Carousel: Improving the Accuracy of Virtual Reality Assessments for Inspection Training Tasks

Jacob Belga, Tiffany D. Do, Ryan Ghamandi, Ryan P. McMahan, Joseph J. LaViola Jr.
jbelga1128@Knights.ucf.edu,tiffanydo@knights.ucf.edu,ryanghamandi1@Knights.ucf.edu,rpm@ucf.edu,jjl@cs.ucf.edu
University of Central Florida
Orlando, FL, USA

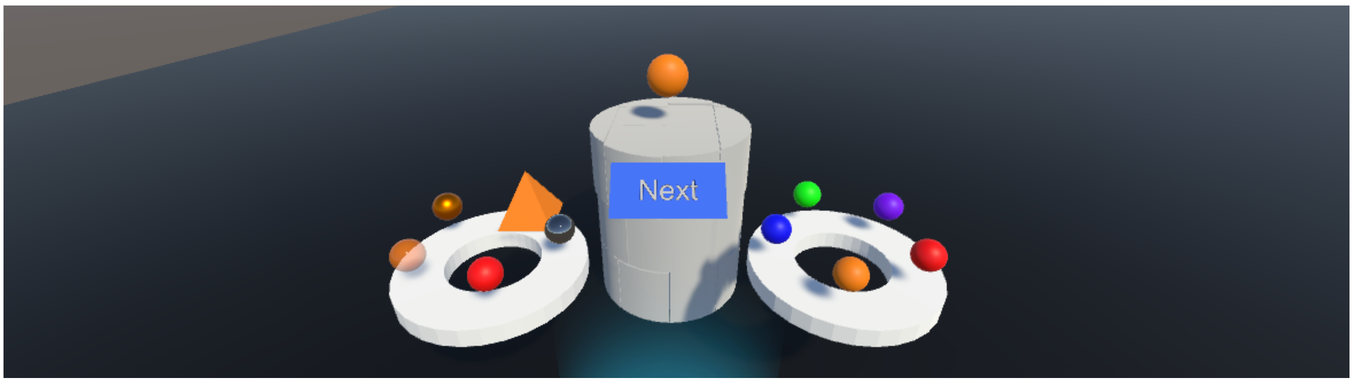

Figure 1: Our new Carousel method is an in-VR assessment method that requires users to actively select the correct state for an inspection point by controlling the breadth of possibilities with the left 3D carousel (e.g., color, material, shape, shininess, and transparency) and the depth of possibilities with the right 3D carousel (e.g., which color).

ABSTRACT

Training simulations in virtual reality (VR) have become a focal point of both research and development due to allowing users to familiarize themselves with procedures and tasks without needing physical objects to interact with or needing to be physically present. However, the increasing popularity of VR training paradigms raises the question: Are VR-based training assessments accurate? Many VR training programs, particularly those focused on inspection tasks, employ simple pass or fail assessments. However, these types of assessments do not necessarily reflect the user's knowledge.

In this paper, we present Carousel, a novel VR-based assessment method that requires users to actively employ their training knowledge by considering all relevant scenarios during assessments. We also present a within-subject user study that compares the accuracy of our new Carousel method to a conventional pass or fail method for a series of virtual object inspection tasks involving shapes and colors. The results of our study indicate that the Carousel method affords significantly more-accurate assessments of a user's knowledge than the binary-choice method.

VRST '22, November 29–December 01, 2022, Virtual/Tsukuba, Japan

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**.

KEYWORDS

Virtual reality, virtual inspections, training assessments.

ACM Reference Format:
Jacob Belga, Tiffany D. Do, Ryan Ghamandi, Ryan P. McMahan, Joseph J. LaViola Jr.. 2022. Carousel: Improving the Accuracy of Virtual Reality Assessments for Inspection Training Tasks. In *VRST '22: ACM Symposium on Virtual Reality Software and Technology, November 29–December 01, 2022, Virtual/Tsukuba, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Inspection training is a crucial element in preparing individuals taking on the role and responsibility of thoroughly understanding and evaluating real-world systems and environments. Given this, researchers have explored training inspection tasks in virtual reality (VR) experiences to provide individuals with accessible and inexpensive alternatives to learning such inspections within real-world contexts. A few examples of such VR-based training solutions include inspecting aircraft cargo bays for defects [10], construction sites for hazards [19], and work zones for deficiencies [1].

In addition to training inspection tasks in VR, researchers have also investigated assessing an individual's knowledge of inspection tasks in VR. Assessment methods in VR have included selecting defects or hazards within the environment [8, 35], passing or failing inspection points [7, 21], selecting the current state of a point from a small set of options [2, 19], or some hybrid combination of these

[27, 36]. However, while researchers have employed these various methods for in-VR assessments of inspection tasks, there has been little research into the accuracy or efficacy of these techniques.

In this paper, we present a new in-VR assessment method inspired by traditional carousel user interfaces [4, 13], which we thereby call the "Carousel" method. Carousel is a novel VR-based assessment method that presents users with the entire breadth and depth of inspection possibilities. Unlike prior in-VR assessment methods, which only require users to passively recognize the correct state from a small subset of possibilities, the Carousel method requires users to actively recall the correct state from all relevant inspection scenarios. Hence, the Carousel method should better reflect users' actual knowledge compared to the prior methods.

To evaluate the efficacy of our new method and to better understand in-VR assessments of inspection tasks, we present a within-subject study comparing our Carousel method to the binary-choice method (i.e., passing or failing each point). To limit potential confounds due to varying degrees of prior context knowledge and to afford tasks of equal difficulty for our within-subject design, we employed a simple inspection task requiring the correct identification of the color and shape of 10 sequential objects. The results of our study (N = 30) indicate that the Carousel method affords significantly more-accurate assessments of user knowledge than the binary-choice method, which yields significantly higher (i.e., inflated) in-VR assessment scores. From our findings, we believe the Carousel method is a novel in-VR assessment technique that provides a more-accurate measure of a user's knowledge than a prior binary method. In the subsequent sections, we discuss related work, the design of Carousel, our comparison study, and its implications.

2 RELATED WORK

To better understand related work, we conducted a traditional literature survey using Google Scholar to find examples of in-VR assessments for inspection-based training tasks. Table 1 provides a summary of the relevant publications that we found. In the following sections, we discuss the four types of in-VR assessments that we found: selection-based, binary-choice, multiple-choice, and hybrid assessments. We also discuss examples of external assessments.

2.1 Selection-based Assessments

Selection-based assessments rely on the user to select areas or objects to identify defects, deficiencies, or hazards in VR.

An early example of a VR training application employing the selection-based assessment method is the aircraft inspection simulation developed by Duchowski et al. [10]. The simulation included a head-mounted display (HMD), a six-degree-of-freedom (6-DOF) mouse for 3D interactions, and a video-based eye tracker for conducting a virtual inspection of an aircraft cargo bay. During the virtual inspection assessment, users used the 6-DOF mouse to point at and select defects within the aircraft environment. This aircraft inspection application and selection-based assessment were employed in numerous studies, including eye movement analysis [10], investigating presence [35], comparison to a desktop-based aircraft inspection simulator [34], evaluating display techniques for representing the gaze of a virtual trainer [23], and investigating the effects of feedforward training [31].

Table 1: Summary of related work pertaining to in-VR inspection assessments.

		In-VR Assessment			
Refs	**Context**	**Selection**	**Binary**	**Multiple**	**External Assessment**
[10]	Aircraft Inspection	✓			
[35]	Aircraft Inspection	✓			
[34]	Aircraft Inspection	✓			
[23]	Aircraft Inspection	✓			
[31]	Aircraft Inspection	✓			
[38]	Construction Hazards	✓			
[8]	Work Zone Deficiencies	✓			
[1]	Work Zone Deficiencies	✓			
[29]	Fire Safety Inspection	✓			
[7]	Vehicle Inspection		✓		
[2]	Building Inspection			✓	
[15]	Mining Hazards			✓	
[33]	Tree Identification			✓	✓
[19]	Construction Hazards			✓	
[25]	Construction Hazards			✓	
[11]	Construction Hazards			✓	
[20]	Conveyor Hazards	✓		✓	
[22]	Conveyor Hazards	✓		✓	✓
[36]	Oil Depot Hazards	✓		✓	
[27]	Construction Hazards	✓		✓	
[28]	Construction Hazards	✓		✓	
[30]	Construction Hazards	✓		✓	✓
[21]	Haul Truck Inspection		✓	✓	
[22]	Haul Truck Inspection		✓	✓	✓

Another selection-based example is the assessment method employed by the construction safety simulation developed by Zhao and Lucas [38]. In their desktop-based simulation, the user can freely navigate a road construction site using first-person or third-person perspective controls. In their VR-based assessment scenario, the user is required to identify hazards by clicking on them. For a similar training context, Chang et al. [8] developed an HMD-based simulation for training how to inspect work zones. For their in-VR assessment, users had to point out any deficiencies in temporary traffic control and signage. This work zone training simulation was eventually tested with Department of Transportation employees [1]. In another HMD-based simulation, Pitana et al. [29] employed a selection-based assessment method for a fire safety inspection application, in which users must select firefighting equipment.

It is important to note that selection-based assessments, like the aforementioned examples, rely on implicit decisions. While users make explicit selections to identify issues that they are aware of, any inspection points not selected are treated as if the user passed the point, even if the user did not observe it during the assessment. These implicit decisions make it difficult to determine if an assessment error is due to the user cognitively deciding that the point is passable or simply due to the user not observing the relevant inspection point. While head tracking [24] or eye tracking [10] can be used to determine whether an inspection point was observable within the user's field of view at some time, it is difficult to determine whether the user mentally assessed the point.

Our new Carousel assessment method is quite different compared to selection-based assessments. First, selection-based assessments function as binary assessments to implicitly pass or explicitly fail each inspection point. On the other hand, our Carousel method requires users to consider each inspection point and to choose its correct state from the breadth and depth of possible scenarios. Hence, our Carousel method affords explicit assessments of each inspection point and requires users to know the correct state, as opposed to potentially guessing whether to pass or fail it.

2.2 Binary-Choice Assessments

Binary-Choice asessments require the user to consider each inspection point and choose whether or not to pass or fail it. By this token, the aforementioned selection-based assessments and binary-choice assessments provide the ability for the user to guess whether to pass or fail an inspection point with a 50% probability of success, regardless of procedural knowledge. In contrast to the selection-based assessments, which rely on implicit decision-making, binary-choice assessments impose explicit decision-making on the user side, which provides better insight into the user's actual knowledge.

An example of the binary-choice assessment is the vehicle inspection simulation developed by Dukes et al. [7]. In their simulation, users were required to evaluate each inspection point and then decide whether to pass or fail a given inspection item. Similarly, McMahan et al. [21] employed a binary choice approach to virtually assessing user's knowledge in a haul truck inspection task. At each inspection point during their simulation, user's were required to denote whether a defect is detected.

Both the binary-choice assessment and our new Carousel method are explicit approaches that require explicit decisions from the user. As mentioned, the binary-choice approach affords a 50% probability of correctly identifying an inspection point. Conversely, our proposed Carousel method affords a much smaller correct selection probability that is dependent on the total number of possible scenarios (i.e., $1/N$). Hence, Carousel method requires users to actively engage and recall the correct state and should provide a more accurate reflection of the user's knowledge.

2.3 Multiple-Choice Assessments

Multiple-choice assessments require the user to consider each inspection point and choose an option for the point from a set of possibilities. These options can be used to indicate the correct state of the point [33], what is wrong with the point [19], or how to address any present issues [36].

Beh et al. [2] employed a multiple-choice assessment approach in their building utility inspection simulation. At each inspection stage, users were prompted with an in-VR quiz with multiple choices. Similarly, Isleyen and Duzgun [15] prompted users to select one of three failure options when inspecting mining hazards. For their tree identification assessment, Vellingiri et al. [33] presented an in-VR quiz consisting of three questions, each with three choices, at each tree. Likewise, Li et al. [19] prompted users with four hazard identification choices when they entered a hazardous region within their construction site virtual environment. Finally, Moore et al. [25] and Eiris et al. [11] have developed a multiple-answer variant of the multiple-choice assessment by employing a checkbox interface for selecting all of the hazards present at an inspection point, as opposed to only one hazard.

Both the multiple-choice approach and our Carousel technique are explicit assessments requiring users to select the correct option from a set of possibilities. The key difference between the approaches is that the multiple-choice assessment employs a subset of likely possibilities (usually three or four options) while our Carousel assessment employs the complete set of possibilities. Hence, the multiple-choice approach affords a higher probability of getting each inspection point correct (e.g., 1/4) than the Carousel approach.

2.4 Hybrid Assessments

Within the literature, researchers have also explored hybrid approaches that combine two of the three assessment approaches discussed above.

The most common hybrid approach, based on our review, is the selection-then-multiple-choice assessment. For example, Lucas et al. [20] developed an in-VR assessment for identifying hazards around conveyor systems and required users to first select hazards and then answer an in-VR multiple-choice quiz pertaining to each hazard. Likewise, Wan et al. [36] had users to select hazards within their virtual oil depot environment and then select an option from an in-VR multiple-choice quiz that describes the correct action to take. Similarly, Pedro et al. [27] required users to select hazards within their virtual construction site environment and then use a drop down list to choose the option that best describes the hazard. In their virtual construction site simulation, Perlman et al. [28] and Sacks et al. [30] required users to select hazards and then assess the risk level of each hazard.

The only other hybrid approach that we found within the literature was a binary-choice-then-multiple-choice assessment employed by McMahan et al. [21]. For their in-VR assessment of haul truck inspections, they first required users to use a binary-choice approach to indicate whether a defect was detected at an inspection point and then prompted the user with multiple choices for what the corrective action should be.

Like our Carousel method, the hybrid assessment approaches can involve the user making multiple decisions for each inspection point. However, the selection-then-multiple-choice hybrid requires an explicit selection decision. For non-selected points, the approach infers an implicit decision to pass the point. On the other hand, the binary-choice-then-multiple-choice assessment approach requires the user to make explicit decisions about each point, much like our Carousel approach. However, it only employs a subset of possibilities while the Carousel method requires the user to consider a complete set of possibilities, which we believe more accurately reflects the user's knowledge.

2.5 External Assessments

In our review of literature pertaining to in-VR assessments for inspection tasks, we found that only a few studies included knowledge assessments external to the VR simulation. In all of these cases, the researchers employed a knowledge test to evaluate the user's knowledge independent of the VR simulation.

For example, McMahan et al. [22] presented the same knowledge test before and after their VR training and assessment modules. For both their conveyor hazards application [20] and their haul truck inspection application [21], they found that post-VR knowledge test scores were significantly higher than pre-VR knowledge test scores, indicating that VR training significantly increased knowledge. Likewise, Vellingiri et al. [33] used pre and post-VR knowledge tests to assess the effectiveness of their tree identification simulation, and they also found significant knowledge increases from pre to post. Sacks et al. [30] also employed pre and post-VR knowledge tests for evaluating their construction hazards simulation, but they also administered a remote knowledge test one month later. They found that VR training significantly increased scores from pre-VR to post-VR and from pre-VR to one month later.

Unlike most of the examples of in-VR inspection assessments, the studies above independently evaluated the effectiveness of their VR training and assessments through external knowledge tests. However, none of these studies compared the outcomes of their in-VR assessments to their external knowledge tests. In this paper, we present results statistically comparing the outcomes of our in-VR assessments to a post-VR knowledge test to advance knowledge with regard to the effectiveness and accuracy of in-VR assessments.

3 THE CAROUSEL INTERFACE

3.1 Carousel Concept

The concept of the Carousel method is to require the user to choose the correct state of an inspection point from all relevant possibilities, as opposed to simply passing or failing the point or choosing an option from a subset of possibilities. We believe by requiring the user to consider all relevant possibilities, the in-VR assessment will better reflect the user's actual knowledge.

3.2 Initial Carousel Design

In early stages of our development, we initially based the Carousel interface on a more familiar and ubiquitous carousel design [4, 13]. As seen in Figure 2, our initial design presented users with left and right arrows to change the current state by cycling through the set of possible states. However, through early testing we found three problems with this design. First, the design required several interactions of using the arrows to find the correct choice among the complete set of options. Based on the GOMS (Goals, Operators, Methods, and Selection) analysis model [16], this approach inherently requires more time and effort than direct manipulation designs [32]. Second, the design did not allow the user to see the entire set of possibilities at any given instance. Finally, the 2D design did not utilize the advantages of interacting in 3D space [18].

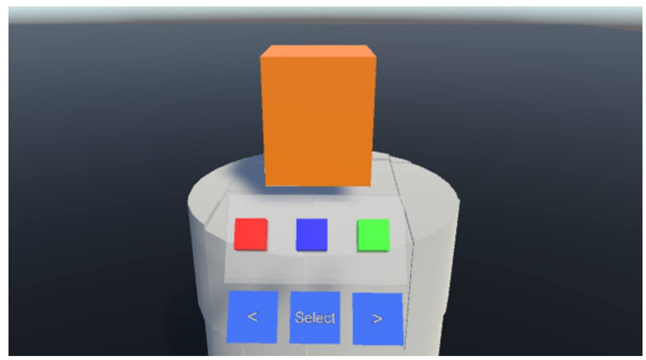

Figure 2: Our initial Carousel design. The left and right arrows allow the user to change the current option (the blue cube) by cycling through the set of options.

3.3 An Improved Carousel Design

To reduce the number of interactions required to make a selection, we redesigned our Carousel interface to use direct manipulation [32] via a simple virtual hand technique [18]. Instead of cycling through the options using the arrows, the user can instead select the option by grabbing it from a 3D circular layout with the virtual hand technique (see Figure **??**). However, this requires all of the options to be visible and within the reach of the user.

In order to present the full set of possibilities, we redesigned the Carousel interface to leverage a breadth and depth design, such as those employed by 3D menu interfaces [5, 26]. The current design presents users with two 3D carousel interfaces: one on the left for selecting the breadth of options (e.g., color, material, shape, shininess, and transparency), and one on the right for selecting the depth of the current breadth selection (e.g., which shape), as seen in Figure 1. The left carousel is populated with variations of the inspection point's current state, altered by a single feature. For example, in Figure 1, the current state is an orange sphere, so the left carousel is populated by a red sphere (color), a textured sphere (material), an orange pyramid (shape), a shiny sphere (shininess), and a semi-transparent orange sphere (transparency). By selecting one of these features, the right carousel is updated with all variations of that feature for the current state. For example, by selecting the red sphere (color) from the left carousel, the right carousel is populated with different color variations, as seen in Figure 1. This new design affords a reduced number of interactions by allowing users to use direct manipulation to select options while also affording a breadth and depth of options that are capable of representing all relevant states or scenarios.

4 EXPERIMENT

To evaluate the efficacy of our new Carousel technique and to better understand in-VR assessments, we conducted a within-subjects study comparing it to the binary-choice method. We chose to compare Carousel to the binary-choice method because both in-VR assessment techniques require explicit decisions from the user. Additionally, because this was an initial investigation into the effects of presenting users with all relevant possibilities for each inspection point, we selected the binary-choice method over the multiple-choice method to provide a more-contrasted experimental design.

To avoid potential confounds related to prior knowledge (e.g., prior knowledge of safety hazards) and to afford tasks of equal difficulty for our within-subject design, we employed a simple object inspection task requiring users to verify the color (5 options) and shape (5 options) of each object. Each inspection assessment required participants to inspect a series of 10 objects. Hence, for the within-subject design, each participant completed two in-VR assessments—one for the Carousel method and one for the binary-choice method—counterbalanced via two cohorts to avoid ordering and learning effects. Table 2 outlines the exact colors and shapes for these two sequences of inspection training tasks.

Table 2: Object sequences for training task A and B

Training Sequence A	Training Sequence B
Red Cube	Blue Cube
Blue Sphere	Green Sphere
Green Cone	Orange Cone
Orange Cylinder	Purple Cylinder
Purple Pyramid	Red Pyramid
Orange Cube	Orange Sphere
Purple Sphere	Green Cube
Red Pyramid	Purple Cone
Blue Cone	Red Cylinder
Green Cylinder	Blue Pyramid

4.1 Research Questions

Given our experimental design, we proposed the following research questions and hypotheses for our study:

RQ1: Which method, Carousel or Binary Choice, affords the highest in-VR assessment scores?

H1: Binary Choice will yield higher in-VR assessment scores due to its 50% probability of correctness, whereas Carousel has a probability of 4% due to selecting one of 25 possibilities (i.e., 5 colors x 5 shapes).

RQ2: Which method, Carousel or Binary Choice, yields the most-accurate assessment scores?

H2: Carousel will be more accurate because it requires individuals to recall the exact state of each inspection point as opposed to simply recognizing the correct state with Binary Choice.

4.2 VR Training and Assessment Tasks

For each condition, participants first completed a standard training task for learning the colors and shapes of the 10 objects and then completed the relevant in-VR assessment after a short break.

4.2.1 Training Task. In the training task, participants were exposed to a sequence of 10 objects of varying colors and shapes in VR. Regarding combination complexity, the color possibilities were *Red, Green, Blue, Purple, and Orange* and the shape type possibilities were *Cube, Cone, Sphere, Pyramid, and Cylinder* for an overall total of 25 possible object variations. Participants were asked to learn each of the 10 objects and their corresponding attributes, in this case color and shape. Participants were only allowed to view and interact with one object at a time.

At each podium, participants were shown a 3D representation of an object with an associated color and shape. Participants were also shown a text box at the podium to explicitly state, in text, what the object's color and shape was (see Figure 3). Once a participant felt they had sufficiently learned the current object, they would press a virtual "Next" button to hide the current object and prompt and to display the object at the next podium. The participant would then use the SteamVR teleportation technique to travel to the next podium, using a newly displayed teleportation waypoint. This provided participants with spatial cues about where objects were located in the virtual environment. The participants would teleport through the environment to visit each podium and object laterally from left to right. We chose to hide previously learned objects to avoid potential confounds in learning strategies (e.g., learning one object at a time vs. learning all the objects at the end).

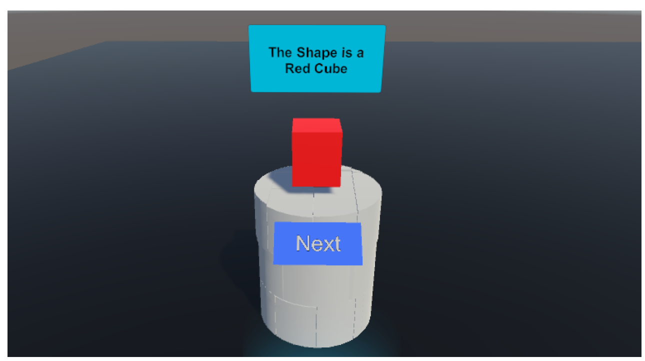

Figure 3: The standard training task interface.

4.2.2 Binary Choice Assessment. In the Binary Choice assessment, participants were exposed to a sequence of 10 podiums and objects in a similar fashion to the training task. The key difference being that at each podium an accept button and a reject button were presented for passing or failing the object. The participant's goal was to inspect each object and determine whether or not it correctly matched the object presented at that podium during the training task. Mismatches were possible in the form of an incorrect color, incorrect shape, or both attributes being incorrect. Once each object was thoroughly inspected, participants would respond via the accept or reject buttons accordingly. Figure 4 shows the Binary Choice interface seen throughout this assessment.

In our study design, we developed predetermined object sequences for the Binary Choice assessment to correspond to object Sequence A and object Sequence B. In the conception of these assessment sequences, we ensured that five of the 10 objects in the assessment were consistently incorrect in the form of being a different color, shape, or both. Table 3 outlines the colors and shapes used in the Binary Choice assessment for Sequences A and B.

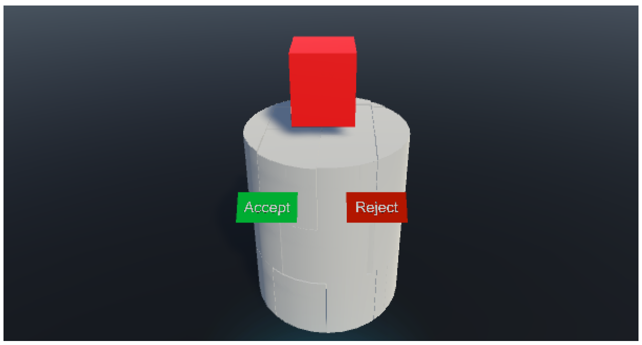

Figure 4: The Binary Choice assessment interface.

Table 3: Object sets for Binary Choice assessments A and B

Binary Choice Sequence A	Binary Choice Sequence B
Red Cube	Blue Sphere
Purple Sphere	Green Sphere
Green Cone	Red Cone
Orange Sphere	Purple Cylinder
Blue Cylinder	Red Pyramid
Orange Cube	Green Cylinder
Purple Sphere	Orange Cube
Red Sphere	Purple Cone
Blue Cone	Orange Cylinder
Red Cylinder	Blue Pyramid

4.2.3 *Carousel Assessment.*
In the Carousel assessment, the goal and workflow was similar to that of the Binary Choice assessment. The exception, however, was that the participant would need to interact with the two carousel interfaces to specify what they believed was the object that they inspected during the initial training task. Since the attributes of the assessment task focused on only color and shape, participants used the left 3D carousel to specify color and the right carousel to specify shape.

4.3 Dependent Variables

Throughout the duration of the study, we gathered several points of data in regard to participant performance and evaluation of their experience with our training and assessment simulations.

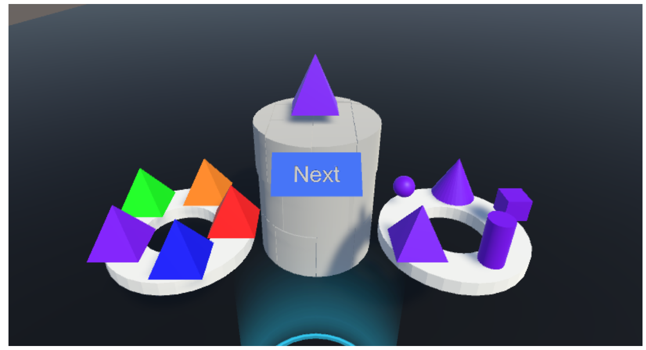

Figure 5: The simplified Carousel assessment interface.

4.3.1 *In-VR Assessment Accuracy.*
For both assessment conditions, participants received 1 point for each correct inspection point and 0 points for each incorrect inspection point. For the Carousel assessment, participants received 1 point if they selected the object with the correct color and shape for a given podium. For the Binary Choice assessment, participants received 1 point for accepting a correct object or rejecting an incorrect object at each podium.

4.3.2 *Post-VR Knowledge Test Accuracy.*
After each instance of the Carousel and Binary Choice assessment, we administered a post-VR knowledge test. This knowledge test was administered via an electronic Qualtrics survey on a nearby desktop computer. The test consisted of 10 questions corresponding to the 10 inspection objects. For each question, the participant would use two drop-down menus to specify the color and shape of each object. In terms of knowledge test accuracy, participants would receive 1 point for each completely correct answer (i.e., both color and shape were correctly specified) and 0 points for any incorrect or partially incorrect answer.

4.3.3 *Simulator Sickness.*
We measured simulator sickness at the beginning of the study and after each in-VR assessment by administering the Simulator Sickness Questionnaire (SSQ) [17].

4.3.4 *Task Load.*
We measured the task load of the in-VR assessments by administering the NASA Task Load Index (NASA-TLX) [14] after each condition. We employed the common Raw TLX (RTLX) variant [14], in which the sub-scale weighting process is eliminated to reduce survey times.

4.3.5 *System Usability.*
We measured the perceived usability of each in-VR assessment method by administering the System Usability Scale [6] after each assessment condition.

4.4 Procedure

Upon arrival, recruited participants were asked to review and sign an informed consent document. We then collected participant demographics (i.e., age and gender) and administered an initial SSQ to serve as a baseline for simulator sickness.

We then introduced participants to the VR system (a Meta Quest 2). We informed them how to adjust the headstrap and lenses to be comfortably worn and how to manipulate the controllers to perform the required VR interactions. We then provided the participants with a verbal overview of our study and what the in-VR training and assessment tasks would entail.

Once familiarized, participants were administered the VR training task for object Sequence A. Participants were allowed to view and learn each of the 10 objects at their own pace and comfort. Once the training task was completed, we encouraged participants to take a break from the VR system before administering the in-VR assessment task for object Sequence A. After the break, participants would don the headset again and were administered one of the in-VR assessments based on their assigned cohort.

Following the in-VR assessment task, participants doffed the headset and were asked to complete the subjective questionnaires on a nearby desktop computer. The questionnaires were administered in the following order: SSQ, NASA TLX, and SUS. After completing the questionnaires, the participants were administered the post-VR knowledge test for object Sequence A.

After completing the post-VR knowledge test, participants were encouraged to take a break before the next set of tasks. After the short break, participants repeated the above procedures for object Sequence B and their second in-VR assessment condition. The time required to complete the study was approximately 60 minutes. Participants were compensated $15 USD cash.

4.5 Participants

For our study, we recruited 30 participants from our university. Participants were required to be 18 years of age or older, have normal or corrected-to-normal vision, and be able to hear, walk, extend both arms, use both hands, and speak and understand English. Participants with any visual, auditory, neurological, or physical disabilities were excluded. Our final participant pool comprised 30 individuals (16 male and 14 female). The ages of our participants ranged from 18 to 32 with a mean age of 21.93.

5 RESULTS

In this section, we highlight and report the findings of our statistical analyses. We first highlight the in-VR Assessment Scores, followed by the Post-VR Knowledge Test scores, and then finally we address subjective measures found throughout the study.

5.1 In-VR Assessment Scores

First, we conducted a paired samples t-test on the in-VR assessment scores for Sequence A and Sequence B to determine whether the training tasks were significantly different. The test, $t(29) = 0.72$, $p = 0.48$, revealed that Sequence A in-VR assessment scores ($M = 0.60, SD = 0.27$) were not significantly different from Sequence B in-VR assessment scores ($M = 0.57, SD = 0.27$), which indicates that the tasks were approximately equivalent in terms of difficulty.

We then conducted a repeated-measures analysis of variance (RM-ANOVA) at a 95% confidence level to determine if the Carousel and Binary Choice methods significantly differed in terms of in-VR assessment scores. A Shapiro-Wilk test indicated that the results were normally distributed. The RM-ANOVA, $F(1, 29) = 70.39, p < 0.01, \eta^2 = 0.71$, revealed that Carousel in-VR assessment scores ($M = 0.39, SD = 0.22$) were significantly lower than Binary Choice scores ($M = 0.77, SD = 0.16$). Participants scored significantly higher in the Binary Choice condition (see in Figure 6).

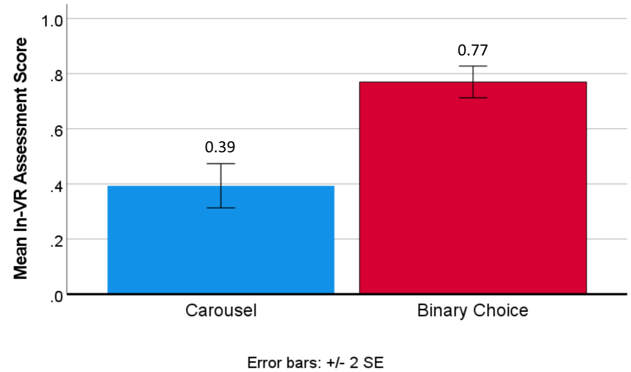

Figure 6: Mean in-VR assessment scores for the Carousel and Binary Choice conditions.

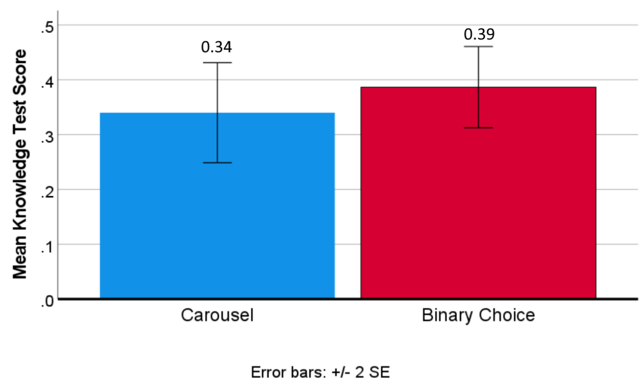

Figure 7: Mean post-VR knowledge test scores for the Carousel and Binary Choice conditions.

5.2 Post-VR Knowledge Test Scores

First, we conducted an RM-ANOVA at a 95% confidence level to determine if the Carousel and Binary Choice methods significantly differed in terms of post-VR knowledge test scores. Again, scores were normally distributed. The RM-ANOVA, $F(1, 29) = 1.13, p = 0.30$, revealed that the Carousel ($M = 0.34, SD = 0.25$) and Binary Choice ($M = 0.39, SD = 0.20$) methods were not significantly different in terms of post-VR knowledge test scores (see Figure 7).

We then analyzed the difference between each in-VR assessment score and its respective post-VR knowledge test score. An RM-ANOVA at 95% confidence level, $F(1, 29) = 65.66, p < 0.01, \eta^2 = 0.69$, revealed that the Carousel assessment-to-test difference ($M = -0.05, SD = 0.11$) was significantly different from the Binary Choice assessment-to-test difference ($M = -0.38, SD = 0.18$). As seen in Figure 8, the difference between the Carousel in-VR assessment score and its post-VR knowledge test score was significantly smaller than the difference between the Binary Choice in-VR score and its post-VR knowledge test score.

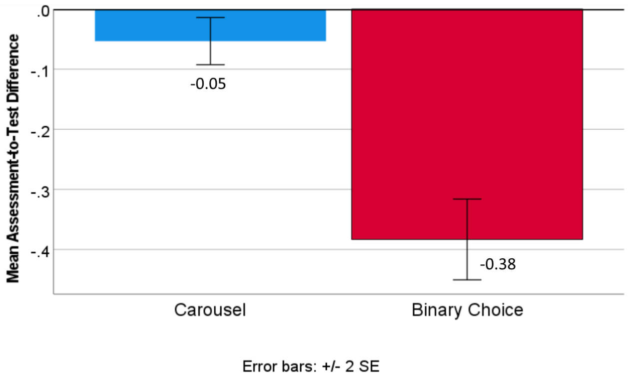

Figure 8: Mean in-VR assessment to post-VR knowledge test differences for the Carousel and Binary Choice conditions.

Table 4: Means and standard deviations for SSQ, TLX, and SUS measures.

	Carousel M (SD)	Conventional M (SD)
SSQ	11.84 (17.21)	16.86 (27.47)
Overall TLX	6.24 (2.91)	5.90 (2.48)
Mental Demand	10.83 (5.57)	8.73 (5.02)
Physical Demand	2.40 (3.25)	1.67 (2.23)
Temporal Demand	2.97 (4.12)	3.10 (4.48)
Performance	7.83 (5.31)	9.97 (5.40)
Effort	9.13 (5.20)	7.60 (5.14)
Frustration	4.27 (5.70)	4.33 (5.22)
SUS	85.33 (14.82)	87.75 (13.46)

5.3 Subjective Measures

For each of the subjective measures below, we conducted RM-ANOVAs at 95% confidence level. The means and standard deviations of each measure can be found in Table 4.

5.3.1 Simulator Sickness. We found that there was no significant difference between the Carousel and Binary Choice methods in terms of total SSQ scores, $F(1, 29) = 0.89, p = 0.35$. SSQ scores were calculated through Kennedy's overall SSQ equation [17].

5.3.2 Task Load. We found that there was no significant difference for Overall TLX scores, $F(1, 29) = 1.32, p = 0.26$. However, upon analyzing subscales, we found that the Carousel assessments had higher mental demand than the Binary Choice assessments, $F(1, 29) = 15.18, p < 0.01, \eta^2 = 0.34$. Furthermore, Carousel assessments trended higher for physical demand, $F(1, 29) = 4.04, p = 0.054$. We did not find any significant differences for temporal demand, performance, effort, or frustration.

5.3.3 System Usability. We found that there was no significant difference between the Carousel and Binary Choice methods in terms of total SUS scores, $F(1, 29) = 2.70, p = 0.11$.

6 DISCUSSION

From our results, we are able to explore and gain insight into the effects of employing our Carousel interface in the assessment stage of a VR inspection learning task. The following sections highlight, in further detail, the implications and conclusions we can derive from these results and whether they align with our initial hypotheses.

6.1 In-VR Assessment Performances

The results from our study showed a significant difference between our Carousel method and the Binary Choice method in terms of in-VR assessment scores. On average, participants scored nearly twice as high using the Binary Choice method as the Carousel method (see Figure 6). This result confirms our H1 hypothesis that Binary Choice would yield higher in-VR assessment scores.

In addition to confirming our hypothesis, this result indicates that the probability of selecting the correct option with a given assessment method has a direct impact on in-VR scores. In our study, the Binary Choice method had a 50% probability of correctness while the Carousel had a probability of 4% due to selecting one of 25 possibilities (i.e., 5 colors x 5 shapes). By extrapolating this result, we can hypothesize that multiple-choice assessment methods would also likely result in significantly higher in-VR assessment scores than Carousel, depending on their number of options (e.g., three choices afford a 33% probability, four choices a 25% probability).

An interesting question for future research is whether there are significant differences in terms of in-VR assessment scores between binary-choice assessments and multiple-choice assessments. There is currently insufficient evidence to determine whether binary-choice assessment methods will result in significantly higher in-VR assessment scores than the multiple-choice methods. Another interesting question is whether binary-choice assessments will perform statistically similar to selection-based assessments, which rely on implicitly passing or explicitly failing each inspection point.

6.2 Accuracy of In-VR Assessments

From the results of the in-VR assessment scores in Figure 6, the natural inclination is to conclude that the Binary Choice method is a better in-VR assessment technique than Carousel, as participants performed significantly better. However, we have demonstrated that the Binary Choice assessment technique is not reflective of a user's knowledge, as seen in Figure 8. On the other hand, our new Carousel method was able to closely approximate a user's post-VR knowledge test score (within 5%). This result confirms our H2 hypothesis that the Carousel method yields a more-accurate assessment than the Binary Choice method.

One obvious implication of this result is that more research needs to be conducted to investigate and validate the efficacy of our new in-VR Carousel assessment method. While the current results strongly indicate that our method is an accurate predictor of actual knowledge, additional studies of different VR training contexts and external evaluations must be conducted to confirm these results. One interesting question is whether the Carousel method is a strong predictor of training transfer to real-world tasks. Our current results rely on a post-VR knowledge test, but a potential follow-up study would involve a post-VR real-world test of the inspection task.

Another implication of this result is that in-VR assessment methods that present all relevant inspection possibilities to users are more accurate than assessment methods that rely on binary options (e.g., pass/fail) or small sets of multiple options. Our Carousel technique is merely one implementation of this design concept. Other interface designs, potentially better ones, are likely to exist. Furthermore, it is unclear whether in-VR assessment methods need to present all of the relevant possibilities or just a larger subset of the possibilities to be more accurate. For example, an interesting question is whether a modified version of the Carousel technique that only presents half of the relevant possibilities would be as effective at predicting knowledge acquisition.

6.3 Potential Use Cases of Carousel

As highlighted in our related work, inspection VR training experiences span multiple disciplines: aircraft inspections, construction hazards, vehicle inspections, and more. Many of these inspection training contexts may benefit from using our Carousel method.

Figure 9: An example of using our new Carousel method for an assessment of a virtual patient examination.

One potential example of employing the Carousel method is for virtual patient examinations (e.g. [9]). As seen in Figure 9, the Carousel method can be employed to better assess a trainee's knowledge of such an examination by returning the virtual patient to a healthy state. For example, if the virtual patient's eyes are yellowish due to jaundice (as seen in the top portion of Figure 9), the trainee would be expected to select the eyeball from the left carousel, which also includes a heart for heart sounds, lung for lung sounds, lips for lip coloration, and a skeleton for patient posture. By selecting the eyeball from the left carousel, the right carousel is then populated with various eye conditions that the trainee should be aware of, such as eye redness and pupil dilation. The trainee can then select the healthy eyes to change the virtual patient. This process requires the trainee to not only recognize that the patient had jaundice, but to understand what healthy eyes look like.

6.4 Limitations and Future Work

One limitation of our research is that it focuses on using VR to train and assess inspection tasks. While there are numerous examples of inspection-based VR training applications within the literature (see Table 1), these applications are obviously a subset of the breadth of VR training applications. Other types of VR training applications often involve more active forms of in-VR assessment, such as using a virtual club to putt a golf ball [12], using virtual micrometers to measure objects [3], or virtually manipulating the arms of a surgical robot [24]. Our Carousel technique clearly does not apply to these non-inspection training tasks. However, for inspection-based VR training simulations, we believe Carousel should be considered.

Another limitation of our research is that the results are derived from a controlled study and inspection task. We employed a simple, highly controlled task that involved users inspecting objects to assess their color and shape. Ecologically valid training tasks, even inspection ones, often involve more complex objects and sensory stimuli than the ones used in our study. However, using a highly controlled study design for our initial investigation of the Carousel technique is logical, as opposed to employing a more ecologically valid design with potential confounds.

Finally, another limitation of our study is that we focused on the comparison of two explicit assessment techniques: Carousel and Binary Choice. As highlighted by Yang et. al [37], there are different physiological factors at play when individuals are exposed to implicit and explicit learning mechanisms. Hence, our results may not apply if we were to compare the Carousel technique to an implicit assessment technique, such as selection-based assessment. With this in mind, future work should include a comparison of our Carousel technique and an implicit assessment technique.

In addition to the above limitations, more research should be conducted to compare the Carousel technique to other forms of in-VR assessment, such as the multiple-choice technique and the selection-based technique.

7 CONCLUSION

In this paper, we have presented the concept, design, and development of Carousel, a novel method for assessing inspection knowledge within VR training applications. The Carousel technique employs two 3D carousels for users to control the breadth and depth of all relevant inspection possibilities. Through our study comparing the Carousel technique to the Binary Choice technique (i.e., a pass/fail method), we have found evidence that the Carousel technique more-accurately reflects a user's inspection knowledge than the Binary Choice method, which artificially inflates in-VR assessment scores due to its higher correctness probability. These results suggest that VR researchers and developers should consider employing in-VR assessment methods that require users to make decisions given more options, as opposed to simplified assessment methods that offer a reduced set of inspection possibilities. Additionally, our results have raised several questions for future research.

ACKNOWLEDGMENTS

This material is based on work partially supported by the National Science Foundation under Grant No. 2021607 – "CAREER: Leveraging the Virtualness of Virtual Reality for More-Effective Training."

REFERENCES

[1] Khaled Aati, Daeyeol Chang, Praveen Edara, and Carlos Sun. 2020. Immersive work zone inspection training using virtual reality. *Transportation research record* 2674, 12 (2020), 224–232.

[2] Huai Jian Beh, Ali Rashidi, Amin Talei, and Yee Sye Lee. 2021. Developing engineering students' capabilities through game-based virtual reality technology for building utility inspection. *Engineering, Construction and Architectural Management* (2021).

[3] Jeffrey Bertrand, David Brickler, Sabarish Babu, Kapil Madathil, Melissa Zelaya, Tianwei Wang, John Wagner, Anand Gramopadhye, and Jun Luo. 2015. The role of dimensional symmetry on bimanual psychomotor skills education in immersive virtual environments. In *2015 IEEE Virtual Reality (VR)*. 3–10. https://doi.org/10.1109/VR.2015.7223317

[4] Christoffer Björkskog, Giulio Jacucci, Bruno Lorentin, and Luciano Gamberini. 2009. Mobile implementation of a web 3D carousel with touch input. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–4.

[5] D.A. Bowman and C.A. Wingrave. 2001. Design and evaluation of menu systems for immersive virtual environments. In *Proceedings IEEE Virtual Reality 2001*. 149–156. https://doi.org/10.1109/VR.2001.913781

[6] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996).

[7] Lauren Cairco Dukes, Amy Ulinski Banic, Jerome McClendon, Toni Bloodworth Pence, James Mathieson, Joshua Summers, and Larry F Hodges. 2013. Evaluation of System-Directed Multimodal Systems for Vehicle Inspection. *Journal of computing and information science in engineering* 13, 1 (2013).

[8] Daeyeol Chang, James Hopfenblatt, Praveen Edara, and Bimal Balakrishnan. 2020. Immersive Virtual Reality Training for Inspecting Flagger Work zones. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 327–330.

[9] Robert Coulter, Linda Saland, T Caudell, Timothy E Goldsmith, and D Alverson. 2007. The effect of degree of immersion upon learning performance in virtual reality simulations for medical education. *InMedicine Meets Virtual Reality* 15 (2007), 155.

[10] Andrew T Duchowski, Eric Medlin, Anand Gramopadhye, Brian Melloy, and Santosh Nair. 2001. Binocular eye tracking in VR for visual inspection training. In *Proceedings of the ACM symposium on Virtual reality software and technology*. 1–8.

[11] Ricardo Eiris, Masoud Gheisari, and Behzad Esmaeili. 2020. Desktop-based safety training using 360-degree panorama and static virtual reality techniques: A comparative experimental study. *Automation in construction* 109 (2020), 102969.

[12] Anton Franzluebbers and Kyle Johnsen. 2018. Performance Benefits of High-Fidelity Passive Haptic Feedback in Virtual Reality Training. In *Proceedings of the Symposium on Spatial User Interaction* (Berlin, Germany) *(SUI '18)*. Association for Computing Machinery, New York, NY, USA, 16–24. https://doi.org/10.1145/3267782.3267790

[13] Pierre Gaucher, Ferran Argelaguet, Jérôme Royan, and Anatole Lécuyer. 2013. A novel 3D carousel based on pseudo-haptic feedback and gestural interaction for virtual showcasing. In *2013 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 55–58.

[14] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909 arXiv:https://doi.org/10.1177/154193120605000909

[15] Ergin Isleyen and H Sebnem Duzgun. 2019. Use of virtual reality in underground roof fall hazard assessment and risk mitigation. *International Journal of Mining Science and Technology* 29, 4 (2019), 603–607.

[16] Bonnie E. John and David E. Kieras. 1996. The GOMS Family of User Interface Analysis Techniques: Comparison and Contrast. *ACM Trans. Comput.-Hum. Interact.* 3, 4 (dec 1996), 320–351. https://doi.org/10.1145/235833.236054

[17] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.

[18] Joseph J LaViola Jr, Ernst Kruijff, Ryan P McMahan, Doug Bowman, and Ivan P Poupyrev. 2017. *3D user interfaces: theory and practice.* Addison-Wesley Professional.

[19] Wanwan Li, Haikun Huang, Tomay Solomon, Behzad Esmaeili, and Lap-Fai Yu. 2022. Synthesizing Personalized Construction Safety Training Scenarios for VR Training. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 1993–2002.

[20] Jason Lucas, R McMahan, R Engle, D Bowman, W Thabet, S Scafrik, and M Karmis. 2008. Improving health and safety through conveyor system training in a virtual environment. In *Abstract accepted for, International Future Mining Conference.*

[21] RP McMahan, DA Bowman, S Schafrik, and M Karmis. 2008. Virtual Environment Training for Preshift Inspections of haul Trucks to improve mining safety. In *International Future Mining Conference & Exhibition*. 520–528.

[22] Ryan P McMahan, Steven Schafrik, Doug A Bowman, Michael Karmis, and J Brune. 2010. Virtual environments for surface mining powered haulage training. In *Proceedings of SME Symposium Celebrating*, Vol. 100. 520–528.

[23] P Mehta, Sajay Sadasivan, Joel S Greenstein, Anand K Gramopadhye, and Andrew T Duchowski. 2005. Evaluating different display techniques for communicating search strategy training in a collaborative virtual aircraft inspection environment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49. SAGE Publications Sage CA: Los Angeles, CA, 2244–2248.

[24] Alec G. Moore, Xinyu Hu, James Coleman Eubanks, Afham Ahmed Aiyaz, and Ryan P. McMahan. 2020. A Formative Evaluation Methodology for VR Training Simulations. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 125–132. https://doi.org/10.1109/VRW50115.2020.00027

[25] H Frank Moore, Ricardo Eiris, Masoud Gheisari, and Behzad Esmaeili. 2019. Hazard identification training using 360-degree panorama vs. virtual reality techniques: a pilot study. In *Computing in Civil Engineering 2019: Visualization, Information Modeling, and Simulation*. American Society of Civil Engineers, 55–62.

[26] Tao Ni, Doug A. Bowman, Chris North, and Ryan P. McMahan. 2011. Design and evaluation of freehand menu selection interfaces using tilt and pinch gestures. *International Journal of Human-Computer Studies* 69, 9 (2011), 551–562. https://doi.org/10.1016/j.ijhcs.2011.05.001

[27] Akeem Pedro, Quang Tuan Le, and Chan Sik Park. 2016. Framework for integrating safety into construction methods education through interactive virtual reality. *J. Prof. Issues Eng. Educ. Pract* 142, 2 (2016), 04015011.

[28] Amotz Perlman, Rafael Sacks, and Ronen Barak. 2014. Hazard recognition and risk perception in construction. *Safety Science* 64 (2014), 22–31. https://doi.org/10.1016/j.ssci.2013.11.019

[29] T Pitana, H Prastowo, and AP Mahdali. 2020. The development of fire safety appliances inspection training using virtual reality (VR) technology. In *IOP Conference Series: Earth and Environmental Science*, Vol. 557. IOP Publishing, 012064.

[30] Rafael Sacks, Amotz Perlman, and Ronen Barak. 2013. Construction safety training using immersive virtual reality. *Construction Management and Economics* 31, 9 (2013), 1005–1017. https://doi.org/10.1080/01446193.2013.828844

[31] Sajay Sadasivan, Joel S Greenstein, Anand K Gramopadhye, and Andrew T Duchowski. 2005. Use of eye movements as feedforward training for a synthetic aircraft inspection task. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 141–149.

[32] Ben Shneiderman. 1997. Direct Manipulation for Comprehensible, Predictable and Controllable User Interfaces. In *Proceedings of the 2nd International Conference on Intelligent User Interfaces* (Orlando, Florida, USA) *(IUI '97)*. Association for Computing Machinery, New York, NY, USA, 33–39. https://doi.org/10.1145/238218.238281

[33] Shanthi Vellingiri, Ryan P McMahan, Vinu Johnson, and Balakrishnan Prabhakaran. 2022. An augmented virtuality system facilitating learning through nature walk. *Multimedia Tools and Applications* (2022), 1–12.

[34] Jeenal Vora, Santosh Nair, Anand K Gramopadhye, Andrew T Duchowski, Brian J Melloy, and Barbara Kanki. 2002. Using virtual reality technology for aircraft visual inspection training: presence and comparison studies. *Applied ergonomics* 33, 6 (2002), 559–570.

[35] Jeenal Vora, Santosh Nair, Anand K Gramopadhye, Brian J Melloy, Eric Medlin, Andrew T Duchowski, and Barbara G Kanki. 2001. Using virtual reality technology to improve aircraft inspection performance: Presence and performance measurement studies. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 45. SAGE Publications Sage CA: Los Angeles, CA, 1867–1871.

[36] Jing Wan, Yue Zheng, Yonghui Li, Huiwei Mei, Lanfang Lin, and Liping Kuang. 2020. Oil depot safety inspection and emergency training system based on virtual reality technology. In *IOP Conference Series: Materials Science and Engineering*, Vol. 782. IOP Publishing, 042018.

[37] Jing Yang and Ping Li. 2012. Brain networks of explicit and implicit learning. (2012).

[38] Dong Zhao and Jason Lucas. 2015. Virtual reality simulation for construction safety promotion. *International journal of injury control and safety promotion* 22, 1 (2015), 57–67.