Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2018, pp. 243–255, doi: http://dx.doi.org/10.1007/978-3-319-92049-8_18.

[38] N. A. Giudice, B. A. Guenther, N. A. Jensen, and K. N. Haase, "Cognitive mapping without vision: Comparing wayfinding performance after learning from digital touchscreen-based multimodal maps vs. embossed tactile overlays," *Frontiers Hum. Neurosci.*, vol. 14, p. 87, Mar. 2020.

[39] H. Palani and N. A. Giudice, "Evaluation of non-visual panning operations using touch-screen devices," in *Proc. 16th Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS)*, 2014, pp. 293–294.

[40] S. Brewster, F. Chohan, and L. Brown, "Tactile feedback for mobile interactions," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2007, pp. 159–162, doi: 10.1145/1240624.1240649.

[41] K. Yatani and K. N. Truong, "SemFeel: A user interface with semantic tactile feedback for mobile touch-screen devices," in *Proc. 22nd Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2009, pp. 111–120.

[42] C. Goncu and K. Marriott, "GraVVITAS: Generic multi-touch presentation of accessible graphics," in *Proc. 13th IFIP TC 13 Int. Conf. Hum.-Comput. Interact. (INTERACT)*, Lisbon, Portugal. Cham, Switzerland: Springer, Sep. 2011, pp. 30–48, doi: 10.1007/978-3-642-23774-4_5.

[43] J. L. Tennison and J. L. Gorlewicz, "Toward non-visual graphics representations on vibratory touchscreens: Shape exploration and identification," in *Proc. 10th Int. Conf. Haptics, Perception, Devices, Control, Appl. (EuroHaptics)*, London, U.K. Cham, Switzerland: Springer, 2016, pp. 384–395.

[44] S. A. Doore, J. Dimmel, T. M. Kaplan, B. A. Guenther, and N. A. Giudice, "Multimodality as universality: Designing inclusive accessibility to graphical information," *Frontiers Educ.*, vol. 8, pp. 1071759–1071774, Apr. 2023, doi: 10.3389/feduc.2023.1071759.

[45] J. Lyons, *Language and Linguistics*. Cambridge, U.K.: Cambridge Univ. Press, 1981.

[46] C. Jung, S. Mehta, A. Kulkarni, Y. Zhao, and Y.-S. Kim, "Communicating visualizations without visuals: Investigation of visualization alternative text for people with visual impairments," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 1095–1105, Jan. 2022.

[47] S. A. Doore, J. K. Dimmel, R. Xi, and N. A. Giudice, "Embedding expert knowledge: A case study on developing an accessible diagrammatic interface," in *Proc. 43rd Annu. Meeting North Amer. Chapter Int. Group Psychol. Math. Educ.*, vol. 43, Philadelphia, PA, USA, Oct. 2021, pp. 1759–1763.

[48] M. Weber and C. Saitis, "Towards a framework for ubiquitous audio-tactile design," in *Proc. Int. Workshop Haptic Audio Interact. Design*, Jan. 2020, pp. 1–6.

[49] K. Yamaguchi and S. Takahashi, "In-the-wild vibrotactile sensation: Perceptual transformation of vibrations from smartphones," 2023, *arXiv:2303.01308*.

[50] J. K. Dimmel and P. G. Herbst, "The semiotic structure of geometry diagrams: How textbook diagrams convey meaning," *J. Res. Math. Educ.*, vol. 46, no. 2, pp. 147–195, Mar. 2015, doi: 10.5951/jresematheduc.46.2.0147.

[51] S. J. Lederman and R. L. Klatzky, "Hand movements: A window into haptic object recognition," *Cogn. Psychol.*, vol. 19, no. 3, pp. 342–368, Jul. 1987, doi: 10.1016/0010-0285(87)90008-9.

[52] R. L. Klatzky and S. J. Lederman, "Touch," in *Handbook of Psychology, History of Psychology*, A. F. Healy and R. W. Proctor, Eds. Hoboken, NJ, USA: Wiley, Apr. 2003, pp. 147–176, doi: 10.1002/0471264385.wei0406.

[53] H. P. Palani, J. L. Tennison, G. B. Giudice, and N. A. Giudice, "Touchscreen-based haptic information access for assisting blind and visually-impaired users: Perceptual parameters and design guidelines," in *Proc. AHFE Int. Conf. Usability User Exp. Hum. Factors Assistive Technol., Adv. Usability, User Exp. Assistive Technol.*, Orlando, FL, USA. Cham, Switzerland: Springer, Jul. 2018, pp. 837–847.

[54] N. Aziz, T. Stockman, and R. Stewart, "Planning your journey in audio: Design and evaluation of auditory route overviews," *ACM Trans. Accessible Comput.*, vol. 15, no. 4, pp. 1–48, Oct. 2022, doi: 10.1145/3531529.

[55] R. E. Landrum, J. R. Cashin, and K. S. Theis, "More evidence in favor of three-option multiple-choice tests," *Educ. Psychol. Meas.*, vol. 53, no. 3, pp. 771–778, Sep. 1993, doi: 10.1177/0013164493053003021.

[56] C. Loudon and A. Macias-Muñoz, "Item statistics derived from three-option versions of multiple-choice questions are usually as robust as four- or five-option versions: Implications for exam design," *Adv. Physiol. Educ.*, vol. 42, no. 4, pp. 565–575, Dec. 2018, doi: 10.1152/advan.00186.2016.

[57] C. C. Serdar, M. Cihan, D. Yücel, and M. A. Serdar, "Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies," *Biochemia Medica*, vol. 31, no. 1, pp. 27–53, Feb. 2021.

[58] K. Charmaz, *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. Thousand Oaks, CA, USA: SAGE Publications, 2006.

[59] R. Thornberg and K. Charmaz, "Grounded theory and theoretical coding," in *The SAGE Handbook of Qualitative Data Analysis*. Thousand Oaks, CA, USA: SAGE Publications, 2014, pp. 153–169, doi: 10.4135/9781446282243.n11.

[60] K. Charmaz, "The genesis, grounds, and growth of constructivist grounded theory," in *Developing Grounded Theory*. Evanston, IL, USA: Routledge, Feb. 2021, pp. 153–187, doi: 10.4324/9781315169170-13.

[61] H. Xie, R. E. Mayer, F. Wang, and Z. Zhou, "Coordinating visual and auditory cueing in multimedia learning," *J. Educ. Psychol.*, vol. 111, no. 2, pp. 235–255, Feb. 2019, doi: 10.1037/edu0000285.

[62] L. F. Cuturi, G. Cappagli, N. Yiannoutsou, S. Price, and M. Gori, "Informing the design of a multisensory learning environment for elementary mathematics learning," *J. Multimodal User Interfaces*, vol. 16, no. 2, pp. 155–171, Oct. 2021, doi: 10.1007/s12193-021-00382-y.

[63] N. Aziz, T. Stockman, and R. Stewart, "An investigation into customisable automatically generated auditory route overviews for pre-navigation," in *Proc. of the 25th Int. Conf. on Auditory Display (ICAD)*, Newcastle-upon-Tyne, U.K., Jun. 2019, pp. 12–19,

STACY A. DOORE (Member, IEEE) received the Ph.D. degree in spatial information science and engineering from The University of Maine, Orono, ME, USA, in 2017. She was a Visiting Assistant Professor with Bowdoin College, from 2018 to 2020. She has been the Clare Boothe Luce Assistant Professor of computer science with the Colby College, Waterville, ME, since 2020. She is the Founder and the Director of the Immersive Navigation Systems and Inclusive Technology Ethics (INSITE) Laboratory. Her current research interests include spatial information access, multimodal interfaces, non-visual navigation systems, emerging assistive technologies, and embedding responsible computing methods into the core computer science curriculum. She is a co-creator of the Computing Ethics Narratives Project and a member of the ACM Taskforce on Computing Ethics.

JUSTIN R. BROWN received the B.S. degree in physics and the M.S. degree in information systems from The University of Maine, Orono, ME, USA, in 2020 and 2023, respectively. He is currently the Laboratory Research Manager of the VEMI Laboratory, The University of Maine. His current research interests include human–vehicle interaction, autonomous vehicles, information access, and multimodal interaction. He has published several papers in these areas and actively orchestrates VEMI's current and future research endeavors with academic and industry collaborators.

SAKI IMAI received the B.A. degree in computer science and mathematical sciences from the Colby College, Waterville, ME, USA, in 2024. She is a Research Intern with IBM Research, Tokyo. She will be pursuing the Ph.D. degree in computer science with Northeastern University. Her research interests include assistive technologies, natural language processing, and speech recognition. She received the third prize in the ACM Student Research Competition at ICSE 2022.

JUSTIN K. DIMMEL received the M.S. degree in mathematics and the Ph.D. degree in mathematics education from the University of Michigan, Ann Arbor, MI, USA, in 2013 and 2015, respectively. He is an Associate Professor of mathematics education and instructional technology with The University of Maine, Orono, ME, USA. He also leads the Immersive Mathematics in Rendered Environments (IMRE) Laboratory, The University of Maine. His research investigates student interactions with mathematical figures, especially those that are rendered in immersive digital spaces.

NICHOLAS A. GIUDICE the Ph.D. degree in psychology from the Cognitive and Brain Sciences Program, University of Minnesota, Twin Cities, MN, USA, in 2004. He is currently a Professor of spatial computing and the Chief Research Officer of the VEMI Laboratory, The University of Maine, where he has been a Faculty Member, since 2008. He is the Chief Research Officer of UNAR Laboratories, a start-up company that he co-founded dealing with multisensory information access. His research program is inherently interdisciplinary, combining principles and methodologies from experimental psychology, spatial perception, cognitive neuroscience, human–computer interaction, and multimodal interface design, focusing on spatial learning and navigation for blind and low vision (BLV) individuals. He has over 150 publications, collaborated on over $17 million in research grants, and advised over 80 students. As a congenitally blind individual, his expertise in assistive technology is invaluable to his work. He serves on the editorial board for *ACM Transactions on Accessible Computing* and *Assistive Technology*.