Development of an Intelligent Tutoring System that Assesses Internal Visualization Skills in Engineering Using Multimodal Triangulation

Hanall Sung, Martina A. Rau, and Barry D. Van Veen

***Abstract*—**In many STEM domains, instruction on foundational concepts heavily relies on visuals. Instructors often assume that students can mentally visualize concepts, but students often struggle with internal visualization skills—the ability to mentally visualize information. In order to address this issue, we developed a formal as well as an informal assessment of students' internal visualization skills in the context of engineering instruction. To validate the assessments, we used data triangulation methods. We drew on data from two separate studies conducted in a small-scale lab experiment and in a larger-scale classroom context. Our studies demonstrate that an intelligent tutoring system with interactive visual representations can serve as an informal assessment of students' internal visualization skills, predicting their performance on a formal assessment of these skills. Our study enriches methodological and theoretical underpinnings in educational research and practices in multiple ways: it contributes to (1) research methodologies by illustrating how multimodal triangulation can be used for test development, (2) theories of learning by offering pathways to assessing internal visualization skills that are not directly observable, and (3) instructional practices in STEM education by enabling instructors determine when and where they should provide additional scaffoldings.

***Index Terms*—Computer aided learning; Educational Technology; Engineering education, Human computer interaction; Intelligent systems; STEM**

H. Sung is with the Learning Sciences and Psychological Studies at University of North Carolina at Chapel Hill, Chapel Hill, NC 27514 USA (e-mail: hanalls@unc.edu).

M.A. Rau is with the Department of Humanities, Social and Political Sciences, Swiss Federal Institute of Technology in Zurich, Switzerland (e-mail: martina.rau@gess.ethz.ch).

B.D. Van Veen is with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI, 53706, USA (e-mail: bvanveen@wisc.edu).

I. INTRODUCTION

MANY concepts in science, technology, engineering, and mathematics (STEM) domains involve visuospatial information [1]. Thus, educational technologies for STEM domains frequently introduce foundational concepts through visual representations[1] such as graphs, figures, and diagrams. The goal in using visuals is to help students construct internal visual representations of the concepts, enabling inference-making and problem-solving [2]. Instruction on more advanced concepts then often assumes that students have *internal visualization skills*, which is the ability to mentally store, manipulate, and integrate visual information [3].

However, students often struggle with internal visualization skills [3]–[5]. For example, when electrical engineering instruction explains signals based on *sinusoids*, a fundamental engineering concept, they typically use a variety of visuals that depict basic concepts such as the amplitude, phase, and frequency of a sinusoid. After introducing foundational concepts through the visuals, instruction transitions to providing equations that describe the sinusoids without their accompanying visuals, presuming that students can *internally* visualize these concepts. However, students often struggle to internally manipulate and integrate concepts related to sinusoids provided through visuals [6] and to transfer their understanding of visual information to symbolic representations of more advanced concepts [7]. Consequently, students' difficulties with internal visualization can severely impede their subsequent learning of more complex concepts that are typically presented with minimal visual aids. While this specific example is taken from the domain of electrical engineering, the issue is broadly relevant because in many STEM domains, concepts are first introduced visually and then described by equations or other more abstract representations.

As the process of internal visualization is not directly observable and occurs within students' minds [3], it is challenging to assess students' internal visualization skills. There is extensive evidence that students' *representational gestures* reflect their internal visualization and the mental operations they use to manipulate internal representations [8]–[11]. Representational gestures are hand movements that "depict action, motion, or shape, or that indicate location or trajectory" [8, p. 245]. Hence, we consider them a well-established measure of internal visualization skills.

However, representational gestures are not a scalable method to assess internal visualization skills because gesture analysis is time consuming and complex. To our knowledge, there are no scalable assessments of internal visualization skills. Yet, scalable assessments are a prerequisite to providing instructional support for internal visualization skills. Therefore, the goal of this paper is to close this gap. We describe our use

[1] In this paper, we use the term visualization to describe a *process,* whereas we use the term representation to describe a *product*.

of data triangulation to develop assessments of students' internal visualization skills. On the one hand, we examine the utility of a potential *informal assessment* of internal visualization skills through the use of log data generated by an educational technology. Specifically, we draw on logs from an intelligent tutoring system (ITS) in which students interact with visual representations. Like many educational technologies, the ITS provides instructional problems that require students to engage with, manipulate, and interpret visual information. The ITS systematically gathers log data that captures students' problem-solving behaviors as they interact with visuals. It is an open question whether log data describing interactions with visuals can serve as an assessment of internal visualization skills. On the other hand, we developed a *formal assessment* of internal visualization skills that could be used as a diagnostic tool to tailor subsequent instruction accordingly.

To validate the informal and formal assessments, we rely on representational gestures, which—as mentioned—are an established measure of internal visualization skills. Given that gesture analysis is only feasible with small samples, Study 1 was a small-scale lab study that served as an initial inquiry into the relationship between the two assessments with representational gestures. Finding that representational gestures correlate with both assessments, we then describe a larger-scale classroom study that focused on the latter two assessments, seeking to address the limitations of Study 1's small sample size and artificial lab context. We note that the main purpose of our research is to demonstrate that a multi-modal approach that combines small-scale lab and large-scale classroom methods is useful for developing assessments for students' internal visualization skills, which can be applied to the variety of domains where internal visualization skills have an impact on students' academic success.

Overall, our work has broad applicability. While our studies are situated in the context of engineering education, we consider our approach of general relevance given the widespread use of as well as students' well-documented difficulties with visuals in STEM domains. For example, in chemistry, students cannot grasp the concept of an atom unless they understand which aspects of a Bohr model are inadequate, or what orbitals in orbital diagrams say about electron [13]. In many other domains, such as physics, arrows can be used to denote various types of information, which has to be understood by students [14].

The main contribution of our work regards its methodological approach. We illustrate how multimodal triangulation can be used to create technology-supported, scalable formative and summative assessment methods. Our approach may be used to inform the development of internal visualization skills assessments for other engineering topics and in other STEM disciplines. A secondary contribution lies in our contribution to instructional practices by revealing that scalable assessments of internal visualization skills are feasible. Having such scalable assessments are necessary, for instance, to determine when students need additional scaffolding, which could enhance instruction in many domains where students often struggle with internal visualization skills. Further, given that many educational technologies include visualizations, the development of assessments of internal visualization skills is of broad relevance to the field of educational software (e.g., intelligent tutors). Finally, our work contributes to future research on educational technologies. By offering pathways to assessing internal visualization skills within an educational technology, future research can shed light into how students acquire internal visualization skills alongside content knowledge as they engage within digital learning tools.

II. THEORETICAL BACKGROUND

A. Internal Visualization Skills

Research suggests that when students encounter external visual representations (e.g., static images such as graphs, diagrams, charts; or dynamic representations, such as animations), they construct internal representations in their mind [15]. *Internal visualization* occurs when students internally store, manipulate, and integrate visual information depicting objects or concepts without viewing external visuals [2], [3]. We use the term *internal visualization skills* to describe the ability to create mental representations of visual information that accurately describe domain-relevant concepts. Because a mental representation is always an abstraction of the original object, it can never be 100% accurate. Further, different individuals may have different mental representations. Finally, mental representations—given that they are internal to the student—cannot be directly observed, which may make it difficult to determine their accuracy. For our study, we define accuracy based on whether we see evidence that the student's mental representation contains the accurate, essential features and relationships inherent in the visual information being represented while learning domain-relevant concepts. In the context of our work, the accuracy of the domain-relevant concepts depicted in the visuals are well defined (e.g., the period of a sinusoid is a concept that is defined as the distance from peak to peak in a time-domain graph).

Internal visualization skills have been studied under different names in various fields. In cognitive neuroscience, internal visualization is referred to as *visual mental imagery*—a set of internal representations that enable individuals to recall, construct, and incorporate mental images in the absence of input [16]. Other researchers focus on *spatial visualization ability*, the ability to mentally imagine the movement of objects [17], [18]. Accordingly, tasks to assess spatial visualization ability involve imagining the result of spatial transformations, such as folding a piece of paper or rotating an object [19]. Building on this research, we conceptualize internal visualization skills as the ability to mentally recall visual information from external information sources, organize the information, activate related prior knowledge, and build coherent and accurate internal representations [20].

Internal visualization skills are distinct from representational competencies, which have received much attention in educational psychology research (for an overview, see [21]). Representational competencies describe the ability to extract correct information from visual representations that are

presented externally [21], [22]. By contrast, internal visualization skills are at play even when external visual representations are absent. As representational competencies are recognized as crucial for effective learning with visuals, these competencies may influence or relate to internal visualization skills. However, there is limited existing research that explores the intricate relationship between representational competencies and internal visualization skills. While many studies examined students' ability to learn with external representations such as a digital puzzle game [23], engineering design tasks [8], and mechanical reasoning problems [10], we are not aware of prior research that differentiates between assessments where visual representations are present versus absent. To our knowledge, there is no formal assessment that evaluates students' internal visualization skills when visuals are not present.

Given that students often struggle to mentally store, manipulate, and integrate the information conveyed by visuals [3]–[5], it is critical to assess their internal visualization skills to support their learning. Since internal visualization skills are not directly observable, much research has investigated how to infer them from observable indicators, such as gestures.

B. Representational Gestures

Gestures that "depict action, motion, or shape, or that indicate location or trajectory" can be conceptualized as *representational gestures* [8, p. 245]. Representational gestures convey semantic content through visual similarity using hand shape or motion [24] and represent perceptual features of objects or concepts [10]. Previous literature related to speech-accompanying representational gestures documents associations between gesture and internal visualization. People more frequently produce co-speech representational gestures when describing spatial transformations or communicating their thought processes [25], [26]. For example, when students verbally describe spatial problem-solving processes, such as the arrangement of gears in a mechanical system, they spontaneously produce hand movements to simulate the rotation of gears (e.g., clockwise or counterclockwise rotation), which are co-speech representational gestures [27]. These co-speech representational gestures are often synchronized with their speech (i.e., redundant gestures), enhancing the comprehensibility of their explanations, but they may also exhibit information that is complementary and not expressed in speech (i.e., non-redundant gestures) [9]. Both gestures are beneficial for acquiring a comprehensive understanding of the internal cognitive processes.

Moreover, spatial skills may influence the frequency or types of gestures. Hostetter and Alibali [28] found that participants with low verbal skills and high spatial visualization skills produced more representational gestures compared to participants with high verbal skills and low spatial visualization skills. On the other hand, Göksun and colleagues [29] found that individuals with low spatial skills generated representational gestures more frequently than those with high spatial skills when conveying solely static information, but less frequently when conveying dynamic information. As shown in this prior research, representational gestures that provide a valuable external manifestation of internal cognitive processes can reveal students' levels of spatial visualization skills and different types of visuospatial information (e.g., static or dynamic) they attempt to convey through gestures. Hence, representational gestures, offering insights into students' manipulation of mental images, can be considered a validated, observable indicator of their internal visualization skills.

However, relying on gestures is not a scalable method for assessing internal visualization skills of large numbers of students. Gesture studies often require labor-intensive and time-consuming manual analyses of video recordings or the use of specialized motion sensing equipment that is difficult to deploy at scale [30]. While some recent studies are beginning to use sensor technology or AI-based machine learning algorithms for automatic gesture (or body movement) detection and gesture classification [31], these methods are too nascent to replace contextualized and situated human observation and interpretation of the meaning and function of gestures in a certain context. Alternatively, another potential scalable indicator is log data that tracks students' problem-solving behaviors.

C. Log data of Problem-solving Behaviors in Learning with Visuals

Many studies have demonstrated the potential of using students' behavioral log data, encompassing timestamps and interaction sequences during learning tasks, to later infer a range of cognitive skills [32], [33]. For example, log data can reveal where students struggle with their understanding during the learning process [34] and require further scaffolding [35]. In a similar vein, Rau [13] used log data on problem-solving steps where students interacted with visuals to examine difficulties in working with the visuals. While Rau [13] reported that log data of students' problem-solving behaviors during learning with visuals could be indicative of difficulties in working on problems in which visuals were present, it is an open question whether these log data can be also used to infer students' internal visualization skills, which presumes that visuals are absent.

In sum, it remains an open question whether log data that tracks students' problem-solving behaviors while learning with external visualizations can serve as an informal assessment of internal visualization skills. If this were the case, log data should predict students' scores on a formal assessment of internal visualization skills. While prior research has shown that log data from ITSs can be used as informal assessments of content knowledge that predicts performance on formal assessments [36], [37], this question has not been examined in the context of internal visualization.

III. THE CURRENT INVESTIGATION

In this paper, we use representational gestures as an established, though not scalable measure of internal visualization skills. We triangulate gesture data with log data and test data to investigate the validity of informal and formal, scalable assessments of internal visualization skills. The use of

multiple measurements provides several advantages. It mitigates the limitations of each measure and presents a more holistic view of internal visualization skills. While representational gestures offer a tangible and interpretable dimension of internal visualization, informal assessment via log data provides a more detailed and nuanced understanding of the cognitive processes involved. Formal assessment via a test is highly interpretable. However, each measure has its limitations; representational gestures are not scalable, log data can be difficult to interpret without context, and taking a test can take up valuable instructional time. By incorporating these measures, we aim to combine their advantages while mitigating their drawbacks. Further, triangulating multiple measures enhances the robustness of our findings while reducing the potential for measurement bias.

A. Research Questions

We conducted two studies (summarized in Figure 1). Systematically grounded in existing literature, our investigation integrates multiple steps between formal and informal measures to explore nuanced relationships, leveraging existing theoretical and empirical knowledge in the field. Study 1 served as an initial inquiry into the validity of our formal and informal assessments of internal visualization skills. It examined the relationship among students' (1) representational gestures as a validated measure with (2) log data of problem-solving behaviors when working with visuals within an ITS, and (3) verbal responses to a post-interview that constituted an early version of our formal assessment of students' internal visualization skills. In line with most prior research, external visuals were present during the post-interview in Study 1. Study 1 addressed the following research questions (RQs):

RQ1: How do representational gestures relate to the post-interview assessment of internal visualization skills?

RQ2: How does log data from the ITS capturing interactions with visuals relate to representational gestures?

Study 2 examined the utility of log data from the ITS as an informal assessment of internal visualization skills. Study 2 was conducted in the context of a larger-scale classroom where—as in most realistic educational settings—gesture analysis was not feasible. It investigated the relationship between (1) students' log data and (2) their responses to a formal post-assessment of their internal visualization skills. To expand prior research, external visuals were absent in the internal visual skills assessment of Study 2. Study 2 investigated:

RQ3: How does log data capturing interactions with visuals relate to the formal assessment of internal visualization skills?

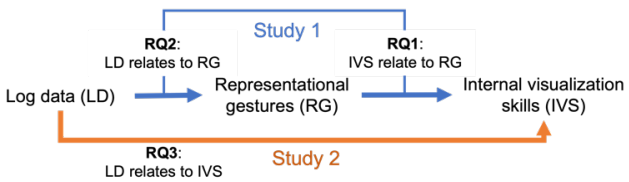


Fig. 1. Research framework (Study 1 and Study 2)

IV. study 1

A. Methods

1) **Participants**

For Study 1, we recruited 19 undergraduate students from a large midwestern university in the United States. Students were recruited through announcements in lectures and recruitment posters. Students were awarded either course credit or $10 per hour for their participation in the study, which lasted up to 3 hours. Five students were excluded from the final dataset because they did not complete the post-intervention interviews. As a result, our analyses included 14 students, seven of whom were STEM majors while the remaining seven were non-STEM majors. None of the students had taken college-level electrical engineering courses relating to signals or signal processing, which were the main topics covered in the problem-solving activities in Study 1.

2) **Materials**

a) Intelligent Tutoring System

Our study was conducted in the context of *Signals Tutor*, an ITS designed and developed by our research team. Signals Tutor offers various interactive problem-solving activities, encompassing multiple-choice options, text box insertions, and interactive visual representations, and covers basic principles of sinusoids using seasonal variations of daylight durations across the globe as a concrete example. Prior to working with the tutoring software, students watched a 10-minute introductory video that provided fundamental background knowledge on seasons and sinusoids. Students then engaged in problem-solving activities with external visual representations using Signals Tutor. Building on prior work [38], [39], Signals Tutor provides problem-solving activities that support students' ability to make sense of visual representations and to practice perceptual fluency in extracting information from the visuals, as detailed in the following. Signals Tutor is representative of ITSs in that domain-relevant visualizations are commonly integrated into various ITSs and other educational technologies. The ITS is designed to assess students' responses, diagnose their misconceptions, and offer targeted feedback and hints that are tailored to the specific issues they are facing. These algorithms use students' problem-solving interactions to draw inferences about the individual's state of domain-specific knowledge as potential misconceptions that the student may hold about the visualizations, as described in our prior work [40], to guide students toward the correct understanding of the concepts being taught. All student interactions with these activities were logged.

Signal Tutor's sense-making activities (Figure 2) aim to help students conceptually explain which visual features map to one another across visuals and how they represent domain-relevant concepts. Sense-making activities provide prompts to reflect on how a given concept is shown by specific features of each visual. Students receive immediate error feedback and hints on demand (see Figure 1). In addition, when students requested hints, they received multiple levels of conceptual hints, adding more scaffolding to solve the problems.

This article has been accepted for publication in IEEE Transactions on Learning Technologies. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TLT.2024.3396393

Perceptual-fluency activities (Figure 3) expose students to many simple classification problems that require rapid translation across different visual representations in order to help students efficiently translate among visuals. Students are given one visual and asked to select one out of four other visuals that shows the same construct (e.g., the same amplitude and time shift of a sinusoid). The four answer choices contrast visual features that could mislead students. Unlike sense-making activities, perceptual-fluency activities provide only correctness feedback, because conceptual feedback can interfere with perceptual processing. Instead, to encourage perceptual processing, students are asked to solve the problems fast and intuitively.

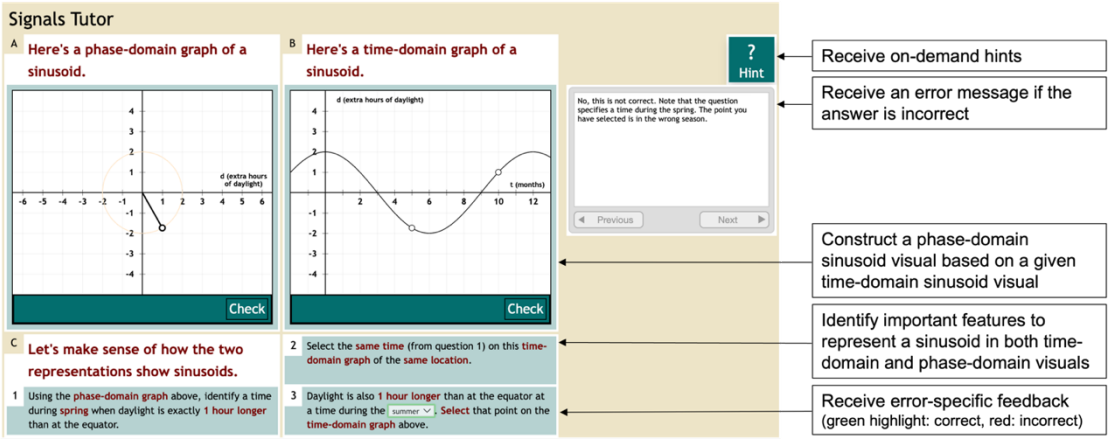


Fig. 2. Example sense-making activity in Signals Tutor

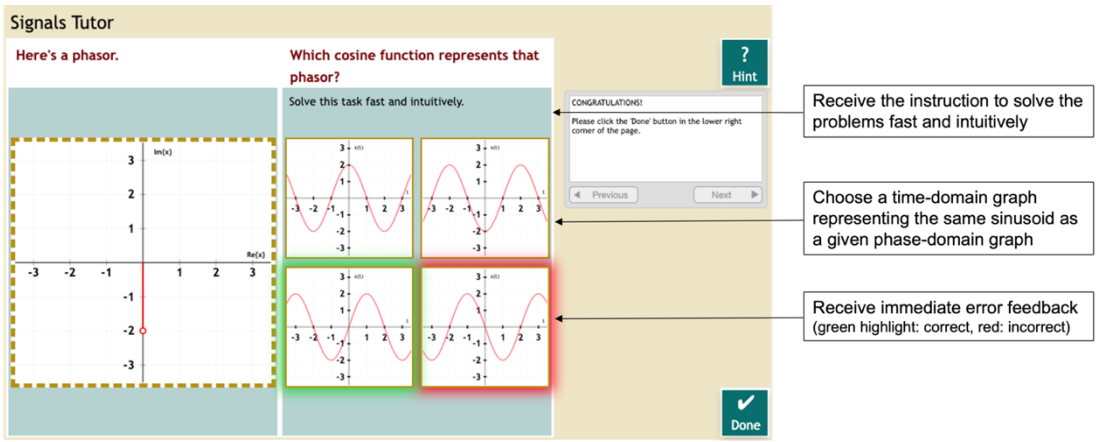


Fig. 3. Example perceptual-fluency activity in Signals Tutor

b) Post-Interview Assessment

Following the ITS, students participated in pre-structured post-interviews employing think aloud protocols, which aimed to assess students' ability to internally visualize the knowledge gained from learning with visuals in Signals Tutor. The interview questions were devised by a content expert in our research team (third author). External visuals were often present during the interviews to specify the contexts of the questions, but the interview questions required students to mentally manipulate the visuospatial information of concepts that were not depicted in external visuals.

Specifically, interview questions were designed to stimulate students' verbal reasoning involved in the construction of mental mappings between visuals and concepts, as well as the mental translation between different visuals. All interview questions required students to recall how domain-relevant concepts were represented in external visuals. Further, some interview questions required students to explain how certain concepts or terms corresponded to specific features of the visuals. Other interview questions asked students to mentally translate one visual to another. In either case, every question had a well-defined correct answer. This design allowed us to evaluate students' responses based on their ability to accurately describe domain-relevant concepts and corresponding visuospatial information learned through the ITS. A detailed list of the interview questions is shown in Appendix A.

For example, one of the interview questions asked: "What does one time period of the sinusoid mean in the phase-domain graph?" (Question #5 in Appendix A; providing a sinusoid on the time-domain (Panel A) and blank phase-domain graph (Panel B), see Figure 4). As illustrated in Figure 4, in order to provide the correct answer, students must (1) recall a concept (i.e., one time period of the sinusoid) related to a visual (i.e., the time-domain graph, Panel A), (2) understand that the same concept can be represented differently in another visual (i.e., the phase-domain graph, Panel B), and (3) integrate the two visuals to transform one into the other (Panel C). This illustrates how the interview questions required students to internally visualize the concepts they had learned during the learning intervention.

Students answered the interview questions verbally. If students did not unpack the steps of their thought processes, they received follow-up questions that probed for further details. During the interviews, a researcher observed students' verbal responses and co-speech gestures (if any) and documented them in field notes as well as video recordings. Students' responses to the interview questions were examined in two different ways: (1) verbal responses were evaluated based on how accurately they recalled and supplied the requested information and based on how promptly they provided their answers; and (2) co-speech gestures were evaluated based on what content-relevant information they conveyed (detailed below).

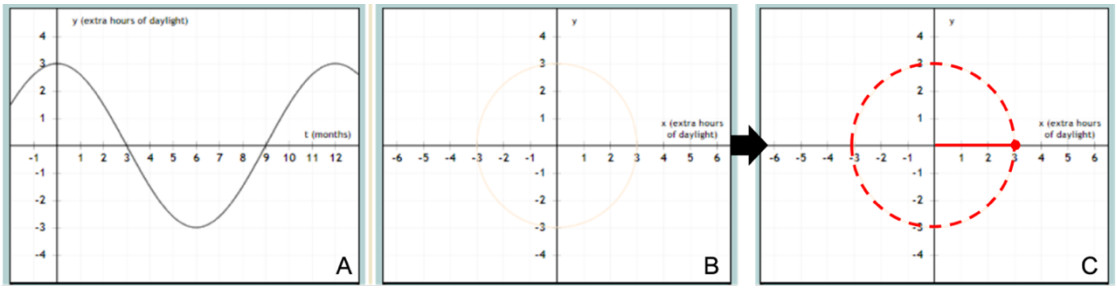


Fig. 4. An example of the visuals that accompanied Interview Question #5 (Panel A and Panel B) and the visual that students were required to internally visualize (Panel C)

3) **Procedure**

The study took at most 3 hours per student. Students first watched the 10-minute introductory video. Then, students worked with Signals Tutor, consisting of sense-making and perceptual-fluency activities, for up to 2.5 hours. Next, students participated in the post-intervention interview. Finally, they received either monetary compensation of up to $30 or a certificate of research participation for extra credit in a course.

4) **Measures**

a) Gesture Coding

Alongside their verbal responses to the post-interviews, students frequently produced co-speech representational gestures. Representational gestures portrayed movement and/or shape information of the time-domain and the phase-domain graphs that were presented during the learning intervention. To assess students' internal visualization skills, we focused on those dynamic representational gestures that reflected visuospatial information provided during the learning intervention. Thus, we coded two types of representational gestures: (a) *directional gestures* depicted specific features or changes of a sinusoid on a time-domain graph (e.g., amplitudes or phase shifts) by moving hands upward and downward or by shifting hands from left to right or vice versa, and (b) *rotational gestures* depicted rotational movements of a phasor on a phase-domain graph. For instance, if a student horizontally shifted their hands from side to side as if illustrating the positive or negative phase shift in a time-domain graph, the gesture was identified as a directional gesture. On the other hand, if a student drew a clockwise circle with their index finger to represent the rotation of a vector in a phase-domain graph, the gesture was identified as a rotational gesture. These categories serve as a standard for gesture coding, which means that if a gesture does not fit into any of them, we did not code it.

During the coding process, two human coders independently coded the interview data (N = 14) to identify whether students produced any kinds of representational gestures (i.e., directional and rotational) in response to a single interview question. Human coder 1 (first author) was the researcher who observed the participants' entire learning processes and posed post-interview questions. As part of the training for gesture coding, human coder 1 and human coder 2 (a research assistant) went through all learning activities together, thereby both shared a common understanding of how to capture the two targeted types of representational gestures. The produced representational gestures were coded binarily (i.e., present or absent) for each interview question, so that if either type of representational gesture was present, it was counted as 1, and if none was present, it was counted as 0. To avoid potential biases in capturing the gestures, both human coders performed independent coding, reconciled any differences by viewing the corresponding segments of video recordings together, and reached a consensus. Interrater reliability was high with Cohen's k = .92 [41].

b) Log Data of Problem-Solving Behaviors: Error Rates and Hint Requests

We extracted time-stamped logs of problem-solving interactions in Signals Tutor. The log data encompassed error rates and hint requests associated with individual micro-steps within problems presented in Signals Tutor. These micro-steps were designed to be detailed and mixed in complexity, including various types of problems such as multiple-choice options, text box insertions, and interactive manipulation of visual representations. The granularity of the data gave comprehensive and fine-grained information about students' cognitive processes at a micro-level while learning with visuals. We computed error rates as the average number of incorrect answers per problem-solving step. We also computed the average number of hint requests per step. Due to the absence of hint requests on perceptual-fluency activities, the total number of hint requests made by students was limited to those made during sense-making activities.

B. Results

We first conducted a qualitative analysis of how students' production of representational gestures relates to their internal visualization skills as assessed based on verbal responses

during the interviews (RQ1). We then quantitatively analyzed how students' log data of problem-solving behaviors (i.e., error rates and hint requests) during the learning process relate to their production of representational gestures (RQ2).

1) **Qualitative Analysis: Representational Gesture and Internal Visualization Skills**

To address RQ1 (how representational gestures relate to the post-interview assessment of internal visualization skills), we qualitatively analyzed the video recordings of interview data. This served to explore the quality and profundity of students' verbal responses to the post-interview questions and these aspects relate to their use of representational gestures. Here, we present two vignettes as representative examples illustrating insights from our qualitative analysis.

For the first vignette, Figure 5 illustrates the case of Student 1 (henceforth S1), a student who produced many representational gestures. S1 made multiple representational gestures ($N = 5$). When responding to the question, "How can we translate time axis on the time-domain graph to the phase-domain graph?" (Question #2 in Appendix A), S1 first provided the correct answer, stating "in this [phase-domain graph], x [x-axis] is extra hours in daylight" (Line 1). S1 then described how the daylight time changes appear on a time-domain graph (Line 2) while moving the right hand upward and downward (*directional gesture*). This response indicates that S1 was able to retain the concept (e.g., daylight time changes) that was visually conveyed during the learning process and promptly translate this information into different visuals (e.g., on a phase-domain and a time-domain graph). S1's response to the subsequent query, which inquired about the time axis on the time-domain graph, was again immediate and accurate (Line 5-6).

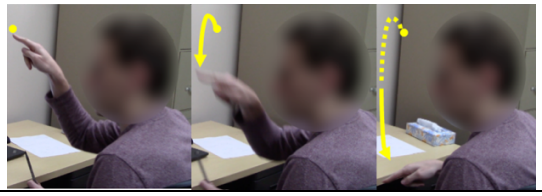
Line	Timeline	Speaker	Utterances
1	03:16 - 03:30	S1	So, in this (phase-domain graph), x (x-axis) is extra hours in daylight, which is basically time above or below that 12 hours. So, at 12 hours, since a day is 24 hours, it's half and half, right?
2	03:30 - 03:39	S1	So, when you have a 1 as your extra hour of daylight, that would be, you have 13 hours of daylight and 11 hours of nighttime. So, that's what x (x-axis) is.
3	03:39 - 03:40	Researcher	X-axis is time-axis?
4	03:40 - 03:41	S1	Yes.
5	03:41 - 03:47	Researcher	Okay. How about the time-domain graph?
6	03:48 - 03:51	S1	Um, in this one (time-domain graph), it is the y (y-axis). The extra hours of daylight.

Fig. 5. Vignette 1: An excerpt of S1's responses in the post-intervention interview

By contrast, for the second vignette, Figure 6 illustrates the case of Student 2 (henceforth S2), a student who produced few representational gestures. S2 made only two representational gestures during the interviews. When responding to similar interview question requiring mental translation between two distinct visuals (e.g., "What does the time period of the sinusoid mean in the phase-domain graph?", Question #5 in Appendix A), S2 was not able to instantly understand what the question meant, saying "Time period? [Paused for 5 seconds] I'm confused. What do you mean?" In order to clarify, the researcher parsed the question into a series of sub-questions. For instance, the researcher asked, "First, here [points at time-domain graph], what's the time period?", and S2 responded "12 months." The researcher replied: "12 months, okay. Can you draw it [points at the time-domain graph]?" Then, S2 drew the correct cosine graph on the time-domain graph with their index finger. S2's verbal responses to these sub-questions revealed that S2 remembered the concepts associated with one visual (i.e., a time-domain graph) but struggled to independently link the visual information with comparable terms. S2's struggled were deteriorating when the researcher subsequently questioned the translation between two visuals ("How can we represent the time period of the sinusoid in this domain [point at a phase-domain graph]?"). Since the given interview question provided a blank phase-domain graph (Appendix A), S2 needed to internally visualize the visuospatial information of the concept to answer this question. In response to the interviewer's query, S2 hesitatingly replied, "How would you present it? Hmm, I don't know. [Paused for 6 seconds]", and then supplied the incorrect answer.

Figure 6 displays S2's reflection after discovering the correct answer to the aforementioned interview question with the researchers' assistance. This example reveals that S2 had several difficulties with internal visualization. One difficulty was that S2 did not recognize the rotational property of the phasor on the phase-domain graph ("I thought, more like, it is just like a point and go for there, I just never thought to go all the way around", Line 1-2). While describing this, S2 produced a rotational gesture for the first and only time during the interview. Moreover, S2's difficulties in understanding the phase-domain graph impeded S2's ability to establish the conceptual linkage between the phase-domain and the time-domain graphs ("I never relate this circle (a phasor representation on the phase-domain graph) to this (a cosine graph on the time-domain graph", Line 3). S2 explicitly acknowledged their unawareness of the connection between two visuals by stating, "Now I do see it, but at that time, I wasn't, I thought that there are two *separate* things." (Line 4)

This article has been accepted for publication in IEEE Transactions on Learning Technologies. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TLT.2024.3396393



Line	Timeline	Speaker	Utterances
1	12:00 - 12:10	S2	I guess I never ever thought to… Hmm, interesting.
2	12:11 - 12:24	S2	I just never had a thought to go it as a circle, I guess. Because I thought, more like, it is just like a point and go for there, I just never thought to go all the way around. …
3	12:41 - 12:45	S2	I don't see like. I see the circle, but I never relate this circle (a phasor representation on the phase-domain graph) to this (a cosine graph on the time-domain graph). …
4	12:58 - 13:07	S2	Now I do see it, but at that time, I wasn't, I thought that there are two separate things.

Fig. 6. Vignette 2: An excerpt of S2's responses in the post-intervention interview

Overall, we found that students who produced several representational gestures, like S1, also provided verbal responses that were characterized by the ability to accurately recall the visual information of linked concepts and to explain quickly how distinct visuals are related. These students were able to autonomously explain their answer without the need for further scaffolding, and they typically provided answers in a short amount of time. In contrast, students who produced few or no representational gestures, like S2, appeared to struggle to recollect how certain terms or concepts related to visuals and to establish connections across visuals. These students frequently asked for clarification of the interview question, stating, "I'm confused, what do you mean?" or took considerable amount of time to respond, or gave up and muttered "I don't know."

Thus, students who produced multiple representational gestures during the post-intervention interviews seemed to exhibit high visual internalization skills; they were capable of accurately and swiftly articulating the retained visuospatial information mapping with related concepts as well as grasping the conceptual connection between different visuals, whereas students who produced few or no representational gestures during the interviews seemed to have more difficulties doing so. This speaks to the validity of the informal and formal assessments.

2) Quantitative Analysis: Representational Gestures and Log data of Problem-Solving Behaviors

To investigate RQ2 (how log data from the ITS capturing interactions with visuals relates to representational gestures), we examined whether the number of representational gestures students produced during the post-interviews is associated with their error rates and hint requests during sense-making and perceptual-fluency activities. As shown in Table 1, students' problem-solving behaviors captured during the learning process within the context of the ITS exhibited considerable variability. Using these variables, we computed three different simple regression models and two hierarchical regression models. Table 2 provides the results from the linear regression analyses.

TABLE 1
Descriptive summary of the variables utilized for quantitative analysis in Study 1

($N = 14$)

Variables	Mean	SD	Min	Max
Number of representational gestures	3.29	2.40	0	8
Error rates during sense-making activities	0.42	0.27	0.15	1.18
Error rates during perceptual-fluency activities	0.40	0.24	0.08	1.07
Total hint requests	43.29	57.33	3	225

Note. Dependent variable: Number of representational gestures

TABLE 2
Summary of linear regression models

($N = 14$)

Model	Predictor variables	β	SE	t	p
1	Intercept		1.05	5.26	0.00
	Error rates during sense-making activities	-0.58	2.12	-2.48	0.03*
2	Intercept		0.93	6.61	0.00
	Error rates during perceptual-fluency activities	-0.72	2.01	-3.56	0.00**
3	Intercept		0.69	6.28	0.00**
	Total hint requests	-0.58	0.01	-2.47	0.03*
4A	Intercept		0.99	6.20	0.00**
	Error rates during sense-making activities	0.71	3.52	0.18	0.86
	Error rates during perceptual-fluency activities	-0.78	3.87	-2.00	0.07
4B	Intercept		1.40	4.45	0.00**
	Error rates during sense-making activities	0.03	4.61	0.06	0.95
	Error rates during perceptual-fluency activities	-0.80	4.42	-1.8	0.10
	Total hint requests	0.07	0.02	0.13	0.90

Note. * $p < .05$, ** $p < .01$
Dependent variable: Number of representational gestures
Model 1: $F(1,13) = 6.16$, $p < .05$, $R^2 = 0.34$, $R^2_{Adjusted} = 0.28$
Model 2: $F(1,13) = 12.66$, $p < .01$, $R^2 = 0.51$, $R^2_{Adjusted} = 0.47$
Model 3: $F(1,13) = 6.10$, $p < .05$, $R^2 = 0.34$, $R^2_{Adjusted} = 0.28$
Model 4A: $F(2,12) = 5.84$, $p > .05$, $R^2 = 0.52$, $R^2_{Adjusted} = 0.43$
Model 4B: $F(3,10) = 3.55$, $p > .05$, $R^2 = 0.52$, $R^2_{Adjusted} = 0.37$

The error rates during sense-making activities (Model 1, t = -2.48, $p < .05$) and perceptual-fluency activities (Model 2, t = -3.56, $p < .01$) had statistically significant predictive power on the number of representational gestures, with 28% and 47% of the model explanation, respectively (Model 1: $R^2_{Adjusted} = 0.28$, Model 2: $R^2_{Adjusted} = 0.47$). Students who had higher error rates during sense-making and perceptual-fluency activities produced significantly fewer representational gestures during the post-interviews.

Similar to Models 1 and 2, Model 3 indicated that the total hint requests during the sense-making activities had statistically significant predictive power on the number of representational gestures (Model 3, t = -2.47, p < .05), explaining 28% of the variance ($R^2_{Adjusted} = 0.28$). Students who requested more hints during the sense-making activities produced significantly fewer representational gestures during the post-interviews.

Following the single-variable models, hierarchical regression models were created by successively adding the second-highest and third-highest predictive variables to the best single-variable model (Model 2) in order to test for a substantial increase in model explanation. The results showed that, when combined, none of the variables in Model 4A and Model 4B had statistically significant predictive power on the number of representational gestures, and there was no significant added value to the model explanation (Model 4A: $R^2_{Adjusted} = 0.43$, Model 4B: $R^2_{Adjusted} = 0.37$) compared to Model 2, which best explained the variance ($R^2_{Adjusted} = 0.47$).

Thus, error rates computed based on log data from an ITS where students interacted with visuals were significant predictors of representational gestures. This speaks to the validity of this measure as an assessment of internal visualization skills.

C. Discussion

By triangulating findings from the qualitative and quantitative analysis, we examined the potential of two types of assessments of students' internal visualization skills. The qualitative findings relating to RQ1 revealed that students' production of representational gestures, a validated although not scalable measure of internal visualization skills, was associated with higher-quality verbal responses to the post-interviews, an early version of our formal internal visualization skills assessment. Students who produced multiple representational gestures exhibited superior visual internalization skills in their verbal responses to the interview questions compared to those who produced few or no representational gestures during the interviews. When verbally responding to the interview questions, students who generated multiple representational gestures were capable of correct recall of the visuospatial information pertaining to visuals and quick translation among the visuals, whereas those who generated few or no representational gestures struggled to do so. Considering that representational gestures are an empirically validated and established indicator of internal visualization [28], [29], the association between students' representational gestures and verbal responses to the interview questions suggests that the early interview version of our formal internal visualization skills assessment indeed assesses what it intends to assess.

The quantitative findings relating to RQ2 indicated that students' log data of interactions with visuals in the ITS were associated with their production of representational gestures. The better students' problem-solving performance, the more representational gestures they produced during the post-interviews. That is, students who had lower error rates and asked for fewer hints throughout the sense-making and perceptual-fluency activities of the ITS were more likely to produce more representational gestures during the interviews, implying better internal visualization skills. These statistically significant relationships provide empirical support for the theoretical underpinnings, enhancing the validity of our approach. This finding implies that students' log data from an ITS with external visual representation has the potential to serve as an informal assessment of their internal visualization skills.

However, Study 1 has several limitations. First, Study 1 had a small sample size. While the focused nature of our investigation, coupled with the qualitative nature of the study, allowed for an in-depth exploration of students' internal visualization skills, the small sample size means that Study 1 had relatively low statistical power. Further, the small sample size implies that our results may not generalize to a broader population. Therefore, future research should be replicated with a larger sample size to validate and extend our findings. Second, Study 1 was conducted in a lab setting. Students' problem-solving behaviors can differ between the lab and classroom setting. Third, students' internal visualization skills were qualitatively assessed via verbal responses to interview questions instead of using a test, which required substantial human input. Fourth, half of the participants were undergraduate STEM majors, whereas the other half were non-STEM majors. Students may have diverse prior knowledge, spatial abilities, and motivations to learn scientific knowledge, depending on their majors. Therefore, it would be helpful to examine internal visualization skills with students enrolled in an engineering course because they would likely be motivated to learn the targeted knowledge and likely have similar prior knowledge. Lastly, in keeping with prior research, students' internal visualization skills were assessed when external visuals were present. However, this makes it difficult to distinguish whether students' ability to retrieve visual information relied solely on internal visualization or on the given external visuals. Therefore, it would be desirable to create an assessment of internal visualization skills that does not present external visuals.

V. STUDY 2

Study 2 was designed to address several limitations of Study 1. First, to address limitations resulting from the small sample size of Study 1, Study 2 included a larger sample of students. Second, to address limitations resulting from Study 1 being conducted in a lab context, we situated Study 2 in a large-scale classroom setting of an introductory undergraduate engineering course on signal processing, offering a real-world educational setting aimed.

At the same time, to ensure continuity from Study 1 to Study 2, both studies employed the same ITS. This has the advantage that we were able to collect the same log data measures of students' interactions with visual elements embedded within the ITS. Additionally, we turned the interview questions from Study 1 into a scalable test in Study 2. Given that Study 1 had validated these measures through gesture analysis, this approach allowed us to leverage the qualitative richness derived from Study 1, increasing the validity of the measures used in Study 2.

In sum, building on Study 1, Study 2 aims to deepen our understanding of the scalable measures we started to explore in Study 1.

A. Methods

1) **Participants**

Participants in Study 2 were 141 undergraduate students who were enrolled in an introductory electrical engineering course titled "Signals, Information, and Computation" at a large midwestern university in the United States. Signals Tutor was incorporated into the course materials. Accordingly, students' participation was graded as a course assignment; completion of all activities in Signals Tutor resulted in full credit, otherwise, no credit was given. After excluding the students who did not complete all activities in Signals Tutor (30 students) and who dropped the course (2 students), our analyses included a total of 109 students.

2) **Materials**

a) Post-Intervention: Internal Visualization Skills Test

To develop the interview questions from Study 1 into a formal assessment, our research team, which included a content expert in electrical engineering (third author), used a multi-step iterative design process. This process resulted in design principles that are embodied by the resulting test items. First, we minimized the use of visuals when developing the test items since the presence of external visuals may have impact on students' internal recollection of visuospatial information [16]. As a result, two thirds of the test items were comprised solely of symbolic representations (e.g., equations). While the remaining one third of the items included external visuals, these visuals were only presented to prompt students to select the corresponding visuals they mentally manipulated or to translate one visual to another. Secondly, each test item was designed to involve little computational effort to motivate students to prioritize visualization strategies over computational strategies. For example, we purposefully chose numbers that are easier to compute if simulated visually in space, as opposed to computing these numbers formulaically, which is technically doable but substantially more complicated and time-consuming due to the large number of decimals involved (e.g., $7/4\pi$ or $5/4\pi$). Lastly, we included test items requiring the translation between one visual to another to evaluate whether students can transfer visuospatial information between visuals [21]. We engaged in several rounds of iterative design where we applied these principles and then reviewed and discussed their implementation. This resulted in a set of multiple-choice test items (21 items) in a formal assessment of students' internal visualization skills (an exemplary test item is shown in Appendix B).

Students had to take the internal visualization skills test as an assignment and received course credit based only on their completion, regardless of their scores. Students were not allowed to use a calculator during the test. The reliability for the internal visualization skills assessment was high with Cronbach's $\alpha = .79$.

3) **Measures**

We extracted time-stamped logs of problem-solving interactions in Signals Tutor in the same way as in Study 1. Additionally, we computed students' scores on the internal visualization skills assessment as the average number of correct responses.

4) **Procedures**

During the electrical engineering course, students individually completed four sections of the instructional activities supplied in Signals Tutor as either in-class activities (Weeks 1, 2, 4) or as homework (Week 5). In Week 5, students were instructed to complete the last section of Signals Tutor and the internal visualization skills test as a homework assignment.

B. Results

To address RQ3 (how log data capturing interactions with visuals relates to the formal assessment of internal visualization skills), we adapted the statistical models developed in Study 1 using the measures listed in Table 3.

TABLE 3
Descriptive summary of the variables utilized in Study 2

($N = 109$)

Variables	Mean	SD	Min	Max
Internal visualization skills test scores	17.28	3.41	7	21
Error rates during sense-making activities	0.30	0.22	0.08	1.46
Error rates during perceptual-fluency activities	0.31	0.27	0.03	1.42
Total hint requests	9.84	20.53	0	123

Note. Dependent variable: Internal visualization skills test scores

Table 4 summarizes the results from the linear regression analyses. We found that students' error rates during sense-making (Model 1, $t = -4.93$, $p < .01$) and perceptual-fluency activities (Model 2, $t = -4.58$, $p < .01$) were statistically significant predictors of their performance on the internal visualization skills test, with 18% and 16% of the model explanation, respectively (Model 1: $R^2_{Adjusted} = 0.18$, Model 2: $R^2_{Adjusted} = 0.16$). Students with lower error rates on sense-making and perceptual-fluency activities achieved significantly higher scores on the internal visualization skills test. In addition to the error rates, the total number of hints requested during sense-making activities (Model 3, $t = -3.29$, $p < .01$) was also a statistically significant predictor of their performance on the

internal visualization skills test, explaining 8% of the variance ($R^2_{Adjusted}$ = 0.08). Students who requested fewer hints performed significantly better on the internal visualization skills test.

In Model 4A, which added the second-highest predictive variable to the best single-variable model (Model 1), both the error rates during the sense-making activities and perceptual-fluency activities showed statistically significant predictive power on the internal visualization skills test scores, with 3.5% increase in model explanations (Model 4A: $R^2_{Adjusted}$ = 0.21). In Model 4B, which included all predictive variables, the error rates during the sense-making activities were the sole statistically significant predictor of test scores, accounting for 21% of the variance, which was the highest model explanation compared to the others with an increase of 1.2% over Model 4A (Model 4B: $R^2_{Adjusted}$ = 0.21).

TABLE 4
Summary of linear regression models in Study 2

(N = 108)

Model	Predictor variables	β	SE	t	p
1	Intercept		0.50	38.51	0.00**
	Error rates during sense-making activities	-0.43	1.35	-4.93	0.00**
2	Intercept		0.47	40.68	0.00**
	Error rates during perceptual-fluency activities	-0.41	1.13	-4.58	0.00**
3	Intercept		0.35	51.25	0.00**
	Total hint requests	-0.30	0.02	-3.29	0.00**
4A	Intercept		0.51	38.33	0.00**
	Error rates during sense-making activities	-0.29	1.64	-2.77	0.01*
	Error rates during perceptual-fluency activities	-0.23	1.35	-2.19	0.03*
4B	Intercept		0.51	38.20	0.00**
	Error rates during sense-making activities	-0.26	1.67	-2.45	0.02*
	Error rates during perceptual-fluency activities	-1.87	1.38	-1.87	0.07
	Total hint requests	-1.31	0.02	-1.31	0.20

Note. * $p < .05$, ** $p < .01$
Dependent variable: Internal visualization skills test scores
Model 1: $F_{(1,107)} = 24.30$, $p < .01$, $R^2 = 0.19$, $R^2_{Adjusted} = 0.18$
Model 2: $F_{(1,107)} = 20.99$, $p < .01$, $R^2 = 0.16$, $R^2_{Adjusted} = 0.16$
Model 3: $F_{(1,107)} = 10.84$, $p < .01$, $R^2 = 0.09$, $R^2_{Adjusted} = 0.08$
Model 4A: $F_{(2,106)} = 14.92$, $p < .01$, $R^2 = 0.22$, $R^2_{Adjusted} = 0.21$
Model 4B: $F_{(3,105)} = 10.63$, $p < .01$, $R^2 = 0.23$, $R^2_{Adjusted} = 0.21$

C. Discussion

The results suggest that students' log data from the ITS is a useful indicator of performance on the formal assessment (RQ3). Students with lower error rates across the sense-making and perceptual-fluency activities and fewer hint requests tended to have higher scores on the internal visualization skills test. This aligns with the quantitative findings of Study 1, which demonstrated the predictive power of students' log data on their production of representational gestures, a validated measure for internal visualization skills (RQ2). Compared to Study 1, Study 2 has higher statistical power to support this claim. Further, Study 2 has higher external validity since the learning intervention was implemented in an actual classroom setting.

More specifically, our results show that, among log data of students' problem-solving behaviors, error rates during sense-making activities showed the highest predictive power of students' internal visualization skills when all variables were incorporated in a single model. Recall that sense-making competencies describe the ability to understand how visual features correspond to domain-relevant concepts and to link multiple visuals based on their conceptual cohesion [38]. Our findings suggest that sense-making competencies, exhibited while students work with external visuals, may be a strong predictor of internal visualization skills, exhibited when external visuals are absent. Altogether, our findings suggest that log data from an ITS with interactive visuals can serve as an informal assessment of internal visualization skills.

Still, Study 2 has several limitations that should be addressed in future research. First, students received no pre- and post-intervention knowledge tests. Although all students were students enrolled in one basic electrical engineering course, their prior knowledge could vary. Future research should include pre-tests to examine correlations with internal visualization skills. Further, it would be interesting to determine to what extent internal visualization skills correlate with content knowledge gains that could be assessed via post-tests. Second, we did not observe students while they took the internal visualization skills assessment. While students were instructed not to use additional equipment, such as a calculator, we do not know if students complied. Third, our study used frequency-based log data of students' problem-solving behaviors. Future research should broaden its analytic scope to incorporate time-based log data of problem-solving behaviors (e.g., the duration of time it took to solve each problem) or deepen analysis into identifying when and where students made errors and asked for hints. This would enable us to investigate whether time-based analytics could deepen our understanding of students' difficulties with internal visualization during the learning process and may yield insights into how to provide additional scaffolding. Fourth, the development of the new test did not involve a stringent validation process. As the purpose of our study was to showcase that it is possible to assess students' internal visualization skills using a combination of formal and informal assessments, we do not propose the developed test as a definitive measure. Finally, our internal visualization skills assessment focuses on a specific engineering concept related to sinusoids. Internal visualization skills are always bound to specific visuals and associated concepts, and although knowledge of sinusoids is a fundamental concept in engineering with extensive applicability to electrical engineering, computer engineering, and biomedical engineering because they allow representing signals and describe many natural and technical phenomena [7]. Nevertheless, future research should expand our work to other

STEM domains to investigate the generalizability of our findings.

VI. General Discussion and Conclusion

Overall, our studies demonstrate the utility of multimodal data triangulation for the development of an assessment of internal visualization skills. Our approach explored associations and patterns among representational gestures, interview data, log data, and test data. Further, because the only established measure of internal visualization skills is not scalable, we combined a small-scale lab study (Study 1) and a larger-scale classroom study (Study 2) to verify the association among the three measures.

Our study illustrates that it is possible to assess students' internal visualization skills using both formal and informal assessments. We provide empirical evidence that log data from an ITS can serve as an informal assessment of students' internal visualization skills. At the same time, we present a new formal assessment of internal visualization skills. Our findings demonstrated the validity of the formal assessment, which focused on situations in which students are required to internally conceptualize visual information and retrieve that information to provide reasoning and solve problems. We found the informal and formal assessments correlated with representational gestures (Study 1) as well as with each other (Study 2).

Demonstrating the feasibility of developing assessments of internal visualization skills is important because these skills are essential in many STEM domains. Any domain where visualizations play a role in early instruction but are later faded out—for instance in favor of mathematical formulas as is common in engineering and mathematics—internal visualization skills play a critical role in students' ongoing learning. Our research contributes to the broader understanding these internal visualization skills, which are not directly observable and occur within students' minds, as well as the complex interplay between internal visualization skills and problem-solving behaviors. The nuanced exploration of these relationships and the development of scalable assessments pave the way for future research endeavors aimed at enabling researchers and content developers to evaluate and support internal visualization skills through instructional interventions and educational technologies.

Our research shows how multimodal triangulation can be used for the development of informal and formal assessments of internal visualization skills. In doing so, we demonstrate the utility of multimodal triangulation methods to examine relationships between informal and formal measures. In the growing line of educational research employing multimodal data to understand learning processes and outcomes, such as multimodal learning analytics (MMLA), triangulation is a common strategy for integrating these diverse data streams. By triangulating how one type of data (often machine-detected data) relates to other data (human-annotated data), researchers endeavor to establish the validity of the potential proxies employed to infer students' cognitive skills [42] or constructs [43]. Our work extends conventional triangulation practices by incorporating a multifaceted array of cross-validation—ranging from human-annotated data to log data, progressing from a small-scale lab setting to a larger-scale classroom setting, and transitioning from informal measures during the learning process to formal measures for assessing learning outcomes. These multi-tiered cross validations not only demonstrate that the informal and formal measures we suggested indeed assess what they intend to assess, but also exemplify the adaptable utility of multimodal triangulation, thereby contributing to the growth of the MMLA field by fostering the assurance of indicator validity. This approach can be applied to the many learning contexts and STEM domains where internal visualization skills are important for student success.

This contribution is relevant to many stakeholders in the field of learning technologies. First, it may help designers of educational technologies to improve the effectiveness of digital learning tools. By offering an informal assessment (e.g., log data) of internal visualization skills, our research can help identify "trouble spots" where students struggle with internal visualization and to determine when scaffolding is needed. The use of log data as an informal assessment is a first step to designing educational technologies that provide timely interventions and tailored support during the learning process in order to improve students' internal visualization skills, which are essential for many STEM disciplines. Given that informal measures of internal visualization skills are highly scalable because they do not require precious classroom time, our research is an important contribution towards helping design educational technologies that support students' acquisition of internal visualization skills relating to visuals of important domain-relevant concepts such as sinusoids.

Second, our approach is relevant to instructors and instructional designers more broadly. By contributing a formal assessment of internal visualization skills, we offer a tool to summatively evaluate how accurately students can mentally represent the visual information they have learned and how well they are prepared for subsequent instruction on more complex, advanced concepts that are typically presented as symbolic representations with limited visual aids. This is useful for both designers of educational technologies to evaluate their effectiveness and for instructors who want to better support their students' learning.

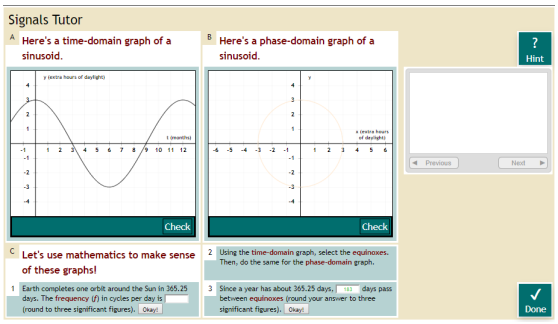
Further, our work yields novel opportunities for educational researchers. Internal visualization skills are of high interest to researchers who examine how students mentally represent information. Yet, this research lacks a scalable means to assess these unobservable skills. Potentially as a consequence of this difficulty, prior research has not sufficiently distinguished between situations when students when external visuals are present versus absent regarding assessments of internal visualization skills. To our knowledge, related research typically focuses on situations where external visuals are present [8], [10], [23] even though this may confound the source of students' mental retention of visual information. An advantage of our formal internal visualization skills assessment is that it minimizes the impact of external visuals on students' internal visualization of concepts. This consideration is

important to accurately evaluate students' ability to internally conceptualize and manipulate the information.

Finally, we believe that our work has is broadly applicable to many STEM fields. Our studies were situated in engineering instruction on sinusoids, which are a fundamental engineering concept with broad applications to electrical engineering, computer engineering, and biomedical engineering [7]. Yet, internal visualization skills play a fundamental role in many other STEM fields, such as chemistry [13] or physics [14]. Thus, supporting internal visualization skills is key to students' success in many domains.

APPENDIX A

An example of provided visuals that accompanied the following interview question (#5).

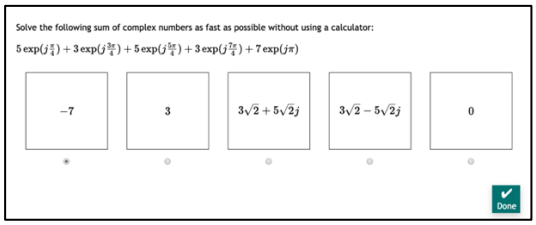


Interview questions

1-1. How did you figure out the vector corresponds to the winter solstice?
1-2. Why does the spring equinox correspond to the vector lying along negative y-axis instead of positive y-axis?
2. How can we translate the time axis on the time-domain graph to the phasor graph?
3. What does the phase shift mean on the phase-domain graph?
4. Why does the phasor corresponding to the equinoxes lie on the y-axis?
5. What does one time period of the sinusoid mean in the phase domain graph?
6. We showed that the frequency of the sinusoid that represents hours of daylight relative to the equator on the Earth is larger than that on Mars. Equivalently, the time period of the sinusoid corresponding to Earth is larger than that of the sinusoid corresponding to Mars. What does this mean in the phase-domain graph?
7. How would you measure the magnitude of a phasor if it did not lie on the x-axis or y-axis?
8. How do you measure a phase to be negative?
9. Can you make connections between the two: Amplitude and the initial phase?

APPENDIX B

An exemplary test item.



ACKNOWLEDGMENT

This work was supported by NSF DUE 1933078.

REFERENCES

[1] D. H. Uttal and C. A. Cohen, "Spatial Thinking and STEM Education.: When, Why, and How?," in *Psychology of Learning and Motivation*, vol. 57, Academic Press., 2012, pp. 147–181.

[2] D. N. Rapp and C. A. Kurby, "The 'Ins' and 'Outs' of Learning: Internal Representations and External Visualizations," in *Visualization: Theory and Practice in Science Education*, Springer Netherlands, 2008, pp. 29–52.

[3] C. A. Cohen and M. Hegarty, "Individual differences in use of external visualisations to perform an internal visualisation task," *Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn.*, vol. 21, no. 6, pp. 701–711, 2007.

[4] K. Atit, K. Gagnier, and T. F. Shipley, "Student Gestures Aid Penetrative Thinking," *J. Geosci. Educ.*, vol. 63, no. 1, pp. 66–72, 2015, doi: 10.5408/14-008.1.

[5] S. Hsi, M. C. Linn, and J. E. Bell, "The Role of Spatial Reasoning in Engineering and the Design of Spatial," *J. Eng. Educ.*, vol. 86, no. 2, pp. 151–158, 1997.

[6] P. Coppens, J. Van Den Bossche, and M. De Cock, "Student understanding of phase shifts, frequency and Bode plots," *Int. J. Electr. Eng. Educ.*, vol. 54, no. 3, pp. 247–261, 2017, doi: 10.1177/0020720916680373.

[7] J. K. Nelson, M. A. Hjalmarson, K. E. Wage, and J. R. Buck, "Students' interpretation of the importance and difficulty of concepts in signals and systems," in *2010 IEEE frontiers in education conference (FIE)*, 2010, no. November, pp. T3G-1, doi: 10.1109/FIE.2010.5673121.

[8] L. English and D. King, "Engineering education with fourth-grade students: Introducing design- based problem solving.," *Int. J. Eng. Educ.*, vol. 33, no. 1, pp. 346–360, 2017.

[9] P. Garber and S. Goldin-Meadow, "Gesture offers insight into problem-solving in adults and children," *Cogn. Sci.*, vol. 26, no. 6, pp. 817–831, 2002.

[10] M. Hegarty, S. Mayer, S. Kriz, and M. Keehner, "The role of gestures in mental animation," *Spat. Cogn. Comput.*, vol. 5, no. 4, pp. 333–356, 2005, doi: 10.1207/s15427633scc0504_3.

[11] M. J. Nathan, M. Wolfgram, R. Srisurichan, C. A. Walkington, and M. W. Alibali, "Threading mathematics through symbols, sketches, software, silicon, and wood: Teachers produce and maintain cohesion to support STEM integration," *J. Educ. Res.*, vol. 110, no. 3, pp. 272–293, 2017, doi: 10.1080/00220671.2017.1287046.

[12] S. Kita, M. W. Alibali, and M. Chu, "How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis," *Psychol. Rev.*, vol. 124, no. 3, pp. 245–266, 2017, doi: 10.1037/rev0000059.

[13] M. A. Rau, "Do Knowledge-Component Models Need to Incorporate Representational Competencies?," *Int. J. Artif. Intell. Educ.*, vol. 27, no. 2, pp. 298–319, 2017, doi: 10.1007/s40593-016-0134-8.

[14] V. López and R. Pintó, "Identifying secondary-school students' difficulties when reading visual representations displayed in physics simulations," *Int. J. Sci. Educ.*, vol. 39, no. 10, pp. 1353–1380, 2017, doi: 10.1080/09500693.2017.1332441.

[15] S. M. Kosslyn, *Image and brain*. MIT press, 1994.

[16] S. M. Kosslyn, "Mental images and the brain," *Cogn. Neuropsychol.*, vol. 22, no. 3/4, pp. 333–347, 2005.

[17] M. G. McGee, "Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences.," *Psychol. Bull.*, vol. 86, no. 5, p. 889, 1979.

[18] T. Lowrie, T. Logan, and M. Hegarty, "The Influence of Spatial Visualization Training on Students' Spatial Reasoning and Mathematics Performance," *J. Cogn. Dev.*, vol. 20, no. 5, pp. 729–751, 2019, doi: 10.1080/15248372.2019.1653298.

[19] M. Hegarty, "Diagrams in the mind and in the world: Relations between internal and external visualizations," in *International Conference on Theory and Application of Diagrams*, 2004, pp. 1–13.

[20] W. Schnotz, "An integrated model of text and picture comprehension," in *The Cambridge handbook of multimedia learning*, 2nd ed., R. E. Mayer, Ed. NY: NY: Cambridge University

Press, 2014, pp. 72–103.

[21] M. A. Rau, "Conditions for the Effectiveness of Multiple Visual Representations in Enhancing STEM Learning," *Educ. Psychol. Rev.*, vol. 29, no. 4, pp. 717–761, 2017, doi: 10.1007/s10648-016-9365-3.

[22] S. Ainsworth, "DeFT: A conceptual framework for considering learning with multiple representations," *Learn. Instr.*, vol. 16, no. 3, pp. 183–198, 2006, doi: 10.1016/j.learninstruc.2006.03.001.

[23] C. H. Lin and C. M. Chen, "Developing spatial visualization and mental rotation with a digital puzzle game at primary school level," *Comput. Human Behav.*, vol. 57, pp. 23–30, 2016, doi: 10.1016/j.chb.2015.12.026.

[24] M. W. Alibali and M. J. Nathan, "Embodiment in Mathematics Teaching and Learning: Evidence From Learners' and Teachers' Gestures," *J. Learn. Sci.*, vol. 21, no. 2, pp. 247–286, 2012, doi: 10.1080/10508406.2011.611446.

[25] M. Chu and S. Kita, "The Nature of Gestures' Beneficial Role in Spatial Problem Solving," *J. Exp. Psychol. Gen.*, vol. 140, no. 1, pp. 102–116, 2011, doi: 10.1037/a0021790.

[26] R. M. Krauss, "Why do we gesture when we speak?," *Curr. Dir. Psychol. Sci.*, vol. 7, no. 2, p. 54, 1998, doi: 10.1111/1467-8721.ep13175642.

[27] M. W. Alibali, R. C. Spencer, L. Knox, and S. Kita, "Spontaneous Gestures Influence Strategy Choices in Problem Solving," *Psychol. Sci.*, vol. 22, no. 9, pp. 1138–1144, 2011, doi: 10.1177/0956797611417722.

[28] A. B. Hostetter and M. W. Alibali, "Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production," *Gesture*, vol. 7, no. 1, pp. 73–95, 2007, doi: 10.1075/gest.7.1.05hos.

[29] T. Göksun, S. Goldin-Meadow, N. Newcombe, and T. Shipley, "Individual differences in mental rotation: What does gesture tell us?," *Cogn. Process.*, vol. 14, no. 2, pp. 153–162, 2013, doi: 10.1007/s10339-013-0549-1.

[30] B. Schneider and P. Blikstein, "Unraveling students' interaction around a tangible interface using multimodal learning analytics," *J. Educ. Data Min.*, vol. 7, no. 3, pp. 89–116, 2015.

[31] X. Ochoa, F. Domínguez, B. Guamán, R. Maya, G. Falcones, and J. Castells, "The RAP system: Automatic feedback of oral presentation skills using multimodal analysis and low-Cost sensors," in *ACM International Conference Proceeding Series*, 2018, pp. 360–364, doi: 10.1145/3170358.3170406.

[32] S. Baker and P. S. Inventado, "Educational data mining and learning analytics: Potentials and possibilities for online education.," in *Emergence and Innovation in Digital Learning*, G. Veletsianos, Ed. Wiley Online Library, 2016, pp. 83–98.

[33] M. Haridas, G. Gutjahr, R. Raman, R. Ramaraju, and P. Nedungadi, "Predicting school performance and early risk of failure from an intelligent tutoring system," *Educ. Inf. Technol.*, vol. 25, no. 5, pp. 3995–4013, 2020, doi: 10.1007/s10639-020-10144-0.

[34] R. Liu, J. C. Stamper, and J. Davenport, "A Novel Method for the In-Depth Multimodal Analysis of Student Learning Trajectories in Intelligent Tutoring Systems," *J. Learn. Anal.*, vol. 5, no. 1, pp. 41–54, 2018, doi: 10.18608/jla.2018.51.4.

[35] I. Roll, V. Aleven, B. M. McLaren, and K. R. Koedinger, "Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system," *Learn. Instr.*, vol. 21, no. 2, pp. 267–280, 2011, doi: 10.1016/j.learninstruc.2010.07.004.

[36] M. Feng and N. Heffernan, "Addressing the assessment challenge with an online system that tutors as it assesses," *User Model User-Adap Inter*, vol. 19, pp. 243–266, 2009, doi: 10.1007/s11257-009-9063-7.

[37] K. Koedinger, E. McLaughlin, and N. Heffernan, "A quasi-experimental evaluation of an On-line formative assessment and tutoring system," *J. Educ. Comput. Res.*, vol. 43, no. 4, pp. 489–510, 2010, doi: 10.2190/EC.43.4.d.

[38] S. Ainsworth, "The educational value of multiple-representations when learning complex scientific concepts," in *Visualization: Theory and practice in science education*, Springer, 2008, pp. 191–208.

[39] P. J. Kellman and C. M. Massey, "Perceptual learning, cognition, and expertise," in *Psychology of learning and motivation*, vol. 58, Elsevier, 2013, pp. 117–165.

[40] J. Rho, M. A. Rau, and B. D. Van Veen, "Preparing future learning with novel visuals by supporting representational competencies," in *International Conference on Artificial Intelligence in Education*, 2022, pp. 66–77, doi: 10.1007/978-3-031-11644-5_54.

[41] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[42] E. L. Starr, J. M. Reilly, and B. Schneider, "Toward using multi-modal learning analytics to support and measure collaboration in co-located dyads," in *Proceedings of International Conference of the Learning Sciences, ICLS*, 2018, vol. 1, pp. 448–455.

[43] H. Sung, C. Shengyang, A. R. Ruis, and D. W. Shaffer, "Reading for breadth, reading for depth: Understanding the relationship between reading and complex thinking using epistemic network analysis," in *A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings, 13th International Conference on Computer Supported Collaborative Learning (CSCL) 2019*, 2019, vol. 1, pp. 376–383, doi: https://doi.dx.org/10.22318/cscl2019.376.

Hanall Sung is a postdoctoral fellow in the Learning Sciences and Psychological Studies at University of North Carolina at Chapel Hill. She received her Ph.D. in Educational Psychology (Learning Sciences) from University of Wisconsin-Madison in 2023. Her research focuses on multimodal ways of knowledge building and sharing in STEM learning with technology. Her research interests include (multimodal) learning analytics, computer-supported collaborative learning, and human-computer interaction.

Martina A. Rau is an Associate Professor in the Department of Humanities, Social and Political Sciences, Swiss Federal Institute of Technology in Zurich, Switzerland. She received her Ph.D. in Human-Computer Interaction from Carnegie Mellon University in 2013. Her research focuses on learning with multiple external representations in educational technologies. She uses a multi-methods approach to integrate learning outcome measures and process-level measures.

Barry D. Van Veen is a Lynn H. Matthias Professor in the Department of Electrical and Computer Engineering and is affiliated with the Department of Biomedical Engineering at the University of Wisconsin-Madison. He received his Ph.D. in Electrical Engineering from University of Colorado in 1986. His research interests involve statistical signal processing and its application that includes problems in adaptive filtering, adaptive beamforming, signal detection, and estimation, equalization, and sensor array signal processing.