Effects of Augmented Reality Enhancements on Students' Scientific Reasoning in an Introductory Physics Laboratory

Abstract

Advances in augmented reality (AR) have been studied to supplement hands-on, physical labs. While prior research has focused on advancing content mastery and science interest, little work has investigated using AR to enhance scientific reasoning in laboratory classrooms. This study introduces an AR-enhanced physics lab to teach scientific reasoning and the understanding of measurement uncertainty. AR overlays of protractors, real-time energy flow graphs, and an automated period counter were created to reduce extraneous cognitive load and illuminate scientific insights throughout a pendulum experiment. Compared to the traditional lab, participants who experienced AR enhancement demonstrated greater insights into data differences, increased trust in their findings based on uncertainty analysis, and a deeper understanding of energy concepts. Results from the between-participant study (n=52) suggest the AR-enhanced lab enhances participants' scientific reasoning skills (Bayes factor of 17.44) and demonstrates AR's potential to improve scientific reasoning while maintaining interest and content mastery.

Keywords: Augmented Reality, Physics, Mobile Learning, Scientific Reasoning

1. Introduction

Extended reality, or a blending of digital and real worlds, has been shown to improve the educational outcomes for K12 laboratory science, with studies primarily focusing on outcomes related to content knowledge, engagement, and motivation (e.g., Brinson (2015)). While these aspects of learning are important, scientific reasoning, or the ability to think critically, form theories, and evaluate conclusions based on data, is another key outcome of laboratory

activities that may be influenced by extended reality and has not been as well studied.

Within the physics domain, Holmes and Bonn created a physics lab called *Pendulum for Pros* to teach scientific reasoning and help participants understand the complexity and ambiguity of empirical work (Holmes, 2014; Holmes and Bonn, 2015). Through this pedagogy, students learn how precision can affect scientific results. Using a timer and protractor, students test if pendulums swinging at 10 or 20 degrees have the same angle. By increasing the precision in their measurements, students can see the limits of the small angle approximation. This paper extends Holmes's work by using augmented reality (AR), a subset of extended reality that places digital information over the user's view of the physical world, integrating real and virtual elements to reduce extraneous cognitive load and allow participants to focus on measurement precision and data evaluation. Previous work has demonstrated AR's ability to embed text and images to decrease extraneous load (Huwer and Seibert, 2018), as well as visualize abstract concepts (Liu et al., 2021). We test our AR variant of Pendulum for Pros with participants using a typical lab set up (control) and an AR-enhanced lab setup and evaluate how well participants understand scientific reasoning and the complexity/ambiguity of empirical work.

The primary objective this study is to evaluate whether an AR-enhanced laboratory can increase scientific reasoning. Our research questions are as follows:

1. Is there a difference in participants' increase in scientific reasoning skills when conducting a lab with AR compared to a lab with traditional materials?
2. How does using AR materials affect participants' learning in terms of scientific reasoning?

The remainder of this paper is organized as follows: Section 2 provides background on the pedagogy used in our study, Section 3 explains our methods, Section 4 presents and discusses the results, and Section 5 describes our conclusions and areas for future work.

2. Background

This section details how scientific reasoning is defined for our study, highlighting the complexity and ambiguity of empirical work. We first situate our

work among the prior work using AR to teach scientific reasoning. Then we discuss theories of distributed cognition and meta-representation, and how we apply them to help learners improve their scientific reasoning.

2.1. Scientific Reasoning

Scientific reasoning is the acquisition and application of skills such as identifying questions, designing and conducting experiments, developing explanations, analyzing alternatives, and constructing and defending arguments (Singer et al., 2005). Learners with scientific reasoning skills can find patterns in data, formulate cause and effect generalizations, build and test models, and connect science to everyday life. They can recognize what assumptions are going into their models, and ponder alternate explanations for their data. Lastly, they can create a conclusion and argue a scientific point of view.

To best support scientific reasoning, we draw upon work from Holmes for teaching scientific reasoning in a physics classroom. This work, dubbed *Pendulum for Pros*, uses a series of contrasting cases to help students best analyze and draw conclusions about their data (Holmes, 2014). Analysis of student lab notebooks during the lab showed both thoughtfulness of collected data, as well as considerations for precision and measurement error. To test the value of AR materials, we use the same pedagogy outlined in *Pendulum for Pros*, but with AR additions to investigate its value to aid in student learning of scientific reasoning.

Previously, AR has shown promise for increasing scientific reasoning. Bakri et al. created an AR worksheet to assist learners in scientific reasoning for an elasticity lab (Bakri et al., 2020). The worksheet asked learners to analyze graphs and imbedded AR overlay videos of a previously conducted lab, and make conclusions and models based on the data presented. However, this work was only evaluated with scienence teachers, but not tested and evaluated on science learners. Additionally, Bakri et al. required a prescriptive lab that followed to pre-made worksheets. In contrast, we use AR as a tool, which can be used for a variety of comparison-based pendulum labs.

Understanding science's ambiguity and complexity—that all measurements have inherent uncertainty—is crucial for learners to become scientists (Singer et al., 2005). Part of conducting an experiment includes troubleshooting equipment and accounting for environmental variables. For example, the pendulum typically studied in lectures makes many assumptions that do not hold exactly in real experiments. Scientific measures, such as pendulum angle

and period are never exact but rather subject to uncertainties due to limitations of instruments, systematic errors, and random fluctuations. Acknowledging and quantifying uncertainties aids learners in critically evaluating the information necessary to make informed decisions in their daily life. These skills can come naturally to students when completing a traditional experiment, which uses real instruments and real measurements. While uncertainty can be randomly simulated in a purely virtual experiment, it is difficult to virtually reproduce uncertainty that arises from experimental procedure or unknown environmental factors. In our study, we highlight these aspects by using a real pendulum with noticeable dampening and graphically show energy leaving the system.

2.2. Distributed Cognition

Distributed cognition focuses on that learning is not only individualized but distributed across peoples and artifacts (Hutchins, 1995). Therefore, both the social and technological structure of the learning environment can influence the learning outcome. In our study, students work in pairs, allowing them to share in knowledge building about the physics concepts during learning. Additionally, we test how changing the technological artifacts, such as the ability to see emerging states of the experiment (See Figure 3), affect the learning gains. Within science education, the physical artifacts do provide constraints to the experimental procedures, but not a prescriptive task (Xu and Clarke, 2012). In our experiment, we evaluate how providing these cognitive artifacts effects students scientific reasoning of energy and experimental uncertainty.

Decreasing Extraneous Cognitive Load: Distributed cognition can aid in learning by reducing the number of tasks a learner must complete in his/her head. This in turn, reduces extraneous load (Sweller et al., 1998). Frank and Kapia used AR as distributed cognition for learning about control algorithms (Frank and Kapila, 2017). When participants implemented their algorithms on a robot, an AR image overlayed the desired position on the physical robot, so participants could see the positional error resulting from different algorithms. In this way, participants did not need to physically measure the positional differences, but could visually see how far the robot deviated from the desired position. In our study, AR distributes cognition by displaying the initial angle of the pendulum, as well as count periods as the pendulum moves.

Meta-Representation: Similarly to physics intuition, learners have an intuitive ability to reason about and connect abstract representations to their physical or metaphysical meaning (diSessa and Sherin, 2000). While traditional representations are presented post data-collection or physically separated from the experiment, AR allows learners to see representations in situ with the experimental procedure (Thees et al., 2020; Yu et al., 2022) which improves learning outcomes. In our experiment, we use graphical representations to highlight energy transformations and analyze how this representation helps students understand energy throughout the experiment. This symbolic representation becomes a shared representation for discussing experimental data, distributing knowledge between the representation and the learning pairs.

3. Methods

This section details the design of the AR environment, as well as the study procedures, measurements, participant demographics, and data analysis. While those in the treatment condition have all components listed below, we use only the same physical pendulum and a variant of the lab notebook (Figure 3) for participants in the control condition.

3.1. AR Environment

We developed an AR mobile application that uses a real pendulum (Figure 1) to visualize the energy transformations real time and reduce extraneous cognitive load as participants collect and evaluate data. The application consists of two screens. The first screen, or Data Collection screen, allows participants to view the pendulum through a camera, and AR overlays provide additional insights (Figure 2). The second screen, or Lab Notebook screen, computes the statistical analysis comprehensive viewing of lab data (Figure 3).

3.1.1. Data Collection Screen

The Data Collection screen shown in Figure 2 distributes the cognition necessary to complete the experiment. The following AR overlays can be seen alongside the physical pendulum: AR Protractor, Energy Graph, Period Counter, and Condition Visualizer.

AR Protractor: The AR protractor allows participants to view the angle of the pendulum without using external tools, as well as see the uncertainty

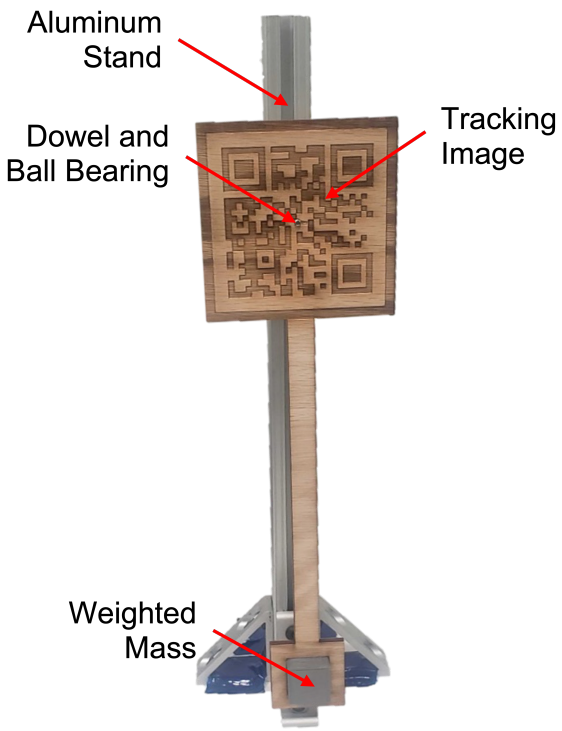


Figure 1: Physical Pendulum used in the study

of the protractor measurement. The protractor is semi-transparent, and anchored to the pendulum pivot. The measurement is displayed digitally along with the instrument's uncertainty, allowing for ease of reading and recalling the current angle. *Energy Graph*: The energy graph shows the transfer of potential, kinetic, and lost energy as the pendulum moves in real time. This allows participants see the loss of energy over time due to outside factors such as friction, as well as the transfer of kinetic, potential, and lost energy as the pendulum moves back and forth. *Period Counter*: The period counter is located just below the timer start and stop controls for the AR condition. This counter counts the period of the pendulum as it oscillates. The counter keeps track of both the number of periods, as well as the average period (time for each swing) in seconds. Because the precision in period measurement increases as the number of swings per trial increases, the counter alleviates participants' need to count a large number of periods. When participants stop the timer, the number of periods and average periods

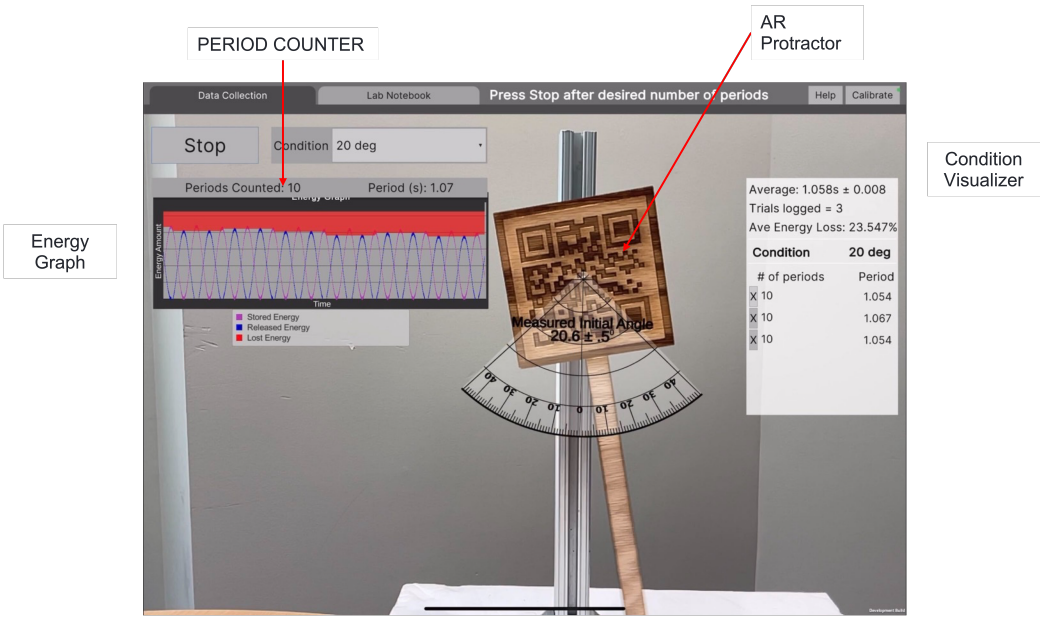


Figure 2: Data Collection Screen for the AR Environment

(along with the total energy dissipation) are displayed, and if confirmed by the participants, saved to the lab notebook.

3.1.2. Lab Notebook

The Lab Notebook used in both conditions of the study stores all of the data participants collected and displays summary statistics. Participants can see the number of trials collected, number of swings or periods/trial, and for the AR condition, the average energy loss. In the background, the system performs a t-test on the pendulum periods under the participant's selected condition. For example, if a participant varied the initial angle when measuring the period, the t-test measured if the change in initial angle brought about a significant difference in the measured period (Figure 3). Note how the AR condition also shows the average energy loss, and places the uncertainty beside the period.

3.2. Experimental Procedure

Participants engage in the experiment in pairs in either the AR or control condition. After signing consent forms, pairs in both conditions watch a pre-experiment video reviewing the definition of a pendulum, a period, energy,

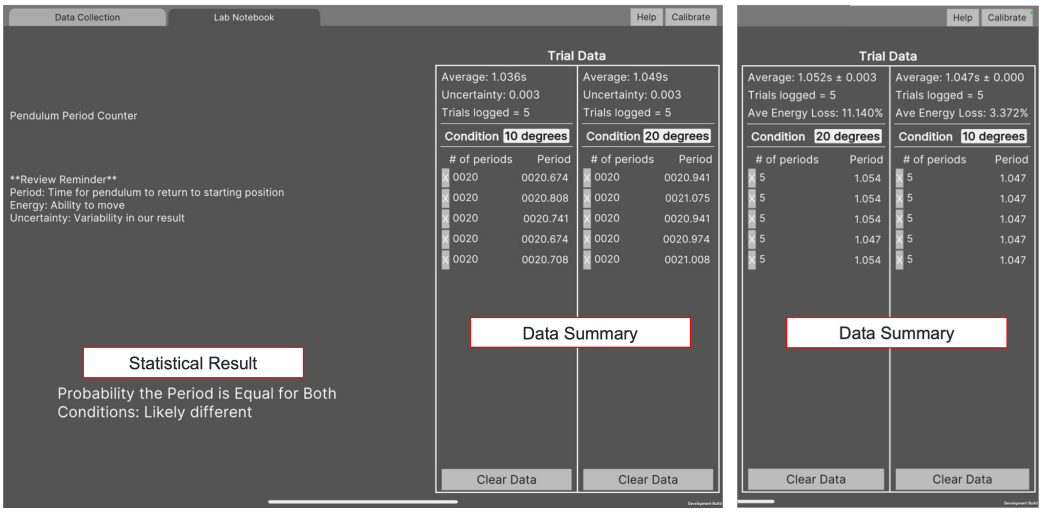


Figure 3: Lab Notebook for Control Group (left). The AR and Control group have similar lab notebooks, and only differ in the top portions of the trial data (display of energy loss). The difference in the AR group is shown on the right and Control Group

and uncertainty and can ask questions for clarity on the video. They then take a pre-test on scientific reasoning. Participants in the control group then conduct the experiments with a digital timer app, the Lab Notebook interface, and the physical pendulum. Participants in the AR group use the AR Collection Screen, Lab Notebook interface, and the physical pendulum. Participants conduct two experiments. In the first experiment, they are familiarizing themselves with the pendulum and the digital interfaces, as well as determining how friction and energy loss effects experimental results. Pairs first hypothesize whether the measured period will change between recording one swing vs. ten swings when averaged over five trials. They then conduct the experiment and discuss why they think the recorded periods are different. The second experiment asks learners to determine if the period is larger, smaller, or the same for ten degrees versus twenty degrees. They can alter the experimental procedure as they see fit, as long as they test 10 degrees and 20 degrees, and perform five trials per condition. After the second experiment, the participants discuss with the researcher the effects of the angle on the period. They also brainstorm ways to increase precision in their experiment, and are instructed to try again, until a difference between the periods is noticed or they can reduce the uncertainty to 0.001 seconds. Throughout both experiments (i.e., testing if the number of swings affects the period and

if the initial angle affects the period), the researcher observes and answers logistical questions. After the experiment, participants complete a post-test and provide demographic information. Two weeks later, participants complete a retention test.

3.3. Measures

The pre-, post-, and retention tests include content, scientific reasoning, and physics interest sections; however, this paper will focus on the scientific reasoning sections of the test. The scientific reasoning tests concepts that do not have a memorized answer, such as how and why to increase precision, as well as how participants would interpret a result given particular data. These questions are adapted from the Physics Laboratory Inquiry in Critical Thinking (Walsh et al., 2019). Before conducting the experiment, we consulted with one of the authors of Walsh et al. to ensure that our adapted questions were still appropriate to measure critical thinking in science labs. Additionally, we piloted the questions with several people with varying levels of physics experience via "think aloud" methods (Ericsson, 2017) and changed the wording to ensure clarity.

3.4. Participants

52 university affiliates (26 randomly assigned per condition) who had not taken physics at a college level were recruited via class announcements, fliers, and emails to take part in this study; this was a convenience sample. Students could sign up with a partner or be paired with another student to perform the study in pairs. Participants came from a variety of backgrounds (42% Asian, 17% White, 13% Mixed, 11% Black, 10% Hispanic, and 6% other), and were an average of 21 years old. While many participants (65%) had completed a physics experiment in high school, the majority (80%) of participants had never completed a pendulum experiment before. All participants were compensated for their time for the initial study and follow-up retention survey.

3.5. Analysis

We focus on analysis of the scientific reasoning section of the assessment, which was given before, after, and two weeks following the study, and completed individually. The analysis of these questions is explained below. All questions in the assessment were adapted from the Physics Lab Inventory of

Critical Thinking (PLIC) (Walsh et al., 2019), which was created for teaching and testing scientific reasoning. We selected questions for the pre-, post-, and retention test from PLIC that covered the following topics:

- How to maximize precision in an experiment and why
- The difference between standard deviation and mean values when analyzing data
- How to determine uncertainty of analog measurements
- Importance of multiple trials in experimental procedure

One grader scored all the free response answers. To ensure reliability, a second grader scored 20% of the answers, and a Cohen's kappa is calculated for those responses. If a question does not receive a score of at least 0.7 (a score of 0.61-0.8 is considered substantial agreement according to McHugh (2012)), all responses were graded by the second grader. If the first and second graders' scores differed by two point, the participant received the average of the two scores for that question. If the scores differed by more than 2 points, a third grader scored that question, and the two scores that were the closest of the three scores were averaged together. While five of the questions needed to be scored by a second grader, less than 2% required a third grader. While it is expected for the traditional lab to generate some learning of scientific reasoning, we hypothesize there will be significantly more learning with the AR group. To evaluate if the change in scientific reasoning was significantly affected by AR, we used Bayesian models. This allows us to see the likelihood that the AR materials affected the final scores on the treatment groups' post test. A model was constructed to determine the effect of the interaction between the treatment condition and the test-retest effect for each participant. If there is an interaction, this indicates that the treatment affected the gain in participants learning on the scientific reasoning portions of the post-test. We then used the Bayes Factor, which can quantify the amount of support for the alternate hypothesis, to compare the following hypotheses:

H0: There is not an interaction between the AR group and pre-post test

H1: There is an interaction between the AR group and pre-post test

We also analyzed the responses the participants gave in pairs throughout the study, along with their free response questions to categorize differences

between the AR and the traditional lab, such as data trust and energy observations.

4. Results and Discussion

Between the pre- and post- tests, there was an greater increase in scientific reasoning for those using AR tools, and students were more likely to use energy and uncertainty when analyzing data. In the retention test, taken two weeks after the experiment, showed a digression back to pre-test levels. For that reason, the analysis focuses on the pre- and post-test, and the retention test is discussed in the limitations and future work.

4.1. AR enhanced labs increase scientific reasoning significantly more than traditional lab materials

Our first research question tests the ability for AR to increase participants' scientific reasoning skills. Particularly, we aimed to see if participants better understand measurement error and precision, and how that related to the acceptance of experimental results.

The six pre-post scientific reasoning questions were analyzed. Figure 4 compares the scores for the control and AR groups, pre- and post-test. While the mean pre-test score for the AR group was lower than the traditional group, the AR group achieved a post-test mean score higher than the traditional group.

We calculated a Bayes' Factor of 18.11, giving a strong preference for the AR condition. Therefore, the AR condition likely had an effect on the learning of science reasoning by the participants. We performed a standard effect size analysis (Cohen's d), analyzing the difference between the pre- and post-test scores. The standardized effect size is 0.63 ± 0.57, indicating that 73.47% of the control group is below the AR group's mean. While this effect size has large variability, it is strictly positive. It is therefore highly likely that participants' science reasoning increased more in the AR group than the control group.

By distributing the cognition necessary to conduct the experiment, such as having a period counter and AR protractor, participants may have been able to focus on evaluating their data, leading to more learning. Additionally, by adding an energy graph to the interface, participants are given more information, and may have been able to develop more informed reasoning about their data. For example, some trials had a small energy loss, while

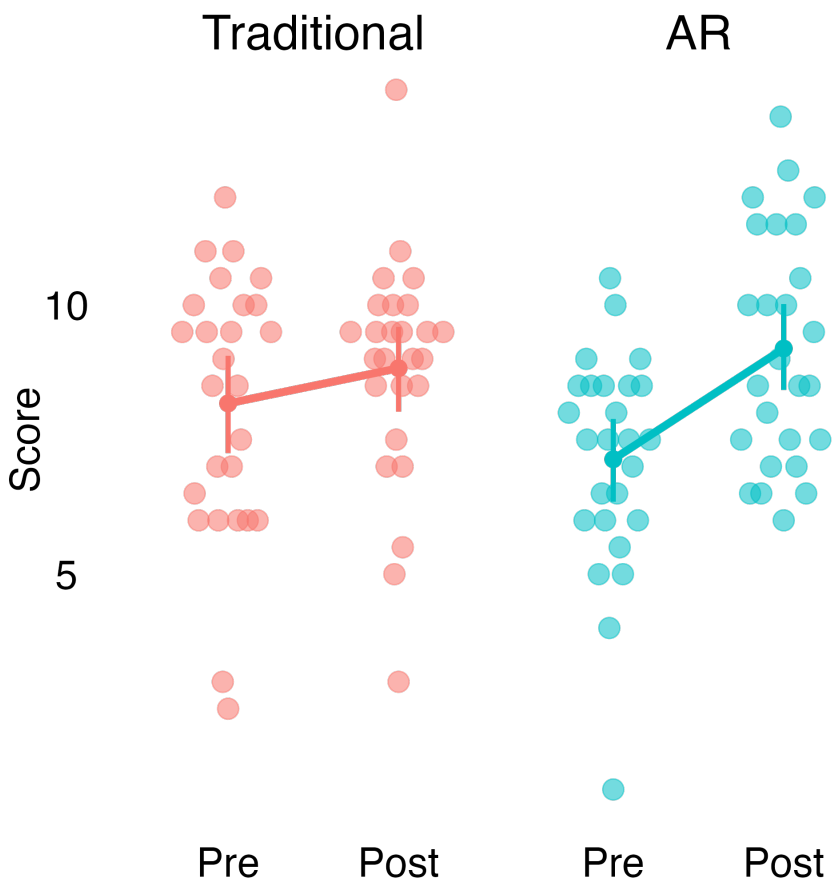


Figure 4: Results of Scientific Reasoning Pre- and Post-Tests

others had a much larger loss. The energy graph allowed the participants to observe in real time the effects of outside factors on data collection, and how they could vary for each experiment. This may have contributed to their understanding of how uncertainty comes into play in each trial individually, and indicated to participants the importance of precision and the effects of uncertainty in evaluating experiments.

4.2. AR Visualizations helped illuminate the uncertainty and energy transfer during the lab.

While students in both conditions in general understood the main idea of the lab (for large angles the period depends on the initial angle), students in

the AR group were more likely to highlight the differences in the data, trust their data, indicate energy as a learned concept/surprising observation.

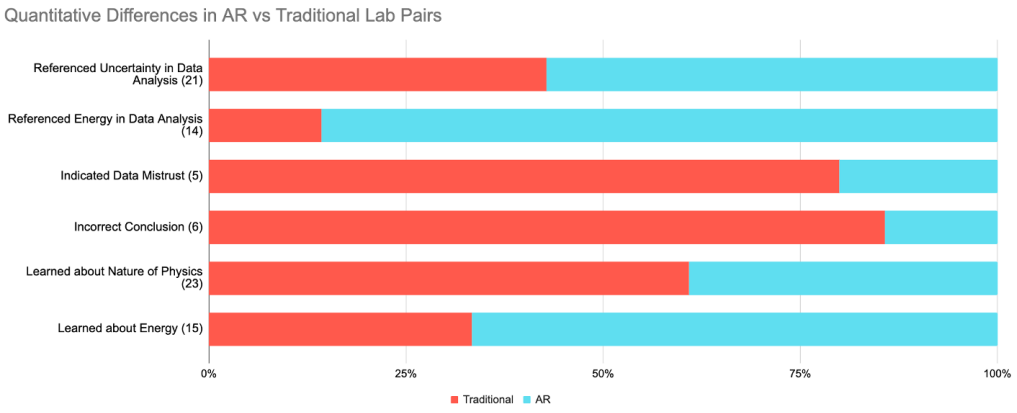


Figure 5: Quantitative difference for AR vs Traditional Lab Pairs. Shown are the percentages of the AR vs Traditional lab pairs for each category. The numbers on the left indicate how many pairs were associated with their repective chategories, while the percentages separates those pairs by condition.

4.2.1. Illuminating Experimental Data Differences

Throughout the experiment, participants were asked questions about their hypothesis and experimental methods. After conducting their first lab (effects of number of swings on the recorded period), we asked what they noticed in the data. A few differences between the conditions are of particular interest.

Those in the AR group discussed energy/dampening differences for the one swing vs ten swings laboratory six times more than the traditional lab group (Figure 5). This makes sense because the AR group had access to the energy loss via AR, while the control group did not. However, while both groups had access to the uncertainty, the AR group was more 33% more likely to notice these differences. While this could be due to the uncertainty variable being placed beside the average period instead of being labeled uncertainty as in the control condition, it could also be an effect of cognitive load. The control group needed to count the period, which may have reduced cognitive resources for evaluating the data.

4.2.2. Data Trust

After the second experiment that tested how the angle relates to the period, participants were asked whether they believed the angle affects the period, and why. A few participants in the control group (31%) did not get that the periods for the two angles were significantly different (Figure 5, fifth row). This may be due to the difference in counting for the control versus AR condition. For those in the control group, hand-counting the periods can sometimes become rhythmic, so that the counting of periods was not based on when they saw the period rise, but according to the rhythm. Therefore, they were more likely to miscount the period, missing the small time difference in period due the initial angle. This was not an issue in the AR condition, as the periods were counted for them. All participants in the AR group were able to get a statistical difference between the periods when conducting the lab.

Participants in the control group indicate some level of mistrust in the data four times more than the AR group (Figure 5). This may be due to the large level of autonomy participants had in the experimental procedures. For the majority of the participants (82%), this was their first time doing a pendulum experiment, and the control condition required they track and control several variables. For this reason, it could have been much harder for those in the control group to have confidence in their experimental procedures. Therefore, when the result comes up differently than expected, they are less likely to trust it.

Additionally, those in the AR group referenced the uncertainty as their basis for a scientific conclusion 2.5 times more than the control group (Figure 5). Therefore, when deciding whether the initial angle affects the period, they based their conclusion on the precision of their measurements and the low uncertainties of their data. This may be been due to their ability to focus on the data analysis and precision more than the control group due to the reduced cognitive load from the AR interface.

4.2.3. Energy Observations

Lastly, participants were asked what they learned, or if they observed anything different from expected at the end of the post test. Those in the AR group were twice as likely to mention energy and its effects on an experiment. This is most likely due to the influence of the energy graph and calculations presented to the participants in the AR group, allowing them to focus and retain information from the data analysis better than those in the control

group. However, participants in the control group were 55% more likely (14 control participants vs 9 AR group participants) to report how the lab helped them to understand the nature of experiments in a physics context. This may be due to the lack of technology included in the control condition, which required participants to be more precise, and therefore felt more similar to the precision and dedication necessary in a lab context.

5. Conclusions and Future Work

The AR-enhanced lab explored the use of augmented reality to teach scientific reasoning in a physics laboratory experiment. The results indicate that the AR lab improved participants' scientific reasoning skills, particularly in understanding the importance of precision in experiments and the role of uncertainty in evaluating experimental results. The AR interface may have allowed participants to focus more on data analysis and interpretation, leading to a better understanding of measurement error and the acceptance of experimental results. Qualitatively the data also supported this, as participants in the AR-enhanced lab were more likely to recognize differences in the data, trust the data, and cite uncertainty as a basis for their conclusions.

Overall, the use of AR materials in the lab setting has shown promising results in enhancing participants' scientific reasoning skills. The use of AR technology can provide participants with real-time data visualization, reducing cognitive load and allowing them to focus on data analysis and precision. This, in turn, can lead to a deeper understanding of scientific concepts and better scientific reasoning skills.

However, it is essential to consider the limitations of this study, such as the controlled lab environment and the short-term retention of knowledge, when interpreting the results. The study was conducted in a lab environment with university participants. Additionally, the effect size has considerable variation (0.63 ± 0.57). While the 95% confidence interval for the effect size never crosses 0, (indicating a statistically significant effect), the confidence interval could be as high as 1.2 or as low as 0.06. It is also possible the observed effects may not replicate in a less controlled environment (i.e., a science lab classroom). Additionally, it is important to note that many of the learning effects were not retained when participants took a retention test two weeks later. Lastly, while a reduction in extraneous cognitive load is hypothesized as the reason for increased learning, it was not explicitly evaluated. Future work could test why learning increases for the AR lab.

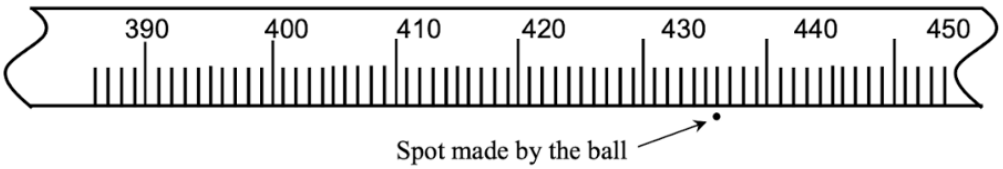
Future research could explore the impact of AR with different populations, such as participants with varying levels of prior physics knowledge and experiences, could provide valuable insights into the broader applicability and effectiveness of AR materials in science education. Further research in this area could contribute to the development of innovative and engaging educational tools that enhance participants' learning experiences and foster a deeper understanding of scientific concepts. Lastly, showing participants an equivalence test as well as a t-test would better show the nuance between statistically significant different, no statistical significant difference, and statistically similar. For the participants that were unable to see a statistical significance in the data, they can look at the equivalence test and see if there is a statistical similarity.

6. Acknowledgments

We thank *Acknowledgment1, masked for blind review* for his help in the measures and statistical analysis, all participants for contributing their time towards this research.

Appendix A. Appendix A: Pre- and Post- Test Questions

1. How would you maximize precision in your experiment?
2. Why would maximizing precision be important?
3. You experimented to test a pendulum's period at different angles. Note there is a high uncertainty (Standard Deviation), but the same average period. What would you do next and why?
    - Conclude periods are the same
    - Conclude periods are different
    - No conclusion, redo/refine the experiment
4. Some physics friends want to determine the distance a ball travels when projected out of a cannon. They shoot the ball from the cannon and obtain the following measurement. Which statement do you agree with most and why?
    - Distance is exactly 436m
    - Distance is approximately 436m
    - Distance is between 435 and 437



- Distance is between 435.5 and 436.5
- Another reason

5. While discussing the result, three ideas arise. Which do you agree with and why?
   - We should shoot the ball one more time again and measure the distance.
   - We don't need to shoot the ball again. We already have the distance.
   - We should shoot the ball several more times and measure the distance each time.
6. They decide to shoot the ball 5 more times and compare it with another group. Which statement do you agree with most and why?

Group A: Average = 435m	Group B: average = 435m
444m	441m
432m	460m
440m	424m

- Group A is better because their readings are 20m apart. Group Bs measurements are 50m apart
- Group A and B are just as good because the average is the same
- Group B is better than A.

References

Bakri, F., Pratiwi, S., Muliyati, D., 2020. Student worksheet with augmented reality technology: media to construct higher order thinking skills of high school students in elasticity topic. Journal of Physics: Conference Series 1521, 022033. URL: https://dx.doi.org/10.1088/1742-6596/1521/

2/022033, doi:10.1088/1742-6596/1521/2/022033. publisher: IOP Publishing.

Brinson, J.R., 2015. Learning outcome achievement in non-traditional (virtual and remote) versus traditional (hands-on) laboratories: A review of the empirical research. Computers & Education 87, 218–237. URL: https://www.sciencedirect.com/science/article/pii/S0360131515300087, doi:10.1016/j.compedu.2015.07.003.

diSessa, A.A., Sherin, B.L., 2000. Meta-representation: an introduction. The Journal of Mathematical Behavior 19, 385–398. URL: https://www.sciencedirect.com/science/article/pii/S0732312301000517, doi:10.1016/S0732-3123(01)00051-7.

Ericsson, K.A., 2017. Protocol Analysis. In: Wilson, R.A., Keil, F.C. (Eds.), A Companion to Cognitive Science, 33, John Wiley & Sons, Ltd, 425–432. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405164535.ch33, doi:10.1002/9781405164535.ch33.

Frank, J.A., Kapila, V., 2017. Mixed-reality learning environments: Integrating mobile interfaces with laboratory test-beds. Computers & Education 110, 88–104. URL: https://www.sciencedirect.com/science/article/pii/S0360131517300325, doi:10.1016/j.compedu.2017.02.009.

Holmes, N.G., 2014. Structured quantitative inquiry labs : developing critical thinking in the introductory physics laboratory. PhD Dissertation. University of British Columbia. URL: https://open.library.ubc.ca/soa/cIRcle/collections/ubctheses/24/items/1.0165566, doi:10.14288/1.0165566.

Holmes, N.G., Bonn, D.A., 2015. Quantitative Comparisons to Promote Inquiry in the Introductory Physics Lab. The Physics Teacher 53, 352–355. URL: https://doi.org/10.1119/1.4928350, doi:10.1119/1.4928350.

Hutchins, E., 1995. Cognition in the Wild. URL: https://direct.mit.edu/books/monograph/4892/Cognition-in-the-Wild.

Huwer, J., Seibert, J., 2018. A New Way to Discover the Chemistry Laboratory: The Augmented Reality Laboratory-License. World Journal of Chemical Education 6, 124–128. URL: http://pubs.sciepub.com/wjce/6/3/4/index.html, doi:10.12691/wjce-6-3-4.

Liu, Q., Yu, S., Chen, W., Wang, Q., Xu, S., 2021. The effects of an augmented reality based magnetic experimental tool on students' knowledge improvement and cognitive load. Journal of Computer Assisted Learning 37, 645–656. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12513, doi:10.1111/jcal.12513. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12513.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia Medica 22, 276–282. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/.

Singer, S., Hilton, M.L., Schweingruber, H., 2005. America's Lab Report: Investigations in High School Science. National Academies Press, Washington, D.C. URL: http://www.nap.edu/catalog/11311, doi:10.17226/11311.

Sweller, J., van Merrienboer, J.J.G., Paas, F.G.W.C., 1998. Cognitive Architecture and Instructional Design. Educational Psychology Review 10, 251–296. URL: https://doi.org/10.1023/A:1022193728205, doi:10.1023/A:1022193728205.

Thees, M., Kapp, S., Strzys, M.P., Beil, F., Lukowicz, P., Kuhn, J., 2020. Effects of augmented reality on learning and cognitive load in university physics laboratory courses. Computers in Human Behavior 108, 106316. URL: https://www.sciencedirect.com/science/article/pii/S0747563220300704, doi:10.1016/j.chb.2020.106316.

Walsh, C., Quinn, K.N., Wieman, C., Holmes, N., 2019. Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. Physical Review Physics Education Research 15, 010135. URL: https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.15.010135, doi:10.1103/PhysRevPhysEducRes.15.010135. publisher: American Physical Society.

Xu, L., Clarke, D., 2012. What Does Distributed Cognition Tell Us about Student Learning of Science? Research in Science Education 42, 491–510. URL: https://doi.org/10.1007/s11165-011-9207-8, doi:10.1007/s11165-011-9207-8.

Yu, S., Liu, Q., Ma, J., Le, H., Ba, S., 2022. Applying Augmented reality to enhance physics laboratory experience: does learning anxiety matter? Interactive Learning Environments 0, 1–16. URL: https://doi.org/10.1080/10494820.2022.2057547, doi:10.1080/10494820.2022.2057547. publisher: Routledge _eprint: https://doi.org/10.1080/10494820.2022.2057547.