Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
ASEE
Paper ID #47019

Study Design and Assessment Framework for Testing Augmented Reality Tools in Engineering Education

Dr. Gimantha N Perera, University of Arizona

Gimantha Perera is a Postdoctoral Scholar/Research Scientist in Systems and Industrial Engineering at the University of Arizona. His professional areas of interest include augmented reality application, healthcare systems engineering, and broadening participation in engineering, particularly at public institutions. Gimantha is focused on assisting a small engineering education task force at UA studying asset-based practices, building communities of practice, industrial engineering education, and the formation of professional identities.

Karen B Chen, North Carolina State University at Raleigh
Dr. Laura Bottomley, North Carolina State University at Raleigh

Dr. Laura Bottomley is the Director of Engineering Education and Senior Advisor to WMEP at NC State University. She has been working in the field of engineering education for more than 30 years, having taught every grade level from kindergarten to engineering graduate school. She is a Fellow of the IEEE and ASEE and has been recognized with the PAESMEM award.

Robert Kulasingam
Emily H Fang, North Carolina State University at Raleigh
Julie Ivy, University of Michigan

Julie Simmons Ivy is a Professor in the Edward P. Fitts Department of Industrial and Systems Engineering and Fitts Faculty Fellow in Health Systems Engineering. She previously spent several years on the faculty of the Stephen M. Ross School of Business a

Study Design and Assessment Framework for Testing Augmented Reality Tools in Engineering Education

Gimantha N. Perera[1*], Emily Fang[2], Robert Kulasingam[2], Laura J. Bottomley[3], Karen B. Chen[2], Julie S. Ivy[4]

[1]Systems and Industrial Engineering, University of Arizona, Arizona, USA
[2]Department of Industrial and Systems Engineering, North Carolina State University, North Carolina, USA
[3]Department of Engineering Education, North Carolina State University, North Carolina, USA
[4]Industrial and Operations Engineering, University of Michigan, Michigan, USA

Abstract

Augmented reality (AR) is gaining traction as a visualization tool for STEM education and professional practice. AR technology can facilitate immersive and interactive learning experiences that cannot be replicated with traditional teaching methods. This methods paper discusses a novel study design and assessment framework that was designed to systematically evaluate an educational AR game-based learning (GBL) application (AR-GBL). Using mixed methods to assess learning outcomes over time, this framework may help address the dearth of longitudinal research on AR in STEM education.

The framework consists of content assessments and adapted surveying tools intended to measure (i) user experience and usability, (ii) deep learning, (iii) knowledge retention, and (iv) educational efficacy. It is presented by this paper in its original context of use: to evaluate the Holographic Applications for Interactive Learning (HAIL) tool. HAIL is an educational AR-GBL application developed to teach conditional probability and the law of total probability to industrial and systems engineering undergraduates. HAIL was piloted with an experimental group and compared to a second control group that received equivalent learning content in a traditional classroom setting.

Overall, the framework was able to provide avenues for (1) assessing the HAIL's usability, (2) comparing learning outcomes and knowledge retention between the experimental and control groups, and (3) generating valuable insights about the efficacy of educational AR-GBL in STEM education. Consequently, this paper reports on the study design and assessment framework utilized for assessing HAILs in combination with future recommendations to provide a methodologically robust foundation for educators and developers seeking to implement AR in learning environments.

Keywords

Augmented Reality, Design, Methods, Assessment, Mixed Methods, Learning Efficacy

Introduction

In STEM education, Augmented Reality (AR) applications may be leveraged to increase student engagement and knowledge retention, effectively mitigating two significant barriers to student learning (Mystakidis et al., 2021). AR environments are highly immersive and interactive simulations (Callaghan et al., 2009; Petrov & Atanasova, 2020) capable of facilitating embodied learning (Alvarez-Marin & Velazquez-Iturbide, 2021; Mystakidis et al., 2022; Yu et al., 2022), which is especially beneficial for STEM education (Macrine & Fugate, 2022). A review of AR's utility in STEM education reported that educational AR applications, specifically those designed with game-based learning (GBL) principles, can teach abstract concepts by providing more tangible and easily understood digital representations (Yu et al., 2022). Furthermore, the review indicated that AR may be especially compatible with instructional strategies and techniques informed by constructivism and experiential learning theory (Yu et al., 2022). Studies investigating AR as an educational tool have reported benefits related to student engagement, focus, motivation, comprehension, cognitive load, and knowledge retention (Bujak et al., 2013; Maisiri & Hattingh, 2022; Radu, 2012; Su et al., 2020). Evidently, AR shows promise in STEM and should be explored in greater depth to determine how it may be applied in different fields (Yu et al., 2022).

Background and Motivation

Despite AR's promise as an educational tool, the affordances unique to the technology remain unclear because of the field's relative immaturity, a concern reiterated by many systematic reviews published over the last decade (Akçayır & Akçayır, 2017; Ibáñez & Delgado-Kloos, 2018; Koutromanos et al., 2015; Wu et al., 2012). Research endeavors on educational AR that integrate and contribute to pedagogical frameworks are especially scarce (Ibáñez & Delgado-Kloos, 2018; Maas & Hughes, 2020), as are works examining learning outcomes over a substantial period of time (Ibáñez & Delgado-Kloos, 2018; Pellas & Vosinakis, 2018), particularly in the context of STEM Education (Vásquez-Carbonell, 2022).

Cheng & Tsai call for researchers to leverage mixed methods analyses (Harvard, 2018) to evaluate the pedagogical efficacy of AR applications for STEM education (Cheng & Tsai, 2013). They also suggest that these studies should orient their designs around salient usability and user experience goals to better understand how learners use and respond to these applications (Cheng & Tsai, 2013). More recent systematic reviews reinforce this sentiment, emphasizing usability's importance in educational technology design, as usability issues may adversely affect learning outcomes (Lu et al., 2022).

Regarding the combination of GBL and educational AR (AR-GBL) specifically, Pellas et al. (2019) argue that AR's technological capabilities complement the pedagogical affordances of GBL. They suggest future research should produce assessment frameworks and examine learning outcomes over time to establish best practices for the design and implementation of educational AR-GBL. Yet, AR-GBL remains relatively understudied, and the methods used to evaluate these applications are inconsistent and often unable to demonstrate educational efficacy empirically (Yu et al., 2022).

Study Design and Assessment Framework Goals

In response to these specific demands for future research initiatives on educational AR, the present paper presents the framework developed for a larger study, Augmented Reality for Engineering Education Advancement (AREEA). This larger study is a research initiative that assists educators in presenting abstract STEM concepts to undergraduate students. Our metrics for success with AR tools are tied to application usability in addition to student's understanding, ease of access to content, knowledge retention, self-efficacy, and enjoyment of their educational experience. The study design and assessment framework use longitudinal analysis and mixed methods instruments to present performance outcomes in tandem with usability metrics. The details of the assessment framework and study design were iteratively designed for robustness and to address the gap in detailed methodologies within the educational AR-GBL application assessment space.

The assessment framework and study design for AREEA are derived from previous study designs developed by the authors and assessment techniques based on engineering education literature. The assessment framework was developed to evaluate the usability of educational AR-GBL applications i.e., Holographic Applications for Interactive Learning (HAIL), via a humanist approach. When coupled with performance outcomes, this approach describes and quantifies the effects of educational AR-GBL applications on users (user experience/usability) in addition to the effectiveness of educational AR-GBL applications in facilitating desired changes to a system. This paper covers i) the iterative process that led to the final study protocol we implemented, ii) the assessment framework for AREEA, and iii) the unexpected adjustments we had to make during rollout.

Framework Design

Conceptual Framework

The development and evaluation of AREEA were primarily framed through the theoretical lens of experiential learning theory. Experiential learning theory broadly establishes how student learning occurs through direct experience, which Kolb (2001) describes as a four-stage, cyclical process (Kolb et al., 2001). These stages consist of (1) concrete experience, (2) reflective observation, (3) abstract conceptualization, and (4) active experimentation (Kolb et al., 2001). Many existing educational AR applications incorporate elements of experiential learning theory because the technology can facilitate meaningful interaction with rich, personalized environments (Goff et al., 2018; Mystakidis et al., 2022). Furthermore, AR is becoming an increasingly promising tool as educators turn to technology and multimedia as a means of promoting conceptual change (Magana et al., 2022; Ozkan & Selcuk, 2015). Informed by the assertions of experiential learning theory, HAILs incorporated gamified elements into their design to engage users, increase their motivation, and produce a student-centered, experiential learning environment that would assist in cultivating conceptual change.

The HAIL is a self-paced supplement to the educational experience. While instructors can curate and guide students through a HAIL module, they are designed with a minimum level of instructional support such that students can access HAIL modules independently and be guided through/ refreshed on a lesson of their choosing. As such, the assessment framework measures and quantifies success through student perceptions and performance. We are aware that this could be measured from the perspective of the teacher (expert service provider), however, this is not the focus of this framework.

We developed an application to supplement conditional probability and the law of total probability (CP & LTP) in an industrial engineering context for AREEA's pilot run. Thus, the HAIL module Mk. 1 is an educational AR-GBL application to visualize the relationships between conditional probabilities and how they can form the total probability of an event given the right conditions.

Instruments

The evaluative framework used: 1) a Post Study System Usability Questionnaire (PSSUQ) for iterative design comparisons; 2) a Structured Bipolar Ladder mixed-methods survey to assess HAIL game elements, design, usability, and learning efficacy; 3) topical content assessments to measure student understanding of the abstract concept being conveyed; 4) a mid-semester self-reported performance evaluation to assess concept retention and learning efficacy; and 5) an end of semester reflective memo detailing the student's educational experience to capture the student's qualitative perception of HAILs and their own performance.

The Post-Study System Usability Questionnaire (PSSUQ) was used as the primary quantitative means of assessing usability. The PSSUQ is an established usability assessment tool for evaluating 2D systems (UIUX Trend, 2016). It has three significant usability categories, namely, system usability, information quality, and interface quality. The PSSUQ assesses most major areas of usability and functionality while offering a limited level of comparability across systems assessed using this instrument (Lewis, 2002; UIUX Trend, 2016). However, 3D applications have more nuanced assessment criteria that need to be assessed.

The structured bipolar ladder (SBLA) is a mixed-methods (qualitative and quantitative) usability assessment method that uses sentiment scores tied to a structured interview to gather user sentiment (Appendix Resource 1). SBLA questions are chosen in advance to ask users about the program's usability, satisfaction, and efficiency (ISO 9241-210:2010) and contextual questions about the learning experience. Interviewees are asked to indicate whether they felt positively or negatively about a topic, the intensity of that sentiment on a scale of one to five (zero for neutral), with one being the lowest intensity, and a description of why the interviewee gave the score they did (Pifarré & Tomico, 2007). The description allows for qualitative verification of the quantitative score being recorded. This not only contextualizes scores but also allows for consistent comparisons of qualitative scores between participants with a higher degree of confidence. SBLA alleviates some of the subjectivity in ranked ordinal scores by tying comparable qualitative feedback to a rating. As SBLA questions on usability are better contextualized than the PSSUQ, they can be used to draw more nuanced conclusions from the PSSUQ, while the PSSUQ allows SBLA scores to be validated.

A demographic survey collected/assessed student information that might affect performance on content assessments, perception, or the AR-GBL experience, including handedness, physical disabilities, sight impairments, previous AR experience, learning preferences, use of corrective lenses, special academic accommodations status, and preferred genre of video games (Appendix Resource 2) (Perera, 2023).

Content assessments were developed to test students' understanding, application, and analysis skills related to CP&LTP. These assessments had a narrative structure similar to the structure of the educational AR-GBL application to ensure that testing matched the content being taught. Questions spanned several knowledge dimensions of Bloom's taxonomy (Krathwohl, 2010). As the HAIL Mk 1 was built for an engineering class, questions on application and analysis using CP&LTP were emphasized (Perera, 2023). Questions in the remember, understand, evaluate, and create knowledge dimensions were also included (Krathwohl, 2010). Point assignments for questions across assessments were proportional to the complexity of, and the work required to answer the question. This ensures knowledge dimensions can be compared across assessments (Appendix Resource 3)(Perera, 2023).

A validated survey on the emotional and cognitive aspects of the student learning experience was adapted for use during the mid-semester review (Appendix Resource 4) (Magana et al., 2022). That tool has been adapted to fit our context. It measures students' understanding of the concept presented in the HAIL Mk 1 (categorized by concept complexity, the student's sense of knowledge enhancement, and clarity of the information provided), engagement, focus, excitement, confusion, frustration, and eureka moments (breakthrough moments in student's knowledge of the subject) (Appendix Resource 4).

A series of short-answer questions and guided memos were developed by adapting previous usability assessment tools through a Delphi process (Linstone & Turoff, 2002) to capture meaningful qualitative feedback from students regarding their learning experience. The short answer questions focused on the students' experience with the AR headset, their level of comfort with the technology, expectations versus reality relating to the AR-GBL experience, and their overall learning experience (Appendix Resource 5). The guided memo was designed for all students, with specific sections for students who used the HAIL Mk 1. The memo aims to ascertain students' expectations versus reality when participating in this experiment; students' levels of confidence, comfort, and confusion regarding CP&LTP; a reflection on the students' level of understanding throughout the semester; a reflection on the student's engagement throughout the semester; and their likes and dislikes regarding their overall learning experience (Appendix Resource 6).

Table 1: Summary table of instruments used in the assessment framework

Instrument	Purpose	Output Type
Post Study System Usability Questionnaire (PSSUQ)	Quantitatively assess usability across three dimensions: systems usability, information quality, and interface quality.	Quantitative Likert-scale scores
Structured Bipolar Ladder (SBLA)	Mixed-methods assessment combining sentiment scoring with qualitative feedback on usability, satisfaction, and efficiency.	Qualitative and quantitative data
Demographic Survey	Collect participant demographics such as age, gender, AR/VR experience, and learning preferences.	Descriptive participant profiles
Content Assessments	Evaluate knowledge retention and understanding using narrative-based and Bloom's taxonomy-aligned questions.	Scores aligned to Bloom's taxonomy levels
Emotional and Cognitive Survey	Measure emotional and cognitive aspects like engagement, focus, frustration, and eureka moments during learning.	Quantitative Likert-scale scores and qualitative feedback
Short Answer Questions	Gather qualitative feedback on AR device comfort, usability, and overall learning experience.	Qualitative short-answer responses
Guided Memo	Provide a reflective account of participants' confidence, comfort, and learning progression.	Qualitative reflective insights

Experimental Plan

The study that used the assessment framework reported in this paper was reviewed and approved by the Institutional Review Board (IRB) of North Carolina State University, ensuring full compliance with ethical research standards. Participants were thoroughly informed about the nature and purpose of the study, provided with detailed consent forms, and assured that their participation was entirely voluntary, with the right to withdraw at any point without penalty. The study design's focus is education, specifically higher education at the undergraduate level, testing AR as a means of improving the student experience. The proposed study design and assessment framework should be transferable to other educational settings, or at least serve as a helpful guide for developing a more tailored study for students in different educational contexts. The assessment framework used a longitudinal observational study design with a test group using an AR practicum and a control group using established teaching practicums. Implementation called for the provision of supplementary classes reviewing conditional probability and the law of total probability (CP & LTP) during the statistics review period of Stochastic Models for Industrial Engineering (an undergraduate course). Participants were to be recruited from this class on a voluntary basis to participate in this experiment. If they meet the eligibility criteria, they would be sorted into two groups using blocking to distribute as many of the demographic markers as evenly as possible between groups. Each group was slated to be given a common supplementary lecture followed by different separate learning practicums. Groups were then assessed using a content assessment, a usability survey, and an interview (if they were in the test group using the HAIL), an assessment of their learning experience, and a reflection on the learning/experimental process.

Depending on the groups students were sorted into, the study protocol called for them to either use the HAIL in lab sessions (Group 1) or receive a traditional form of classroom instruction as their practicum (Group 2). More details on this point are discussed in the following section on implementation. While screening participants for the study, standard demographics, in addition to AR/VR experience, gaming experience, perceived preferred learning style, use of corrective lenses, academic accommodations, and preferred academic accommodations, were recorded. In recording this data, we aimed to investigate whether results would vary between those with additional experiences and students' own perceived learning styles.

Framework Implementation

Maintaining a clear logistical process is an important aspect of implementing the study design effectively. A progress spreadsheet was utilized throughout, allowing for easier participant information, status, and data tracking. This spreadsheet was updated often, as each phase of the study was completed.

All students attended a supplementary practicum (Session 1) to review the concepts of CP & LTP in a traditional classroom setting as they took the initial demographic survey. In Session 2, we screened the participants and used blocking to separate them into two groups as evenly as possible. Group 1 participants were emailed with 75-minute sign-up appointment slots powered by Google Calendar. We created these appointment slots based on the availability of the HoloLens and the Virtual and Augmented Reality Lab. Group 2 participants were sent different emails containing multiple time slots for a second practicum session. While each subject in Group 1 experienced the HAIL during 75-minute time slots as their Session 2 practicum, Group 2 received a second traditional classroom instructional practicum that lasted 1 hour. We distributed the same content assessment to both groups, aiming to measure the students' knowledge of CP & LTP thus far. An emotional/cognitive survey was also completed during this stage.

Content assessments were created to assess the understanding and retention of CP & LTP. Questions covering remembering content to evaluate analyses performed by applying CP & LTP were included to compare across student groups at different levels of learning. Once completed, assessments were completely anonymized (of their names and group numbers) before grading to eliminate any bias that may occur when grading. Grading was done using a formalized key that awarded no credit, partial credit, or full credit based on specific conditions. When an unexpected answer was found, it was discussed amongst graders and added to the key for consistency. Assessments were graded by two independent graders according to the key and were checked for consistency by a third grader. An emotional/cognitive survey was also administered alongside the first content assessment. This survey serves as the incorporation of mixed methods to measure the usability of the HAIL game elements, design, usability, and learning efficacy.

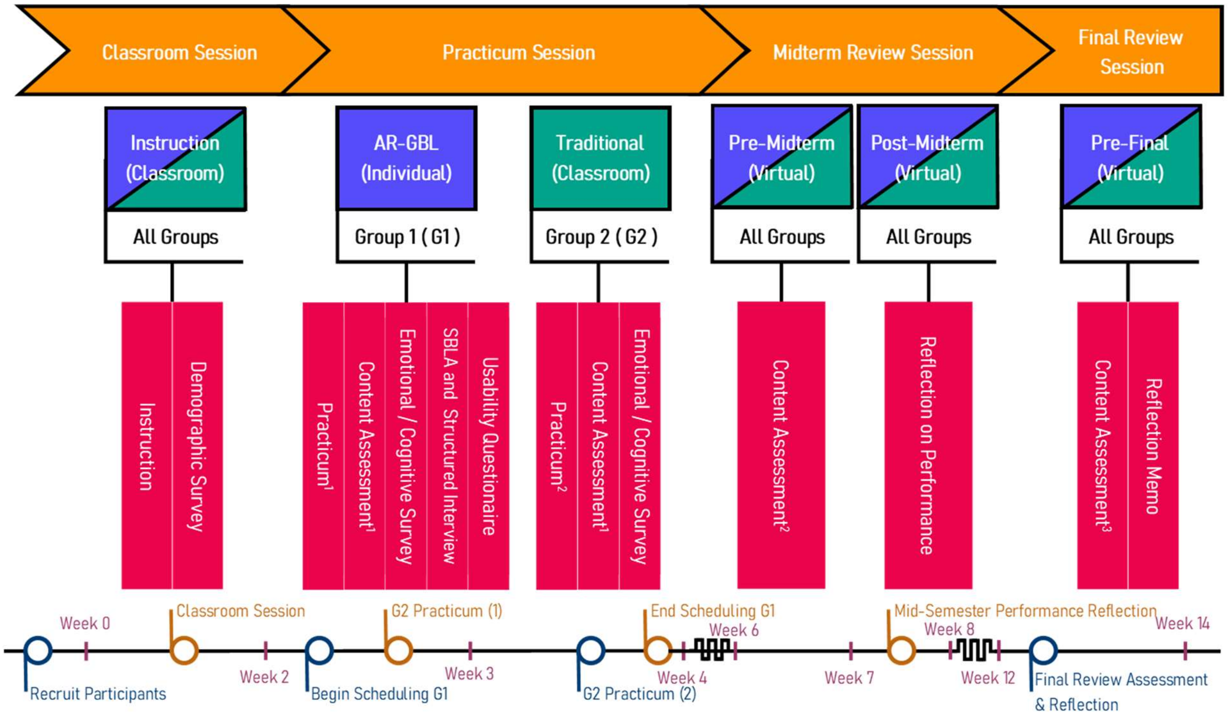

Figure 1: AREEA Experimental Plan and Timeline
The volunteer student participants were split into two groups. Group 1 was the test group that used an AR game to learn and review conditional probability and the law of total probability, while Group 2 had a traditional classroom session reviewing conditional probability and the law of total probability. Both groups took the same content assessments immediately after the review session and provided feedback at the same times later on in the semester.

Two to three weeks after Session 2 concluded, all participants were given a second content assessment covering CP & LTP to complete before taking their in-class midterm exam as a form of review content tied to the midterm. The research team coordinated with the course instructor to include a problem that was directly related to, or utilized CP&LTP, on the midterm. After completing their midterm exam, both groups were given a reflection survey on their performance on the exam and the CP&LTP problem it contained. The final session (3) involved the last content assessment as a form of review content tied to the final exam and a reflection memo before the final exam of the class.

Participants and Groups

Forty-two undergraduate students, juniors, and seniors in industrial engineering, textiles engineering, and statistics participated in the AREEA pilot run. The screening criteria for the study required participants to have not previously taken Stochastic Models in Industrial Engineering (the class the study was partnered with, ISE 362) and be between 18 and 64 years of age, in addition to agreeing to comply with COVID-19 safety protocols. Individuals in the AR treatment group could not have injuries that would limit their ability to make hand gestures such as a thumbs up or the okay sign, be able to wear a large pair of goggles that weighs 1.2lb for at least 10 minutes, not be prone to motion sickness, and not have had Lasik surgery (so that they could fully utilize the AR hardware chosen for the study).

Participant ages ranged from 19-23 years of age, with a gender distribution of 66.6% identifying as male and 33.4% identifying as female. Participate grade point averages (GPAs) ranged from 1.9 to 4.0, averaging 3.36. 39.0% of participants did not have any past experience with AR or VR, 2.4% had used AR before, 12.2% had used both AR and VR before, and 46.4% had used only VR before. Participants were split up via unblinded dispatching, balancing GPA, gender, and race as evenly as possible to form representative blocks (Groups 1 and 2).

Analysis Plan Enabled by this Framework

Understanding, Retention, and Learning Experience

Given the design of this experiment, several analysis paths are available. The design of the AREEA assessment framework erred on the side of gathering more data than less. Thus, while we present a general analysis plan here, a deeper and more nuanced analysis is possible, given the depth of and interaction between the qualitative and quantitative data gathered. Aggregated variables were created for each level of Bloom's taxonomy (Krathwohl, 2010) associated with questions on the content assessment. Thus, in addition to longitudinal data on learning efficacy, confidence levels, engagement with the content, focus during lectures, and cognition, aggregated scores for remembering, understanding, application, analysis, and evaluation (creation was not tested) can also be derived from the data gathered.

There exist 86 comparable variables across surveys and assignments. Seventy-nine quantitative variables (73 directly from data gathering instruments and 6 aggregated variables) are comparable between the AR and control groups. For consistency, all ranked ordinal Likert-scale questions are coded such that higher numbers indicate positivity or improvement. The ordinal scores for the 79 comparable quantitative variables can be compared with parametric t-tests and non-parametric median tests (Kruskal-Wallis). The aggregated scores for Bloom's taxonomy levels for each student can be considered as a single observation when making comparisons between the groups. Any outliers could be noted and examined to identify potential causes for the deviation by referencing related qualitative data from outlying participants.

Longitudinal learning and retention profiles can be generated for each student in the AR and control groups to directly compare the understanding of concepts taught in AR vs. traditional classroom settings. Direct comparisons of understanding captured by assessments immediately after the Practicum Session (Figure 1) allow us to quantify the immediate impact of the different teaching modalities overall and broken down by learning taxonomy stages. Retention can similarly be compared across the groups using the time-staggered assessments. Additionally, data on learning efficacy, confidence levels, engagement with the content, focus during lectures, and cognition can be used to generate a profile for non-performance-related quality of education received using the different teaching modalities.

Standardized Usability

Averages on aggregated PSSUQ categories (the four categories span 19 items on a 1 to 7 Likert-like scale, with 1 being strongly agree [positive sentiment] and 7 being strongly disagree [negative sentiment]) can be used as a standardized report on usability. The convention is to report aggregated PSSUQ scores for System Usability, the Quality of the Information in the System, the Quality of the User Interface, and the single "Overall Satisfaction with the System" (Lewis, 2002) question. These values can be compared directly to, previous iterations of your application, other studies that used similar applications, or averages across studies that use the PSSUQ in your domain.

Higher Resolution Contextual Usability

For higher resolution analysis, distributions of the SBLA (on a (-5) to 5 scale, with -5 being strongly negative sentiment and 5 being strongly positive sentiment) scores can be reported alongside qualitative keywords contextualizing SBLA scores. Individual scores tied to a description can be adjusted to match the highest frequency of scores that have the same description for higher accuracy in quantitative representations of usability and learning (e.g./ if Participant A scores the extent to which the education modality helped them visualize CP&LTP as a 0 with the description of "the visualization tools used to represent probabilities are unintuitive", and a majority of participants that gave the same or similar descriptions instead gave a score of (-1), Participant A's score can be adjusted to a (-1) based on their description). For lower resolution, but more concise analysis, averages for the SBLA sentiment scores can be reported alongside the keywords and themes used to describe those scores to contextualize the AR application's performance. The SBLA explores more specific components of the AR-GBL experience than the PSSUQ, which checks for general usability. There are three SBLA questions in this framework related to learning modalities that can be compared across groups using parametric t-tests or non-parametric median tests (Kruskal-Wallis), depending on sample size. These results can function as direct comparisons of learning environments between groups in addition to usability data on students' educational experiences.

Qualitative Responses

Any qualitative text data gathered via the SBLA (researcher notes on student description of their scores, and direct responses from students) can be transcribed using NVivo (Version 14) www.lumivero.com) or similar qualitative analysis programs to establish codes and themes. Transcribed transcripts, students' free response feedback to open-ended survey questions, and the final reflection memo, once checked for errors, can be coded into separate themed groups by independent analysts and analyzed to identify the emergent themes from the coding of responses. Themes can be generalized labels for recurring ideas or trends in student responses too complex to categorize in a single code. The frequency distributions of codes from responses can be compared across groups for six question categories from the surveys and memos to contextualize the differences between the teaching modalities used in each group. A master code table containing representative quotes defining each code and the ungrouped frequency of codes across all responses can be reported to assist in interpreting the between-group codes if higher resolution is needed. Compound ideas and recurring themes in the students' responses can also be used to support quantitative scores for learning efficacy, confidence levels, engagement with the content, focus during lectures, and cognition.

Discussion and Limitations

Contribution

This paper introduces a robust study design and mixed-methods assessment framework tailored for evaluating an educational AR-GBL application in STEM education, with a particular focus on usability, learning, knowledge retention, and learning efficacy. A key contribution of the study design and framework is their ability to collect longitudinal data, allowing for the evaluation of learning retention over time. This addresses a primary research objective described by a previous systematic review on AR-GBL (Pellas & Vosinakis, 2018) to evaluate the application's efficacy as a tool for tertiary ISE instruction. In doing so, the assessment framework aims to provide future researchers with a principled, theoretically-grounded methodology that fulfills the directive for empirical research established by prior works on AR for STEM education(Cheng & Tsai, 2013; Mystakidis et al., 2022; Pellas & Vosinakis, 2018).

The framework uses a robust suite of assessment instruments, each targeting distinct aspects of the educational experience. These include usability (PSSUQ and SBLA), emotional and cognitive engagement, and knowledge retention using taxonomy-aligned content assessments. The combination of quantitative and qualitative measures ensures that the findings are robust, performance measures are contextualized, and reflective of both the system's usability and its educational impact. The PSSUQ and SBLA, both integrated into the assessment framework, offer a comprehensive and nuanced perspective on usability. The SBLA's mixed-methods design enables the contextualization of quantitative scores through participant narratives, while the PSSUQ offers reliable comparative data that can be compared to future iterations of this educational AR-GBL application or other applications that use the PSSUQ. In the absence of such usability assessments, researchers would be unable to conclude if an application that performed poorly, failed to impact the students' learning experience due to the teaching modality (AR-GBL in the case of this paper), or the poor usability of the application precluding students from benefitting from the unique features teaching modality offers.

The assessment framework not only measures immediate learning outcomes but also tracks knowledge retention over time. By employing Bloom's taxonomy in the content assessments, learning and retention can be reported granularly to assess whether educational AR-GBL and traditional education impact understanding, remembering, application, analysis, or evaluation differently. The framework's inclusion of emotional and cognitive surveys adds an additional success metric to the data by capturing engagement, frustration, and "eureka moments." These insights can be weighed against academic performance depending on the priorities of the research study.

Including overlapping quantitative and qualitative measures ensures the framework's internal validity, enabling a deeper exploration of factors like usability and educational efficacy. With context-specific alterations to content assessments and tailored experience questions, this assessment framework can be adapted to evaluate tools for teaching in broader engineering contexts. As personalized education becomes more prevalent, assessment methodologies such as the framework proposed here will increase in value.

Limitations

The initial study design for project AREEA was to have eligible participants be separated into three separate groups. In addition to the two groups mentioned in the Experimental Plan, we intended to have a third group test a 2D version of the HAILs in a computer lab setting. This group would be a control for GBL, as they would not experience augmented reality but only game-based learning. Unfortunately, due to time constraints, we could not finish the 2D version of the HAIL in time. We also recommend random dispatching to minimize bias among groups in future experiments.

The availability and accessibility of the Virtual and Augmented Reality Lab and the HoloLens 2 should be considered more carefully. While students in Group 1 had the opportunity to experience the HAIL, that experience was contained within a 75-minute appointment session, while students in Group 2 had lecture notes to take home and reference while they completed content assessments. To address this disparity in resources, Group 1 may be given lecture notes based on the HAIL experience or given additional time outside of the 75-minute HAIL session to revisit the application as the study progresses. Thus, controlling for equal access to review materials related to a participant's group's educational modality will result in a fairer comparison of content retention. Regarding forming the groups themselves, while unblinded-dispatching (blocking) can ensure group demographics are balanced, with larger samples, random assignment of participants to groups will ensure less bias in the experimental design.

The order in which assessments were presented can be more aligned with the progression of the study. Content assessments 1, 2, and 3 evolved to be more specific over time, with assessment 3 being the most relevant to the content covered in the Practicum Sessions. Future work should involve reversing the order in which these assessments were administered to allow students to start out with more targeted assessment questions and gradually introduce more application-heavy assessments as students grow more practiced with CP&LTP as they apply it in their courses.

Conclusions

This paper presents a robust study design and mixed-methods assessment framework designed to evaluate an educational AR-GBL application's usability, learning efficacy, and knowledge retention. The study design is contextualized in this paper by its use in assessing the HAIL Mk1, the results of which will be analyzed and validated in future work (the performance results of which will be published elsewhere). The assessment framework addresses methodological gaps in AR educational research by measuring longitudinal data and combining quantitative and qualitative evaluation methods to provide a comprehensive understanding of user experience and educational impact.

Key strengths of the framework include its integration of validated instruments such as the PSSUQ and the SBLA, which collectively provide a detailed and contextual view of usability and learner satisfaction. Additionally, the inclusion of Bloom's taxonomy in content assessments allows for a granular analysis of cognitive outcomes across different levels of learning, from basic understanding to higher-order thinking skills. Emotional and cognitive surveys are also included to capture engagement metrics and the learner's subjective experience, providing a holistic perspective on the educational experience of groups subject to different teaching modalities. Combining the information from the methods in the assessment framework can provide a fair comparison of AR-GBL to traditional classroom experiences across several metrics of success.

While logistical challenges and resource constraints impacted the study's design and deployment, these limitations can be mitigated or designed around in future iterations of the assessment framework. By presenting a detailed and adaptable methodology, this paper adds to the literature on AR-GBL assessment, offering a template for a robust evaluation of the usability and educational effectiveness of immersive learning technologies.

References

Akçayır, M., & Akçayır, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review*, *20*, 1–11. https://doi.org/10.1016/j.edurev.2016.11.002

Alvarez-Marin, A., & Velazquez-Iturbide, J. A. (2021). Augmented Reality and Engineering Education: A Systematic Review. *IEEE Transactions on Learning Technologies*, *14*(6), 817–831. https://doi.org/10.1109/TLT.2022.3144356

Bujak, K. R., Radu, I., Catrambone, R., MacIntyre, B., Zheng, R., & Golubski, G. (2013). A psychological perspective on augmented reality in the mathematics classroom. *Computers & Education*, *68*, 536–544. https://doi.org/10.1016/J.COMPEDU.2013.02.017

Callaghan, M. J., McCusker, K., Lopez Losada, J., Harkin, J. G., & Wilson, S. (2009). Engineering Education Island: Teaching Engineering in Virtual Worlds. *Innovation in Teaching and Learning in Information and Computer Sciences*, *8*(3), 2–18. https://doi.org/10.11120/ITAL.2009.08030002

Cheng, K.-H., & Tsai, C.-C. (2013). Affordances of Augmented Reality in Science Learning: Suggestions for Future Research. *Journal of Science Education and Technology*, *22*(4), 449–462. https://doi.org/10.1007/s10956-012-9405-9

Goff, E. E., Mulvey, K. L., Irvin, M. J., & Hartstone-Rose, A. (2018). Applications of Augmented Reality in Informal Science Learning Sites: A Review. *Journal of Science Education and Technology*, *27*(5), 433–447. https://doi.org/10.1007/s10956-018-9734-4

Harvard. (2018). *Mixed Methods Research*. https://catalyst.harvard.edu/community-engagement/mmr/

Ibáñez, M.-B., & Delgado-Kloos, C. (2018). *Augmented reality for STEM learning: A systematic review*. https://doi.org/10.1016/j.compedu.2018.05.002

Kolb, D. A., Boyatzis, R. E., & Mainemelis, C. (2001). Experiential Learning Theory: Previous Research and New Directions. In *Perspectives on Thinking, Learning, and Cognitive Styles*. Routledge.

Koutromanos, G., Sofos, A., & Avraamidou, L. (2015). The use of augmented reality games in education: A review of the literature. *Educational Media International*, *52*(4), 253–271. https://doi.org/10.1080/09523987.2015.1125988

Krathwohl, D. R. (2010). *Theory Into Practice A Revision of Bloom's Taxonomy: An Overview*. https://doi.org/10.1207/s15430421tip4104_2

Lewis, J. R. (2002). Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human-Computer Interaction*, *14*, 463–488. https://doi.org/10.1080/10447318.2002.9669130

Linstone, H., & Turoff, M. (2002). *The Delphi Method—Techniques and Applications*.

Lu, J., Schmidt, M., Lee, M., & Huang, R. (2022). Usability research in educational technology: A state-of-the-art systematic review. *Educational Technology Research and Development*, *70*(6), 1951–1992. https://doi.org/10.1007/s11423-022-10152-6

Maas, M. J., & Hughes, J. M. (2020). Virtual, augmented and mixed reality in K–12 education: A review of the literature. *Https://Doi.Org/10.1080/1475939X.2020.1737210*, *29*(2), 231–249. https://doi.org/10.1080/1475939X.2020.1737210

Macrine, S. L., & Fugate, J. M. B. (2022). *Movement Matters: How Embodied Cognition Informs Teaching and Learning*. MIT Press.

Magana, A. J., Hwang, J., Feng, S., Rebello, S., Zu, T., & Kao, D. (2022). Emotional and cognitive effects of learning with computer simulations and computer videogames. *Journal of Computer Assisted Learning*, *38*(3), 875–891. https://doi.org/10.1111/JCAL.12654

Maisiri, W., & Hattingh, T. (2022). Integrating game-based learning in an industrial engineering module at a South African University. *2022 IEEE IFEES World Engineering Education Forum - Global Engineering Deans Council, WEEF-GEDC 2022 - Conference Proceedings*. https://doi.org/10.1109/WEEF-GEDC54384.2022.9996240

Mystakidis, S., Berki, E., & Valtanen, J.-P. (2021). Deep and Meaningful E-Learning with Social Virtual Reality Environments in Higher Education: A Systematic Literature Review. *Applied Sciences*, *11*(5), Article 5. https://doi.org/10.3390/app11052412

Mystakidis, S., Christopoulos, A., & Pellas, · Nikolaos. (2022). *A systematic mapping review of augmented reality applications to support STEM learning in higher education*. *27*, 1883–1927. https://doi.org/10.1007/s10639-021-10682-1

Ozkan, G., & Selcuk, G. S. (2015). Effect of Technology Enhanced Conceptual Change Texts on Students' Understanding of Buoyant Force. *Universal Journal of Educational Research*, *3*(12), 981–988. https://doi.org/10.13189/ujer.2015.031205

Pellas, N., & Vosinakis, S. (2018). The effect of simulation games on learning computer programming: A comparative study on high school students' learning performance by assessing computational problem-solving strategies. *Education and Information Technologies*, *23*(6), 2423–2452. https://doi.org/10.1007/s10639-018-9724-4

Perera, G. N. (2023). *The Effects of Technological Innovations in Technical And Nebulous Service Systems [TITANS2].* https://www.lib.ncsu.edu/resolver/1840.20/41487

Petrov, P. D., & Atanasova, T. V. (2020). The Effect of Augmented Reality on Students' Learning Performance in Stem Education. *Information 2020, Vol. 11, Page 209*, *11*(4), 209. https://doi.org/10.3390/INFO11040209

Pifarré, M., & Tomico, O. (2007). Bipolar laddering (BLA): A participatory subjective exploration method on user experience. *Proceedings of the 2007 Conference on Designing for User eXperiences, DUX'07*. https://doi.org/10.1145/1389908.1389911

Radu, I. (2012). Why should my students use AR? A comparative review of the educational impacts of augmented-reality. *ISMAR 2012 - 11th IEEE International Symposium on Mixed and Augmented Reality 2012, Science and Technology Papers*, 313–314. https://doi.org/10.1109/ISMAR.2012.6402590

Su, C., Enrui, L., Yang, S., Changhao, L., Shuhui, L., & Yihua, S. (2020). Probability learning in mathematics using augmented reality: Impact on student's learning gains and attitudes. *Interactive Learning Environments*, *28*(5), 560–573. https://doi.org/10.1080/10494820.2019.1696839

UIUX Trend. (2016). *PSSUQ (Post-Study System Usability Questionnaire)—UIUX Trend*. https://uiuxtrend.com/pssuq-post-study-system-usability-questionnaire/

Vásquez-Carbonell, M. (2022). A Systematic Literature Review of Augmented Reality in Engineering Education: Hardware, Software, Student Motivation & Development Recommendations. *Digital Education Review*, *41*, Article 41. https://doi.org/10.1344/der.2022.41.249-267

Wu, W. H., Hsiao, H. C., Wu, P. L., Lin, C. H., & Huang, S. H. (2012). Investigating the learning-theory foundations of game-based learning: A meta-analysis. *Journal of Computer Assisted Learning*, *28*(3), 265–279. https://doi.org/10.1111/J.1365-2729.2011.00437.X

Yu, J., Denham, A. R., & Searight, E. (2022). A systematic review of augmented reality game-based Learning in STEM education. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-022-10122-y

Appendix

Resource 1: Structured Bipolar Ladder (SBLA)
The SBLA is a mixed-methods assessment that combines sentiment scoring with qualitative feedback on usability, satisfaction, and efficiency. This includes quantitative Likert-scale scores and qualitative feedback for the interaction, HAIL Mk1 application, and learning categories.

Resource 2: Participant Demographic Questionnaire
The participant demographic questionnaire is used to collect participant demographics such as age, race, gender, GPA, AR/VR experience, video game experience, and learning preferences. This creates descriptive profiles of each participant.

Resource 3: Representative Assessment Questions
The representative assessment questions are a sample of the full content assessment questions used to evaluate participants' knowledge retention and understanding using narrative-based and Bloom's taxonomy-aligned questions.

Resource 4: Cognitive Survey
The cognitive survey measures emotional and cognitive aspects like engagement, focus, frustration, and eureka moments during learning. This survey includes quantitative Likert-scale scores and qualitative feedback.

Resource 5: Reflection on HAILs
The reflection gathers qualitative feedback on AR device comfort, usability, and overall learning experience. This reflection includes qualitative, short-answer responses.

Resource 6: Memo on Your Learning Experience
The memo provides a reflective account of participants' confidence, comfort, and learning progression. In particular, it asks participants to describe how their confidence level evolves throughout this study. This memo includes qualitative, short-answer responses.

The following link and QR code connect to the full documents of these resources. Select the "*Augmented Reality for Engineering Education Advancement*" option and scan the "*ReadMe - Guide to Resources In This Folder*" file before using the instruments:
https://ise.ncsu.edu/vr/downloads/