Immersive virtual environment for scale cognition and learning: Expert-based evaluation for balancing usability versus cognitive theories

<author information removed
for blind review>

While scale cognition and learning is a crosscutting concept that pervades science and can aid students in making connections across disciplines, students struggle to conceptualize and consider scales that go far beyond their everyday world experience. Virtual reality technology affords embodied learning experiences, which enable students to physically engage in learning activities in an environment with rich information. Scale Worlds is a virtual learning environment implemented in an immersive CAVE, which portrays scientific entities of a wide range of sizes. A user can scale themselves up or down by powers of ten, in order to experience entities from an atom to the Sun. This paper reports on an expert-based usability evaluation of Scale Worlds, including three sets of A/B testing, by five usability experts. Outcomes of the usability evaluation will inform the refinement of Scale Worlds. The evaluation provides insights for usability evaluation and design in immersive virtual environments.

INTRODUCTION

The US science standards posit "scale, proportion, and quantity" as a crosscutting concept that pervades science and can aid students in making connections across disciplines (National Research Council, 2012). US mathematics standards state that fourth grade students should be able to generalize their understanding of place value to 1,000,000 (10^6) and the relative sizes of numbers in each place, and eighth graders should be able to use a single digit times an integer power of ten to represent very large or very small numbers (Common Core, 2000). However, research shows that learners of all ages hold inaccurate ideas about the size of scientifically relevant entities (Delgado, 2013; Magaña et al., 2012; Tretter et al., 2006). Learners often confuse cells and atoms (Flores et al., 2003; Harrison & Treagust, 1996), not realizing there are huge *relative size differences* among entities too small to see.

Theory on educational practice in virtual reality (VR) emphasizes the coupling between the use of body motions and cognitive activities (Arroyo et al., 2017; Skulmowski & Rey, 2018). Embodied cognition theory posits that mind, body, and environment are interrelated and mutually dependent; cognition is not only a capacity of the brain, but the ability to coordinate mind, body and interactions in an environment into a dynamic system geared to a specific purpose where sensory perception and movement are important (van der Schaaf, Bakker, and ten Cate 2019). VR affords embodied learning experiences, which enable students to physically engage in learning activities in an environment with rich information (Dalgarno & Lee, 2010; Johnson-Glenberg et al., 2014). The use of virtual and augmented reality for science education has revealed increased learning gains relative to traditional instruction (Johnson-Glenberg et al., 2014).

Our research team has developed Scale Worlds, which is a virtual learning environment encompassing scientific entities of a wide range of sizes. A user can scale themselves up or down by powers of ten, in order to experience entities from an atom to the Sun, using a numeric panel they adjust with a handheld controller. The prototype was implemented in an immersive Cave Automatic Virtual Environment (CAVE) designed to support effective user perception and interaction (Figure 1). The development of Scale Worlds was guided by theory on visual representations (Peterson et al., 2021) and scale cognition (Delgado, 2013; Magaña et al., 2012).

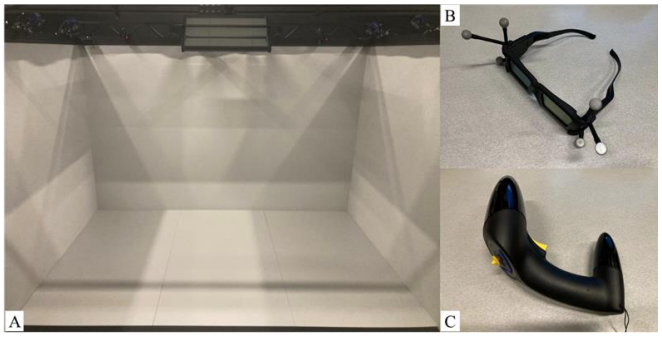

Figure 1. The CAVE and accessories: (A) The CAVE consists of four projectors and their respective projection surfaces, including three screens or "walls" and a floor, (B) 3D shutter glasses with positional markers (the reflective balls), and (C) a hand-held wand (controller).

During development, we encountered conflicts between usability and the cognitive theories. Three cases of these theory–usability conflicts suggested the need for A/B testing to assess the cost to usability. In the first case, *posts* distributed in Scale Worlds are intended to help users assess relative size (Delgado, 2009), because posts near the user are repeated near large entities, turning them into relative units (see Lamon, 1994, on unitizing). The posts function as an *armatur*e, i.e., non-depictive visual elements that assist in the interpretation of depictive entities (Peterson et al., 2021). However, posts may increase extraneous cognitive load by increasing the degree of element interactivity (Sweller 2010). Three versions were tested: a "forest" of posts repeated in all directions, a "path" of posts connecting entities, and a "plain" lacking any posts.

The second case of potential theory–usability conflict involves *interactions with the numeric panel* that controls the scaling animation. We compared a button click on the controller—a straightforward interaction, familiar from prior experience with media—to embodied interactions using the controller to move the decimal in standard notation and the exponent in scientific notation. Embodied cognition suggests

that directional gestures will better align with embodied conceptualizations. Teachers routinely ask students to "move the decimal point" left or right in standard notation, so a hand motion left or right to shrink or grow may better align with students' conceptualizations, supporting learning. In scientific notation, the exponent increases or decreases, suggesting an up or down motion aligned with the embodied conceptual understanding of "more is up" (Lakoff & Johnson, 1980).

Third, standard usability practices suggest minimizing delay between a user's action and the system's response, in this case the embodied interactions using the controller to move the decimal or exponent. Animations showing the resultant changes in decimal or exponent could coincide with the *scaling animation*, with no delay between user action and navigation, but this would likely result in the user overlooking the numeric animations. Thus, a staggered animation is hypothesized to better support learning, as well as implying cause and effect by guiding the user's attention to whatever is moving at any given moment. For instance, increasing the exponent causes the decimal to move right, and represents (or causes) a size increase.

The purpose of this initial formative evaluation performed by usability experts was to examine the usability of the Scale Worlds VR system, including the likeability of different designs of interactions through A/B testing, and to help resolve any contradictions between theory-driven design features and various aspects of usability. Outcomes of the usability evaluation will inform the refinement of the next iteration of Scale Worlds.

METHODS

Participants

Five usability experts participated in the heuristic usability evaluation session. Five was considered a favorable cost-benefit ratio following usability literature (Faulkner, 2003). Criteria for participation were: at least two years of experience in user interface design and evaluation; or a graduate degree with relevant experience in human factors. Individuals with history of epileptic seizure or blackout, tendency for motion sickness, or sensitivity to flashing lights were excluded. Each evaluation session lasted approximately two hours.

Equipment

The Scale Worlds prototype (Figure 2) was developed and implemented in a CAVE (Viscube, Visbox, St. Joseph, IL), which is a room-size, 3D projection-based immersive virtual space (Cruz-Neira et al., 1992). The CAVE has three walls and a floor, and it uses four stereoscopic projectors with a resolution of 1920×1800 (Barco F50, Barco) to create an image on its corresponding wall or floor (Figure 1A). A motion tracking system (DTrack 2, ART GmbH) is used to monitor the real-time position and orientation of the active shutter 3D glasses of the user (Figure 1B). A user operates a hand-held wand (i.e., controller) to manipulate and interact with virtual objects and the user interface (Figure 1C). Participants were encouraged to think aloud, which was audio recorded using a SONY IC Recorder.

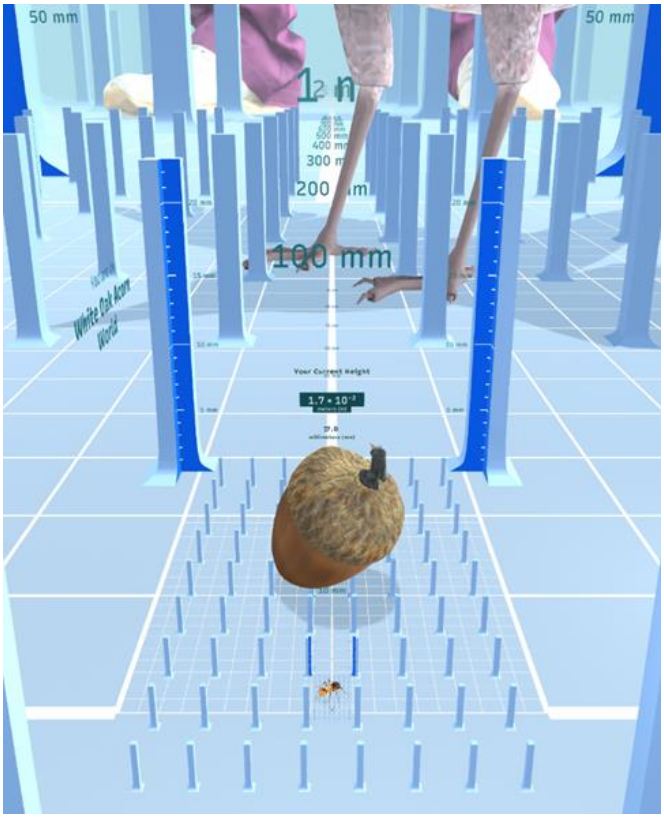

Figure 2. The user view of Scale Worlds at 1.7×10^{-2}, which is the White Oak Acorn World. A user has been "shrunk" 100x and their size is similar to that of the acorn. Also visible are other entities that have their own respective worlds (e.g., an ant, 1.7×10^{-3}, from one world smaller). The posts are scale armatures and their sizes are relative to their respective worlds (e.g., posts in one world are 10 times shorter than in the next world up). The numbers (e.g., 100 mm, 200 mm) indicate the length of each gridline.

Description of Scale Worlds

Scale Worlds encompasses scientific entities at tenfold increments in size. A user interface (UI) in the form of a panel with numeric information is placed in front of the entities (Figure 3A). A user can point at the exponent on the numeric panel using the ray cast from the wand, then perform a "swipe" gesture while holding the trigger to go to another scale world (Figure 3C). Upon swipe, a series of animations occurs: (1) the exponent of the scientific numeric notation flips to the next larger or smaller value; (2) the decimal point of the standard notation moves left or right; (3) the whole world scales as scientific entities grow or shrink; and (4) the unit updates to reflect a convenient unit (e.g., 1 km rather than 1,000 m). A user can also move between worlds by pointing at the decimal and moving the wand left or right while holding the trigger.

Experiment procedure and usability evaluation task

Before entering the CAVE, the facilitator explained the purpose of the usability evaluation and provided safety precautions. The usability evaluation protocol and informed consent were approved by <institution name removed for blinded review> Institutional Review Board. Informed consent was obtained from all participants. Each participant experienced a practice scenario first, to familiarize them with

the CAVE, 3D shutter glasses, and wand. The practice scenario had a clear sky background and various interactive 3D shapes that could be aimed at, clicked on, and moved using the wand.

Figure 3. The UI of Scale Worlds. (A) The UI in Human World, which is 1.7×10^0 m, with (B) a user holding the wand without the ray hitting any interactable UI component, (C) the color of the exponent changing from white to green when the ray hit the exponent, and (D) the decimal enlarging in size when the ray hit the decimal.

The complete usability evaluation consisted of the four primary user tasks (Table 1) followed by the A/B testing (Table 2). Participants were encouraged to think aloud throughout, and were audio recorded. Rest breaks outside the CAVE were offered between each task. One rest break was enforced after Task 2, during which the participant sat in a chair, removed the glasses, and completed a demographic questionnaire. Upon completing the four user tasks, the participant completed the Post-Study System Usability Questionnaire (PSSUQ; Lewis 2002) on paper and the NASA Task Load Index (NASA-TLX; Hart and Sta 1988) on a computer.

Table 1. The purposes and the descriptions of the four primary user tasks.

Task purpose	Detailed task instruction
Task 1: General exploration of Scale Worlds.	You are a first-time user of Scale Worlds. Suppose you are exploring Scale Worlds and trying to use it to learn about size, scale, and numbers. First, describe what you see in this Scale World. Next, describe what you would like to do in Scale Worlds.
Task 2: Evaluation of UI elements.	Suppose you want to learn about size, scale, and numbers. Describe to me anything related to scale and numbers that you see in the environment you believe is helping you to achieve the goal (of learning scale).
Task 3: Examination of scale armatures.	Suppose you would like to compare sizes of different entities in this World. Show me how you would do that and explain why.
Task 4: Interactions with scaling elements and animations.	Scale Worlds aim to help you learn size, scale, and numbers through visual rendering of scientific entities of different sizes in different Worlds. Suppose you want to go to a world with entities that are larger (or smaller) than the entities shown now. Tell me which world you are in (how do you know) and what you see.

Table 2. Three features with alternatives that were tested during A/B testing.

Feature	Alternatives
Scaling interactions between scale worlds	Gesture (hand motion) – swipe up/down over the exponent or swipe left/right over the decimal.
	Button clicks – click corresponding buttons on the controller while pointing at exponent or decimal.
Scale armature posts	Forest – a series of graduated posts extend to all directions and form the feature of armature like a forest.
	Path – graduated posts arranged linearly like a path.
	Plain – no posts in the environment.
Scaling animations	Staggered – the animations of exponent flip, decimal slide, world scaling, and unit change happen one after another (i.e., are staggered) after the user has scaled.
	Simultaneous - the animations of exponent flip, decimal change, world scaling, and unit change happening simultaneously after the user has scaled.

A/B testing for design alternatives of Scale Worlds

Alternatives of Scale Worlds features were created that pose potential conflicts between cognitive theories and user experience (UX) (Table 2, Figure 4). Therefore, A/B testing was performed to compare the various alternatives of the three features of interest. Bipolar laddering (Fonseca, 2015), a participatory subjective exploration method on UX, was conducted after A/B testing to understand user perception and preferences for the alternatives. Participants reported "outlook" (positive/negative), indicating if they liked each alternative. Then participants rated (on a scale of 0–10) the extent to which they liked or disliked the alternative. Finally, they reported "notes" to justify their scores and sentiment.

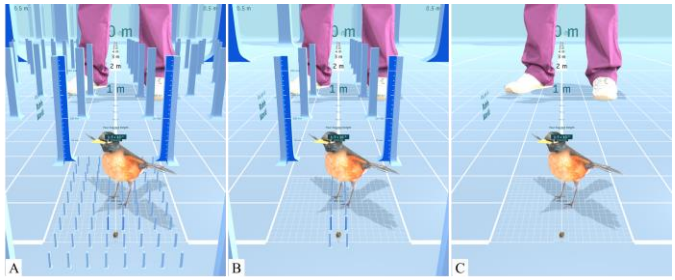

Figure 4. The three scale armature alternatives shown in A/B testing: (A) a high density "forest" of posts, (B) fewer posts that form a "path" connecting entities, and (C) a "plain" with no posts visible.

Variables and analysis

Quantitative evaluation of Scale Worlds and subjective workload. Quantitative data were collected from the PSSUQ to evaluate the system's usefulness, information quality, and interface quality on a 7-point Likert scale (1=strongly agree, 7=strongly disagree). Any PSSUQ item that received a 4 (neutral) or greater than 4 (disagree) is reported in the Results, as they suggest indifference or dissatisfaction. The NASA-TLX was administered to assess the workload of using Scale Worlds.

Themes from thinking aloud. Qualitative data were collected during think aloud. Specific usability comments were extracted from verbalizations, then grouped into themes.

A/B testing results from bipolar laddering. Scores and explanatory feedback were collected based on the participants' experience and likeability of the three features of Scale Worlds during A/B testing.

RESULTS

Post-study system usability questionnaire. Three items from the PSSUQ received a rating of 4 or higher from more than two participants (Table 4), indicating usability problems. All other items received ratings below 4 or N/A. Three participants were neutral or disagreed with "The system gave error messages that clearly told me how to fix problems"; two with "Whenever I made a mistake using the system, I could recover easily and quickly"; and two with "This system has all the functions and capabilities I expect it to have." Participants stated that the presence of error message and information would have been useful when they reached the extremes of scale and could not scale further—they thought the system was broken or stopped. One participant indicated that the system did not have all the functions and capabilities they expected. Specifically, they suggested providing a chart with all the scale worlds with the current world highlighted and being able to skip worlds.

NASA-TLX. Given the small number of usability experts (n=5), only descriptive statistics (mean and standard deviation) of the NASA-TLX weighted scores were computed (Figure 5).

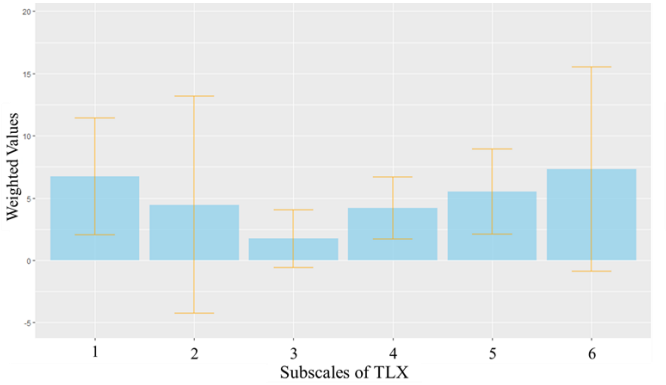

Figure 5. Mean NASA-TLX score for categories. 1: Mental demand (mean=6.73, sd=4.67), 2: Physical demand (mean=4.67, sd=8.70) 3: Temporal demand (mean=1.73, sd=2.30), 4: Performance (mean=4.20, sd=2.49), 5: Effort (mean=5.53, sd=3.41), 6: Frustration (mean=7.33, sd=8.22).

Participants stated the staggered animations for decimal, exponent, convenient unit, and scale world were difficult to follow and resulted in increased mental demand, and more so the simultaneous animations. Two features contributed to a high frustration rating: difficulty aiming at the small interactable components on the UI, and the inability to skip worlds.

Think aloud. We identified recurring themes from participants' verbalized thought process during the primary four tasks (Table 3). Two of the participants stated that being able to see their own body reduced immersion and the feeling of shrinking or growing. On the other hand, four participants thought using their own body was useful to compare size. Three participants mentioned that the position of UI helped guide them to where they should stand but the location was restrictive.

Bipolar laddering (BLA) for A/B testing. We determined the aspects mentioned at higher rates by the users and their respective scores. The mention index shows the proportion of users who talked about the element (Table 4). For armatures, the forest armature was preferred by three participants because it provided "great depth perception" and was perceived as the "most immersive," while two other participants preferred the path armature because it was "less distracting." For the interaction mode of scaling, four participants indicated preference for button clicks because it was more "accurate" or it was the "expected" interaction. While two participants felt the gesture interaction to be "novel" and "interesting," the remaining participants felt negatively about the interaction because it was "hard to be accurate" and there were "lots of steps." For the type of animation, two participants preferred the staggered animation because it was "good to see more details." Three participants expressed the advantage of the simultaneous animation to be "faster, less delay."

Table 3. Count of usability challenges that were identified by participants.

Themes from UI usability comments	Frequency
The position of UI tells where to stand but is somehow restricted in CAVE.	2
Not sure what to expect from UI.	3
Being able to see participant's own body reduces immersion.	2
Should be able to skip worlds.	3
Interactable components difficult to aim, should be larger.	2
Some of the columns, text, and numbers were blocked by the entity in the current world.	2
Need instructions for how to interact with UI.	4

Table 4. BLA positive common (PC), negative common (NC), positive particular (PP), and negative particular (NP) elements for A1 option (scale armature columns, forest), A2 option (scale armature columns, path), A3 option (scale armature columns, plain), B1 option (scaling interaction, gesture), B2 option (scaling interaction, button clicks), C1 option (scaling animation, staggered), C2 option (scaling animation, simultaneous).

Item ID	Description	Avg. Score	Mention Index (%)
1PC(A1)	Depth and scale cues	7.5	40
2PP(A1)	Immersion	10	20
1PP(A2)	Depth and scale cues	1	20
2PC(A2)	Immersion	3.5	40
3PP(A2)	Distraction	8	20
1NC(A3)	Depth and scale cues	7	80
1NP(A3)	Immersion	10	20
1PP(B1)	Novelty	6	20
2PP(B1)	Natural	8	20
1NC(B1)	Accuracy	3	40
1PC(B2)	Accuracy	5	40
2PP(B2)	Functionality	8	20
1PC(C1)	Clarity	6.7	60
1NC(C1)	Tediousness	5	40
1PC(C2)	Faster	5.25	80
1NP(C2)	Clarity	10	20

DISCUSSION

Five usability experts performed four representative user tasks in Scale Worlds and compared several design alternatives in A/B testing. Three main usability problems were identified using PSSUQ, leading to 3D UI design recommendations for VR experiences in a CAVE. Two problems were associated with the information quality. We identified that error messages and information should be added to the VR system in the future. One problem was associated with interface quality that a navigational map should be added to support the learning of scale in an immersive virtual environment.

As for workloads on the users, the NASA-TLX revealed that participants perceived relatively higher mental demand and frustration. Since Scale Worlds was developed to be a scale learning environment, we expected a higher score in the mental workload subscale. To reduce frustration, we will increase the size of interactable elements in future versions of Scale Worlds.

The present study also aimed to understand and examine theory–usability tradeoffs. Seeing their own body in the CAVE reduced participants' immersion but allowed them to use their own body to compare sizes in the system. This result revealed the conflict and trade-offs between the usability factor of immersion and the educational factor of embodied cognition. Restricting access to the UI helped guide users to where they should stand for optimal feeling of shrinking or growing during scaling, but they felt this was restrictive. A wider range of positions for participants to stand to interact with the UI should be implemented. Results from A/B testing revealed additional theory–usability tradeoffs, including the depth perception and immersion of the forest armature versus the "less distracting" path armature. User selection of this armature in future versions will be implemented. The gesture mode of interaction is inspired by embodied cognition theory; improving the aiming accuracy by increasing the size of the interactable components—as mentioned by two participants—might sufficiently address issues with gesture interactions for future versions. A final trade-off concerns the type of animation, where the staggered animation was designed to support learning by implying cause and effect and by guiding the user's attention to whatever is moving at any given moment, and participants validated this by noting that it allowed them to see more details; yet it was perceived to be "slow" and "tedious." Enhancing scale learning and subjective user satisfaction are in conflict. A potential solution is allowing users to speed up the animation once they have gained familiarity with Scale Worlds and acquired knowledge about size and scale.

Limitations. This study involved usability experts whose perspective might be different from students who are target users of Scale Worlds. Thus, general system usability problems were identified and a subsequent study will be conducted to understand target users' perspectives.

CONCLUSION

The virtual learning environment, Scale Worlds, enabled users to scale themselves up or down by powers of ten by interacting with the 3D UI. It has shown promise for utilizing VR in the enhancement of scale cognition and learning. Usability problems were identified by quantitative data and cross-validated by qualitative data from think aloud, providing insight into usability issues. Along with participant preferences for the alternatives revealed during A/B testing, conflicts between cognition theories and usability aspects were identified and will inform the refinement of the next iteration of Scale Worlds in an immersive CAVE.

ACKNOWLEDGMENTS

<Funding agency removed for blind review.>

REFERENCES

Arroyo, I., Micciollo, M., Casano, J., Ottmar, E., Hulse, T., & Rodrigo, M. M. (2017). Wearable learning: Multiplayer embodied games for math. CHI PLAY 2017 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play, 205–216. https://doi.org/10.1145/3116595.3116637

Common core. (2000). Mathematics Standards | Common Core State Standards Initiative. http://www.corestandards.org/Math/

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V, & Hart, J. C. (1992). The CAVE: audio visual experience automatic virtual environment. Communications of the ACM, 35, 64–72. https://doi.org/10.1145/129888.129892

Dalgarno, B., & Lee, M. J. W. (2010). What are the learning affordances of 3-D virtual environments? British Journal of Educational Technology, 41, 10–32.

Delgado, C. (2009). Development of a research-based learning progression for middle school through undergraduate students' conceptual understanding of size and scale [University of Michigan, Ann Arbor]. https://deepblue.lib.umich.edu/handle/2027.42/64794

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

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behavior Research Methods, Instruments, & Computers, 35(3), 379-383.

Flores, F., Tovar, M. E., & Gallegos, L. (2003). Representation of the cell and its processes in high school students: An integrated view. International Journal of Science Education, 25(2), 269–286. https://doi.org/10.1080/09500690210126793

Fonseca, D., Redondo, E., & Villagrasa, S. (2015). Mixed-methods research: a new approach to evaluating the motivation and satisfaction of university students using advanced visual technologies. Universal Access in the Information Society, 14(3), 311–332. https://doi.org/10.1007/s10209-014-0361-4

Harrison, A. G., & Treagust, D. F. (1996). Secondary students' mental models of atoms and molecules: Implications for teaching chemistry. Science Education, 80(5), 509–534. https://doi.org/10.1002/(SICI)1098-237X(199609)80:5<509::AID-SCE2>3.0.CO;2-F

Hart, Sandra G., and Lowell E. Sta. 1988. "HUMAN MENTAL WORKLOAD P.A. Hancock and N. Meshkati (Editors) Elsevier Science Publishers." The Journal Of San Jose State University 52(Human Mental Workload):381.

Johnson-Glenberg, M. C., Birchfield, D. A., Tolentino, L., & Koziupa, T. (2014). Collaborative embodied learning in mixed reality motion-capture environments: Two science studies. Journal of Educational Psychology, 106(1), 86–104. https://doi.org/10.1037/a0034008

Lakoff, G., & Johnson, M. (1980). Metaphor we live by. In Chicago/London.

Lamon, S. J. (1996). The Development of Unitizing: Its Role in Children's Partitioning Strategies. Journal for Research in Mathematics Education, 27(2), 170–193. https://doi.org/10.2307/749599

Lewis, James R. 2002. "Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies." International Journal of Human-Computer Interaction 14(3–4):463–88. doi:10.1080/10447318.2002.9669130.

Magaña, 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

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

Peterson, M., Delgado, C., Tang, K.-S., Bordas, C., & Norville, K. (2021). A taxonomy of cognitive image functions for science curriculum materials: Identifying and creating 'performative' visual displays. International Journal of Science Education, 43(2), 314–343. https://doi.org/10.1080/09500693.2020.1868609

Skulmowski, A., & Rey, G. D. (2018). Embodied learning: introducing a taxonomy based on bodily engagement and task integration. In Cognitive Research: Principles and Implications (Vol. 3, Issue 1, p. 6). Springer. https://doi.org/10.1186/s41235-018-0092-9

Sweller, John. 2010. "Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load." Educational Psychology Review 22(2):123–38. doi: 10.1007/s10648-010-9128-5.

Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., & Minogue, J. (2006). Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. Journal of Research in Science Teaching, 43(3), 282–319. https://doi.org/10.1002/tea.20123

van der Schaaf, Marieke, Arthur Bakker, and Olle ten Cate. 2019. "When I Say … Embodied Cognition." Medical Education 53(3):219–20. doi: 10.1111/medu.13678.