International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2024)
S. Hasegawa, N. Sakata and V. Sundstedt (Editors)

An Exploration of the Effects of in-VR Assessment Format on User Performance and Experience

Pedro Acevedo[1] Angela L. Jimenez[1] Alejandra J. Magana[2] Bedrich Benes[3] Christos Mousas[1]

[1]Department of Computer Graphics Technology, Purdue University, West Lafayette, Indiana 47907, U.S.A.
[2]Department of Computer and Information Technology, Purdue University, West Lafayette, Indiana 47907, U.S.A.
[3]Department of Computer Science, Purdue University, West Lafayette, Indiana 47907, U.S.A.

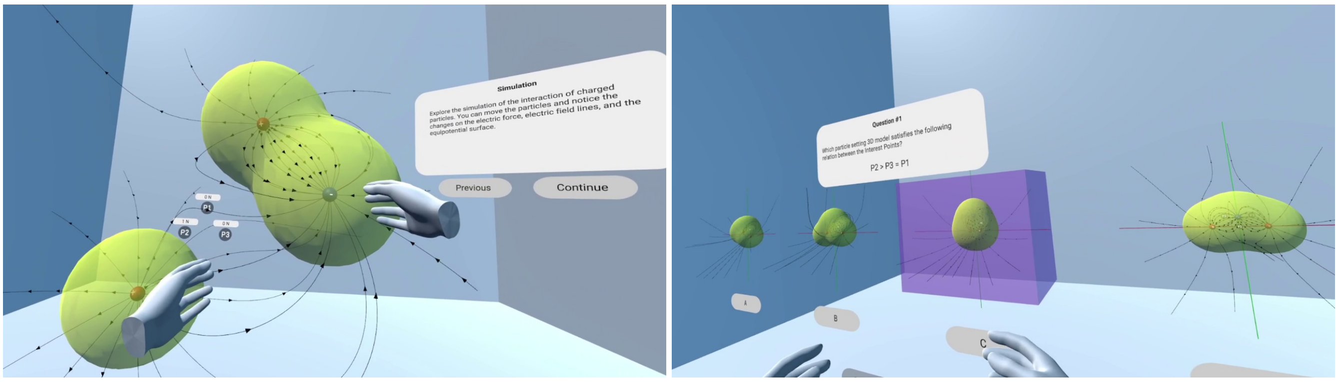


Figure 1: *We developed a VR application with embedded assessments. Here, we illustrate the designed charged particle lesson (left) and an example of a multiple-choice question (right).*

Abstract

For virtual reality (VR) training and learning applications, post-intervention assessment serves as a means to validate the effectiveness of the designed practice. These assessments can occur in the virtual environment by embedding questionnaires and necessary response mechanisms. Researchers have explored embedded VR (in-VR) assessment to minimize disruption to immersion and interference with the user's sense of presence compared to 2D screen-based (out-VR) surveys. However, the influence of in-VR assessment formats on user experience and performance still needs to be explored. Therefore, we conducted a within-group study ($N = 25$) to compare three assessment formats on task load, usability, user experience, self-efficacy, and performance metrics (i.e., completion time, movement, and response correctness). Using an educational application focused on charged particles and electric fields, we observed no significant differences in self-reported user experience metrics across the in-VR assessment formats. However, participants achieved higher scores when interacting with the 3DStatic assessment. This preference for 3DStatic assessment highlights the advantages of 3D visualizations in VR over traditional 2D user interfaces.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual reality

1. Introduction

Virtual Reality (VR) technologies have become increasingly popular in recent years. Thus, researchers and practitioners in various fields have been using them extensively. As a result, the market of head-mounted displays (HMDs) expects sales to rise in the following five years due to investments in immersive experiences like VR (https://www.marketsandmarkets.com/Market-Reports/head-mounted-display-hmd-market-729.html). These events show increased interest in VR applications in different fields and objectives. Among those focuses is the usage of VR for training and learning. VR designers aimed to develop applications that enhance guidance and replicate real-world tasks [FTJ*20; XLA*21]. However, evaluating knowledge transfer

DOI: 10.2312/egve.20241370

between virtual and real-world tasks is essential to validate the effectiveness of the virtual environment design [QMPY20a]. Different metrics can be used, such as self-reported measures or knowledge assessments [APB*20; KKZ*23; XLA*21]. One method for these assessments is to integrate questions directly into the virtual environment. Researchers [FKTK20; RSSS18] have shown the benefits of embedded VR (in-VR) assessment, including minimizing disruption to immersion and interference with the user's sense of presence by reducing instructor interaction. A standard in-VR questionnaire layout resembles the questions using 2D displays through user interface (UI) elements and pointer interactions [APB*20], differently from 2D screen-based (out-VR) questionnaires which are presented on a desktop computer out of any VR interface.

The in-VR assessment format could affect participants' experiences differently based on the implemented interactions [APB*20; PAP*20; SHKP21]. Therefore, we examine the in-VR assessment format and its potential impacts on user experience and performance. Considering the complexity and importance of electromagnetism-related concepts, given their abstract nature and challenges in learning through traditional media [FGAC03; SMN*17], we developed a short VR experience centered around charged particles and electric fields (see Figure 1). We conducted our experiment with an educational application because post-instruction assessment is necessary, and agentic interactions may have benefits. [CBD21; RT11]. Within this experience, we designed an interactive simulation enabling users to manipulate charged particles, visualize electric field behavior, and analyze resultant forces from particle interactions. We think exploring various in-VR assessments, such as through UI elements and ray casting (2DUI), 3D model observation (3DStatic), or interactive simulation manipulation (3DInteractive), can help designers determine the type of evaluation to include in their educational virtual reality application, considering better assessing user performance and enhancing user experience.

In this paper, we organize the content as follows. In Section 2, we discuss related works. In Section 3, we provide details on the methodology. In Section 4, we present our results. In Section 5, we discuss our findings, limitations, and implications. Finally, in Section 6, we conclude and discuss potential future work.

2. Related Works and Research Design

2.1. Assessment Formats

Assessments can be categorized as formative or summative [MBM19; Sad89]. Formative assessments are designed to diagnose students' understanding during a lesson, aiming to enhance learning outcomes, motivation, and conceptual change [SYA*08]. In contrast, summative assessments are conducted after a learning intervention to evaluate instructional effectiveness [May11]. Computer-based assessments can support the measurement of complex competencies [SR17], and the choice of response mechanism and modality can significantly impact students' final scores [CS16; MM06]. Traditional multiple-choice questions are commonly used in virtual learning environments, including VR applications [MBM19; CBD21]. Instructors and designers often simplify assessment modalities due to the complexity of integrating VR design and implementation into academic practices [CBD21]. Other assessment methods include integrated feedback, adaptive progress monitoring, and open-ended questions [MSA03; ZSHC20].

Reeve's and Tseng's [RT11] framework for learning engagement emphasizes the active role of students in their learning process, including assessments, mainly through agentic engagement, where students contribute to their educational experiences by expressing preferences, asking questions, seeking help, and making suggestions. Adapting this approach, we aimed to validate whether allowing students to engage with their assessments in an agentic manner could lead to different outcomes compared to traditional multiple-choice questions. In our study, we provided agency by offering an interactive visualization that students could edit and manipulate to respond to the assessment. We also used summative assessments to explore participants' conceptual understanding after the VR lesson, incorporating multiple-choice questions, varying interactions, and visuals.

2.2. In-VR Questionnaire

Researchers have utilized in-VR questionnaires to capture participants' emotions, opinions, and ratings [APB*20]. Designers have proposed toolkits to integrate questionnaires in VR, such as Regal et al. [RSSS18], who developed VRate, a Unity Engine package to integrate surveys for VR. They used a 2D UI to place questionnaire content with multiple choices and rankings featuring buttons and slider components. The authors employed JSON and CSV files to control the input and output of the questionnaire data. Feick et al. [FKTK20] developed the VRQuestionnaireToolkit to gather subjective data such as system usability, workload, and simulation sickness metrics commonly used in VR and human-computer interaction (HCI) research [KKZ*23]. They employed pointing as a selection technique and world-anchoring to position the questionnaire content.

Safikhani et al. [SHKP21] designed an in-VR questionnaire toolkit that connects with a survey web service called LimeSurvey. They offer two methods for questionnaire response: a 2D menu with pointer interaction and a 3D interaction involving object selection and lever manipulation. They conducted a brief user study comparing the usability and workload of different in-VR formats (2D vs. 3D) with out-VR questionnaires. Their results suggest a preference for in-VR questionnaires over out-VR ones, with no significant difference between 2D and 3D design layouts regarding workload and usability, indicating that dimensionality does not impact. However, they conducted their study with a small sample size (16 participants), so further exploration is needed to validate their conclusions. Considering that explorations of questionnaire formats give us insights into the common approaches to integrated questionnaire response systems in VR, in this study, we used UI elements, 3D models, and interactive simulations for the in-VR assessment formats.

2.3. In-VR Assessment

Petersen et al. [PMM21] integrated a brief assessment of UI elements with four multiple-choice questions about viruses after an immersive museum tour. Their assessment covered factual knowledge, virtual agents' appearance, cognitive load, and enjoyment. They did not analyze in-VR questionnaire preferences; instead, the authors chose to integrate all elements into a single VR application due to the study's remote nature. Similarly, Alexandrovsky et al. [APB*20] conducted a literature review exploring the usage of in-VR questionnaires. They found that 2D layout is the most common approach between the reported papers and the interviewed VR experts. Additionally, the authors conducted a user study comparing four different versions in-VR questionnaire versions of the interface: world-pointer, world-trackpad, body-pointer, and body-trackpad. They observed significant differences in usability, with participants finding the world pointer layout the easiest to use. However, the authors included a small sample size (ten participants) for that part of the study, warranting rigorous methods to validate their conclusion.

Belga et al. [BDG*22] reviewed previous in-VR assessments for inspection-based training, classifying them based on interaction type: selection, binary-choice, and multiple-choice. Selection assessments evaluate users' decisions in choosing the correct answer, such as identifying deficiencies in traffic control [CHEB20] or choosing the correct firefighting equipment [PPM20]. Binary-choice assessments involve a 50% probability of passing or failing the inspection. Multiple-choice assessments require users to choose from different possibilities, like short quizzes after each inspection [BRTL21] or checking identified hazards in a list [EGE20]. These formats involve simpler interactions, such as laser pointing or hovering over elements. Additionally, Belga et al. [BDG*22] proposed a carousel assessment method, where users consider each inspection point and choose its correct state from various scenarios. This interaction requires users to grab an item with their hand, displaying its possibilities in another circle, allowing for better visualization. They compared their method with a binary-based assessment and found that the carousel method resulted in significantly more accurate assessments, as evidenced by higher in-VR assessment scores. Following their comparison approach, we aimed to validate whether the utilized formats can affect user experience and performance on an educational VR platform.

2.4. Immersive Educational Applications

Regarding using immersive tools for education, Pittman and LaViola [PL20] employed augmented reality to visualize physics-related concepts such as Coulomb's law, parallel circuits, and the Doppler effect. Participants reported positive experiences but encountered issues with the gestures required for interacting in the augmented reality (AR) environment. Acevedo et al. [AMM*22] designed an immersive virtual reality experience focused on Coulomb's law, including haptic feedback. Their results indicated higher learning gains in out-VR assessments after the VR intervention, with no significant difference between the haptic and non-haptic conditions. They considered that using out-VR assessments could have impacted participants' experiences with the head-mounted display (HMD).

Compared with traditional instructions, Checa et al. [CMB21] developed a VR game for teaching computer hardware assembly. Students needed to grab and place specified hardware components in their application to complete the assessment. Results showed that VR and desktop games increased student satisfaction compared to lectures. Johnson-Glenberg et al. [JBK21] conducted a study on different degrees of embodiment and interaction using VR and PC b. Their results show a lower knowledge test performance for participants with low embodiment and interaction conditions than the ones with higher due to lower agency.

Another of the explored aspects when using immersive technology is presence. Presence refers to the "sense of being there" as the perceptual illusion evoked by the experienced synthetic environments [Mik06], as can be a VR interaction. Qian et al. [QMPY20b], relate the higher sense of presence when interacting with a virtual-reality fusion application for chemistry learning, in which they used an interaction with real and virtual objects when immersed in the virtual environment. They found a significant difference in presence in VR conditions than in desktop. Lønne et al. [LKLS23] validated the sense of presence on a VR 360-video application, in which participants confronted a situation of interpersonal conflicts. They found a higher positive effect and presence on the VR conditions than the desktop group. These studies highlight the relevance of investigating educational VR applications and their potential to leverage learning experiences.

2.5. Research Questions

We conducted a within-group study to explore the effect of in-VR assessment formats (i.e., 2DUI vs. 3DStatic vs. 3DInteractive) on workload usability, user experience, self-efficacy, and performance (i.e., completion time, movement, and response correctness). We asked the participants to respond to three different in-VR assessments around charged particles and Coulomb's law (`https://byjus.com/jee/coulombs-law/`). Based on the logged data collection and self-reported surveys, we aimed to respond to the following research questions:

- **User performance**
  - **RQ1:** Does the in-VR assessment format affect participants' responses?
  - **RQ2:** How does the in-VR assessment format affect completion time?
- **User experience**
  - **RQ3:** Which is the most usable in-VR assessment format?
  - **RQ4:** Does the in-VR assessment format affect the self-reported task load?
  - **RQ5:** How does the in-VR assessment format affect the participants' self-efficacy?
  - **RQ6:** How does the in-VR assessment format affect user experience?

2.6. Contributions

In this study, we aim to explore the influence of the in-VR assessment format on an educational application. Previous studies have treated in-VR questionnaires merely as a mechanism, overlooking their potential impact on user actions and outcomes, especially concerning the format used. Researchers have focused on studies involving in-VR questionnaires, favoring this method over out-VR survey responses. In contrast, our goal was to explore which form of in-VR assessment could benefit users and determine their preference for these types of assessments, similar to Belga et al. [BDG*22] study, who validated the used formats in user-perceived metrics as task load or usability. With this objective in mind, assuming a VR designer has already incorporated an in-VR assessment, we would like to provide insights into the chosen format and its effects on the end users.

3. Materials and Methods

3.1. Participants

For our within-group study, we conducted an *a priori* power analysis to determine the sample size using G*Power v.3.10 software [FELB07]. Based on a medium effect size of $f = .30$ [Coh13], one group with three repeated measures and a non-sphericity correction $\epsilon = .75$, to achieve an 80% power ($1-\beta$ error probability), the analysis recommended a minimum of 24 participants. We recruited 25 participants (age: $M = 21.96$, $SD = 4.03$) from our university. Of the sample, 17 were male, and eight were female; among them, 52% did not have previous VR experience, and others had some (20%) or were very experienced (28%). The participants volunteered to take part in this study without receiving any monetary compensation.

3.2. VR Application

We developed a VR application for our study in Unity game engine (version 2021.3.30f1) with the Oculus XR plugin (version 3.3.0) deployed for a Meta Quest 2 HMD (`https://github.com/PedroAcevedo/In-VR-assessment-format-charged-particle.git`). We designed an application that considered charged particles, Coulomb's law, and the electric field. Specifically, we designed a simulation scenario where the user can observe and move the charged particles and visualize the different particle settings based on their inputs. The simulation elements include spheres representing charged particles and 3D indicators showing the exerted electric force on a specific coordinate position called Interest Points (indicated as P1, P2, and P3).

The VR application featured three sections: a VR actions tutorial, a lesson module, and in-VR assessments. In the introduction, users familiarized themselves with the controllers, UI button interactions, grabbing actions, and locomotion through the joystick. Prior research indicated that tutorials significantly improve performance and user experience [KMM21]. In the lesson module, we included a short lecture about Coulomb's law and charged particles. Finally, the in-VR assessments consisted of three conditions for this study, resembling the following formats:

- **2DUI:** We placed the questionnaire anchored to the world and on a UI panel (see Figure 2a). We represented the multiple-choice options as images in a grid layout. The participant should press a UI button on the image to respond to the question.
- **3DStatic:** We placed the questionnaire content in front of the participants, using UI elements for the question statement and 3D models extracted from a simulation setting as multiple-choice options (see Figure 2b). We placed the 3D models (four in total, one per option) around the user's view. To answer the question, the participant should press a UI button below the 3D model.
- **3DInteractive:** We presented the questionnaire content to the participants using a UI element and an interactive simulation with movable charged particles for the option choice (see Figure 2c). The participant could grab and move the charged particles within the interactive simulation to form a new configuration in response to the question. We also included Interest Points, but their exerted forces were not displayed—only their 3D representations were visible, without showing any values.

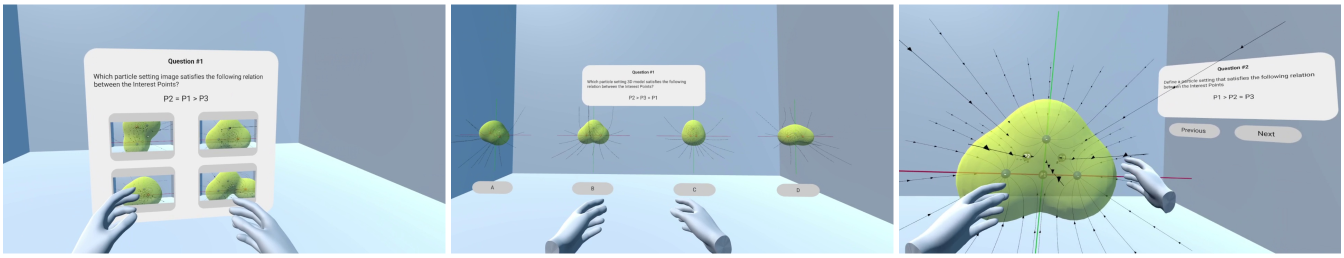
3.3. Assessment

We included two questions per assessment, asking participants to choose a particle setting that satisfies the relation between the Interest Points. The question included in the assessment is *"Which particle setting [image/model] satisfies the following relation between the Interest Points? P1 ($<$ or $>$ or $=$) P2 ($<$ or $>$ or $=$) P3."* For the 3DInteractive condition, we adjusted the question due to the different response mechanisms: *"Define a particle setting that satisfies the following relation between the Interest Points. P1 ($<$ or $>$ or $=$) P2 ($<$ or $>$ or $=$) P3."* A physics professor from our university validated the selected set of combinations and questions and its conceptual difficulty.

To maintain consistency in the questions across conditions, we randomized the particle settings, consisting of three charged particles' elements placed in the same position, by changing the sign of a particle (positive or negative) and the requested relation (e.g., P1 $>$ P2 $=$ P3 or P1 $=$ P2 $<$ P3). These properties on the particle settings required different options for the multiple-choice conditions (2DUI and 3DStatic) and other valid particle positions (3DInteractive) while keeping the same question. In Table 1, we present an example of a participant assessment per format and questions. In our experiment, we balanced the conditions using the Latin squares method [Wil49] to avoid any possible carry-over (residual) effects or learning effects across the examined assessments.

3.4. Measurements and Ratings

In our study, we collected data on users' performance and experiences. For the performance, the application records objective data from the participant's interactions, such as completion time (TIME), movement (MOVE), and response grading (SCORE) for each condition. The TIME denotes the time the participant needs to respond to an in-VR questionnaire. Movement refers to the total distance that the user moves in the virtual environment, so if the user does not change their initial position using the joystick, the MOVE value will be equal to zero. We calculated the SCORE by comparing the participant's choice and the correct answer per question. We made a different validation for the 3DInteractive, where the correct response is determined by evaluating the relation between the Interest Points. We capture the values on the points in the last setting delimited by the participant.



(a) (b) (c)

Figure 2: *We designed three in-VR assessment formats: (a) questions displayed in UI and using images for responses, (b) using 3D models for the option selection method, and (c) interactive simulation as a response mechanism.*

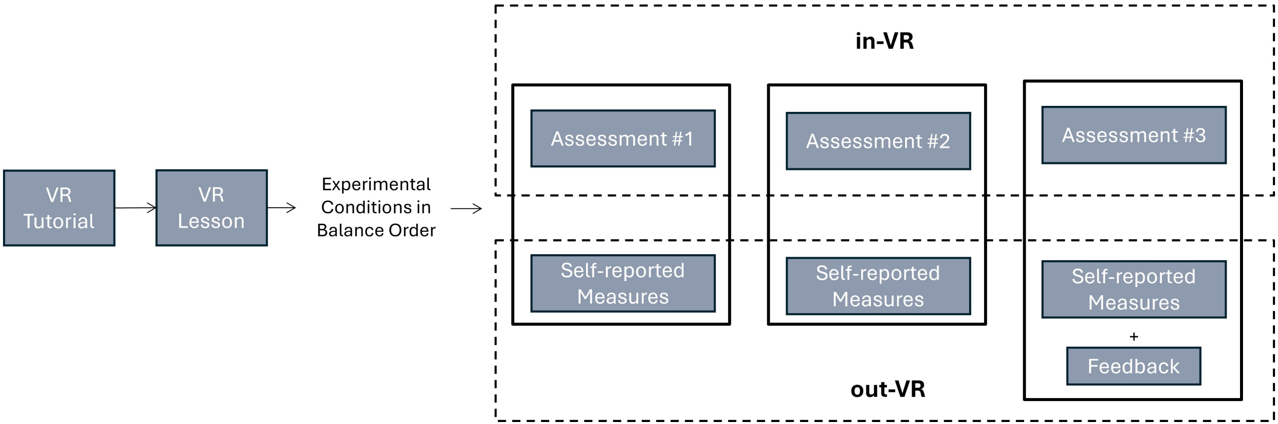


Figure 3: *The procedure for the data collection process.*

Table 1: *Assessments questions example.*

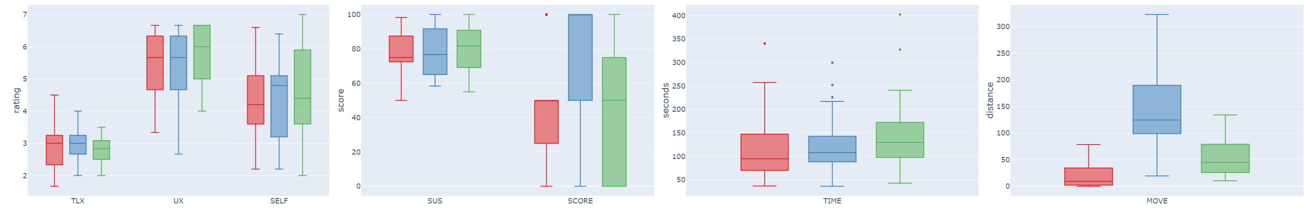
Condition	Question	Relation	Setting
2DUI	Which particle setting image satisfies the following relation between the Interest Points?	P2 = P1 > P3	Particle 1: - Particle 2: + Particle 3: -
		P3 > P1 > P2	Particle 1: + Particle 2: - Particle 3: -
3DInteractive	Define a particle setting that satisfies the following relation between the Interest Points.	P3 > P1 = P2	Particle 1: + Particle 2: - Particle 3: +
		P1 > P2 = P2	Particle 1: - Particle 2: - Particle 3: -
3DStatic	Which particle setting model satisfies the following relation between the Interest Points?	P2 > P3 = P1	Particle 1: + Particle 2: + Particle 3: -
		P1 > P3 > P2	Particle 1: + Particle 2: - Particle 3: +

Regarding user experience, we employed self-reported measurement surveys from the literature. We evaluated the task load using NASA's task load index (TLX) scale [Har06], the system usability using the System Usability Scale (SUS) [Bro*96], user experience (UX) with questions around enjoyment (*"Did you enjoy the VR experience?"*), frustration (*"I felt frustrated when answering the questions in the virtual environment."*), and willingness for future use (*"Are you willing to answer questions in the way you answered them in this virtual environment again?"*). Moreover, we used a five-question self-efficacy (SELF) survey [SJ95]. Our participants rated each statement on a 7-point Likert scale based on their perceptions of these surveys. We conducted an out-VR survey using the Qualtrics online survey tool because our objective was to evaluate the assessment method rather than the VR experience in general.

3.5. Procedure

We followed the guidelines and practices approved by our university's institutional review board (IRB) to conduct the user study. To begin, we provided a consent form to participants who volunteered for the study and waited for their signatures before commencing the intervention. Subsequently, we administered demographic questions, including age, gender, and VR experience. We explained to participants unfamiliar with the HMD (Meta Quest 2) how to use the device. Consequently, we proceeded with the VR application following the procedure presented in Figure 3. Upon entering the virtual environment, the participants started on an introduction scene. Then, they engaged in a brief lesson (tutorial session) on the charged particles and Coulomb's law, summarizing the basics of the concepts and progressively presenting simulation el-

Proceedings published by Eurographics - The European Association for Computer Graphics.



(a) (b) (c) (d)

Figure 4: *Box plots of our results: (a) TLX, UX, and SELF, (b) SUS and SCORE, (c) TIME, and (d) MOVE. A thick horizontal line denotes the median. In the plots we have the conditions as 2DUI (), 3DStatic (), and 3DInteractive ().*

ements. We designed a step-by-step lesson based on an instruction book on electricity and physics [LSM21] and included four steps: (1) illustrating particle representation signs (red for positive and blue for negative), (2) explaining how an Interest Point works, (3) presenting a scenario with two particles and one Interest Point, and (4) demonstrating the interactive simulation similar to the assessment options (see Figure 1). Participants could spend as much time as needed until they felt confident about the concept, with an average duration of approximately six minutes.

Once ready, we presented the experimental conditions. Participants responded to an in-VR assessment with two questions in the specified format before removing the HMD to complete self-reported measures surveys on a web browser. The participants repeated this procedure two more times. After the final questions, we invited our participants to provide comments or suggestions about the application. Finally, we thanked the participants and allowed them to leave the lab. Each participant spent no more than 40 minutes completing the study.

4. Results

We utilized IBM SPSS v.28.0 statistical analysis software for all analyses. We used the one-way repeated measures analysis of variance (ANOVA) to analyze user performance (TIME, MOVE, and SCORE) and user experience metrics (TLX, SUS, UX, and SELF). We assessed the normality assumption using Shapiro-Wilk tests at the 5% significance level and visually inspected Q-Q plots of the residual. A $p < .05$ was considered statistically significant. In Figure 4, we illustrate the boxplots of our results. For the statistically significant results, we used Bonferroni-corrected estimates for our post hoc comparisons.

4.1. User Performance

Our analysis did not reveal a statistically significant result of the in-VR assessment format on TIME (Wilk's $\Lambda = .889$, $F[2,23] = 1.435$, $p = .259$, $\eta_p^2 = .111$). However, we found a statistically significant result of the in-VR assessment format on MOVE (Wilk's $\Lambda = .209$, $F[2,23] = 43.455$, $p < .001$, $\eta_p^2 = .161$). Pairwise comparisons showed that participants move significantly more when interacting with the 3Dstatic assessment ($M = 145.29$, $SD = 69.75$) than on the 2DUI ($M = 21.34$, $SD = 23.18$) at $p < .001$ and 3DInteractive ($M = 56.38$, $SD = 37.77$) at $p < .001$. We also found that participants move significantly more when interacting with the 3DInteractive assessment than the 2DUI at $p = .001$. Lastly, we found a statistically significant effect of the in-VR assessment format on SCORE (Wilk's $\Lambda = .487$, $F[2,23] = 12.137$, $p < .001$, $\eta_p^2 = .513$). Pairwise comparisons showed that participants had significantly better scores when interacting with the 3DStatic assessment ($M = 76.00, SD = 29.30$) than 2DUI ($M = 48.00$, $SD = 33.79$) at $p < .001$ and 3DInteractive ($M = 42.00$, $SD = 40.00$) at $p < .001$.

4.2. User Experience

Our analysis did not reveal statistically significant results of the in-VR assessment format on TLX (Wilk's $\Lambda = .910$, $F[2,23] = 1.134$, $p = .339$, $\eta_p^2 = .090$), SUS (Wilk's $\Lambda = .966$, $F[2,23] = .410$, $p = .669$, $\eta_p^2 = .034$), UX (Wilk's $\Lambda = .951$, $F[2,23] = .595$, $p = .560$, $\eta_p^2 = .049$), and SELF (Wilk's $\Lambda = .976$, $F[2,23] = .288$, $p = .752$, $\eta_p^2 = .024$).

4.3. Participants Suggestions

We asked participants for feedback regarding both the developed in-VR assessments. Overall, participants showed their interest in the VR experience and considered it "*really good*" (PA2), "*cool and easy to use*" (PA19), and "*very fun to interact with*" (PA17). In addition, participants felt that the application helped them to "*learn the concepts a lot better*" (PA9) and to "*easily complete the task*" (PA14) after the instructions. Concerning the different in-VR assessments, participants found the 3D conditions particularly enjoyable, with references to the 3DInteractive format as "*enjoyable to use*" (AP19) and interactions with the 3DStatic format described as "*wonderful*" (PA23). Conversely, one participant likened the 2DUI format to "*reading a textbook in VR*" (PA23) and identified it as a drawback. Participants offered various suggestions for improving the VR application, with common themes including incorporating a feedback component to indicate correct or incorrect answers (PA10, PA13, and PA25), implementing an initial practice quiz before the actual assessment (PA22 and PA25), increasing the pace of the lesson by integrating more interactions with the simulation (PA12) and adding voice-over to the text for enhanced accessibility in VR (PA24).

5. Discussion

5.1. User Performance

To understand the performance of our participants, we captured their interaction with each in-VR assessment, including their responses, movement, and completion time. Concerning user responses (**RQ1**), participants scored higher when interacting with the 3DStatic condition than 2DUI and 3DInteractive. Researchers have found that leveraging the 3D perspective offered by VR benefits students' performance and spatial understanding [DJR17; WUM*23]. Moreover, other researchers have reported that visualizing data from different 3D perspectives aids comprehension [TKAM06; VST05]. We also found that our participants preferred interacting with 3D models over plain 2D images when navigating the in-VR assessment. These score results align with the findings of Belga et al. [BDG*22], where participants scored higher in the binary-choice assessment than the carousel method, mainly due to the higher probability of success—50% versus 4%, respectively. In our study, the likelihood of obtaining a correct answer was similarly lower in the 3DInteractive condition, where users had control over the simulation and needed to place three particles on a correct setting, compared to the 3DStatic condition, which only required choosing from four options.

Furthermore, in the 3DStatic condition, our participants exhibited the most movement in the virtual environment, indicating an interest in exploring the 3D model from various perspectives to solve the assessment. Regarding completion time (**RQ2**), participants spent more time interacting with the 3DInteractive assessment, though we did not find significant differences. These results suggest that the time spent on the assessment was similar across conditions, with no influence of the in-VR format; the complexity of questions and understanding of concepts likely played a role. Similar to the findings from Safikhani et al. [SHKP21], no difference was observed in the time spent on 2D and 3D in-VR assessment formats.

5.2. User Experience

For the self-reported measurements, our participants rated their interaction with each in-VR assessment based on task load, usability, user experience, and self-efficacy. Regarding usability (**RQ3**), we did not find significant differences between experimental conditions; interestingly, our participants rated the 3DInteractive assessment as the most intuitive of the three. Similar results were found in the study by Safikhani et al. [SHKP21], in which they included an in-VR questionnaire with grabbing interaction as a response mechanism, yielding no difference in SUS ratings compared to the 2DUI questionnaire. This suggests that our participants perceived and rated the UI menu press and grabbing actions as more intuitive. It is worth noting that the mean scores of SUS are classified into the good and acceptable range [BKM09].

Regarding task load (**RQ4**), we found no significant differences between the three in-VR assessments, indicating a similar workload required to interact with them. Similarly as previous assessment user studies [BDG*22], where TLX have reported no significant result across interactive and UI conditions. Regarding self-efficacy (**RQ5**), our participants rated similarly for all three experimental conditions. Our participants may have experienced some difficulty with the assessment due to their initial interaction with the concepts and possible insecurities with the assessed topic. This could have influenced their ratings, resulting in a neutral rating regarding self-efficacy for all assessments.

For user experience (**RQ6**), we did not find significant differences. Our participants rated 3DInteractive the highest, suggesting that they enjoyed it and would be willing to use this assessment format again, but no conclusive results are reported from this rating. This finding revealed no necessary impact regarding the leverage assessment regarding the system experience, giving each condition as acceptable and usable as the others. The noticed trend can be reflected in the reported ratings in terms of UX and SUS between the positive scales (see Figure 2).

5.3. Limitations

We acknowledge some limitations in our study. These limitations do not invalidate our findings but provide context for their interpretation and suggestions for future research.

One limitation concerns participants' time completing the questionnaire; the average time per condition was at most 2.5 minutes. This brief exposure could have influenced their ratings. However, extending this duration might entail a tradeoff between assessment complexity and user experience within the VR format, determining which factor influenced participants' ratings. Regarding the instructional content presented, although delivering a lesson in VR may offer benefits, we should validate its effectiveness. Our participants expressed concerns about reading excessive text in VR, which may be uncomfortable for some people. Despite our attempt to segment the lesson into concise "slides," our participants found the pace of the activity difficult. Exploring lectures or brief lessons in VR for knowledge acquisition and experiential learning could be a promising avenue for future research. Moreover, we scored the 3DInteractive condition based on exact matches of Interest Point values to the requested relation, which may have lowered participants' scores, making equality matching particularly difficult. To address this, we suggest incorporating a threshold range, allowing approximate matches to be considered correct and making it easier for participants to identify suitable particle settings. Finally, since our findings are limited to the explored conceptual topic of charged particles, we recommend validating our findings with different educational- or training-focused assessments.

5.4. Implications

Further research is necessary to validate the potential effects of in-VR assessment on users, mainly when the VR application requires an evaluation component. For designers, integrating an embedded assessment into their VR application may require additional development effort due to the inclusion of more interactions. Nevertheless, some insights support the integration of assessment into the virtual environment to prevent disruption of immersion. This study explored which format for in-VR assessment integration could offer benefits rather than impose additional burdens on users. In this regard, we propose using 3D models instead of 2D images for multiple-choice questions where feasible. In our study, integrating

3D models of simulation settings helped participants' conceptual understanding, as evidenced by their higher scores and expressed preferences. However, given the convenience of the design and development process, standard UI menu interactions for surveys or assessments, similar to those used outside VR, cannot be discarded due to the lack of significant differences in participants' ratings reported in this paper.

6. Conclusions and Future Work

We conducted a study to explore the effects of in-VR assessment formats on user experience and performance. We have developed an application to assess an electromagnetism-related concept. Participants engaged in a short lesson about charged particles before responding to the embedded evaluation. We defined three in-VR assessment formats: 2DUI, 3DStatic, and 3DInteractive. Our results indicate that participants indicated no differences among the in-VR assessments regarding self-reported user experience metrics. However, regarding user performance, participants scored higher when interacting with the 3DStatic assessment and exhibited more movement in the virtual environment. This preference elicits the benefits of 3D visualizations enabled by VR and HMD, reflecting users' preference toward this format over 2DUI.

Future studies should validate the potential impacts of using in-VR assessments. To minimize the influence of concepts, we should include more straightforward questions separate from specific activities to focus users' attention on interacting with embedded content rather than their confidence in exam responses. Another option is to implement a feedback mechanism allowing participants to retry questions until they reach a 100% score, guaranteeing multiple interactions based on their knowledge. Further research is needed to evaluate the effectiveness of the interactions used for in-VR assessments, as well as the context, such as learning topics, gender-balanced samples, or types of assessments (e.g., open-ended questions). We recommend that designers and researchers embed assessments in VR, incorporating both 3D and 2D interactions, which are intuitive and user-friendly.

References

[AMM*22] Acevedo, Pedro, Magana, Alejandra, Mousas, Christos, et al. "Effects of Tactile Feedback on Conceptual Understanding of Electromagnetism in a Virtual Reality Experience". *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 2022, 588–593. DOI: 10.1109/ISMAR-Adjunct57072.2022.00122 3.

[APB*20] Alexandrovsky, Dmitry, Putze, Susanne, Bonfert, Michael, et al. "Examining Design Choices of Questionnaires in VR User Studies". *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. ACM, Apr. 2020. DOI: 10.1145/3313831.3376260. URL: http://dx.doi.org/10.1145/3313831.3376260 2, 3.

[BDG*22] Belga, Jacob, Do, Tiffany D., Ghamandi, Ryan, et al. "Carousel: Improving the Accuracy of Virtual Reality Assessments for Inspection Training Tasks". *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. VRST '22. ACM, Nov. 2022. DOI: 10.1145/3562939.3565618. URL: http://dx.doi.org/10.1145/3562939.3565618 3, 4, 7.

[BKM09] Bangor, Aaron, Kortum, Philip, and Miller, James. "Determining what individual SUS scores mean: Adding an adjective rating scale". *Journal of usability studies* 4.3 (2009), 114–123 7.

[Bro*96] Brooke, John et al. "SUS-A quick and dirty usability scale". *Usability evaluation in industry* 189.194 (1996), 4–7 5.

[BRTL21] Beh, Huai Jian, Rashidi, Ali, Talei, Amin, and Lee, Yee Sye. "Developing engineering students' capabilities through game-based virtual reality technology for building utility inspection". *Engineering, Construction and Architectural Management* 29.7 (July 2021), 2854–2877. ISSN: 0969-9988. DOI: 10.1108/ecam-02-2021-0174. URL: http://dx.doi.org/10.1108/ECAM-02-2021-0174 3.

[CBD21] Castaneda, Lisa M., Bindman, Samantha W., and Divanji, Riddhi A. "Don't forget to assess: How teachers check for new and deeper learning when integrating virtual reality in the classroom". *Journal of Research on Technology in Education* 55.2 (Aug. 2021), 210–229. ISSN: 1945-0818. DOI: 10.1080/15391523.2021.1950083. URL: http://dx.doi.org/10.1080/15391523.2021.1950083 2.

[CHEB20] Chang, Daeyeol, Hopfenblatt, James, Edara, Praveen, and Balakrishnan, Bimal. "Immersive Virtual Reality Training for Inspecting Flagger Work zones". *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, Dec. 2020. DOI: 10.1109/aivr50618.2020.00066. URL: http://dx.doi.org/10.1109/AIVR50618.2020.00066 3.

[CMB21] Checa, David, Miguel-Alonso, Ines, and Bustillo, Andres. "Immersive virtual-reality computer-assembly serious game to enhance autonomous learning". *Virtual Reality* 27.4 (Dec. 2021), 3301–3318. ISSN: 1434-9957. DOI: 10.1007/s10055-021-00607-1. URL: http://dx.doi.org/10.1007/s10055-021-00607-1 3.

[Coh13] Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, May 2013. ISBN: 9781134742707. DOI: 10.4324/9780203771587. URL: http://dx.doi.org/10.4324/9780203771587 4.

[CS16] Cohen, Donita and Sasson, Irit. "Online quizzes in a virtual learning environment as a tool for formative assessment". *JOTSE* 6.3 (2016), 188–208 2.

[DJR17] Dede, Christopher J., Jacobson, Jeffrey, and Richards, John. "Introduction: Virtual, Augmented, and Mixed Realities in Education". *Virtual, Augmented, and Mixed Realities in Education*. Springer Singapore, 2017, 1–16. ISBN: 9789811054907. DOI: 10.1007/978-981-10-5490-7_1. URL: http://dx.doi.org/10.1007/978-981-10-5490-7_1 7.

[EGE20] Eiris, Ricardo, Gheisari, Masoud, and Esmaeili, Behzad. "Desktop-based safety training using 360-degree panorama and static virtual reality techniques: A comparative experimental study". *Automation in Construction* 109 (Jan. 2020), 102969. ISSN: 0926-5805. DOI: 10.1016/j.autcon.2019.102969. URL: http://dx.doi.org/10.1016/j.autcon.2019.102969 3.

[FELB07] Faul, Franz, Erdfelder, Edgar, Lang, Albert-Georg, and Buchner, Axel. "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences". *Behavior Research Methods* 39.2 (May 2007), 175–191. ISSN: 1554-3528. DOI: 10.3758/bf03193146. URL: http://dx.doi.org/10.3758/bf03193146 4.

[FGAC03] Furió, C., Guisasola, J., Almudí, J. M., and Ceberio, M. "Learning the electric field concept as oriented research activity". *Science Education* 87.5 (July 2003), 640–662. ISSN: 1098-237X. DOI: 10.1002/sce.10100. URL: http://dx.doi.org/10.1002/sce.10100 2.

[FKTK20] Feick, Martin, Kleer, Niko, Tang, Anthony, and Krüger, Antonio. "The Virtual Reality Questionnaire Toolkit". *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20. ACM, Oct. 2020. DOI: 10.1145/3379350.3416188. URL: http://dx.doi.org/10.1145/3379350.3416188 2.

[FTJ*20] Franzluebbers, Anton, Tuttle, Alexander James, Johnsen, Kyle, et al. "Collaborative Virtual Reality Training Experience for Engineering Land Surveying". *Advances in Intelligent Systems and Computing*. Springer International Publishing, Aug. 2020, 411–426. ISBN: 9783030525750. DOI: 10.1007/978-3-030-52575-0_34. URL: http://dx.doi.org/10.1007/978-3-030-52575-0_34 1.

[Har06] Hart, Sandra G. *Nasa-task load index (NASA-TLX); 20 years later: (577632012-009)*. 2006. DOI: 10.1037/e577632012-009. URL: http://dx.doi.org/10.1037/e577632012-009 5.

[JBK21] Johnson-Glenberg, Mina C, Bartolomea, Hannah, and Kalina, Elena. "Platform is not destiny: Embodied learning effects comparing 2D desktop to 3D virtual reality STEM experiences". *Journal of Computer Assisted Learning* 37.5 (2021), 1263–1284. DOI: 10.1111/jcal.12567 3.

[KKZ*23] Kosch, Thomas, Karolus, Jakob, Zagermann, Johannes, et al. "A Survey on Measuring Cognitive Workload in Human-Computer Interaction". *ACM Computing Surveys* 55.13s (July 2023), 1–39. ISSN: 1557-7341. DOI: 10.1145/3582272. URL: http://dx.doi.org/10.1145/3582272 2.

[KMM21] Kao, Dominic, Magana, Alejandra J., and Mousas, Christos. "Evaluating Tutorial-Based Instructions for Controllers in Virtual Reality Games". *Proceedings of the ACM on Human-Computer Interaction* 5.CHI PLAY (Oct. 2021), 1–28. ISSN: 2573-0142. DOI: 10.1145/3474661. URL: http://dx.doi.org/10.1145/3474661 4.

[LKLS23] Lønne, Tuva Fjærtoft, Karlsen, Håvard R., Langvik, Eva, and Saksvik-Lehouillier, Ingvild. "The effect of immersion on sense of presence and affect when experiencing an educational scenario in virtual reality: A randomized controlled study". *Heliyon* 9.6 (2023), e17196. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2023.e17196. URL: https://www.sciencedirect.com/science/article/pii/S2405844023044043 3.

[LSM21] Ling, Samuel J, Sanny, Jeff, and Moebs, William. *University physics*. OpenStax, Rice University, 2021 6.

[May11] Mayer, Richard E. "Applying the science of learning". (2011) 2.

[MBM19] Makransky, Guido, Borre-Gude, Stefan, and Mayer, Richard E. "Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments". *Journal of Computer Assisted Learning* 35.6 (July 2019), 691–707. ISSN: 1365-2729. DOI: 10.1111/jcal.12375. URL: http://dx.doi.org/10.1111/jcal.12375 2.

[Mik06] Mikropoulos, Tassos A. "Presence: a unique characteristic in educational virtual environments". *Virtual Reality* 10.3–4 (Sept. 2006), 197–206. ISSN: 1434-9957. DOI: 10.1007/s10055-006-0039-1. URL: http://dx.doi.org/10.1007/s10055-006-0039-1 3.

[MM06] Myers, Carrie B. and Myers, Scott M. "Assessing Assessment: The Effects of Two Exam Formats on Course Achievement and Evaluation". *Innovative Higher Education* 31.4 (Sept. 2006), 227–236. ISSN: 1573-1758. DOI: 10.1007/s10755-006-9020-x. URL: http://dx.doi.org/10.1007/s10755-006-9020-x 2.

[MSA03] Mislevy, Robert J., Steinberg, Linda S., and Almond, Russell G. "Focus Article: On the Structure of Educational Assessments". *Measurement: Interdisciplinary Research & Perspective* 1.1 (Jan. 2003), 3–62. ISSN: 1536-6359. DOI: 10.1207/s15366359mea0101_02. URL: http://dx.doi.org/10.1207/S15366359MEA0101_02 2.

[PAP*20] Putze, Susanne, Alexandrovsky, Dmitry, Putze, Felix, et al. "Breaking The Experience: Effects of Questionnaires in VR User Studies". *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. ACM, Apr. 2020. DOI: 10.1145/3313831.3376144. URL: http://dx.doi.org/10.1145/3313831.3376144 2.

[PL20] Pittman, Corey and LaViola, Joseph J. "PhyAR: Determining the Utility of Augmented Reality for Physics Education in the Classroom". *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Mar. 2020. DOI: 10.1109/vrw50115.2020.00231. URL: http://dx.doi.org/10.1109/VRW50115.2020.00231 3.

[PMM21] Petersen, Gustav Bøg, Mottelson, Aske, and Makransky, Guido. "Pedagogical Agents in Educational VR: An in the Wild Study". *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. ACM, May 2021. DOI: 10.1145/3411764.3445760. URL: http://dx.doi.org/10.1145/3411764.3445760 2.

[PPM20] Pitana, T, Prastowo, H, and Mahdali, A P. "The Development of Fire Safety Appliances Inspection Training using Virtual Reality (VR) Technology". *IOP Conference Series: Earth and Environmental Science* 557.1 (Aug. 2020), 012064. ISSN: 1755-1315. DOI: 10.1088/1755-1315/557/1/012064. URL: http://dx.doi.org/10.1088/1755-1315/557/1/012064 3.

[QMPY20a] Qian, Jingcheng, Ma, Yancong, Pan, Zhigeng, and Yang, Xubo. "Effects of Virtual-real fusion on immersion, presence, and learning performance in laboratory education". *Virtual Reality & Intelligent Hardware* 2.6 (Dec. 2020), 569–584. ISSN: 2096-5796. DOI: 10.1016/j.vrih.2020.07.010. URL: http://dx.doi.org/10.1016/j.vrih.2020.07.010 2.

[QMPY20b] Qian, Jingcheng, Ma, Yancong, Pan, Zhigeng, and Yang, Xubo. "Effects of Virtual-real fusion on immersion, presence, and learning performance in laboratory education". *Virtual Reality & Intelligent Hardware* 2.6 (2020), 569–584. ISSN: 2096-5796. DOI: 10.1016/j.{VR}ih.2020.07.010 3.

[RSSS18] Regal, Georg, Schatz, Raimund, Schrammel, Johann, and Suette, Stefan. "VRate: A Unity3D Asset for integrating Subjective Assessment Questionnaires in Virtual Environments". *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, May 2018. DOI: 10.1109/qomex.2018.8463296. URL: http://dx.doi.org/10.1109/qomex.2018.8463296 2.

[RT11] Reeve, Johnmarshall and Tseng, Ching-Mei. "Agency as a fourth aspect of students' engagement during learning activities". *Contemporary Educational Psychology* 36.4 (Oct. 2011), 257–267. ISSN: 0361-476X. DOI: 10.1016/j.cedpsych.2011.05.002. URL: http://dx.doi.org/10.1016/j.cedpsych.2011.05.002 2.

[Sad89] Sadler, D. Royce. "Formative assessment and the design of instructional systems". *Instructional Science* 18.2 (June 1989), 119–144. ISSN: 1573-1952. DOI: 10.1007/bf00117714. URL: http://dx.doi.org/10.1007/BF00117714 2.

[SHKP21] Safikhani, Saeed, Holly, Michael, Kainz, Alexander, and Pirker, Johanna. "The Influence of in-VR Questionnaire Design on the User Experience". *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. VRST '21. ACM, Dec. 2021. DOI: 10.1145/3489849.3489884. URL: http://dx.doi.org/10.1145/3489849.3489884 2, 7.

[SJ95] Schwarzer, R. and Jerusalem, Matthias. *General Self-Efficacy Scale*. 1995. DOI: 10.1037/t00393-000. URL: http://dx.doi.org/10.1037/t00393-000 5.

[SMN*17] Shaikh, Uzma A. S., Magana, Alejandra J., Neri, Luis, et al. "Undergraduate students' conceptual interpretation and perceptions of haptic-enabled learning experiences". *International Journal of Educational Technology in Higher Education* 14.1 (May 2017). ISSN: 2365-9440. DOI: 10.1186/s41239-017-0053-2. URL: http://dx.doi.org/10.1186/s41239-017-0053-2 2.

[SR17] Shute, V.J. and Rahimi, S. "Review of computer-based assessment for learning in elementary and secondary education". *Journal of Computer Assisted Learning* 33.1 (Jan. 2017), 1–19. ISSN: 1365-2729. DOI: 10.1111/jcal.12172. URL: http://dx.doi.org/10.1111/jcal.12172 2.

[SYA*08] Shavelson, Richard J, Young, Donald B, Ayala, Carlos C, et al. “On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers”. *Applied Measurement in Education* 21.4 (Sept. 2008), 295–314. issn: 1532-4818. doi: 10.1080/08957340802347647 2.

[TKAM06] Tory, M., Kirkpatrick, A.E., Atkins, M.S., and Moller, T. “Visualization task performance with 2D, 3D, and combination displays”. *IEEE Transactions on Visualization and Computer Graphics* 12.1 (Jan. 2006), 2–13. issn: 1077-2626. doi: 10.1109/tvcg.2006.17. url: http://dx.doi.org/10.1109/tvcg.2006.17 7.

[VST05] Velez, M C, Silver, D, and Tremaine, M. “Understanding Visualization through Spatial Ability Differences”. *VIS 05. IEEE Visualization.* IEEE, 2005. doi: 10.1109/visual.2005.1532836. url: http://dx.doi.org/10.1109/VISUAL.2005.1532836 7.

[Wil49] Williams, EJ. “Experimental Designs Balanced for the Estimation of Residual Effects of Treatments”. *Australian Journal of Chemistry* 2.2 (1949), 149. issn: 0004-9425. doi: 10.1071/ch9490149. url: http://dx.doi.org/10.1071/ch9490149 4.

[WUM*23] Won, Mihye, Ungu, Dewi Ayu Kencana, Matovu, Henry, et al. “Diverse approaches to learning with immersive Virtual Reality identified from a systematic review”. *Computers & Education* 195 (Apr. 2023), 104701. issn: 0360-1315. doi: 10.1016/j.compedu.2022.104701. url: http://dx.doi.org/10.1016/j.compedu.2z022.104701 7.

[XLA*21] Xie, Biao, Liu, Huimin, Alghofaili, Rawan, et al. “A Review on Virtual Reality Skill Training Applications”. *Frontiers in Virtual Reality* 2 (Apr. 2021). issn: 2673-4192. doi: 10.3389/frvir.2021.645153. url: http://dx.doi.org/10.3389/frvir.2021.645153 1, 2.

[ZSHC20] Zainuddin, Zamzami, Shujahat, Muhammad, Haruna, Hussein, and Chu, Samuel Kai Wah. “The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system”. *Computers & Education* 145 (Feb. 2020), 103729. issn: 0360-1315. doi: 10.1016/j.compedu.2019.103729. url: http://dx.doi.org/10.1016/j.compedu.2019.103729 2.