Check for updates

*Research Article*

# Effects of Virtual Reality and Cueing Elements on Performance, Workload, and Visual Behavior in Scale Cognition

Human Factors
2026, Vol. 0(0) 1–22
© 2026 Human Factors and Ergonomics Society
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00187208261506560
journals.sagepub.com/home/hfs
Sage

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

## Abstract

**Objective:** This study aims to investigate how the use of virtual reality (VR) and different cueing elements within VR affect individuals' scale cognition performance, workload, and visual behavior.

**Background:** Scale cognition refers to the ability to conceptualize spatial, temporal, and quantitative dimensions for measuring objects and processes. Although VR has shown promise in supporting various cognitive tasks, there remains a lack of research examining its effects on scale cognition. Moreover, VR-based learning may have adverse effects on users when learning environments are not carefully designed.

**Method:** The research team developed a VR-based learning environment, Scale Worlds, which integrated three types of cueing elements to support learning: graphic armatures, multimodal cues, and numeric measures. A between-subjects experiment was conducted with 40 engineering students across five conditions: three conditions featuring individual cueing elements, one without any cueing elements, and one incorporating all cueing elements. The pre- and post-test scores of scale cognition assessment, workload, eye-movement data, and semi-structured interview data were collected and analyzed.

**Results:** The VR-based learning environment significantly improved participants' scale cognition, particularly in the subscales of absolute and relative reasoning. Cueing was found to affect distinct dimensions of scale cognition, as well as cognitive engagement, visual attention, and frustration levels during learning.

**Conclusion:** VR-based learning environments can improve scale cognition, and embedded cueing elements play a modulatory role in shaping users' learning experiences.

**Application:** These findings highlight the potential of using VR to support cognitive tasks and carefully designed cues in VR to shape learning processes.

[1]Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, USA

**Corresponding Author:**
Karen B. Chen, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695, USA.
Email: kbchen2@ncsu.edu

**Keywords**
virtual reality, scale cognition, cue, workload, visual behavior

## Introduction

Scale cognition refers to an individual's ability to conceptualize spatial, temporal, quantitative, or analytical dimensions to measure and study objects and processes (Delgado, 2009), which is integral to science and engineering, as highlighted by the Next Generation Science Standards (National Research Council, 2012) and Common Core mathematics standards (Common Core, 2010). It has been conceptualized as a set of cognitive processes, including grouping or ordering objects by size, reasoning about relative size differences, and estimating absolute size values (Harper-Gampp et al., 2025; Magana et al., 2012). Scale cognition enables individuals to understand and reason about objects and phenomena across vastly different magnitudes. It plays a critical role in science, technology, engineering, and mathematics (STEM) education, as STEM disciplines require individuals to reason across multiple scales to address complex problems (Magana, 2014). However, research has shown that students often develop "compressed" mental representations of scale (Delgado, 2013; Longo & Lourenco, 2007; Magana et al., 2012), unable to differentiate objects such as molecules and atoms at the microscopic scale, or planets and stars at the macroscopic scale (Swarat et al., 2011).

Virtual reality (VR) offers immersive, three-dimensional (3D) experiences distinct from traditional settings. These affordances support cognitive tasks involving abstract concepts by making otherwise intangible phenomena more concrete and experiential, as demonstrated in studies on 3D molecular models (Pietikäinen et al., 2021) and geometric forms (Medina Herrera et al., 2024). These findings suggest that VR may also hold promise for scale cognition. Despite these potential benefits, the effectiveness of VR is not inherent to the technology itself but depends on the design and presentation of virtual content. Prior research has demonstrated that VR implementation does not always yield positive outcomes and may even negatively affect users (Makransky et al., 2019; Matovu et al., 2023; Parong & Mayer, 2018), potentially due to distractions and increased cognitive workload caused by the design and presentation of virtual content (Albus et al., 2021; Frederiksen et al., 2020). This highlights a research opportunity in VR-based learning environment design that not only supports scale cognition but also accounts for other critical human factors during learning, such as workload, attention, and user experience.

Prior research supports the effectiveness of *cueing* in learning (Schneider et al., 2018). Cueing refers to instructional signals designed to help learners understand the elements or organization of materials (Albus et al., 2021; Mautone & Mayer, 2001; Schneider et al., 2018). Cueing elements can take various forms, including color (e.g., highlighting), auxiliary graphics (e.g., arrows), and textual annotations. The use of cueing has been shown to direct learner attention, aiding in selecting, integrating, and processing essential information in traditional learning contexts (Li et al., 2023; Vogt et al., 2021). From a human factors perspective, cues have also been used in VR to enhance focus (Liu et al., 2022), recall, and attention management (Albus et al., 2021), suggesting they may help modulate and shape behaviors in immersive learning. However, research comparing different cue types on cognitive tasks and their effects on workload and visual behavior in VR remains limited. Furthermore, given the contextual nature of cueing, the same cue type may vary in form and function across content, and its effectiveness should therefore be examined on a case-by-case basis (Radianti et al., 2020).

The research team developed Scale Worlds (Figure 1), a virtual learning environment incorporating scientific entities across scales to support conceptualization of scale, size, and numbers, and studied its implementation in middle school classrooms (Chen et al., 2024). Designed in alignment with Magana et al.'s (2012) scale cognition framework, Scale Worlds scaffolded scale learning through two approaches. First, embodied cognition theory was implemented through two mechanisms: (1) users embodying differently sized entities from a first-person viewpoint and (2) embodied interactions linking bodily actions to scale reasoning, with "upward" controller movements triggering transitions to "larger scale" worlds. Experiencing differently sized entities from a first-person viewpoint has been shown to recalibrate perceived scale, potentially by rescaling multisensory body representations that modulate visual processing and space perception (van der Hoort et al., 2011). Performing bodily actions that align with conceptual meanings has been shown to facilitate comprehension of abstract concepts (Cook et al., 2008; Skulmowski & Rey, 2018), potentially by grounding cognitive processing in sensorimotor simulations. Second, a series of *constant elements* was integrated: *stacking entities* enabling size comparison by counting how many smaller entities align with a larger one; a *metric key* providing a visual reference for unitized measurement; and a *progress navigation panel* indicating the user's position within the scale continuum from microscopic to astronomic level.

To further enhance scale cognition, the team designed three types of cueing elements based on a structural framework of visual displays (Peterson et al., 2021): (1) *graphic armatures* to assist users in visually measuring and comparing entities; (2) *multimodal cues* to provide multi-channel information that supports intuitive scale navigation; and (3) *numeric measures* to provide explicit numbers and units that situate entities within a base-ten

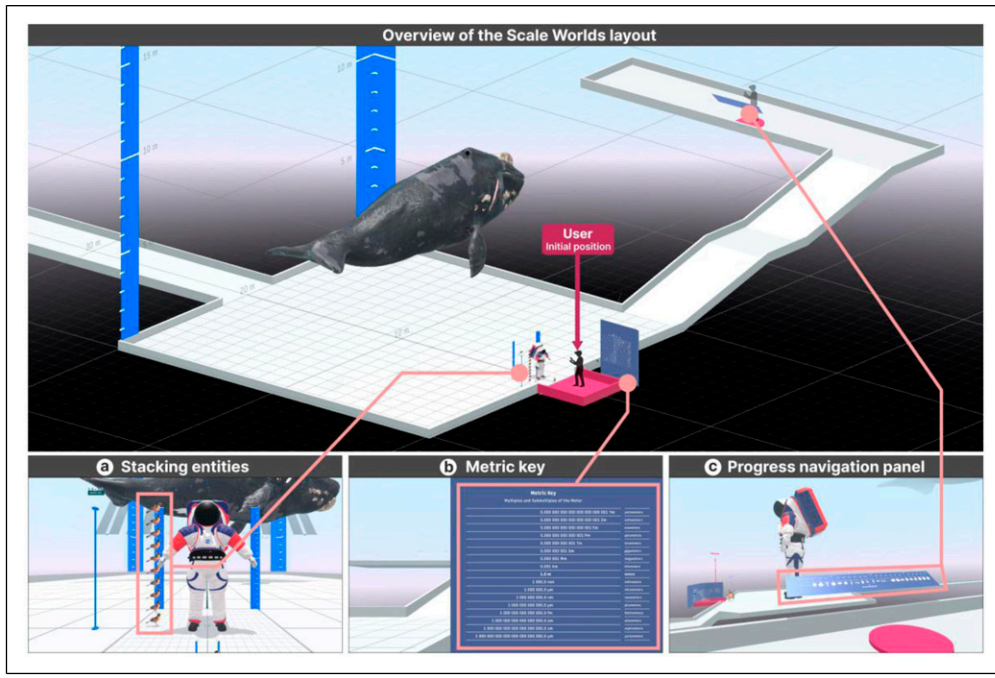


**Figure 1.** The top panel shows a holistic view of Scale Worlds' layout and the placement of constant elements, including (A) stacking entities, (B) metric key, and (C) progress navigation panel

number system. It is important to note that these cueing elements function as supportive design features that operate in conjunction with the primary learning materials, with Scale Worlds and its constant features providing the foundational learning context for these cueing elements.

The present study was guided by two research questions (RQs): RQ1 focused on the overall learning effects of Scale Worlds. Our prior work aimed to first ensure the usability of Scale Worlds (Wu et al., 2022), and thus its effects on scale cognition outcomes have not yet been empirically examined. Therefore, we asked: What is the effect of Scale Worlds on scale cognition performance? RQ2 focused on the effects of cueing on both learning outcomes and processes. Specifically, it asked: What are the effects of three designed cueing elements embedded within Scale Worlds on scale cognition performance, workload, and visual behavior?

The following hypotheses were developed for the RQs: Scale Worlds would improve learners' overall scale cognition (H1.a) and performance on its subdimensions (i.e., ordering, grouping, relative reasoning, and absolute reasoning) (H1.b); cueing elements were expected to improve overall scale cognition (H2.a) and performance on its subdimensions (H2.b); cueing elements were expected to reduce perceived overall workload (H3.a), as well as perceived workload across subscales (H3.b); cueing elements were expected to increase visual attention allocated to certain constant elements in Scale Worlds (H4); and cueing elements were expected to reduce cognitive demand during the learning process (H5).

## Methods

### Participants

A power analysis performed using G*Power (Faul et al., 2007) indicated a required sample size of 34 ($\alpha = .05$, power = .80) based on a moderate effect size from a preliminary study (Cohen's d = 0.50; Cheng et al., 2024). Therefore, 40 participants were recruited for the study for an equal number of participants per condition. Inclusion criteria specified first-year engineering undergraduates with no prior exposure to advanced discipline-specific instruction, no susceptibility to VR discomfort, and no prior experience with Scale Worlds. This research complied with the APA Code of Ethics and was approved by the Institutional Review Board of North Carolina State University. Informed consent was obtained from each participant.

Participants had a mean age of 18.97 years (SD = 1.49), including 24 males, 15 females, and one non-binary individual. Racial composition included 23 White, 11 Asian, two Black or African American, one American Indian or Alaska Native, and three undisclosed. Ethnic composition included 34 non-Hispanic, five Hispanic or Latino/Latina, and one undisclosed. Most participants had prior VR experience ($n$ = 34, 85.0%), though use was infrequent: 18 (52.9%) used VR once a year, seven (20.6%) once every 6 months, five (14.7%) once every 3 months, three (8.8%) at least once a month, and one (2.9%) at least once a week.

### Equipment and Virtual Environment

Scale Worlds was created and rendered utilizing the Unity Engine (2018.4.28f1, https://unity3d.com/), and presented to the participants using a head-mounted display (HMD; Vive, HTC, and Valve Corporation; Figure 2) with a handheld controller. The controller supported teleportation and scaling up or down interaction in Scale Worlds. An eye-tracking device (Pupil Labs GmbH, https://pupil-labs.com/) was mounted inside the HMD to capture eye movements. The eye-tracking data were recorded for subsequent analysis in Pupil Player (Pupil Labs GmbH, https://pupil-labs.com/), a software that provides multiple functions for eye-tracking data preprocessing and analysis. Using the corresponding Unity plugin, participants' gaze data could be temporally synchronized

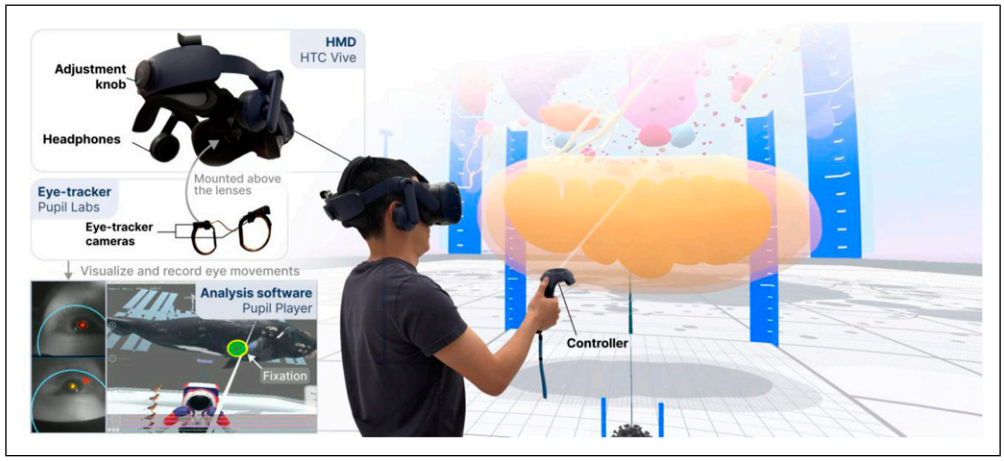


**Figure 2.** Overview of participant and equipment setup of the study, including the HMD, eye-tracking setup, and the depiction of a participant holding a controller interacting within Scale Worlds

and mapped onto the virtual environment for data analysis.

Scale Worlds encompassed 24 scientific entities arranged along an axis receding from the participants' starting position, each representing a "world" with scales varying in *tenfold increments* from $10^{-12}$ to $10^{12}$. The embedded cueing elements in Scale Worlds are illustrated in Figure 3: *Graphic armatures* are non-semantic elements, such as lines and arrows, that contribute to meaning by relating to other visual elements (Peterson, 2022; Peterson et al., 2021). In Scale Worlds, graphic armatures included grid lines on the ground and hash marks on vertical rulers, indicating a length in sub-multiples of the current scale. *Multimodal cues* provide multi-channel sensory information that supports navigation of scale. Five background color schemes provided visual cues across scale ranges: grayscale ($10^{-12}$–$10^{-7}$), pink ($10^{-6}$–$10^{-4}$), daylight ($10^{-3}$–$10^{3}$), dark blue ($10^{4}$–$10^{6}$), and black ($10^{7}$–$10^{12}$). Auditory cues were delivered through transition sound effects whose pitch was adjusted by a predefined scaling parameter $\Delta$: multiplied by $1 - \Delta$ when scaling up (lower frequency, longer duration), and by $1 + \Delta$ when scaling down (higher frequency, shorter duration). *Numeric measures* provide explicit numbers and units that help situate entities within a base-ten number system. In Scale Worlds, they included a numeric panel on the sides of entities, indicating the current world's scale in meters, such as "$1.6 \times 10^{-1}$ m," with additional numeric marking for distances or heights in units pertinent to the current world on the ground and behind entities, such as "100 mm."

## Procedure

The overall experimental procedure is shown in Figure 4. Upon providing informed consent, participants completed a demographic questionnaire, followed by a pre-test using the validated Assessment of Size and Scale Cognition (ASSC; Harper-Gampp et al., 2025) and a grouping activity. After researchers demonstrated how to wear the HMD and interact with the virtual environment, participants entered Scale Worlds and completed eye-tracking calibration. Then, they were randomly assigned to one of the experimental conditions. Participants first familiarized themselves with Scale Worlds for 2 minutes and then completed

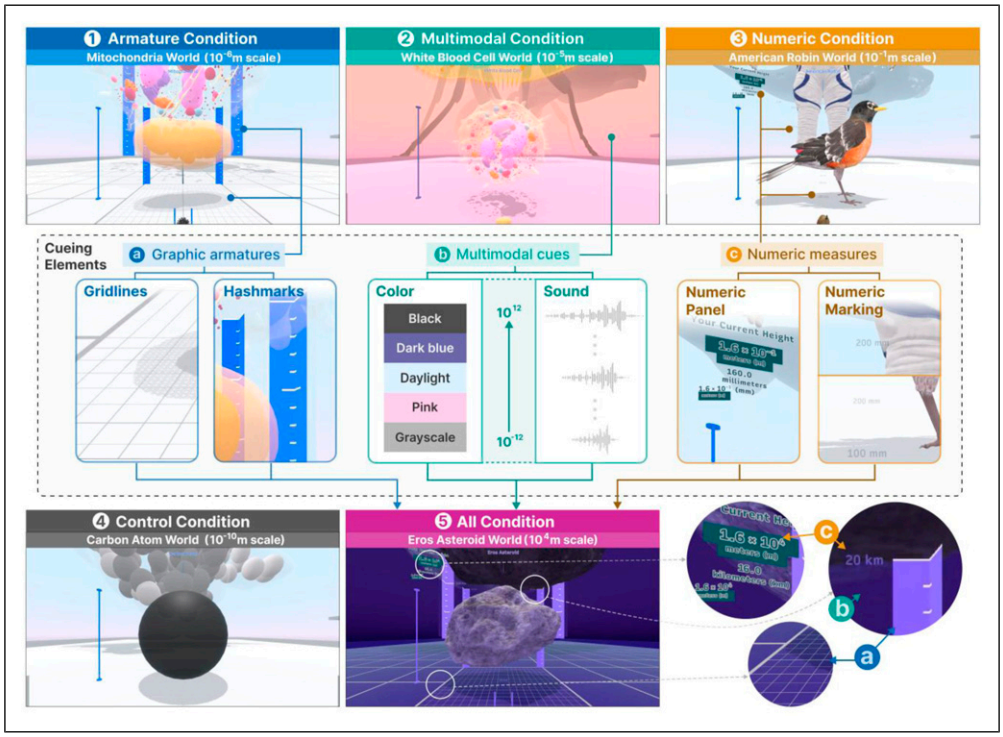


**Figure 3.** Screenshots of the Scale Worlds across five conditions: (1) Armature, (2) Multimodal, (3) Numeric, (4) Control, and (5) All, using three types of cueing elements: (A) Graphic armatures, (B) Multimodal cues, and (C) Numeric measures. Five examples of entities are used to illustrate the use of cueing elements across magnitudes

specific scale comparison tasks as instructed, such as selecting an entity and finding another that was 1,000 times larger (see Appendix A for the full list of prompts). During this process, they could freely teleport within Scale Worlds. Afterward, participants completed the NASA-TLX to report their perceived workload, followed by a post-test consisting of the ASSC and the grouping activity. Finally, they participated in a semi-structured interview to share feedback (Appendix B). Participants were compensated $15 per hour for 90 minutes of participation.

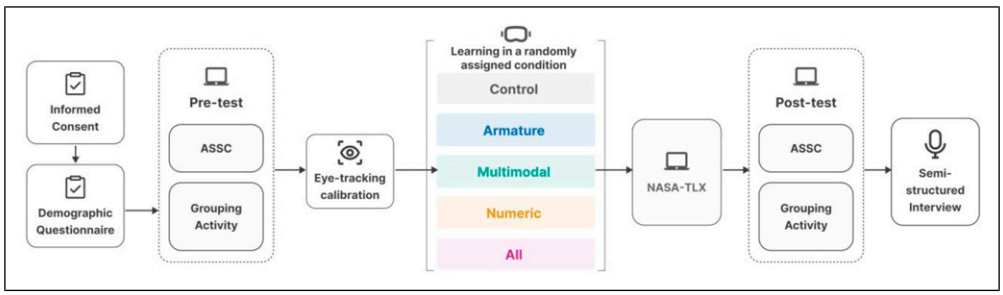


**Figure 4.** Overview of the experimental procedure

## Variables and Analysis

A between-subjects design with five conditions was used in this study, including four experimental and one control (Figure 3). The experimental conditions were graphic armatures only ("Armature"), multimodal cues only ("Multimodal"), numeric measures only ("Numeric"), and the condition with all three elements combined ("All"). The control condition ("Control") was devoid of all three virtual elements. The dependent measures included scale cognition performance, workload, and eye movements. All hypotheses were evaluated using null-hypothesis significance testing ($\alpha = .05$) for frequentist models and Bayes factors ($BF_{10}$) for Bayesian models. Additionally, semi-structured interviews were conducted to collect qualitative feedback.

*Scale Cognition Performance.* Scale cognition performance was assessed based on participants' responses to the ASSC (Harper-Gampp et al., 2025) and a grouping activity. To reduce reliance on recall and ensure that responses reflected scale reasoning, the entities used in both tests were different from those presented in Scale Worlds. The ASSC includes three subscales, each targeting a fundamental dimension of scale cognition (Magana et al., 2012): ordering (scored 0–20), absolute reasoning (0–40), and relative reasoning (0–40). A total score was computed by summing across all subscales (Harper-Gampp et al., 2025). Normality assumptions were verified before statistical analysis using the Shapiro–Wilk test. A paired-sample t-test was conducted to compare pre- and post-test ASSC total scores across all participants. An ANCOVA was used to examine the effect of conditions, controlling for the pre-test score as a covariate.

For the ASSC subscales, a Bayesian multivariate multilevel model was used because it allowed multiple dependent variables to be examined jointly while accounting for repeated measures and individual differences, making it more suitable than null-hypothesis significance testing (NHST) methods (e.g., ANOVA). The model included fixed effects of assessment time point (pre- and post-test; hereafter referred to as "Time"), condition, and their interaction, with participant-level random effects accounting for repeated observations. The model was fitted using Markov Chain Monte Carlo sampling (chains = 4, iter = 4,000, warmup = 1,000, family = "gaussian") implemented in the *brms* package (version 2.22.0) in R, with weakly informative priors applied (Bürkner, 2017). Model diagnostics indicated sufficient convergence ($\widehat{R} = 1.00$), and effective sample sizes exceeded 3,000, suggesting stable posterior inference. $BF_{10}$ was reported as the strength of evidence favoring the alternative model over the null model. Model evidence was quantified using posterior estimates ($\beta$), 95% credible intervals (*CrI*), and the posterior probability of direction (*pd*; a *pd* of 0.99 reflects 99% certainty in the effect's direction). Cohen's *d*-type effect sizes ($d_\Delta$) were computed from the posterior draws by standardizing the estimated pre-post change scores using the pooled standard deviation (Dankel & Loenneke, 2021; Kruschke, 2013). The posterior mean and 95% credible interval of $d_\Delta$ were reported.

For the grouping activity, participants categorized 20 scientific entities into self-defined groups. Since multiple valid groupings exist, accuracy was not scored; instead, pass/fail was determined by whether the participant's groups could be formed by partitioning their ordered list at any cut point. For example, given the ordered list [Saturn, Mercury, school bus, human], splitting after Mercury yields {Saturn, Mercury} and {school bus, human} (pass), whereas {Saturn, school bus} and {Mercury, human} cannot be formed by any such partition (fail). Responses were independently coded by the first author and coauthor T. H.-G., a science education expert, yielding perfect inter-rater reliability (Cohen's $\kappa = 1.00$; Landis & Koch, 1977). Each participant's outcome was then classified into one of four pre-post categories (pass-pass, pass-fail, fail-pass, or fail-fail), and a Chi-square test examined whether these patterns differed across conditions.

*Eye Movements.* Eye-tracking data were collected to provide insights into learning processes in addition to performance measures (RQ2). Specifically, fixations within areas of interest (AOIs) were used to assess visual attention distribution across virtual elements, and blinks served as an indicator of cognitive demand (Biondi et al., 2023; Hou et al., 2025). Data were preprocessed in Pupil Player and filtered using established criteria: fixation duration 100–200 ms, maximum dispersion 1° (Salvucci & Goldberg, 2000), blink duration 100–400 ms (Hollander & Huette, 2022), and pupil diameter 2–8 mm (Watson & Yellott, 2012). Data outside the specified ranges, as well as missing and incomplete data, were removed. Two researchers independently coded fixations on five AOI types in Scale Worlds, including three constant features (stacking entities, metric key, and progress navigation panel), scientific entities (e.g., white blood cell), and other elements (e.g., ground). These AOIs were present across all conditions to support fair comparison. This process yielded an inter-rater reliability of Cohen's $\kappa = .93$, indicating almost perfect agreement. An ANOVA was then conducted to examine the effects of conditions on fixation patterns and blinks.

*Workload.* Workload was assessed using the NASA Task Load Index (NASA-TLX), which includes six subscales (i.e., mental demand, physical demand, temporal demand, performance, effort, and frustration), each rated on a 0–100 scale (Hart & Staveland, 1988). Pairwise comparisons were then administered to weight each subscale's relative contribution, and an overall weighted composite score was calculated following Hart and Staveland's (1988) procedure. An ANOVA examined condition effects on total workload and subscale scores.

*Subjective Feedback.* Audio recordings of interviews were transcribed and deductively coded according to two pre-established analytic questions: Q1. How did participants find the features of Scale Worlds useful for their learning process? Q2. How did participants perceive changes in their scale cognition after using Scale Worlds? Two researchers independently coded responses into themes using NVivo (Lumivero, 2025) with an inter-rater agreement score of Cohen's $\kappa = .75$, indicating substantial agreement (Landis & Koch, 1977).

## Results

### Scale Cognition Performance

A paired t-test revealed a significant increase in ASSC total scores from pre-test ($M = 50.09$, $SD = 7.93$) to post-test ($M = 56.85$, $SD = 12.27$), $t\ (33) = 4.87$, $p < .0001$, Cohen's $d = 0.84$, indicating that Scale Worlds effectively improved overall scale cognition with a large effect size (Figure 5(A)). An ANCOVA was conducted to compare ASSC post-test total scores across conditions while adjusting for pre-test baseline scores (Figure 5(B)). It showed no significant effect of conditions ($F\ (4, 24) = 0.63$, $p = .64$, $\eta_p^2 = .095$), but a significant association between pre- and post-test scores ($F\ (1, 24) = 30.53$, $p < .001$, $\eta_p^2 = .56$).

For ASSC subscales, the Bayesian model comparison favored the full model including the Condition × Time interaction ($BF_{10} = 3.73 \times 10^{11}$), as well as the model including Condition ($BF_{10} = 3.35 \times 10^{8}$) and Time ($BF_{10} = 1.99 \times 10^{4}$) over their respective reduced models. Posterior estimates indicated that participants in the Multimodal condition showed greater improvement in ordering than those in the Control condition ($\beta = 4.14$, *CrI* [0.08, 8.22], $pd = 0.999$; $d_\Delta = 1.11$, *CrI* [0.02, 2.21]; Figure 6). No significant differences were found in pre-to-post changes in absolute or relative reasoning across conditions. Collapsing across conditions, the model supported pre-post increases in both relative reasoning ($\beta = 3.27$, *CrI* [1.04, 5.50], $pd = 0.998$; $d_\Delta = 0.90$, *CrI* [0.04, 1.75]) and absolute reasoning ($\beta = 3.0$, *CrI* [0.90, 5.09], $pd = 0.997$; $d_\Delta = 0.40$, *CrI* [−0.53, 1.28]). No significant difference in grouping activity performance patterns was found across conditions ($\chi^2 = 13.23$, $p = .35$; Figure 7).

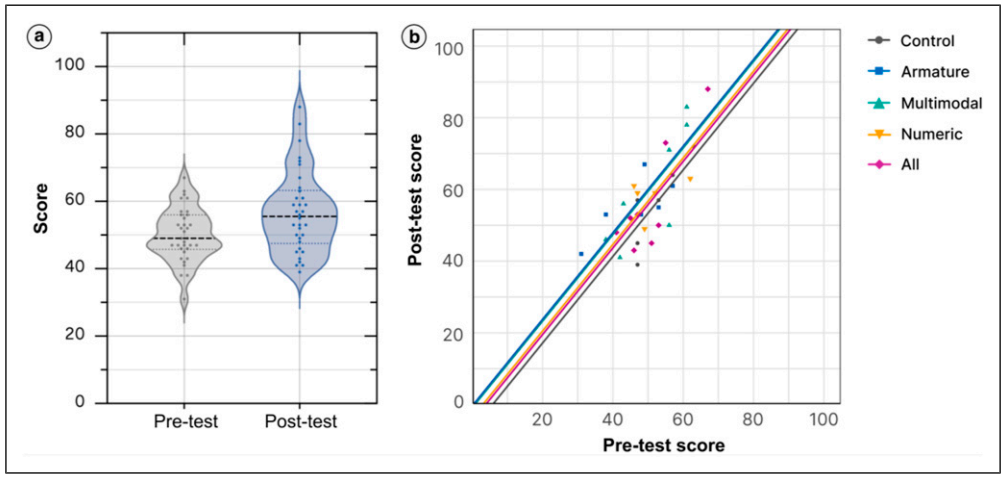


**Figure 5.** (A) Violin plot for ASSC pre- and post-test total scores, where wider regions indicate where scores are more concentrated; (B) Scatter plot showing fitted regression lines for pre- and post-test total scores across five conditions

## *Workload*

There was no significant difference in total workload across conditions (Table 1), but frustration differed significantly across conditions, $F(4, 35) = 2.90, p < .05, \eta_p^2 = .25$, with participants in the All condition reporting lower frustration than those in the Control condition.

## *Eye Movements*

The ANOVA revealed a significant effect of condition on blink count ($F(4, 20) = 3.97$,

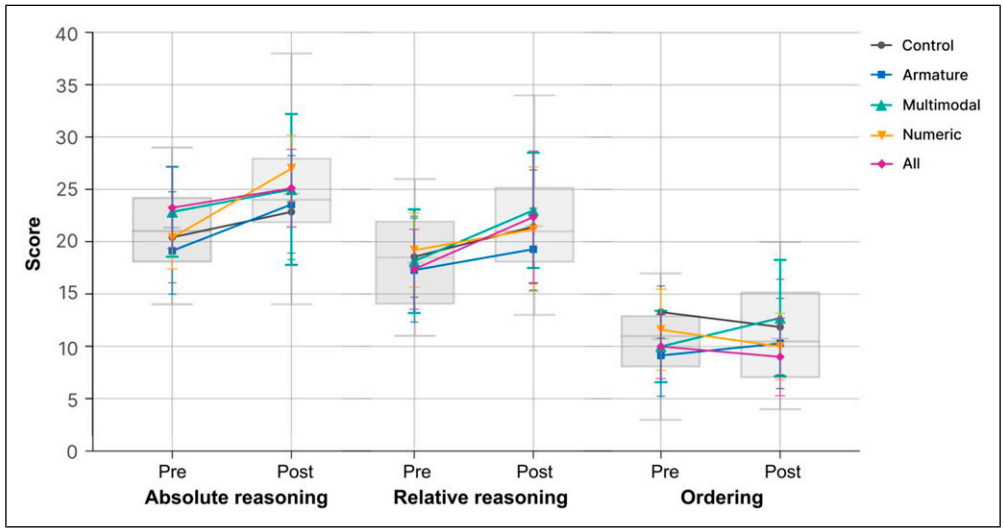


**Figure 6.** Pre- and post-test scores for ASSC subscales. Lines connect pre- and post-test means within each condition, with error bars representing standard deviations. The shaded gray boxes represent the interquartile range (25$^{th}$ to 75$^{th}$ percentile), and whiskers extend to the minimum and maximum of the scores recorded

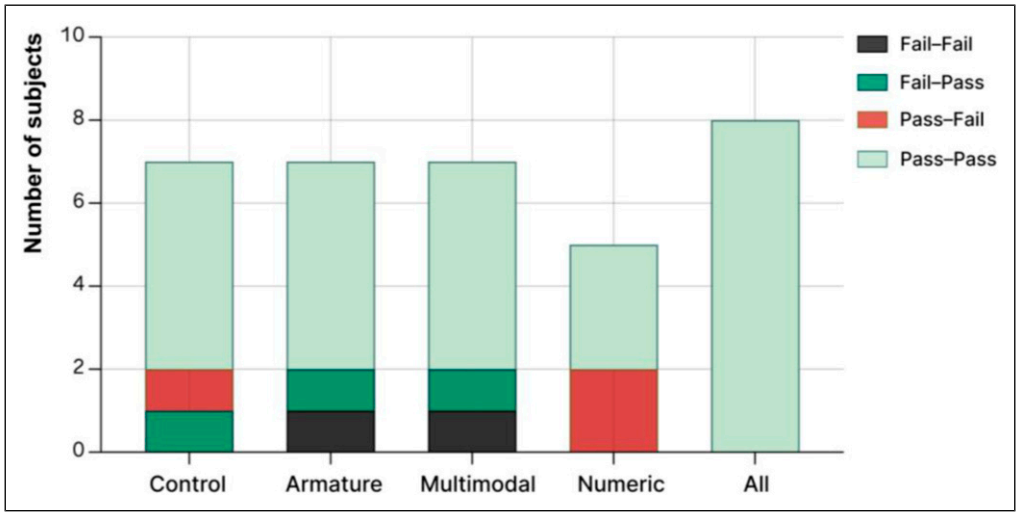


**Figure 7.** Number of participants in each condition by performance pattern on the grouping activity, rated in terms of pass or fail. For example, "Fail–Pass" indicates participants who failed the grouping activity in the pre-test but passed it in the post-test

$p < .05$, $\eta_p^2 = .43$), with the Armature condition ($M = 449.25$, $SD = 399.30$) showing significantly more blinks than the Control ($M = 121.43$, $SD = 81.06$), Multimodal ($M = 73.0$, $SD = 69.19$), Numeric ($M = 105.50$, $SD = 69.30$), and All conditions ($M = 45.50$, $SD = 27.31$), all $p$ s $< .05$. Condition also significantly affected fixations on the progress navigation panel ($F(4, 20) = 3.60$, $p < .05$, $\eta_p^2 = .43$; Table 2), with Tukey post-hoc comparisons indicating significantly more fixations in the Numeric than the Control condition ($F(1, 20) = 10.67$, $p < .01$).

## Subjective Feedback

Three themes were identified regarding Q1 (useful features for learning; Table 3). First, the *affordances of VR*, including immersion, embodiment, and three-dimensional representation,

**Table 1.** Descriptive statistics and ANOVA results for NASA-TLX scores across experimental conditions

| Measures (NASA-TLX) | Mean (SD) | | | | | F-value | *p*-value |
|---|---|---|---|---|---|---|---|
| | Control | Armature | Multimodal | Numeric | All | | |
| Total | 55.83 (13.64) | 50.17 (12.47) | 42.92 (22.15) | 45.79 (14.01) | 37.71 (15.65) | 1.50 | 0.22 |
| Mental demand | 66.25 (13.82) | 54.38 (21.62) | 49.38 (29.21) | 53.12 (26.45) | 42.5 (21.88) | 1.18 | 0.36 |
| Physical demand | 28.12 (11.0) | 31.25 (27.35) | 26.88 (15.34) | 29.38 (23.52) | 18.75 (7.44) | 0.54 | 0.71 |
| Temporal demand | 41.88 (21.2) | 40.62 (16.57) | 33.75 (23.57) | 38.12 (20.86) | 26.25 (21.34) | 0.74 | 0.57 |
| Performance | 56.88 (20.69) | 45.0 (21.55) | 34.38 (29.69) | 34.38 (24.41) | 31.88 (26.98) | 1.41 | 0.25 |
| Effort | 63.12 (23.14) | 60.62 (16.57) | 52.5 (27.65) | 45.62 (21.95) | 41.88 (19.07) | 1.40 | 0.25 |
| Frustration | 35.62 (17.0) | 29.38 (11.48) | 18.75 (14.08) | 26.88 (30.7) | 15.0 (19.46) | 2.90 | **0.036*** |

*Note.* *$p < .05$.

**Table 2.** Descriptive statistics and ANOVA results of fixation counts on AOIs across conditions

| AOIs | Mean (SD) | | | | | F-value | *p*-value |
|---|---|---|---|---|---|---|---|
| | Control | Armature | Multimodal | Numeric | All | | |
| Stacking entities | 9.43 (7.72) | 13.67 (13.5) | 7.75 (7.09) | 6.67 (9.03) | 2.25 (2.63) | 0.92 | 0.47 |
| Metric key | 2.57 (3.55) | 29.67 (48.81) | 3.5 (6.35) | 0.83 (1.17) | 0.5 (1.0) | 1.97 | 0.14 |
| Entities | 40.14 (37.2) | 51.33 (38.11) | 65.0 (53.54) | 50.5 (57.51) | 18.5 (20.89) | 0.61 | 0.65 |
| Progress navigation panel | 0.43 (0.79) | 0.0 (0.0) | 1.0 (1.41) | 5.5 (5.13) | 0.75 (1.5) | 3.60 | **0.02*** |
| Others | 28.57 (33.4) | 31.0 (36.5) | 41.5 (26.71) | 28.5 (28.49) | 18.75 (23.54) | 0.29 | 0.88 |

*Note.* *$p < .05$.

were most frequently mentioned by participants. These affordances were perceived to support immediate perception of size and scale through immersive visual, spatial, and embodied experiences, providing an experiential basis for higher-level reasoning central to scale cognition. Second, participants identified *stacking entities, tenfold increments, and numeric measures* as useful design components in Scale Worlds. These features share the characteristic of providing countable units: stacking entities allow visual counting of how many smaller objects align with a larger one; tenfold increments offer a consistent powers-of-ten scaling pattern; and numeric measures provide absolute size in standardized units (e.g., meters), enabling participants to quantify and anchor entities across scales. Third, *sequential narration* structured the scalar progression by guiding participants step by step across scales, helping them perceive gradual transitions and understand the system's scaling mechanism.

Four themes were identified regarding Q2 (perceived changes in scale cognition; Table 4). First, participants reported an expanded "comfort zone" for recognizing entity scale, indicating *greater familiarity with extreme magnitudes*. Second, they perceived improved ability to *compare object sizes* relative to themselves and to one another. They reported being better able to use "anchors," such as the size of the Earth or an ant, to determine relative scale. Third, they were better able to *apply scale knowledge to novel situations*. They reported being able to mentally place unfamiliar entities along the ordering structure developed during the VR experience. Lastly, participants became aware of how "way off" their previous understanding of certain scales had been, noting that Scale Worlds offered an intuitive first-person perspective for observing entities of different sizes, thereby *correcting their previous misunderstandings*.

## Discussion

### *VR-Based Learning Environment Enhances Scale Cognition*

Scale Worlds effectively enhanced users' overall scale cognition, supporting H1.a. The improvement in post-test scores aligned with participants' self-reported experiences, indicating that Scale Worlds helped correct prior misconceptions (Theme 7; hereafter T7) and facilitated application of scale understanding to newly encountered entities (T6). This improvement can be attributed to the following four key factors.

First, VR's *immersive and three-dimensional representations* enabled participants to "put things into perspective," thereby correcting inaccurate mental models of familiar objects. As Participant 32 ("P32") remarked, *"If I hadn't seen it myself, I wouldn't have imagined the ISS to be so much smaller than I previously thought!"* Second, Scale Worlds' narrative structure, which scaled along a linear continuum in tenfold steps, helped resolve ambiguity

**Table 3.** Thematic coding results of interview responses for Q1: How did participants find the features of Scale Worlds useful for their learning process?

| Themes (T) | Codes | Representative quotes | Freq.* |
|---|---|---|---|
| T1. Anchoring scale perception through VR affordances | **Three-dimensional representation** | *"You can see it in a fluid environment, where you're surrounded by other things, can view it from multiple angles, and get a sense of its scale in relation to yourself."* (P3*) | 34 |
| | | *"Slides and videos are 2D, whereas here you can actually move around the objects and see them as they are."* (P27) | |
| | **Embodiment** | *"Being able to experience something helps you learn it better than just trying to understand it from text."* (P10) | 7 |
| | | *"Viewing things from a first-person perspective allows you to actually experience the scale."* (P28) | |
| | **Immersion** | *"It was very immersive, and I think it gave me a better sense of size when it came to comparing the things."* (P32) | 5 |
| | | *"It was definitely really immersive. That's the thing about VR. you're just in it, like it becomes your whole world. So, you aren't really like second guessing what you see around you."* (P2) | |
| T2. Grounding scale understanding through countable and unit-based representations | **Tenfold increments** | *"I'd say keeping the growth with the factors of 10 is definitely a good idea."* (P19) | 24 |
| | | *"Knowing that, like when I scaled up and scaled down, I was going up by 10, or down by 10. That also kind of helped put things into perspective."* (P2) | |
| | **Stacking entities** | *"Seeing things in relation to each other, especially with the stacking function, helps makes up and puts it in perspective a lot better."* (P19) | 21 |
| | | *"One thing I did look at is when it would line up the item, like 10 of them in a row, and then you could kind of see how big it was compared to the little items."* (P7) | |
| | **Numeric measures** | *"I used, mostly use the height feature where I read the height of what I am right now to compare with what the object is."* (P17) | 16 |
| | | *"The display on the left that showed you the current reading and the power of 10 based on meters[…] it was helpful in gauging the scale of everything."* (P16) | |

*(continued)*

**Table 3.** (continued)

| Themes (T) | Codes | Representative quotes | Freq.* |
|---|---|---|---|
| T3. Structuring scalar progression through sequential narration | **Sequential narration** | *"The textbooks would struggle to convey the sense of scale—like when you start really small and then move gradually up, and you get to see the transition. Textbooks and slides just can't really do that."* (P38)<br>*"I liked how it showed the like the two levels larger it would show them in the background, like appearing like further away, so you could really see the scale compared to how the size you were."* (P12) | 10 |

*Note.* *Freq. = frequency of codes within each theme. "P#" = Participant ID (e.g., P3 = Participant 3).

associated with numerical representations of unfamiliar scales. As P22 described, although they knew the number of the Sun's diameter, it was only by placing that number within a broader scale-ordering context and going through the narrative by themselves that they came to understand its true significance relative to human scales. By reducing misunderstandings and ambiguity, participants developed a more accurate internal reference frame for scale.

Third, entities in Scale Worlds may have served as representative category members ("prototypes"; Bowman et al., 2020; Rosch, 1975) to help infer the scale range of similar entities in the post-test, through two mechanisms: activating pre-existing prototypes for familiar categories (e.g., using the bird in Scale Worlds as a scale reference for a squirrel in the post-test) and establishing new prototypes for less familiar scale ranges (e.g., matching the Brooklyn Bridge in the post-test to the Golden Gate Bridge in Scale Worlds). Lastly, the *embodied* nature of Scale Worlds may have supported knowledge retrieval during the post-test. As P23 noted, they were able to mentally revisit the Scale Worlds scene and simulate the scale up and down activity to retrieve a mental image of scale. This aligns with embodied learning theories (Macedonia, 2019; Nathan, 2021; Skulmowski & Rey, 2018), which emphasize that first-person, bodily engaged experiences facilitate mental simulations that support knowledge retrieval. By reinforcing prototypes and fostering embodied cognition, Scale Worlds supported the recall and application of improved scale understanding to novel objects.

The Bayesian model provided evidence for pre-to-post increases in both relative and absolute scale reasoning, partially supporting H1.b. Relative reasoning involves proportional comparisons, such as determining how many times larger one entity is than another or reasoning with analogies like "A is to B as C is to D" (Delgado & Peterson, 2018; Magana et al., 2012). Scale Worlds supported this by enabling intuitive comparisons between users (as humans) and virtual entities, as well as across entities of different scales (T5). First, participants could *observe multiple entities simultaneously* in each "world." For example, in Right Whale World, participants could see a human, a robin, the ISS, and the Eros asteroid from the whale's perspective. This enabled direct size comparisons across a shared spatial context. Second, each scale up or down operation represented a consistent tenfold factor, allowing participants to translate visual comparisons into a structured framework by *counting scale steps* (T2, T3). Third, participants may have used familiar entities as mental referents ("anchors") for relative scale estimation. Entities such as the human, ant,

**Table 4.** Thematic coding results of interview responses for Q2: How did participants perceive changes in their scale cognition after using Scale Worlds?

| Themes (T) | Codes | Representative quotes | Freq.* |
|---|---|---|---|
| T4. Expanding familiarity with extreme magnitudes | - | *"I learned more from the micro-world; I think it's a lot easier to conceptualize big things. Whereas smaller things you don't really have anything to reference against."* (P3*) | 30 |
| | | *"With the macro world, I kind of got how big things are. But the micro world just kept getting smaller- and once you hit picometers, you think it can't go further. Then you see carbon atoms, and they're tiny compared to everything else."* (P6) | |
| T5. Promoting relative reasoning | **Compare among entities** | *"I can see how the COVID-19 virus compared to the cell, and then the antibody compared to COVID- that was really interesting."* (P5) | 23 |
| | | *"Especially with some of the smaller things that I wasn't quite sure how they compared in size, such as like the virus and the cells. I think that it did improve how I could see the difference in scale between those."* | |
| | **Compare to oneself** | *"It's about being able to see things in relation to my own size."* (P4) | 11 |
| | | *"I guess it's about changing your size and seeing how a human compare to different kinds of things that exist - like, for example, how much larger we are than certain biological components."* (P3) | |
| T6. Enhancing knowledge application | - | *"I mean, when I was taking the post-test, I was trying to recall like, how many times did I scale up or down?"* (P2) | 16 |
| | | *"I definitely when I was taking the test the second time, I felt more knowledgeable."* (P6) | |
| | | *"When I remember this now I can visualize and actually be there to scale up or down."* (P32) | |
| T7. Correcting misconception | - | *"The whale was smaller in comparison to me than I was expecting it to be. Because it was only like nine times larger. I expected the whale to be a lot larger compared to me."* (P1) | 7 |
| | | *"There's no way there are that many white blood cells in an ant. […] wow, those cells are actually bigger than I thought."* (P34) | |

*Note.* *Freq. = frequency of codes within each theme. "P#" = Participant ID (e.g., P3 = Participant 3).

and Earth served as reference points due to their strongly canonical real-world sizes, enabling reasoning about nearby entities' scale. For instance, P2 reported using Earth as a reference and recalled how many steps away another entity was in terms of scale.

Absolute reasoning, on the other hand, refers to the ability to provide specific numerical representations of an entity's size by encoding perceived scale into abstract numerical systems. The *countable units* in Scale Worlds, such as stacked entities, tenfold scale increments, and numeric measures, may have facilitated absolute reasoning (T2). These features provided users with concrete, repeating units that supported quantification of scale and helped bridge visual perception with symbolic number systems, thereby fostering more accurate

internalization of abstract magnitudes. This aligns with prior studies in distance estimation (Petzschner et al., 2012) and spatial learning (McCrink & Galamba, 2015), which highlight the role of symbolic cues or quantity in grounding perceptual experiences in cognitive representations.

### *Cueing Affects Scale Cognition Performance, Workload, and Visual Behavior*

Although cueing did not significantly affect overall scale cognition (rejecting H2.a) or workload (rejecting H3.a), it subtly influenced specific subdimensions of scale cognition, as reflected in ASSC subscale results (supporting H2.b), and shaped learning processes as indicated by the frustration subscale (partially supporting H3.b), fixation patterns (supporting H4), as well as blink behavior, albeit in the opposite direction from what was predicted (rejecting H5). These findings suggest that cueing functioned as a modulating factor during the VR-based learning.

The Multimodal condition was associated with better ordering performance in the post-test. Notably, a downward trend in ordering scores was observed in the Control, Numeric, and All conditions, possibly due to retroactive interference (Dewar et al., 2007; Sosic-Vasic et al., 2018): experiencing Scale Worlds may have formed a new scale continuum that reshaped existing ordering structures in memory, disrupting previously accurate conceptions before they were fully consolidated. In contrast, the Multimodal condition's background color scheme provided visual segmentation that helped participants "chunk" entities based on their sizes into buckets, simplifying the task of ordering 24 entities to ordering five buckets and thus reducing memory demands (Chekaf et al., 2016; Kahn et al., 1990; Thalmann et al., 2019). Variations in auditory cue duration further enabled differentiation of scale spans between adjacent entities. Together, these cues reinforced categorical distinctions and interval judgments, facilitating understanding of the scale's sequential structure. The inclusion of additional cues, such as numeric measures, however, may have diminished this effect by drawing attention away from the scale continuum.

Graphic armatures (gridlines and hash marks) did not demonstrate a clear supportive effect on scale cognition and were associated with increased cognitive engagement, as indicated by higher blink counts (Biondi et al., 2023). The effect likely stemmed from the absence of directly relevant information on the armatures, particularly clear scale measurements. This was supported by participant feedback, as six out of eight participants in the Armature condition expressed a desire for numerical labels to be added. However, this increased cognitive engagement was not reflected in self-reported workload ratings, possibly because the cognitive changes elicited during learning were not salient enough to be captured by post-experience measures, as participants were not required to fully rely on armatures to understand scale. Consequently, participants may not have consciously recognized these cognitive shifts in their retrospective evaluations.

Contrary to intuition, the inclusion of numeric measures did not lead to better performance for any of the scale cognition dimensions. One possible explanation is that certain constant features may have prompted mental representations or cognitive activities that yielded effects similar to those produced by explicit numerical cues. For example, by counting stacked entities, users may have developed their own internal numeric measures. However, visually presenting numeric measures may have diverted attention from other environmental elements, as indicated by increased fixations on the progress navigation panel, which displayed a visual overview of all worlds and their relative positions along the scale continuum. Numeric measures may have acted as visual anchors, prompting cross-referencing of the progress panel for spatial orientation and scale tracking. This heightened awareness of structural cues provided guided navigation for referencing and organizing scale-related information,

suggesting that cueing elements can shape attention distribution and influence how other virtual elements are used.

Having all three cueing elements was associated with reduced frustration. The availability of multiple cueing elements could give users a greater sense of agency in navigating the content, reducing feelings of helplessness and frustration (Makransky & Petersen, 2021). Their combined use may contribute to increased cognitive coherence, helping users establish a consistent internal representation of scale (Mayer & Fiorella, 2014). Moreover, richly layered sensory environments can foster positive emotional states that buffer against frustration (Lønne et al., 2023). These findings suggest that integrating multiple cueing types can exert a synergistic effect on affective experience in VR-based learning.

The findings of this study offer practical implications for both STEM education and the design of VR-based learning environments. Specifically, Scale Worlds promoted scale cognition, underscoring its potential as a learning tool across STEM disciplines that involve understanding and reasoning about scale, such as astrophysics and nanotechnology. In practice, Scale Worlds could complement classroom instruction by serving as an interactive activity alongside lectures or laboratory activities (Chen et al., 2024). This study also identified factors that facilitated effective scale learning, including VR affordances and design strategies such as grounding perceptual experiences in countable quantities and structuring content through tenfold sequential narration. It further showed that cueing elements can shape learning and should be carefully designed to support instructional goals and user experience. These insights provide practical considerations for the design of future VR-based learning environments.

### *Limitations*

Several limitations should be acknowledged for this study. First, participants were first-year engineering students from a single university with predominantly prior VR experience (85.0%), limiting generalizability to other age groups or learners from different academic backgrounds, such as K–12 students or adults with less digital familiarity. Additionally, the small sample size limits the generalizability of the findings for a between-subjects design. Second, scale cognition was measured only immediately after the VR session, so it remains unclear whether improvements reflect lasting conceptual gains or are confounded by practice-related and acute visual processing effects; future work should include delayed post-tests to better distinguish from such confounds. Third, while the design allowed separate and combined evaluation of the three cueing elements, it could not determine how particular subsets may support scale cognition. Future work could examine the interplay among cueing elements to identify the most effective configuration. Fourth, while pre-test scores accounted for baseline scale cognition, they may not have captured participants' prior familiarity with specific entities, which may have influenced learning outcomes.

## Conclusion

Scale cognition is a foundational cognitive ability that underpins learning across multiple scientific domains. Leveraging VR as a promising educational medium, this study introduced Scale Worlds, a virtual learning environment designed to support scale cognition that integrates three types of cueing elements to enhance scale cognition. A between-subjects experiment involving 40 participants demonstrated that Scale Worlds significantly enhanced their overall scale cognition performance, particularly in the subscales of absolute and relative reasoning. The embedded cueing elements were found to affect distinct dimensions of scale cognition, as well as cognitive engagement, visual attention, and frustration levels during the learning process. These findings provide promising evidence for the use of VR in supporting abstract concept learning and inform scalable strategies for the design of immersive learning environments.

## Key Points

The VR-based learning environment, Scale Worlds, significantly improved participants' scale cognition, particularly in the subscales of absolute and relative reasoning.

Having only Graphic Armatures was associated with higher blink counts, potentially indicating increased cognitive engagement in the absence of accompanying explanatory information. Having only Multimodal Cues improved participants' scale-ordering performance, potentially by providing visual segmentation and auditory differentiation along the scale continuum. Having only Numeric Measures was associated with an increase in visual attention to structural cues in the virtual environment.

The combined use of all three cueing elements reduced participants' frustration levels.


### ORCID iDs

Fangyuan Cheng https://orcid.org/0009-0001-9306-8471

Karen B. Chen https://orcid.org/0000-0003-2904-1394



### Funding

The authors 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).


### Declaration of Conflicting Interests

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

### References


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

Biondi, F. N., Saberi, B., Graf, F., Cort, J., Pillai, P., & Balasingam, B. (2023). Distracted worker: Using pupil size and blink rate to detect cognitive load during manufacturing tasks. *Applied Ergonomics, 106*, Article 103867. https://doi.org/10.1016/j.apergo.2022.103867

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife, 9*, Article e59360. https://doi.org/10.7554/eLife.59360

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chekaf, M., Cowan, N., & Mathy, F. (2016). Chunk formation in immediate memory and how it relates to data compression. *Cognition*, *155*, 96–107. https://doi.org/10.1016/j.cognition.2016.05.024

Chen, K. B., Harper-Gampp, T., Wu, L., Delgado, C., & Peterson, M. (2024). Learning scale in virtual reality: Experiences and perception of immersive technology at a public middle school. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 270–275. https://doi.org/10.1177/10711813241265650

Cheng, F., Harper-Gampp, T., Planchart, R., Dunning, M., Peterson, M., Delgado, C., & Chen, K. B. (2024). Study of graphic armatures, multimodal cues, and numeric measures in virtual reality on learners' performance and workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 1229–1234. https://doi.org/10.1177/10711813241269236

Common Core. (2010). *Mathematics standards – Common core state standards initiative*. https://www.corestandards.org/Math/

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058. https://doi.org/10.1016/j.cognition.2007.04.010

Dankel, S. J., & Loenneke, J. P. (2021). Effect sizes for paired data should use the change score variability rather than the pre-test variability. *The Journal of Strength & Conditioning*

*Research*, *35*(6), 1773–1778. https://doi.org/10.1519/JSC.0000000000002946

Delgado, C. (2009). *Development of a research-based learning progression for middle school through undergraduate students' conceptual understanding of size and scale* [Doctoral dissertation, University of Michigan]. Deep Blue. 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

Delgado, C., & Peterson, M. (2018). An enhanced framework for scale cognition leveraging visual metaphor theory and analogical reasoning theory. In J. Kay & R. Luckin (Eds.), *Rethinking learning in the digital age: Making the learning sciences count* (Vol. 3, pp. 1607–1608). International Society of the Learning Sciences. https://doi.org/10.22318/cscl2018.1607

Dewar, M. T., Cowan, N., & Della Sala, S. (2007). Forgetting due to retroactive interference: A fusion of early insights into everyday forgetting and recent research on anterograde amnesia. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, *43*(5), 616–634. https://doi.org/10.1016/s0010-9452(08)70492-1

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

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., You, H., Peterson, M., & Chen, K. B. (2025). Measuring size and scale: The development and validation of the Assessment of Size and Scale Cognition (ASSC). *Research in Science Education*, *56*(2), 427–450. https://doi.org/10.1007/s11165-025-10274-w

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.), *Human mental workload* (pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hollander, J., & Huette, S. (2022). Extracting blinks from continuous eye-tracking data in a mind wandering paradigm. *Consciousness and Cognition, 100*, Article 103303. https://doi.org/10.1016/j.concog.2022.103303

Hou, J., Lee, J. F., & Doherty, S. (2025). State-of-the-Art of eye tracking in mobile-assisted learning studies: A review of twenty years of empirical studies. *Journal of Computer Assisted Learning, 41*(1), Article e13118. https://doi.org/10.1111/jcal.13118

Kahn, M. J., Tan, K. C., & Beaton, R. J. (1990). Reduction of cognitive workload through information chunking. *Proceedings of the Human Factors Society Annual Meeting*, *34*(19), 1509–1513. https://doi.org/10.1177/154193129003401919

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

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*, Article 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*, Article 792069. https://doi.org/10.3389/fpsyg.2021.792069

Longo, M. R., & Lourenco, S. F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia*, *45*(7), 1400–1407. https://doi.org/10.1016/j.neuropsychologia.2006.11.002

Lønne, T. F., Karlsen, H. R., Langvik, E., & Saksvik-Lehouillier, I. (2023). 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), Article e17196. https://doi.org/10.1016/j.heliyon.2023.e17196

Lumivero. (2025). *NVivo [computer software]*. https://lumivero.com/products/nvivo/

Macedonia, M. (2019). Embodied learning: Why at school the mind needs the body. *Frontiers in Psychology, 10*, Article 2098. https://doi.org/10.3389/fpsyg.2019.02098

Magana, A. J. (2014). Learning strategies and multimedia techniques for scaffolding size and scale cognition. *Computers & Education*, *72*, 367–377. https://doi.org/10.1016/j.compedu.2013.11.012

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., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, *33*(3), 937–958. https://doi.org/10.1007/s10648-020-09586-2

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

Matovu, H., Ungu, D. A. K., Won, M., Tsai, C.-C., Treagust, D. F., Mocerino, M., & Tasker, R. (2023). Immersive virtual reality for science learning: Design, implementation, and evaluation. *Studies in Science Education*, *59*(2), 205–244. https://doi.org/10.1080/03057267.2022.2082680

Mautone, P. D., & Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology*, *93*(2), 377–389. https://doi.org/10.1037/0022-0663.93.2.377

Mayer, R. E., & Fiorella, L. (2014). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 279–315). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.015

McCrink, K., & Galamba, J. (2015). The impact of symbolic and non-symbolic quantity on spatial learning. *PLoS One, 10*(3), Article e0119395. https://doi.org/10.1371/journal.pone.0119395

Medina Herrera, L. M., Juárez Ordóñez, S., & Ruiz-Loza, S. (2024). Enhancing mathematical education with spatial visualization tools. *Frontiers in Education, 9*, Article 1229126. https://doi.org/10.3389/feduc.2024.1229126

Nathan, M. J. (2021). *Foundations of embodied learning: A paradigm for education*. Routledge. https://doi.org/10.4324/9780429329098

National Research Council. (2012). *A framework for K-12 science education: Practices, cross-cutting concepts, and core ideas*. 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

Peterson, M. (2022). How to do things with pictures: Imagery in visual media as workspace. *Visual Communication Quarterly*, *29*(1), 3–16. https://doi.org/10.1080/15551393.2021.2021083

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

Petzschner, F. H., Maier, P., & Glasauer, S. (2012). Combining symbolic cues with sensory input and prior experience in an iterative Bayesian framework. *Frontiers in Integrative Neuroscience*, *6*, 58. https://doi.org/10.3389/fnint.2012.00058

Pietikäinen, O., Hämäläinen, P., Lehtinen, J., & Karttunen, A. J. (2021). VRChem: A virtual reality molecular builder. *Applied Sciences, 11*(22), Article 10767. https://doi.org/10.3390/app112210767

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*, Article 103778. https://doi.org/10.1016/j.compedu.2019.103778

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78). Association for Computing Machinery. https://doi.org/10.1145/355017.355028

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

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

Sosic-Vasic, Z., Hille, K., Kröner, J., Spitzer, M., & Kornmeier, J. (2018). When learning disturbs memory – Temporal profile of retroactive interference of learning on memory formation. *Frontiers in Psychology*, *9*, Article 82. https://doi.org/10.3389/fpsyg.2018.00082

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

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(1), 37–55. https://doi.org/10.1037/xlm0000578

Van Der Hoort, B., Guterstam, A., & Ehrsson, H. H. (2011). Being Barbie: The size of one's own body determines the perceived size of the world. *PLoS One, 6*(5), Article e20195. https://doi.org/10.1371/journal.pone.0020195

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

Watson, A. B., & Yellott, J. I. (2012). A unified formula for light-adapted pupil size. *Journal of Vision*, *12*(10), 12. https://doi.org/10.1167/12.10.12

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

## Author Biographies

**Fangyuan Cheng** is a Ph.D. student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. She earned her master's degree from Tongji University in 2023 and her bachelor's degree from Hunan University in 2020. Her research focuses on human information processing in extended reality.

**Tyler Harper-Gampp** is a postdoctoral researcher at Texas State University. He received his Ph.D. in learning and teaching in STEM from North Carolina State University in 2024. He also earned an M.A. in teaching secondary science and a B.S. in chemistry from the University of North Carolina Wilmington in 2018 and 2016, respectively. His research

focuses on the intersection of learning sciences and learning technology, with the aim of advancing STEM education through innovative and equitable instructional methods.

**Rebecca Planchart** is a graduate student in the Master of Graphic and Experience Design program at North Carolina State University, where she also received her undergraduate degree in 2018. Her research interests include user experience design, data visualization, and human-centered applications of artificial intelligence.

**Mitchell Dunning** is a graduate student at North Carolina State University, where he earned his bachelor's degree in computer science in 2023. He works as a technical and 3D artist, specializing in the creation of representative botanical 3D models. His work supports the development of virtual reality applications and educational tools, with a focus on high-fidelity visual assets for immersive learning environments.

**Quinn Hilligoss** is an undergraduate student in industrial and systems engineering at North Carolina State University, with a minor in economics. His interests include data analysis, engineering design, and the application of industrial and systems engineering methods to problems at the intersection of engineering, technology, and business. He has experience in team-based research and engineering projects, with a focus on analytical problem solving and decision making.

**Matthew Peterson** is an associate professor in the College of Design at North Carolina State University. He received his Ph.D. in design from North Carolina State University in 2011, where he also received his master's and bachelor's degrees. His research focuses on visual representation and interaction, particularly in the design of novel user interfaces for emerging and evolving technologies.

**Cesar Delgado** is an associate professor in the Department of Science, Technology, Engineering, and Mathematics Education at North Carolina State University. He received his Ph.D. in science education from the University of Michigan, Ann Arbor, in 2009. He also earned an M.S. in chemistry and an M.A. in learning technologies from the University of Michigan, Ann Arbor, in 2008, and an M.A. in secondary education from the University of Alabama, Tuscaloosa, in 2003. He researches students' knowledge and learning of the magnitude of variables, including spatial and temporal scale.

**Karen B. Chen** is an associate professor in Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. Karen Chen received her PhD degree from Department of Biomedical Engineering at University of Wisconsin-Madison in 2015. Her interest is in human performance and behavior in virtual reality, and human-computer interaction.

# Appendix

## *Appendix A*

*Instructions for Participants During the Scale Worlds Experience in the Experiment*

(1) After you scale up or down, could you tell me what your current height is?
(2) What is the diameter of a white blood cell? How do you know it?
(3) Identify an entity that is 10,000 m in size. How do you know it?
(4) Choose one entity other than a human and estimate or determine its size. Why do you make this assumption?
(5) What entity is 100 times smaller than the diameter of a white blood cell?
(6) Choose one entity and identify another that is 1,000 times larger than it.
(7) How many times smaller is the diameter of a white blood cell compared to your current height? Please try to provide a precise estimate.
(8) At your current height, how many "you" would it take to line up with the next entity? You may use either the height along the $x$-axis or $y$-axis as the estimation criterion. Please try to provide a precise estimate.

(9) At your current height in Scale Worlds, there is one entity about your size, two smaller than you, and two larger than you. Sort these five visible entities into two groups. Briefly explain the reasoning behind your categorization.
(10) Sort the visible entities into two groups based on size. Briefly explain why you made this categorization. You are free to scale to other worlds to help you decide.

## Appendix B

*Semi-structured Interview Protocol*

(1) Could you describe your overall experience in Scale Worlds, the immersive virtual environment to learn about size and scale? What is the most useful feature for your learning?
(2) Did you use any numeric information found in Scale Worlds to determine the size of the scientific entity and/or understand size?
    (a) If yes, what numeric information have you used, what were you using it for, how did you use it during the Scale Worlds experience? Please explain.
    (b) If not, why did you not use it? Please explain.
(3) Did you use any armatures, for instance, grid lines and hash marks, found in Scale Worlds to determine the size of the scientific entity and/or understand size? Please explain.
    (a) If yes, what armatures have you used, what were you using them for, how did you use them during the Scale Worlds experience?
    (b) If not, why did you not use them? Please explain.
(4) Did you use any multi-sensorial cues, for instance, visual and audio cues, found in Scale Worlds to determine the size of the scientific entity and/or understand size?
    (a) If yes, what multi-sensorial cues have you used, what were you using them for, how did you use them during the Scale Worlds experience? Please explain.
    (b) If not, why did you not use them? Please explain.
(5) Did the use of VR technology aid in your understanding and retention of abstract concepts, such as scale, size, and numbers? How does this compare to traditional learning methods? Please explain.
(6) Did Scale Worlds help build or expand your existing knowledge about size, scale, and numbers? Why or why not?
    (a) If yes, did you benefit more from the micro world or the macro world? Please explain.
(7) Before using Scale Worlds, what were the scientific entities that you were most familiar with? After using Scale Worlds, did your "comfort range" of scientific entities change? Why or why not?
(8) Were there any challenges or constraints related to the scale of objects or the spatial arrangement in the VR learning environment?
(9) In general, how should the learning experiences in virtual reality be improved? In terms of learning size, scale, and numbers in virtual reality, what improvements could there be?
(10) Is there anything else you would like to tell us?