

Leveling Up Physics Education with Gamified Virtual Reality Simulations

José Graça and Pamela Teubig

UNIDCOM/IADE - Research Unit in Design and Communication, Universidade Europeia, Lisbon, Portugal

jose.graca@universidadeeuropeia.pt

pamela.teubig@universidadeeuropeia.pt

Abstract: Traditional physics education often struggles to convey abstract, three-dimensional concepts, leading to low student engagement and high cognitive load. While Virtual Reality (VR) offers immersive visualization, passive VR environments can still suffer from novelty wear-off, limiting their long-term pedagogical value. To address this, this study introduces PhysiVerse, an immersive VR experience structured around a task-based progression style that guides students through a hypothesis-testing learning loop. In this environment, students commit to a prediction before acting, test the prediction by running the experiment, and compare the results against their initial hypothesis. This cycle is reinforced by a gamified scoring system where users win points based on the closeness of their prediction to the empirical truth, incentivizing precise conceptual understanding over trial-and-error interaction. Focusing on concepts related to Mechanical Energy, this study evaluates the PhysiVerse experience through psychometric questionnaires, pre/post knowledge assessments, and in-experience behavioural metrics, including prediction accuracy as a measure of prior understanding.

Keywords: Education, Experimental physics, Virtual reality, Gamified learning

1. Introduction

The natural sciences occupy a uniquely challenging position in higher education, demanding both theoretical rigour and hands-on experimentation that has historically required dedicated physical infrastructure and specialised equipment. The rapid expansion of digital pedagogies has prompted a re-evaluation of how experiential learning is delivered, yet physical laboratory access remains a persistent bottleneck, particularly in resource-constrained institutions.

This tension is pronounced in experimental physics, where abstract theory must be reconciled with measurable reality. Hestenes et al. (1992), through the Force Concept Inventory, revealed deficits in students' understanding of Newtonian mechanics even after formal instruction, while practical laboratory instruction faces high equipment costs, scheduling limitations, and safety hazards.

Consumer-grade immersive virtual reality (VR) presents a promising avenue, with meta-analyses reporting moderate to strong positive effects of VR-based instruction on learning outcomes (Merchant et al., 2014; Wu et al., 2020). More broadly, gamification and game-based learning have been shown to yield positive effects on cognitive and motivational learning outcomes (Hamari et al., 2014; Sailer and Homner, 2020), though their integration with VR has been characterised as a 'double-edged sword': while gamification may enhance immersion and intrinsic motivation, it also risks aggravating cognitive load, potentially undermining the very outcomes it is intended to support (Radianti et al., 2020; Krath et al., 2021), a design challenge that remains insufficiently explored in undergraduate physics education.

This study investigates the potential of gamified immersive VR as a pedagogical tool for undergraduate experimental physics education, focusing on three research questions:

RQ1: To what extent do users experience a sense of presence and environmental fidelity within a gamified VR physics laboratory?

RQ2: How does a gamified VR environment influence student engagement, motivation, and willingness to repeat experimental activities?

RQ3: What are the effects of a predict-then-verify VR learning cycle on students' knowledge acquisition of core physics concepts?

Together, these questions address the phenomenological quality of the virtual experience, the affective and behavioural impact of gamification, and the cognitive outcomes of the prediction-execution cycle.

2. Background

Several digital platforms exist for experimental science education, yet each has limitations. PhET Interactive Simulations provides browser-based simulations, but their two-dimensional format limits the interactivity and immersion needed for spatial intuitions about three-dimensional phenomena. Labster offers more immersive environments, but its predefined experiment library precludes the customisation necessary to align with specific curricular objectives.

These limitations are consequential in physics, where the representational gap between two-dimensional diagrams and three-dimensional phenomena has been identified as a source of cognitive load (Madden et al., 2013). Existing platforms also lack the natural interaction modalities of real laboratory work, such as tactile manipulation and proprioceptive feedback, dimensions recognised as important for scientific reasoning.

Immersive VR offers distinctive advantages: three-dimensional interactive spaces where students manipulate virtual apparatus using natural hand gestures, observational perspectives impossible or dangerous in real laboratories, and real-time feedback on parameter changes. Shin (2017) demonstrated that immersion and presence positively influence perceived usability, and Makransky and Lilleholt (2018) established that the emotional value of VR experiences predicts learning outcomes and intrinsic motivation.

Beyond VR, gamification offers an engagement strategy. Sailer and Homner's (2020) meta-analysis reported positive effects on cognitive and motivational outcomes, while Hamari et al. (2014) emphasised their context-dependency. Empirical VR studies in physics (Pirker et al., 2017; Canright and White Brahmia, 2024) report motivational gains, yet Krath et al. (2021) noted gamification's theoretical foundations remain fragmented and gamified VR in undergraduate physics remains scarce, leaving open whether these gains translate into learning.

3. Methodology

The aim of this study is to evaluate PhysiVerse, a gamified VR physics learning environment, along three dimensions: the sense of presence and environmental fidelity experienced by users (RQ1); the impact on student engagement, motivation, and willingness to repeat experimental activities (RQ2); and the effects of a predict-then-verify learning cycle on knowledge acquisition (RQ3). A quantitative approach was adopted, employing four validated psychometric instruments: the Igroup Presence Questionnaire (IPQ) (Schubert et al., 2001), the System Usability Scale (SUS) (Brooke, 1996), the Technology Acceptance Model 3 (TAM3) (Venkatesh and Bala, 2008), and the User Experience in immersive Virtual Reality questionnaire (iUXRV) (Tcha-Tokey et al., 2016). Knowledge acquisition was assessed by administering questions on the theoretical concepts underlying each experiment both before and after the VR experience.

Beyond self-report and knowledge-assessment data, the PhysiVerse software automatically logs behavioural metrics throughout each session. These capture the time spent on each experimental phase, the frequency and duration of aid usage (helping visuals, formula overlays, or highlighted object indicators), the number of correct and incorrect predictions, the total configurations explored, the final score achieved, and the direction in which the user's head is facing to measure attention to different objects. By recording these traces in real time and without relying on participant recall, the software provides objective data that complement the subjective perceptions captured by the survey instruments and the learning gains measured through pre/post knowledge questions.

3.1 PhysiVerse Experience

To conduct the evaluation, a dedicated VR experience entitled PhysiVerse (Figure 1) was developed using the Unity 6.3 engine for consumer-grade head-mounted displays. The environment resembles a fully explorable classroom, with a table where experiment materials are instantiated upon selection. The experience is structured around three sequential phases: orientation, prediction and execution, and free exploration.



Figure 1: PhysiVerse mechanical energy experiment setup

The first phase (Figure 2) serves as a tutorial and orientation hub. Participants are guided through the core interaction mechanics: navigation, object grasping and placement, button and slider manipulation, ensuring that subsequent experimental activities are not confounded by unfamiliarity with the interface. Users can select from different experiments, view a global leaderboard and consult the achievements available for each experiment. The leaderboard and achievement system establish a gamification context from the outset, activating the motivational and competitive dimensions measured by the TAM3 and iUXRV instruments.



Figure 2: PhysiVarse tutorial phase, with user interface and animations explaining possible actions

The second phase constitutes the core instructional and assessment sequence. After selecting an experiment, participants predict the outcome of a given configuration before executing it (Figure 3). In the Mechanical Energy experiment, they might predict the final velocity of an object descending an incline. Only after committing to a prediction are they permitted to run the experiment and observe the actual result. This predict-then-verify structure creates conditions for cognitive dissonance when predictions diverge from observations and provides a direct measure of prior knowledge. During execution, participants configure the apparatus by adjusting parameters, placing objects, and operating instruments, following either a guided mode with visual prompts or an unassisted mode. An adaptive scaffolding mechanism progressively increases visual assistance if the system detects that the participant has not advanced after a defined period.



Figure 3: Prediction system for the mechanical energy experiment, for a car moving 25cm on a 20 degree angled slope

The third phase is a sandbox mode in which participants freely explore the experiment without constraints (Figure 4). Having completed the structured sequence, they may adjust any parameter at will and observe the outcomes in real time. Pedagogically, this allows students to test hypotheses generated during the structured phase and develop intuitive understandings through self-directed experimentation. Evaluatively, it generates behavioural data on exploration breadth and sustained engagement that the structured second phase alone would not capture.

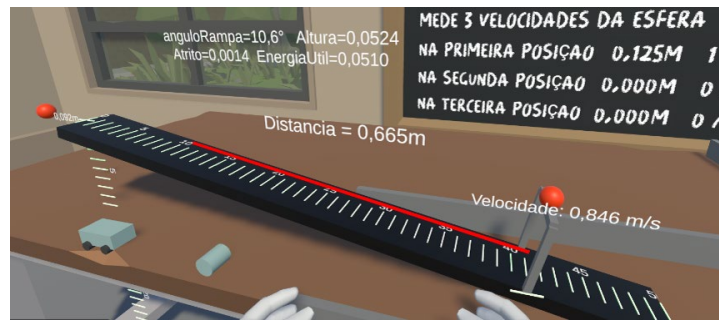


Figure 4: PhysiVarse sandbox mode with visual aid enabled

3.2 Procedure and Participants

The study involved 23 students ($N = 23$) enrolled in various bachelor's programmes that include physics-based curricular units, who volunteered to participate without receiving incentives or academic credit. All participants provided informed consent prior to participation and were informed of their right to withdraw at any time, in compliance with GDPR requirements. Prior to the VR session, participants read the experiment protocol and began the questionnaire, which collected demographic information and assessed baseline knowledge through pre-experiment questions. Participants then completed the three-phase PhysiVarse sequence: tutorial and experiment selection, prediction and execution, and free exploration. Afterward, they returned to the questionnaire to answer the same theoretical questions (post-experiment), followed by the full psychometric survey across the IPQ, SUS, TAM3, and iUXRV dimensions. The entire protocol took approximately 30 minutes.

3.3 Instrument and Data Collection

Rather than administering each instrument in full, a selection of items was drawn from each to reduce respondent fatigue while preserving coverage of the constructs most relevant to the present study. From the TAM3, items were selected from six constructs: Perceived Usefulness, Perceived Ease of Use, Behavioral Intention, Perceptions of External Control, Computer Self-Efficacy, and Perceived Pleasure. The final instrument comprised 55 statements: 10 from the IPQ, 10 from the SUS (administered in full as prescribed by its validation protocol), 18 from the TAM3, and 17 from the iUXRV. Although this selective approach represents a departure from full instrument administration, the retained items were chosen to preserve the construct definitions of each scale, and the SUS was administered in full to maintain its benchmark comparability. This compromise between instrument fidelity and respondent burden implies that construct scores should be interpreted as indicators rather than fully validated scale scores, a limitation that is revisited in the Discussion. Responses were collected on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), except for the SUS, which used its standard 5-point scale. Data were collected via Google Forms and processed according to each instrument's prescribed scoring procedure.

4. Results

This section presents the findings from the three data sources: the psychometric questionnaires (SUS, IPQ, TAM3, and iUXRV), the pre/post knowledge assessments, and the in-experience behavioural metrics, organised by construct domain.

4.1 System Usability

The System Usability Scale (SUS) yielded a mean score of 83.04 ($SD = 6.71$), well above the established SUS benchmark of 68 (Bangor et al., 2009), indicating high perceived usability. The iUXRV Usability in VR subscale corroborated this with a mean of 6.27 ($SD = 0.47$) on a 7-point scale.

4.2 Presence

The IPQ results indicate a strong sense of presence. Spatial Presence (SP) received the highest mean at 6.06 ($SD = 0.52$), followed by Experienced Realism (REAL) at 5.46 ($SD = 0.76$), and Involvement (INV) at 5.07 ($SD = 0.60$). All three subscale means fell above the scale midpoint.

4.3 Technology Acceptance

The TAM3 results reveal a pattern in which constructs related to ease of use and personal capability received substantially higher ratings than those related to perceived utility and external support. Computer Self-Efficacy

registered the highest mean at 6.32 (SD = 0.46), notable given that only 39% of participants had prior VR experience. Perceived Ease of Use was similarly elevated at 6.07 (SD = 0.58), consistent with the SUS and iUXRV usability findings. Behavioural Intention rated highly at 5.91 (SD = 0.58), and Perceived Pleasure reached 5.61 (SD = 0.69).

In contrast, Perceived Usefulness yielded a mean of only 4.26 (SD = 0.81), and Perceptions of External Control registered 4.35 (SD = 0.79), both near the scale midpoint, indicating ambivalence rather than endorsement.

4.4 Immersive User Experience

The iUXRV results indicate a predominantly positive experiential quality. VR Sickness registered a mean of only 1.43 (SD = 0.78) on a 7-point scale, indicating minimal adverse physical effects. Usability in VR was rated at 6.27 (SD = 0.47), the highest iUXRV subscale score. Emotions received a mean of 5.43 (SD = 0.73), indicating positive affective states, while Aesthetics received the lowest rating at 4.70 (SD = 0.41), just above the midpoint.

4.5 Summary of Psychometric Findings

Table 1 consolidates the questionnaire results. The overall pattern is one of high usability, strong presence, and positive emotional and motivational responses, tempered by moderate perceptions of pedagogical usefulness and external support.

Table 1: Summary of questionnaire results (N = 23).

Instrument	Construct	Mean	SD
SUS	Usability	83.04	6.71
IPQ	Spatial Presence	6.06	0.52
IPQ	Involvement	5.07	0.60
IPQ	Experienced Realism	5.46	0.76
TAM3	Perceived Usefulness	4.26	0.81
TAM3	Perceived Ease of Use	6.07	0.58
TAM3	Behavioural Intention	5.91	0.58
TAM3	Perceptions of External Control	4.35	0.79
TAM3	Computer Self-Efficacy	6.32	0.46
TAM3	Perceived Pleasure	5.61	0.69
iUXRV	VR Sickness	1.43	0.78
iUXRV	Usability in VR	6.27	0.47
iUXRV	Aesthetics	4.70	0.41
iUXRV	Emotions	5.43	0.73

4.6 Knowledge Acquisition

To assess learning gains, two multiple-choice questions addressing concepts directly related to the experimental activities were administered before and after the VR experience. Table 2 summarises the results.

Table 2: Pre- and post-experiment knowledge assessment results (N = 23)

Question	Pre-test correct	Pre-test %	Post-test correct	Post-test %
If two objects with different masses roll down the same inclined plane, starting from rest at the same height and with no significant friction, which one will reach the bottom with greater speed?	9	39.1	18	78.3
In an experiment in which a body moves down an inclined plane, its velocity is measured at various points, and its kinetic energy is calculated. If friction is not negligible, what relationships are expected to hold between the change in kinetic energy and the change in gravitational potential energy?	5	21.7	8	34.8

Question 1 showed a substantial improvement, with correct responses doubling from 9 to 18 participants, suggesting that the VR experience was effective in conveying the underlying concept. Question 2 exhibited a more modest gain of 13.0 percentage points (from 5 to 8 correct responses), indicating that the concept was less effectively addressed or posed greater difficulty regardless of the instructional medium.

One participant regressed from correct to incorrect on Question 1; this is discussed further in Section 5.3. The overall pattern is one of improvement, particularly on Question 1.

4.7 Behavioural Metrics

The in-experience behavioural metrics, logged automatically by the PhysiVerse software, provide an objective account of how participants distributed their time and effort. Table 3 summarises the key metrics.

Table 3: In-experience behavioural metrics (N = 23).

Metric	Mean	SD
Time on tutorial (s)	76.61	24.15
Time on prediction and execution (s)	387.08	148.08
Time on free exploration (s)	291.03	204.91
Aid time during prediction and execution (s)	204.14	116.75
Total predictions made	2.74	0.85
Correct/close-to-correct predictions	1.26	0.99
Configurations explored	3.57	1.74

Participants spent a mean of 76.61 seconds (SD = 24.15) on the tutorial phase. The prediction and execution phase consumed the largest proportion of session time at 387.08 seconds (SD = 148.08), though the substantial standard deviation indicates considerable variability in completion time.

The free exploration phase had a mean duration of 291.03 seconds (SD = 204.91), approximately 4.9 minutes, with the highest variability of any metric: some participants spent well over 10 minutes freely manipulating parameters, while others completed only a cursory exploration.

Aid usage during the prediction and execution phase was substantial: participants spent a mean of 204.14 seconds (SD = 116.75) with visual aids active, approximately 52.7% of the prediction and execution time.

Participants made a mean of 2.74 predictions (SD = 0.85), with 1.26 (SD = 0.99) classified as correct or close to correct, yielding an overall prediction accuracy rate of approximately 46%.

In the sandbox phase, participants explored a mean of 3.57 different configurations (SD = 1.74), indicating moderate exploratory behaviour with notable variability.

5. Discussion

The results offer a multifaceted picture of how undergraduate physics students experience and learn from a gamified VR environment structured around a prediction-execution learning loop. The findings are discussed below in relation to the three research questions, the key tension that emerged, and the practical implications for educational VR design.

5.1 Simulation Fidelity and Presence

The first research question addressed simulation fidelity. While a direct comparison was not conducted, the presence data offer self-report evidence of the virtual environment's phenomenological quality. The high Spatial Presence score (see Table 1) indicates that participants strongly perceived themselves as physically situated within the virtual laboratory, consistent with Pirker et al. (2017), who reported that fully immersive systems produced significantly higher presence scores than mobile alternatives. This suggests that PhysiVerse achieved a level of spatial immersion that is a necessary, though not sufficient, condition for a credible VR laboratory surrogate.

Experienced Realism was also above the midpoint (see Table 1) but notably lower than Spatial Presence, likely reflecting the inherent limitations of consumer-grade VR in replicating tactile and kinesthetic feedback. This gap has been documented in prior work (Schubert et al., 2001) and may represent a persistent constraint of VR-based laboratory surrogates.

Involvement received the lowest IPQ subscale score (see Table 1), likely attributable to the structured, task-oriented nature of the prediction-execution phase, consistent with findings that involvement tends to be lower in goal-directed than in open-ended VR tasks (Schubert et al., 2001). Future studies might administer presence measures at multiple points to disentangle phase-specific effects.

5.2 Student Engagement and Motivation

The second research question examined how the gamified VR environment influences student engagement, motivation, and willingness to repeat experimental activities.

The engagement and motivational indicators are consistently strong (see Table 1). The ease-of-use, pleasure, and behavioural intention constructs all rated well above the scale midpoint, and the extremely low VR Sickness score further supports sustained usability. These positive motivational effects are consistent with the broader gamification literature, where Sailer and Homner's (2020) meta-analysis reported small-to-moderate positive effects on cognitive and motivational outcomes, and Krath et al. (2021) emphasised that such effects are theoretically grounded in motivation and behaviour theories. The gamification elements appear to have contributed positively to motivation without triggering the cognitive overload identified as a risk of combining gamification with VR (Radianti et al., 2020).

However, the most striking finding is the gap between ease-of-use and pleasure on the one hand, and Perceived Usefulness on the other (see Table 1). Within the TAM framework, Perceived Usefulness is a stronger predictor of sustained adoption than Perceived Ease of Use (Venkatesh and Bala, 2008). An experience that is enjoyable but not perceived as educationally valuable risks being treated as a novelty. The low Perceptions of External Control score (see Table 1) compounds this concern, suggesting that participants did not feel adequately supported to use VR as a regular learning tool. This gap must be interpreted in light of the item-selection approach: because only a subset of TAM3 items was administered, the construct score may not fully capture the multi-dimensional nature of perceived usefulness as operationalised in the original instrument. The behavioural metrics offer partial corroboration: the substantial time spent on aid (see Table 3) suggests that participants engaged seriously with the learning content, but the absence of a control group prevents disentangling the effect of the gamified VR experience from novelty or task engagement.

5.3 Knowledge Acquisition

The third research question addressed the effects of VR-based instruction on students' knowledge acquisition. As shown in Table 2, the pre/post assessment showed improvement on both questions, with Question 1 exhibiting a substantial gain and Question 2 a more modest one.

The prediction-execution cycle appears to have been particularly effective for the concept assessed by Question 1, where committing to a prediction, observing a discrepant result, and resolving the dissonance may have produced a robust learning effect. The more modest gain on Question 2 suggests that this mechanism may be less effective when the prediction is too uninformed to create meaningful dissonance.

The single instance of knowledge regression warrants attention. While it would be inappropriate to generalise from one case, it raises the possibility that VR's visual representation can introduce or reinforce misconceptions. Makransky and Lilleholt (2018) cautioned that the emotional richness of VR can lead to overconfidence in incorrect inferences without sufficient scaffolding, a boundary condition that future studies should systematically track.

In the absence of a control group, these knowledge gains cannot be attributed exclusively to the VR medium. Furthermore, the assessment relied on only two multiple-choice items, which provide a partial picture of knowledge acquisition and preclude formal statistical significance testing. The observed proportion gains (Question 1: 39.1% to 78.3%; Question 2: 21.7% to 34.8%) correspond to medium-to-large effect sizes (Cohen's h of approximately 0.79 and 0.28 respectively), but the small sample ($N = 23$) and the absence of a control group mean these gains may reflect the experimental activity itself, pre-test priming, or a combination of factors rather than the VR medium specifically. The pre/post design establishes that learning occurred, but not that VR was the cause. Future research should address these limitations through comprehensive knowledge assessments and randomised controlled comparisons against physical laboratories and non-immersive simulations.

5.4 The Ease–usefulness Gap: Implications for Adoption

Theoretically, this gap suggests that the TAM3 relationships may operate differently in educational VR. In the original TAM, Ease of Use is hypothesised to influence Usefulness indirectly, a system that is easy to use frees

cognitive resources for the primary task, thereby increasing perceived effectiveness. The present results suggest that this pathway may be disrupted: novelty, immersion, and gamification may make the experience subjectively easy and pleasurable without conveying a clear sense of educational value, creating a decoupling between ease of use and perceived usefulness that the standard TAM model does not predict. The behavioural data provide additional nuance: participants who spent more time in free exploration did not necessarily rate the experience as more useful, and the substantial aid usage during the prediction phase suggests that engagement with the learning content was mediated by scaffolding rather than by independent discovery. This pattern is consistent with the contextual dependency of gamification effects reported by Hamari et al. (2014).

Practically, designers of educational VR should not assume that high usability and enjoyment automatically translate into perceived educational value. Explicit metacognitive scaffolding, post-experiment reflection prompts, explicit links between the virtual activity and underlying theory, or comparative summaries showing how the VR experience maps onto traditional laboratory outcomes, may be necessary to bridge the gap. The prediction-execution cycle already creates a natural moment for reflection that could be leveraged more deliberately through guided prompts.

5.5 Behavioural Patterns and Engagement Strategies

The tutorial completion time (see Table 3) aligns with the high usability and self-efficacy scores, suggesting that the onboarding phase was effective, consistent with prior findings that well-designed VR tutorials can rapidly establish baseline competence (Pirker et al., 2017), and supporting the methodological decision to include the tutorial as a confound-control measure.

The most revealing behavioural finding concerns the distribution of time across the structured and free exploration phases (see Table 3). The variability in the sandbox phase was dramatically higher than in the prediction-and-execution phase, suggesting that the structured, scaffolded phase produced consistent engagement, whereas the self-directed sandbox amplified individual differences in curiosity and persistence. This implies that merely making a sandbox available may be insufficient; structured prompts within the free-play phase, such as challenge questions or suggested parameter combinations, may be needed to sustain engagement.

The substantial aid usage during the prediction and execution phase (see Table 3) raises questions about the balance between scaffolding and independent problem-solving. The high aid usage is consistent with the adaptive scaffolding mechanism, but the possibility that it was too aggressive, activating before participants had sufficient opportunity to struggle productively, cannot be ruled out. Productive struggle is a well-established component of effective science instruction (Warshauer, 2015), and premature scaffolding may short-circuit the cognitive processes that lead to robust learning. The relationship between aid usage and learning outcomes warrants further investigation.

6. Conclusion and Future Work

This study evaluated PhysiVerse, a gamified VR physics learning environment structured around a prediction-execution learning cycle, along three dimensions. Regarding presence (RQ1), participants reported strong spatial immersion, though the gap between Spatial Presence and Experienced Realism highlights a persistent limitation of consumer-grade VR in replicating tactile feedback. Regarding engagement and motivation (RQ2), results were positive but nuanced: usability, pleasure, and behavioural intention were high, yet Perceived Usefulness lagged substantially behind, suggesting that an enjoyable VR experience does not automatically translate into perceived educational value, an ease-usefulness gap that challenges the standard TAM pathway. Regarding knowledge acquisition (RQ3), the pre/post assessment showed meaningful improvement, particularly for the concept targeted by the prediction-execution cycle, though the absence of a control group and the limited two-item assessment preclude attributing these gains exclusively to the VR medium.

The behavioural metrics enriched these self-report findings by revealing that structured, scaffolded phases produced consistent engagement, whereas the self-directed sandbox phase amplified individual differences. The substantial aid usage raises the question of whether adaptive scaffolding may, in some cases, short-circuit productive struggle. Together, these findings suggest that gamified VR can create engaging and usable learning experiences, but that realising its full pedagogical potential requires deliberate design choices: metacognitive scaffolding to bridge the ease-usefulness gap, structured prompts within free exploration phases, and carefully calibrated adaptive assistance.

Several directions for future work emerge. First, randomised controlled experiments comparing PhysiVerse against physical laboratories and non-immersive simulations are needed to isolate VR's specific contribution to learning. Second, comprehensive knowledge assessments with statistical significance testing should replace the limited two-item measure used here. Third, longitudinal studies would clarify whether the ease-usefulness gap narrows with familiarity. Fourth, the relationship between aid usage and learning gains should be examined at the individual level. Finally, expanding beyond Mechanical Energy and recruiting larger, more diverse samples would strengthen the generalisability of these findings.

Acknowledgement

This work is supported by FCT - Fundação para a Ciência e Tecnologia, I.P., in the framework of the Project UID/00711/2025- Research Unit in Design and Communication - UNIDCOM/IADE -, with the DOI identifier 10.54499/UID/00711/2025 (<https://doi.org/10.54499/UID/00711/2025>).

We want to thank co-supervisors Daniel Galaviz and Sara Cruz for their support on the development of this study.

Ethics Declaration This study did not require formal ethical clearance. All participants provided informed consent prior to their involvement.

AI Declaration: Grammarly, an AI-powered writing assistant, was used during the preparation of this paper. The tool supported grammar, spelling, and punctuation issues and suggesting improvements for clarity and conciseness.

References

- Bangor, A., Kortum, P., and Miller, J. (2009) Determining what individual SUS scores mean: Adding an adjective rating scale, *Journal of Usability Studies*, Vol. 4, No. 3, pp 114-123.
- Brooke, J. (1996) SUS: A quick and dirty usability scale, In P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland (Eds.), *Usability Evaluation in Industry*, pp 189-194, London: Taylor & Francis.
- Canright, J., and White Brahmia, S. (2024) Modeling novel physics in virtual reality labs: An affective analysis of student learning, *Physical Review Physics Education Research*, Vol. 20, 010146, <https://doi.org/10.1103/PhysRevPhysEducRes.20.010146>
- Hamari, J., Koivisto, J., and Sarsa, H. (2014) Does gamification work? A literature review of empirical studies on gamification, In *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*, pp 3025-3034, IEEE. <https://doi.org/10.1109/HICSS.2014.377>
- Hestenes, D., Wells, M., and Swackhamer, G. (1992) Force concept inventory, *The Physics Teacher*, Vol. 30, pp 141-158, <https://doi.org/10.1119/1.2343497>
- Krath, J., Schürmann, L., and Von Korflesch, H. F. (2021) Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, *Computers in Human Behavior*, Vol. 125, 106963, <https://doi.org/10.1016/j.chb.2021.106963>
- Madden, M. E., Baxter, M., Beauchamp, H., Bouchard, K., Habermas, D., Huff, M., Ladd, B., Pearson, J., and Plague, G. (2013) Rethinking STEM education: An interdisciplinary STEAM curriculum, *Procedia Computer Science*, Vol. 20, pp 541-546.
- Makransky, G., and Lilleholt, L. (2018) A structural equation modeling investigation of the emotional value of immersive virtual reality in education, *Educational Technology Research and Development*, Vol. 66, No. 5, pp 1141-1164, <https://doi.org/10.1007/s11423-018-9581-2>
- Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., and Davis, T. J. (2014) Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis, *Computers & Education*, Vol. 70, pp 29-40. <https://doi.org/10.1016/j.compedu.2013.07.033>
- Pirker, J., Lesjak, I., and Gütl, C. (2017) An educational physics laboratory in mobile versus room scale virtual reality - a comparative study, *International Journal of Online Engineering (iJOE)*, Vol. 13, pp 106-120, <https://doi.org/10.3991/ijoe.v13i08.7371>
- Radianti, J., Majchrzak, T. A., Fromm, J., and Wohlgenannt, I. (2020) A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda, *Computers & Education*, Vol. 147, 103778, <https://doi.org/10.1016/j.compedu.2019.103778>
- Sailer, M., and Homner, L. (2020) The gamification of learning: A meta-analysis, *Educational Psychology Review*, Vol. 32, pp 77-112, <https://doi.org/10.1007/s10648-019-09498-w>
- Schubert, T., Friedmann, F., and Regenbrecht, H. (2001) The experience of presence: Factor analytic insights, *Presence: Teleoperators and Virtual Environments*, Vol. 10, No. 3, pp 266-281, <https://doi.org/10.1162/105474601300343603>
- Shin, D.-H. (2017) The role of affordance in the experience of virtual reality learning: Technological and affective affordances in virtual reality, *Telematics and Informatics*, Vol. 34, No. 8, pp 1826-1836.
- Tcha-Tokey, K., Loup-Escande, E., Christmann, O., and Richir, S. (2016) A questionnaire to measure the user experience in immersive virtual environments, *International Journal of Virtual Reality*, Vol. 16, pp 33-48, <https://doi.org/10.20870/IJVR.2016.16.1.2880>

- Venkatesh, V., and Bala, H. (2008) Technology acceptance model 3 and a research agenda on interventions, *Decision Sciences*, Vol. 39, No. 2, pp 273-315.
- Warshauer, H. K. (2015) Productive struggle in middle school mathematics classrooms, *Journal of Mathematics Teacher Education*, Vol. 18, No. 4, pp 375-400.
- Wu, B., Yu, X., and Gu, X. (2020) Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis, *British Journal of Educational Technology*, Vol. 51, No. 6, pp 1991-2005, <https://doi.org/10.1111/bjet.13023>