Proceedings of the Human Factors and Ergonomics Society Annual Meeting
1–6
Copyright © 2024 Human Factors and Ergonomics Society
DOI: 10.1177/10711813241269236
journals.sagepub.com/home/pro
Sage

Study of Graphic Armatures, Multimodal Cues, and Numeric Measures in Virtual Reality on Learners' Performance and Workload

Fangyuan Cheng[1], Tyler Harper-Gampp[1], Rebecca Planchart[1], Mitchell Dunning[1], Matthew Peterson[1], Cesar Delgado[1], and Karen B. Chen[1]

Abstract

Virtual reality (VR) is increasingly utilized in education, yet its effectiveness can vary due to potential distractions and excessive workload. Prior research suggests that virtual signaling elements can enhance learning in VR environments. However, the effectiveness of different design elements for specific learning content and their impact on learner workload remain understudied. This study examines the influence of graphic armatures, multimodal cues, and numeric measures on scale learning in Scale Worlds, a VR learning environment for exploring scientific entities across multiple scales. Preliminary results indicate that numeric measures notably enhance learning outcomes by providing direct scale representations. It shows that different virtual elements can variably affect learners' scale learning outcomes and behaviors and can lead to varying levels of workload. This study underscores the importance of aligning the design of virtual elements with educational objectives and ensuring they induce an appropriate level of workload for learning in VR learning environment.

Keywords

virtual reality, virtual elements, performance, workload, scale learning

Introduction

Virtual reality has increasingly been applied in various disciplines for educational purposes (Radianti et al., 2020), offering immersive, three-dimensional learning environments distinct from traditional settings (Pellas et al., 2021). However, recent empirical studies revealed that VR-based learning might not always yield positive learning outcomes and might even negatively affect learners (Makransky et al., 2019; Parong & Mayer, 2018). This may be due to distractions and cognitive overload caused by the design and presentation of VR content (Albus et al., 2021; Frederiksen et al., 2020), highlighting an opportunity for further research.

Research across traditional learning media supports the effectiveness of "signaling" in improving learning (Schneider et al., 2018). Signaling refers to instructional cues designed to help learners understand the elements or organization of instructional materials (Albus et al., 2021; Mautone & Mayer, 2001; Schneider et al., 2018). These cues can take various forms, such as color, auxiliary graphics like arrows, and textual annotations. The signaling principle have been proven to effectively direct learner attention, aiding in the selection, integration, and processing of essential information in traditional learning contexts (Li et al., 2023; Vogt et al., 2021). Signaling elements have also been used in VR to enhance focus (Liu et al., 2022), recall, and attention management (Albus et al., 2021). However, there is a lack of research comparing the effects of different types of signaling elements on VR learning outcomes and whether various signaling elements might cause different levels of workload for learners. Furthermore, due to the contextual nature of signaling, the same type of signaling elements may vary in form and function across different learning environments, and therefore the effectiveness of these elements should be re-evaluated for specific learning content (Radianti et al., 2020), including abstract concepts such as size and scale.

Scale is integral to science and engineering education, as highlighted by the Next Generation Science Standards

[1]North Carolina State University, Raleigh, USA

Corresponding Author:
Karen B. Chen, North Carolina State University, 915 Partners Way, Raleigh, NC 27606, USA.
Email: kbchen2@ncsu.edu

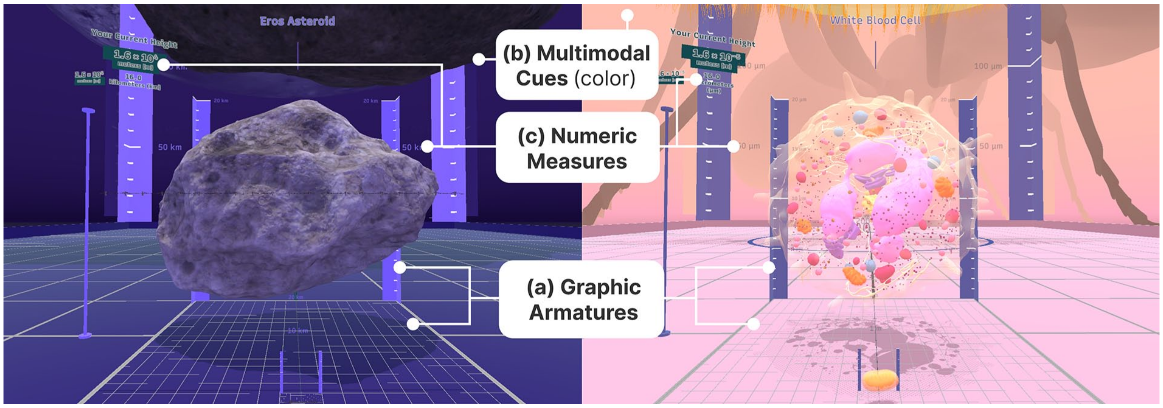

Figure 1. A screenshot of the Eros Asteroid world (left) and the White Blood Cell world (right) from the Scale World with three types of virtual elements: (a) graphic armatures, (b) multimodal cues, and (c) numeric measures.

(National Research Council, 2012) and Common Core mathematics standards (Common Core, 2000). Research revealed a significant gap in students' understanding of scale (Delgado, 2013; Magana et al., 2012), particularly in accurately grasping the sizes of molecules, cells, and atoms, as well as conceptualizing large numerical values and dimensions of various scales (Swarat et al., 2011).

The research team has developed Scale Worlds, a virtual learning environment incorporating different signaling elements and scientific entities across multiple scales (ranging from an atom to the Sun) to help students learn scale, size, and numbers (Wu et al., 2022). Three types of virtual elements were included in Scale Worlds to support learning: (1) graphic armatures to assist in measuring entities, (2) multimodal cues to support intuitive navigation through size transitions and aid cognitive processes such as grouping and ordering, and (3) numeric measures such as numbers and units for indicating size and scale. The present study aimed to examine how these three kinds of virtual elements affect learners' performance, in terms of scale cognition and associated behaviors, and workload.

Methods

Participants

Fifteen participants (5 females, 9 males, and 1 non-binary), aged 18 to 23 years ($M = 18.8$, $SD = 1.27$) with no prior experience with Scale Worlds, were enrolled in this study with informed consent, which was approved by the North Carolina State University Institutional Review Board. The inclusion criteria specified that participants must be first-year undergraduate students in the College of Engineering and have no tendencies for VR discomfort. First-year engineering students were included for their fundamental understanding of scale without being influenced by discipline-specific scale knowledge.

Equipment and Virtual Environment

Scale Worlds was created and rendered utilizing a game engine (Unity 2018.4.28f1, https://unity3d.com/) and provided to the participants using a head-mounted display (HMD; Vive, HTC, and Valve Corporation) and handheld controllers. Scale Worlds encompassed 24 scientific entities arranged vertically facing the participants, each representing a "world" with scales varying in tenfold increments from 10^{-12} to 10^{12}. Participants could explore different worlds to experience perspectives at various scales through interactions of teleportation and scaling up or down via controllers.

The virtual elements included in this study are shown in Figure 1. *Graphic armatures* include gridlines on the ground and hash marks on rulers, indicating a length of 20 units of the current scale (e.g., 20 cm in the Acorn World). *Multimodal cues* include visual and auditory elements: five color schemes (grayscale, pink, daylight, dark blue, and black) and sounds that change in duration and frequency with scaling. *Numerical measures* include digits displayed on an information panel on the sides of entities, indicating the current world's scale in meters, such as "1.6×10^{-5} m," with additional numbers for distances or heights in units pertinent to the current world on the ground and behind entities, such as "20 pm."

Experiment Procedure

Upon providing informed consent, participants first completed a validated pre-test, the Assessment of Size and Scale Cognition (ASSC; Harper-Gampp et al., 2023a, 2023b), to gauge their basic perspective on scale. Researchers demonstrated how to wear the HMD and interact with Scale Worlds. Participants then explored Scale Worlds and performed specific size and scale comparisons as instructed (e.g., pick an entity and then find another entity that is 1,000 times larger). Participants' behaviors in

VR were recorded from a first-person perspective through screen recording, and their verbalized thoughts were audio-recorded. After experiencing Scale Worlds, participants reported their perceived workload level, followed by the ASSC post-test. Finally, participants participated in a semi-structured interview to share their thoughts and feedback on the VR experience. Participants will be compensated $15/hr for approximately 90 min of participation.

Variables and Analysis

A between-subject design with five conditions was used in this study, including four experimental and one control group. The experimental groups were: graphic armatures only (GA), multimodal cues only (CS), numerical measures only (NM), and the group with all three elements combined (AL). The control group was devoid of all three virtual elements. The dependent variables included learning outcomes, workload, and verbalization and observation.

Learning Outcome. The ASSC was used for both pre- and post-tests to assess learners' scale cognition abilities. The ASSC comprises five sections, each designed to evaluate scale cognition abilities in specific areas (i.e., ordering, grouping, absolute and relative reasoning). Sample questions included "Please place the following objects in order from largest to smallest according to their size" and "Please group the objects based on their size." The entities asked in the ASSC were different from those shown in Scale Worlds to avoid memorization. Changes in learning outcomes of different conditions were compared by taking the difference (denoted as Δ) between total scores and individual abilities scores from pre-tests and post-tests.

Workload. Workload was assessed using the NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988) to examine participants' workload during their learning experience in Scale Worlds. A weighted average was employed to assess workload levels. This study analyzed participants' weighting of workload subscales, overall workload levels, and subscale workload levels across different experimental conditions.

Verbalization and Observation Data. The think-aloud and interview data were transcribed and then coded deductively using pre-established codes related to the usage of different virtual elements. Additionally, screen recordings of participants' behaviors were used to contextualize their verbalizations, providing a deeper understanding of their interactions with virtual elements in scale learning.

Results

Due to the small sample size in this preliminary study, a descriptive analysis of the ASSC scores and NASA-TLX weighted average scores (did not meet normality conditions) and the analysis of variance (ANOVA) at $\alpha = .05$ significance level for the NASA-TLX weight allocation (satisfied parametric test criteria) are reported.

Learning Outcome

Overall, there was an increase in the mean ASSC score from pre-test across all conditions (pre-test: $M(SD) = 48.09(9.8)$; post-test: $M(SD) = 55(12.87)$; $\Delta = 6.91$). The NM condition exhibited the greatest improvement ($\Delta = 16.5$), followed by the CS condition ($\Delta = 8.83$, $SD = 4.95$), while the GA condition showed the least improvement ($\Delta = 1.0$, $SD = 4.24$). The AL condition ($\Delta = 2.33$, $SD = 7.23$) and the control condition ($\Delta = 2$, $SD = 10.6$) exhibited similar degrees of learning change.

For the subsections of the ASSC, the CS condition showed an improvement in ordering ability ($\Delta = 4.34$, $SD = 3.51$), while the other conditions scored lower on their post-test, with the control condition experiencing the largest decrease ($\Delta = -2.66$, $SD = 3.06$). All conditions improved in absolute reasoning ($\Delta = 5.03$, $SD = 4.92$), with the NM condition showing the highest improvement ($\Delta = 12.0$). Regarding relative reasoning, the NM condition had the best improvement ($\Delta = 8.5$), whereas the GA ($\Delta = -4.5$, $SD = 6.36$) and control ($\Delta = -0.33$, $SD = 7.78$) conditions decreased. Grouping ability showed minimal improvement ($\Delta = 0.13$, $SD = 1.25$), with slight effects in the control ($\Delta = 0.67$, $SD = 1.52$) and the GA ($\Delta = -0.67$, $SD = 1.52$) conditions.

Workload

ANOVA results revealed a statistically significant difference in weight allocation among the NASA-TLX subscales ($F(5, 84) = 13.93$, $p < .001$). Tukey post-hoc analysis showed higher weights for mental demand ($M = 0.25$, $SD = 0.07$) over physical demand ($M = 0.04$, $SD = 0.07$), effort ($M = 0.16$, $SD = 0.08$), and frustration ($M = 0.16$, $SD = 0.11$), and for performance ($M = 0.22$, $SD = 0.09$) over physical demand ($M = 0.04$, $SD = 0.07$) and frustration ($M = 0.16$, $SD = 0.11$). Considering the effect of conditions, the NM condition was associated with the lowest allocation of temporal demand ($M = 0.09$, $SD = 0.07$). The control group received the highest weight for frustration ($M = 0.27$, $SD = 0.07$), while the AL condition received the lowest ($M = 0.04$, $SD = 0.08$).

For the weighted average scores across conditions, the AL condition demonstrated the lowest total workload ($M = 31.22$, $SD = 14.63$), while the NM condition exhibited the highest ($M = 44.56$, $SD = 11.70$). The control condition showed the least effort ($M = 4.0$, $SD = 0.67$) and the highest temporal demand ($M = 8.0$, $SD = 8.11$) and frustration ($M = 8.89$, $SD = 2.78$). The GA condition reported the lowest mental demand ($M = 9.67$, $SD = 4.97$) and the poorest performance ($M = 3.78$, $SD = 3.67$). The CS condition reported the best performance ($M = 3.78$, $SD = 3.67$) and highest level of effort ($M = 10.89$, $SD = 12.51$). The NM condition showed the lowest temporal demand (M=13.33, SD=7.42) and the highest mental demand (M=4.22, SD=4.02). The AL condition showed the lowest frustration (M=0.22, SD=0.38).

Verbalization and Observation

Researchers synthesized and interpreted the semantic codes, leading to three key findings related to each virtual element as follows:

Numeric Measures as Direct Answers. Numeric measures were identified as the most frequently used virtual elements. Semi-structured interview revealed that all six participants who were assigned to a condition with numeric measures, participants reported utilizing them during the learning process. Among the two types of numeric measures, the information panel was found to be particularly useful for answering instructional questions related to absolute reasoning. For instance, P2: "I found it really helpful to have the numbers there. Knowing that when I scaled up and down, it was by a factor of 10 helped put things into perspective." Groups with numeric measures tended to naturally accept the digit of the numeric measures as the answer and understood the 10-step scale change between worlds more intuitively. On the other hand, participants in the GA (P5) and CS (P8) conditions had to independently discover the 10-step scale change to answer questions. Additionally, participants in conditions lacking numeric measures suggested that adding numerical values to directly indicate the scale of each world would help them better understand scale (P12: "I think maybe there is a way to see the true length, like they had a number, instead of just being relative to the previous object.").

Graphic Armatures as Tools for Precise Measurements. While only two of six participants who had graphic armatures in their condition reported utilizing them, they were considered useful primarily when precise measurements were needed (P3: "I referred to them a few times for sake of the questions."). When participants needed to provide a precise measurement of an entity, they located the entity's edges relative to the gridlines and calculated the number of grid cells. Two out of three participants in the GA condition noted that the armatures alone were insufficient to support their acquisition of adequate information for answering questions and learning scales (P5: "I didn't really use the blue ruler that much, mostly since I didn't realize what it was measuring.").

Multimodal Cues as Experiential Enhancers. Color and sound were rarely noticed by participants as useful for learning about scale. In all six observations involving color and sound, only two participants noticed changes in color schemes, and three noticed sound effects. None of the participants indicated directly using color and sound for learning scale. Only with one participant mentioned, "it might not be something that you would exactly notice, but it would subconsciously help you" (P4). While not directly linked to scale learning, there are three participants noted that color and sound enhanced the overall immersion and helped them stay more engaged.

Discussion

The analysis of ASSC scores found that while using Scale Worlds showed a trend toward enhancing overall learning outcome, different elements variably affected scale cognition abilities (e.g., ordering, grouping). In particular, numeric measures as a virtual element had the greatest positive influence in the total score and both absolute and relative reasoning abilities. Verbalization and observation data revealed that learners frequently used numeric measures on the information panel as references when responding to instructional questions, suggesting that the direct numerical representation of scale aids learners in understanding both the absolute sizes of entities and their relative sizes to each other.

Not all scale cognition abilities improved after experiencing Scale Worlds. Interestingly, only the CS condition showed improved ordering ability, while other conditions saw declines. Although participants' self-reports indicated they did not actively use color and sound for learning scales, the results suggest that these elements may still serve as subtle learning supports, providing an unconscious context for understanding and remembering ordering relationships between entities. Furthermore, declines were more pronounced in the GA and control conditions, with decreases in ordering, relative reasoning, and grouping abilities in the GA condition, and in ordering and relative reasoning abilities in the control condition. The reasons for these declines might be complex and were not captured in the verbalization and observation data. Despite this, the insight underscores the importance of deliberately selecting virtual elements to ensure they align with specific educational goals and learning cognitive process, as inappropriate designs could detrimentally affect learning outcomes.

The NASA-TLX weight allocation indicated that mental demand and performance were key workload factors in learning abstract concepts of size and scale in virtual environments. Differences in weight allocation suggested virtual elements were able to influence perceived workload weighting. Additionally, there was consistency between the weight allocation for the subscales and the weighted average scores that both metrics for temporal demand were lowest in the NM condition. Integrating verbalization and observation data, it can be inferred that the NM condition provided answers to instruction questions more directly by offering digits representations, reducing the time needed for calculations and, therefore, the perceived temporal demand (i.e., the extent of time pressure felt due to the rate or pace at which tasks or task elements occurred). Additionally, both metrics for frustration were lowest in the AL and highest in the control condition. This was likely because the AL condition provided more extensive reference information, allowing learners to resolve their frustration from multiple virtual elements. These findings may aid researchers or developers in designing learning environments with a more careful consideration of the virtual elements' impact on users' workload.

Cross-referencing results from the ASSC with the NASA-TLX provided insights into the relationship between learning outcomes and workload. The GA condition, with the least mental demand, demonstrated the poorest performance evaluations and ASSC scores in total, ordering, and grouping and relative reasoning abilities. In contrast, the NM condition, with the most mental demand, showed the lowest temporal demand and the highest scores in ASSC total, absolute and relative reasoning. A possible explanation is that a certain level of mental demand was beneficial for learning scales, whereas graphic armatures did not sufficiently stimulate it. However, high mental demand did not always lead to better outcomes. For example, the control condition had the second-highest mental demand but the second worst ASSC total score. This may be due to the corresponding mental demand being applied to irrelevant elements rather than learning scales. Therefore, it could not be asserted that greater mental demand necessarily improves learning effectiveness, indicating the need for future research to focus on a more granular differentiation of the use of mental demand to elucidate its precise relationship with learning outcomes.

This study offers insight for the future development of VR learning environments, suggesting that the design of virtual elements should align with learning objectives and the ability of the environment with the presence of the elements to induce an appropriate level of mental demand to achieve those objectives. Furthermore, it suggests that balancing the design of virtual elements to support learning while managing their impact on cognitive workload across various learning contexts warrants further investigation.

Limitations

This preliminary study currently encompasses only a small sample size ($n=15$), it was not feasible to use parametric tests for some variables to obtain statistically significant results. Further study with larger sample size is warranted to substantiate the findings. Furthermore, more data sources of the learners' states and behaviors are needed to fully explain the variations in learning performance and workloads. Moreover, a measurement tool with more detailed sub-components was needed to determine the source of workload in learning activities.

Conclusion

While virtual reality (VR) has been extensively employed in educational settings, the impact of various virtual elements in specific learning context remains underexplored. This study examined the influence of graphic armatures, multimodal cues, and numeric measures in supporting scale learning. Preliminary findings indicate that these elements differentially affected learners' performance and workload. Numeric measures notably enhanced learning outcomes by providing direct scale representations. However, the results suggested that the introduce of virtual elements did not always enhance performance and may have adverse effects on learning. This study underscores the importance of aligning the design of virtual elements with educational objectives and ensuring they induce an appropriate level of mental demand for learning in the VR learning environment.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Science Foundation (DRL-2055680).

ORCID iDs

Fangyuan Cheng https://orcid.org/0009-0001-9306-8471
Karen B. Chen https://orcid.org/0000-0003-2904-1394

References

Albus, P., Vogt, A., & Seufert, T. (2021). Signaling in virtual reality influences learning outcome and cognitive load. *Computers & Education*, *166*, 104154. https://doi.org/10.1016/j.compedu.2021.104154

Common Core. (2000). *Mathematics standards – Common Core state standards initiative*. http://www.corestandards.org/Math/

Delgado, C. (2013). Navigating deep time: Landmarks for time from the big bang to the present. *Journal of Geoscience Education*, *61*(1), 103–112. https://doi.org/10.5408/12-300.1

Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., & Andersen, S. A. W. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: A randomized trial. *Surgical Endoscopy*, *34*(3), 1244–1252. https://doi.org/10.1007/s00464-019-06887-8

Harper-Gampp, T., Delgado, C., Peterson, M., & Chen, K. B. (2023a, April). *Designing and developing an instrument to assess scale cognition* [Paper presentation]. National Association for Research in Science Teaching Annual Conference, Chicago, IL, United States.

Harper-Gampp, T., Delgado, C., Peterson, M., & Chen, K. B. (2023b, April). *Refining a panel of experts validation methodology for instrument development* [Roundtable presentation]. American Education Research Association, Chicago, IL, United States.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in* *psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Li, W., Feng, Q., Zhu, X., Yu, Q., & Wang, Q. (2023). Effect of summarizing scaffolding and textual cues on learning performance, mental model, and cognitive load in a virtual reality environment: An experimental study. *Computers & Education*, *200*, 104793. https://doi.org/10.1016/j.compedu.2023.104793

Liu, R., Xu, X., Yang, H., Li, Z., & Huang, G. (2022). Impacts of cues on learning and attention in immersive 360-degree video: An eye-tracking study. *Frontiers in Psychology*, *12*, 792069. https://doi.org/10.3389/fpsyg.2021.792069

Magana, A. J., Brophy, S. P., & Bryan, L. A. (2012). An integrated knowledge framework to characterize and scaffold size and scale cognition (FS2C). *International Journal of Science Education*, *34*(14), 2181–2203. https://doi.org/10.1080/09500693.2012.715316

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, *60*, 225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Mautone, P. D., & Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology*, *93*(2), 377.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas* (p. 13165). National Academies Press. https://doi.org/10.17226/13165

Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology*, *110*(6), 785–797. https://doi.org/10.1037/edu0000241

Pellas, N., Mystakidis, S., & Kazanidis, I. (2021). Immersive virtual reality in K-12 and higher education: A systematic review of the last decade scientific literature. *Virtual Reality*, *25*(3), 835–861. https://doi.org/10.1007/s10055-020-00489-9

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, *147*, 103778. https://doi.org/10.1016/j.compedu.2019.103778

Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educational Research Review*, *23*, 1–24. https://doi.org/10.1016/j.edurev.2017.11.001

Swarat, S., Light, G., Park, E. J., & Drane, D. (2011). A typology of undergraduate students' conceptions of size and scale: Identifying and characterizing conceptual variation. *Journal of Research in Science Teaching*, *48*(5), 512–533. https://doi.org/10.1002/tea.20403

Vogt, A., Albus, P., & Seufert, T. (2021). Learning in virtual reality: Bridging the motivation gap by adding annotations. *Frontiers in Psychology*, *12*, 645032. https://doi.org/10.3389/fpsyg.2021.645032

Wu, L., Sekelsky, B., Peterson, M., Gampp, T., Delgado, C., & Chen, K. B. (2022). Immersive virtual environment for scale cognition and learning: Expert-based evaluation for balancing usability versus cognitive theories. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1972–1976. https://doi.org/10.1177/1071181322661094