Game-based assessment framework for virtual reality, augmented reality and digital game-based learning

Article *in* International Journal of Educational Technology in Higher Education · June 2023

DOI: 10.1186/s41239-023-00405-6

4 authors:

Chioma Udeozor
Newcastle University
12 PUBLICATIONS **93** CITATIONS
SEE PROFILE

Philippe Chan
CHENEXT Technologies
6 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Fernando Russo Abegão
Newcastle University
19 PUBLICATIONS **253** CITATIONS
SEE PROFILE

Jarka Glassey
Newcastle University
97 PUBLICATIONS **1,412** CITATIONS
SEE PROFILE

Game-Based Assessment Framework for Virtual Reality, Augmented Reality and Digital Game-Based Learning

Chioma Udeozor[1], Philippe Chan[2], Fernando Russo Abegão[1], Jarka Glassey[1]*

[1]School of Engineering, Newcastle University

[2]Department of Chemical Engineering, KU Leuven

*Corresponding author: jarka.glassey@newcastle.ac.uk

Game-Based Assessment Framework for Virtual Reality, Augmented Reality and Digital Game-Based Learning

Abstract

Immersive learning technologies such as virtual reality (VR), augmented reality (AR) and educational digital games offer many benefits to teaching and learning. With their potential to immerse learners in realistic environments and facilitate higher-order cognitive learning, these technologies could be used to complement current classroom pedagogical practices. However, given that these learning environments differ from conventional classroom learning activities, current assessment practices may be insufficient for assessing learning in immersive environments. This paper develops the concept of a game-based assessment framework (GBAF) for educators interested in the assessment of learning in digital games, VR or AR. Importantly, this paper also presents the application of the framework to the design and implementation of assessments for a VR game during the game design phase. Grounded in the principles of Constructive Alignment and the Evidence-Centred Design (ECD) framework, this assessment framework describes the steps to consider for assessments and outlines the components that must be aligned for the design of assessments. To illustrate the application of the GBAF to the design of assessments for immersive learning environments, a stepwise design of assessments for a VR game is presented. The results of the outcome of the assessment of laboratory health and safety competencies of six engineering students is also presented. The GBAF offers simple and useful guidelines for the design of assessments around game tasks. It could serve as a structured basis for educators and researchers to design assessments to measure lower and higher-order cognitive learning in complex immersive environments.

Keywords: Assessment, Immersive learning, Virtual reality, Assessment design

1 Background

The applications of immersive technologies such as virtual reality (VR), augmented reality (AR) and digital games (DGs) span a wide range of industries and sectors, including education, healthcare, tourism, military and aviation. For education, immersive technologies provide situated and interactive learning environments that enhance learning experiences and foster deeper learning (Shute, *et al*., 2017). Grounded in constructivism (Piaget, 1973), immersive learning environments offer interactive space for learners to construct knowledge through meaningful interactions. Some rationales behind the growing interest in immersive technologies as tools to facilitate learning are grounded in their ability to immerse students in realistic interactive environments and foster engagement and motivation (Gee, 2003; Plass *et al*., 2015; Squire, 2003). These qualities of immersive digital worlds have been shown to enhance longer knowledge retention following a VR intervention (Chittaro & Buttussi, 2015). The potential of VR, AR and DGs to improve learning experiences, and foster deeper and higher-order cognitive learning make these tools extraordinary complements to current teaching tools. However, there is still a lack of clarity on the implications of the use of immersive learning tools for classroom pedagogical practices. One of the questions still at the heart of this relatively novel learning and teaching (L&T) approach is how to assess learning (Connolly *et al*., 2009; Razak *et al*., 2012). As a major challenge highlighted by educators, there are calls for the development of new assessment methodologies relevant to immersive learning environments (Kumar *et al*., 2021; Razak *et al*., 2012; Shute, *et al*., 2017). The goal of this paper is to develop the concept and to introduce a robust but simple game-based assessment framework that is educator-friendly and useful for designing and implementing assessments for immersive learning. The application of the framework to the design of assessment for a VR game is also described.

1.1 Assessments with immersive learning technologies

Assessment of learning is at the heart of what educators do and is crucial to the learning process of students. In higher education (HE), assessments are often designed following the principle of Constructive Alignment which requires adequate connections between L&T activities, assessment tasks and intended learning outcomes (Biggs & Tang, 2010). Assessments provide opportunities to measure progress made by learners on the learning objectives, evaluate teaching strategies, as well as pass judgement on the abilities of students based on performance in assessment tasks. Conventionally, assessments are administered as open/closed book exams in the forms of multiple-choice tests, short answer questions, essays, reports or portfolios designed to measure specific learning objectives. However, there are concerns about the validity of the use of some of these assessment types in HE with increasing demand for more authentic assessments that measure higher-order level cognitive processes (Villarroel *et al*., 2019). Assessments are thought to be authentic when they are based on the activities of students that replicate real-world work or tasks (McArthur, 2022; Svinicki, 2004). According to Ashford-Rowe, *et al.* (2014), authentic assessments should be challenging, performance or product-outcome based, must ensure knowledge transfer, enhance self-reflection and self-assessment, be contextual, accurate and encourage discussions and collaboration. These elements considered critical for authentic assessments are not always easy to apply to classroom assessments but can be achieved through problem-based learning/assessment (Merrett, 2022).

Another set of tools that are increasingly used for authentic assessments in HE and professional settings is immersive technologies. The use of DGs for graduate recruitment is on the rise with large multinationals like McKinsey, Shell, Unilever and Deloitte incorporating these into recruitment processes (Bina *et al*., 2021; Kashive *et al*., 2022). In various HE disciplines, there are considerable numbers of studies reporting on the use of DGs for teaching and assessment as reviewed in several systematic reviews (e.g. Gorbanev *et al*., 2018; Udeozor *et al*., 2022). In addition to providing an active learning environment, VR, AR and DGs offer complex learning environments that challenge students, and that are also problem-based, and realistic, offering students the opportunity to apply knowledge and skills to different real-world contexts while supporting collaboration and self-assessments. All these elements, considered critical for authentic assessment, are achievable with immersive technologies. The interactions of students within immersive environments generate large amounts of cognitive and non-cognitive data that are captured in log files, providing an estimate of their knowledge and skills (Shute, Rahimi, *et al*., 2017). With their abilities to simulate realistic environments, immersive technologies provide accurate representations of world-of-work environments that give students a sense of the real-world applications of the knowledge learned in the classroom. These technologies also allow educators to assess students based on how they apply multiple cognitive skills to complex tasks. Importantly, such assessment is based on what *students do* (performance-based) rather than what *they say* or choices made in multiple choice tests. Nonetheless, immersive technologies might be insufficient for measuring certain intended learning outcomes. In such cases, traditional assessment methods such the use of multiple choice tests and essay questions should be considered.

To inform the design of valid assessments for immersive learning environments, a few frameworks and systems like the information trails (Loh, 2012) and the Evidence Centred Design (ECD) framework (Almond *et al*., 2003) have been used. Of these, the most commonly used embedded assessment design framework for immersive environments is the ECD framework (see Jaffal & Wloka, 2015; Kerr & Chung, 2012; Shute & Rahimi, 2021) The ECD framework is also the basis for the design of Stealth Assessment which is popular in the field of game-based learning and assessment (Alcañiz et al., 2018; Min et al., 2020; Smith et al., 2019). Although successfully used to design assessments for immersive environments, the design process is often complex and time-consuming and requires advanced statistical and machine-learning skills (Kim *et al*., 2016; Westera, 2019; Westera *et al*., 2020). It is also most effective for use during the design phase of the immersive learning environment making it less useful for assessment designs around a pre-existing game. These make ECD particularly challenging to use by educators who lack both the time and resources to develop new immersive tools and assessments for classroom use.

For wider adoption and broader impact of immersive learning technologies in education, particularly in HE, the process of designing and implementing assessment for the classroom measurements of learning in VR, AR and DGs must be simple and educator-friendly. There is currently no established assessment design framework relevant to educators interested in the use of immersive technologies for classroom teaching and assessment. Hence, this paper proposes an educator-targeted assessment framework for designing assessment tasks for immersive learning. To do this, this paper aims to answer the following research question:

> ***RQ***: How can educators design and implement assessment in a robust, structured and relatively simple way in order to measure learning in immersive environments such as games, and virtual and augmented reality applications?

2 Conceptual Frameworks

The assessment framework proposed in this paper is the Game-Based Assessment Framework (GBAF). The GBAF is underpinned by two established pedagogical conceptual frameworks: the Evidence Centred Design (ECD) framework (Mislevy, Steinberg, *et al*., 2003) and the principle of Constructive Alignment (CA) (Anderson, 2013).

2.1 Evidence-Centred Design Framework (ECD)

The ECD is a framework for designing assessments based on evidentiary reasoning. It is well known for ensuring the validity of evidence collected for assessment, and for its suitability for measuring complex competencies (Arieli-Attali *et al*., 2019). The ECD enables the linking of competencies assessed with assessment tasks (Mislevy, Steinberg, *et al*., 2003). It is a product of the Educational Testing Services (ETS), the organisation known for developing, administering and scoring standardized tests such as TOEFL and GRE. This framework was developed to enable the design of assessments for a broad range of assessment types, from standardised tests to portfolios and simulation-based tests (Mislevy, Almond, *et al*., 2003). The ECD consists of models that specify the operational elements of an assessment and their interdependencies. As shown in Figure 1, the ECD is made up of the Student Models, the Evidence Models, Task Models, Assembly Model and Presentation Model.

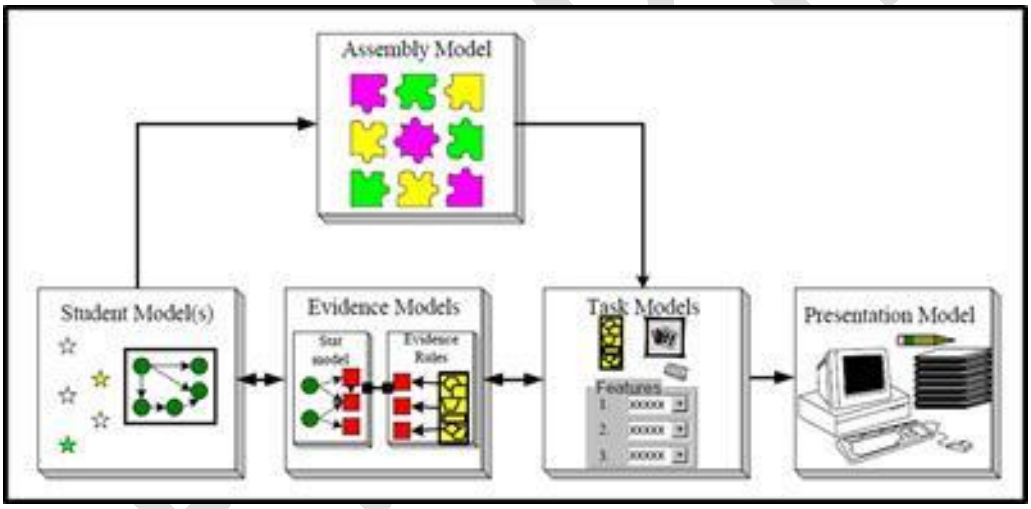


Figure 1: The ECD Framework (adapted from Mislevy, Almond, *et al*., 2003).

Student Models: the student models answer the question '*what are we measuring?*', and are also referred to as competency or proficiency models (Behrens *et al*., 2012; Shute & Ventura, 2013). The student model defines the variables associated with the skills, knowledge and abilities being measured. The values or competencies of students on these variables are initially unknown and are updated at every point in time during interaction with the immersive environment. Each value is expressed by a probability distribution (Mislevy, Almond, *et al*., 2003). In the case of multidimensional student models, Bayesian networks provide graphical language for showing multidimensional associations (Almond, 2015).

Evidence Models: these models answer the question '*how do we measure it?*', and describe how to update information about the student model variables based on evidence produced by students in a given task (Mislevy, Almond, *et al*., 2003). These provide evidence of the competencies of students by linking what they do to the competencies measured. An evidence model is made up of two parts: evidence rules and measurement models. Evidence rules describe how the performance of a student in a given task is summarised from observable variables. In a standardised test, evidence rules guide the response scoring procedure. The measurement model on the other hand provides details of the relationships between the student model variables and the observable variables. It contains statistical models for the accumulation and synthesis of evidence across tasks, and thus guides the summary scoring procedure (Mislevy, Almond, *et al*., 2003). Bayesian Inference Networks are the preferred statistical approach used by many scholars due to their graphical underpinning that aligns well with the principles of ECD (Behrens *et al*., 2012).

Task Models: the task models describe a group of tasks that are presented to students to assess proficiency in a given subject. The task model answers the question '*where do we measure it*?'. Groups of tasks or activities in the task model elicit observable evidence of unobservable competencies in the student model (Shute & Ventura, 2013). In each group of tasks, there are typically several tasks measuring the same variable (Mislevy, Almond, *et al*., 2003). In a standardised test, for instance, each measured competency will generally require different task models and different sets of items/questions would be needed to assess them.

Assembly Model: this model answers the question '*how much do we need to measure?*'. It describes how much evidence or how many tasks are needed to make valid inferences about the students (Almond, 2015). It also ensures that multiple possible forms of tasks presented to students are comparable, especially in computer adaptive testing where students receive unique test forms (Mislevy *et al*., 2012). For automatic scoring, assessment designers must construct a mathematical realisation of the student model and an evidence model for each task option (Kim *et al*., 2016).

Presentation Model: This describes how the assessment tasks are presented to students. It provides specifications for how the other models are initiated in the delivery system (Mislevy, Steinberg, *et al*., 2003).

The applications of the ECD framework to game-based assessment designs have largely focused on the first three models, with less emphasis on the assembly and presentation models. The ECD framework has been used to design unobtrusive game-based assessments sometimes referred to as stealth assessments (Shute, *et al*., 2017). It has been used for game-based assessments in subjects like physics (Kim *et al*., 2016), calculus (Smith *et al*., 2019) and 21st-century skills (Sweet & Rupp, 2012). The design of assessments for immersive environments using ECD is nontrivial, which is a possible reason for the limited adoption and wider preference for traditional assessment types.

2.2 Constructive Alignment

Constructive Alignment is the second principle upon which the GBAF was designed. Grounded in the constructivist learning theory, constructive alignment works on the idea that students learn by constructing knowledge through active engagement in the learning environment (Biggs, 2003). The fundamental principle of constructive alignment is that intended learning outcomes are aligned with learning activities and assessment tasks as shown in Figure 2.

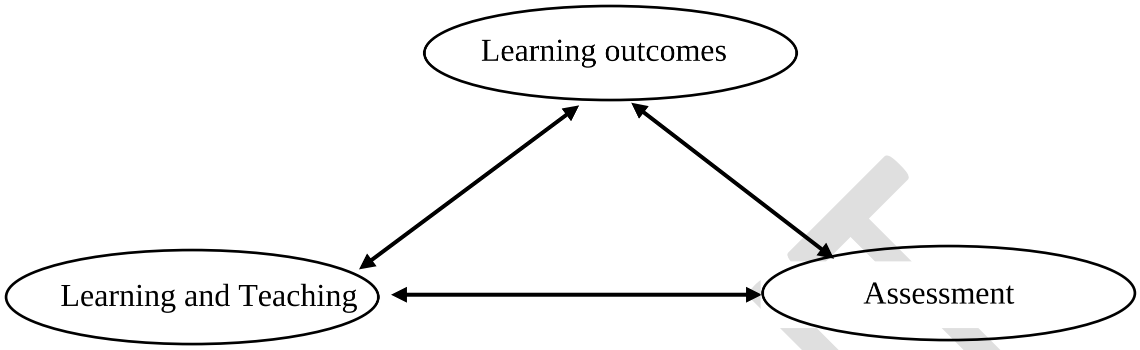


Figure 2: Construct Alignment principle (adapted from Biggs, 1999).

Biggs proposed four steps to ensure the alignment of all components of the system. First, intended learning outcomes (ILOs) have to be defined following an appropriate taxonomy such as the structure of observed learning outcomes (SOLO) taxonomy or Bloom's taxonomy (Biggs & Tang, 2010). It is also considered important to distinguish between the types of knowledge to be assessed. Two fundamental types that need to be distinguished are declarative knowledge, which is not a function of the actions of students, and functioning knowledge which is a result of the actions of the students. Each ILO is to be written with appropriate verbs to indicate standards of achievement (Biggs & Tang, 2010). The second step involves the choice of L&T activities. Although lectures and tutorials, which mostly require passive listening from students, are commonplace in HE, other L&T activities that offer active and engaging environments are recommended. The third step is to design the assessment tasks. Assessment tasks should be aligned with one or more ILOs. These tasks should require students to use the operative verbs in the ILO. An L&T activity that is itself the assessment (such as in games or problem-based learning) offers the best form of alignment (Biggs & Tang, 2010). Biggs and Tang also argue that assessment tasks are best when they are authentic to the discipline. The last of the four steps in designing assessments using the constructive alignment principle is the development of grading criteria.

The constructive alignment principle is widely used by educators in HE for curriculum and instructional designs. It is the main principle required for programme specification, assessment criteria and statement of learning outcomes (Ali, 2018). It provides a logical, effective and familiar principle for assessment design that would allow academic practitioners with or without game-based learning experience to design valid assessments around game tasks. The principles of constructive alignment complement the ECD framework by emphasising strong connections between ILOs, assessment tasks and L&T activities, factors that are crucial to the design of classroom instructional activities. Whereas constructive alignment offers educators familiar guidelines on assessment design, the ECD has a wider application in the field of games-based assessment.

3 Game-Based Assessment Framework (GBAF)

The GBAF draws on the principles of the ECD framework and constructive alignment to provide educators, researchers and game designers with a user-friendly assessment framework that can be applied to immersive learning environments. It proposes a set of steps for assessment design and implementation when using immersive technologies. As shown in Figure 3, the GBAF is made up of five components that must be aligned for a valid assessment.

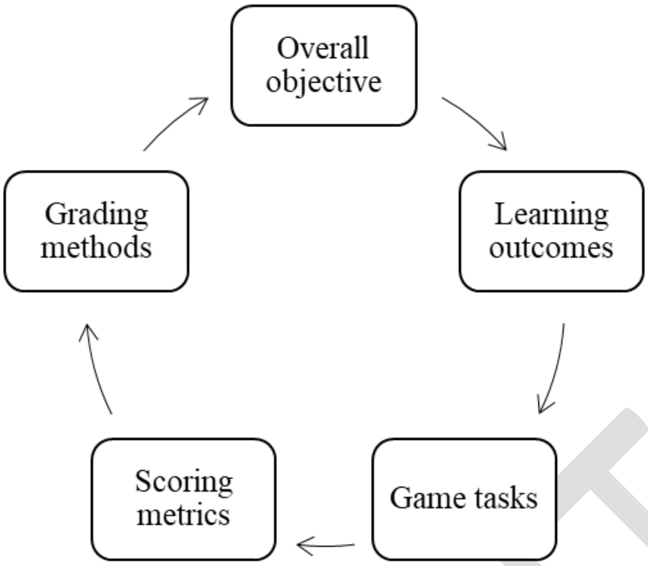


Figure 3: Components of the game-based assessment framework

The GBAF can be applied to AR, VR or DGs. For any of these applications, assessment design begins by outlining the objective of the assessment. These objectives are often broad, encompassing more than one ILO. Next is the careful description of the ILOs to be assessed. Each ILO should be closely linked to the tasks within the immersive environment as shown in Figure 4. It is also crucial that appropriate metrics that can be used to infer the knowledge levels of students from each task are identified. All these components must be aligned for a valid and effective assessment of competencies in immersive environments.

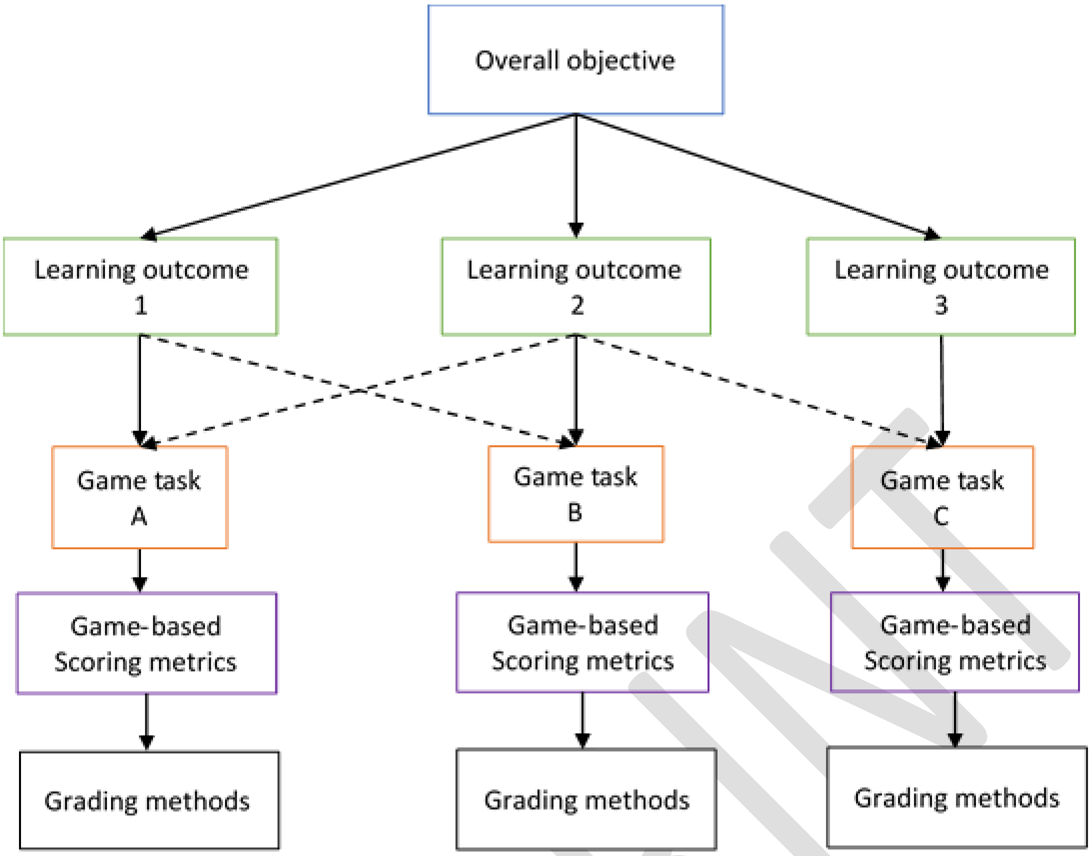


Figure 4: Game-based assessment framework.

3.1 Overall objective

Overall objective describes broadly the purpose of the assessment and in general terms, the competence assessed (see section 4.3 for examples). A clear articulation of the objective of an assessment in an immersive environment is critical to identifying an appropriate immersive tool. This can also provide students with information about the learning and performance expectations, in addition to demonstrating the purpose and relevance of the immersive tool.

3.2 Learning outcomes

This describes the specific knowledge, skills, or expertise that students are expected to acquire from a learning activity. It also informs students about the competencies that will be assessed. For performance-based assessments applicable to immersive environments, statements of ILOs should focus on how students will be able to apply their knowledge in the simulated real-world context. When designing assessments around immersive learning tasks, the ILOs to be assessed should inform expectations in terms of what students ought *to do* to be considered successful. Each ILO should be written with adequate consideration for the available tasks in the immersive environment to be used. This is often less complicated when designing a new environment because of the flexibility to design tasks around ILOs of interest. When an existing game is used, available game tasks should sufficiently address the ILOs outlined.

3.3 Game Tasks

Immersive technologies offer active learning environments where students interact with game elements or collaboratively with other students to complete given tasks. Unlike the conventional assessment tasks that require students to respond to questions, game assessment tasks are performance-based. Game tasks constitute activities that require students *to do*, that is, to perform actions in realistic settings. Game-based assessments are effective for assessing higher-order cognitive processes due to the complexity and authenticity of the environments (Kim *et al*., 2016). One game task could elicit numerous competencies of students and hence could be used to measure more than one ILO. Adequate alignment between the ILOs and game tasks is crucial to the design of a valid game-based assessment. Available game tasks should sufficiently measure the ILOs by requiring students to perform actions that would elicit their level of knowledge on the outlined ILOs (this is illustrated in section 4.1.4).

3.4 Scoring metrics

Another factor to take into consideration in the design of assessments for immersive environments is the game metrics that provide evidence of the knowledge and skills of students. Scoring metrics here refer to those game metrics that can be used to assess the performance of students given the ILOs. VR, AR and DGs collect and store the process or telemetry data containing information about the actions carried out by students and that can be used for assessment purposes. This means that real-time in-game performance data of students can be used to infer competencies. These data can be as general as ‘tasks completed’, ‘time spent on task’, ‘levels completed’, or ‘correct/incorrect answers’, or they can be detailed as ‘numbers of retries’, ‘materials/hints accessed’, ‘locations visited’, or ‘logic sequences of steps’. In a new game, the scoring metrics can be determined and integrated into the game mechanics, however, when a pre-existing game is used, it is necessary to identify available scoring metrics and ascertain whether they cover all ILOs. It is worth noting that the quantity and quality of data collected in immersive environments differ. This is can pose a major challenge for educators as fewer data might limit outcomes assessed, while too much unnecessary data might make it challenging for assessment design and increase data processing times (Loh, 2009). This challenge is common in pre-existing games designed with no assessment considerations from the outset. Meanwhile, when assessment is embedded during the design phase of the game, data collection can be limited to information pertinent to the ILOs assessed.

Nonetheless, identifying relevant metrics in a pre-existing immersive environment of choice could require reverse engineering, that is working backwards starting with identifying available game metrics to determining learning outcomes to measure. Figure 5 outlines a decision-making process that can be useful for determining the appropriateness of an immersive environment for assessment. For an educator interested in using a game or other immersive learning application for assessment, identifying an appropriate application that meets the instructional objectives is the first step. The educator would then have to play the game to determine whether it sufficiently covers the ILOs. If the game meets the requirements, the next step would be to look at the data captured in the game and determine whether they can be used to measure the specified learning outcomes. If the game captures some but not all of the data thought relevant for the ILOs, it would be useful to find out from the game developers if additional data can be captured. If this is not possible and the data collected during gameplay is not useful for the intended purpose, a different game should be sought. Lastly, it is also important to find out whether the gameplay log data can be made available to the educator for analysis. This is essential for grading the performance of students. For more details on the application of the GBAF to a pre-existing game, see Authors (n. d.).

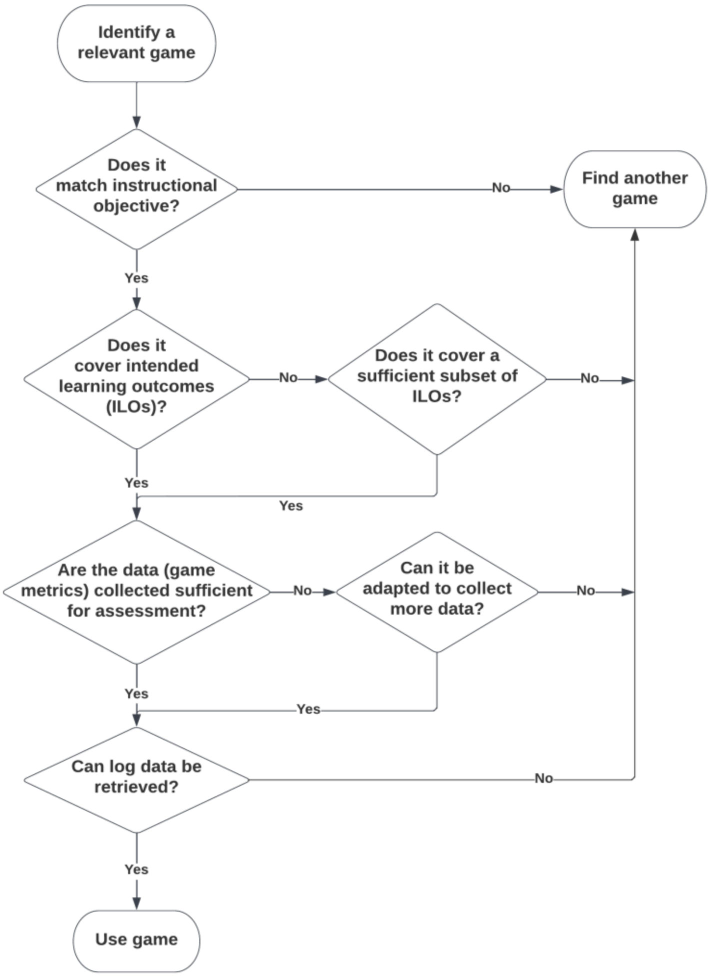

Identify a relevant game
Does it match instructional objective?
No
Find another game
Yes
Does it cover intended learning outcomes (ILOs)?
No
Does it cover a sufficient subset of ILOs?
No
Yes
Yes
Are the data (game metrics) collected sufficient for assessment?
No
Can it be adapted to collect more data?
No
Yes
Yes
Can log data be retrieved?
No
Yes
Use game

Figure 5: Decision-making process for using an existing game for assessment.

3.5 Grading methods

As with traditional classroom assessments, determining grading methods is necessary if grades are to be awarded. The grading methods are the criteria or formula for determining the competency level of students based on their performance on the scoring metrics. They are used to award marks or grades to students. In game-based assessments where competencies are assessed following actions (or inactions) in complex environments, grading criteria must account for these complexities. Instead of simply grading by correct or incorrect answers chosen or given by students, speed of response, the efficiency of solutions, errors made, hints requested and other variables that enable authentic assessments could be used for grading.

One of the last steps of designing assessments for immersive learning environments is a consideration for feedback integration. Feedback is crucial for learning and thus should be an integral part of all assessments. For immersive learning environments, feedback can be immediate, by offering hints or performance metrics to students soon after an action is completed. Where a pre-existing game is used, delayed feedback should be considered soon after gameplay sessions. This can be provided personally to each student as individual feedback or to groups of students during debriefing sessions. However, it has been found that immediate feedback has a higher positive impact on learning compared to delayed feedback (Tsai *et al*., 2015).

4 Application of the Game-Based Assessment Framework to VR LaboSafe Game

To show the practical application of the GBAF to a new immersive learning environment, VR LaboSafe game, an assessments were designed and embedded into a VR health and safety (H&S) game used for chemical engineering education. The assessment tasks in this game were designed during the development of the VR LaboSafe game (https://github.com/PhilippeChan/VRLaboSafeGameDemo) to match the ILOs of interest as depicted in Figure 6. These ILOs were then assessed while students played the game. The VR LaboSafe Game is a training game that utilises VR technology to train students and professionals in chemical laboratory safety risks. The gameplay has problem-solving characteristics requiring players to explore a realistic virtual chemical laboratory to find and eliminate safety risks. Unresolved or incorrectly eliminated safety risks could result in accidents that would negatively affect the performance of players. This VR game provides an environment to train students on safety awareness and practices by simulating dangerous scenarios, which cannot be easily replicated in real-life. The VR LaboSafe game was designed following sound instructional design principles with careful considerations for learning information, motivational elements and VR-induced simulator sickness symptoms. For detailed information on the VR LaboSafe game design, see Chan, Van Gerven, Dubois, & Bernaerts (2021).

4.1 Method

4.1.1 Participants

Seven undergraduate chemical engineering students from Newcastle University, UK took part in this study. Convenience sampling method was used to recruit participants for the study (Creswell, 2011). Participants were made up of six male and one female student in their 2nd year of study. 2nd year students from two faculties were invited to take part in the study but only seven students turned up resulting in the very small sample size. The participants had a general awareness of laboratory safety rules from previous laboratory sessions but had not been assessed on this subject. One participant did not finish the two levels of gameplay for technical reasons, therefore only the data of six participants is presented here. Two of these participants indicated having used VR applications for entertainment.

4.1.2 Procedure

Students were divided into two groups due to the limited number of head-mounted devices (HMDs) available. Each experiment included a description of the aim of the session, a declaration of the potential risks associated with the use of VR HMD and the signing of consent forms by the participants. The Meta Quest 2 HMDs, also known as Oculus Quest 2, with hand controllers which are some of the affordable yet sophisticated HMDS available in the market were used in this study. To familiarise students with the devices, initial training sessions had to be completed before beginning the gameplay and assessment tasks. The training included how to grip objects, open doors, teleport, and how to put on personal protective equipment (PPE). In total, all activities, including training, gameplay and assessment, lasted for up to 2 hours depending on each student.

4.1.3 Data collection

The performance of students was measured using their gameplay data and integrated multiple-choice questions (MCQs). The GBAF was used to design this assessment as shown in Figure 6. For risk spotting, the attempts made by students at correctly taking pictures of potentially risky scenarios were logged and this was used to assess this competence. MCQs were used to assess risk identification and consequence predictions. Students were presented with questions such as "which hazard types are present in this photo?", What are the possible consequences of this risk?". Lastly, for risk elimination and safety protocols, data on the actions of students were collected as detailed in subsection 4.1.4 (scoring metrics). Performance data were collected anonymously with no personal information of students acquired in the process. Necessary ethical approval was obtained from the Ethics Committee of the university before the experiment.

4.1.4 Assessment design/items

As previously mentioned, the assessment design for this VR H&S game happened during the development of the game. The ILOs informed the game design including the tasks and game metrics. Adopting the GBAF, the assessment design components of the VR LaboSafe Game are shown in Figure 6.

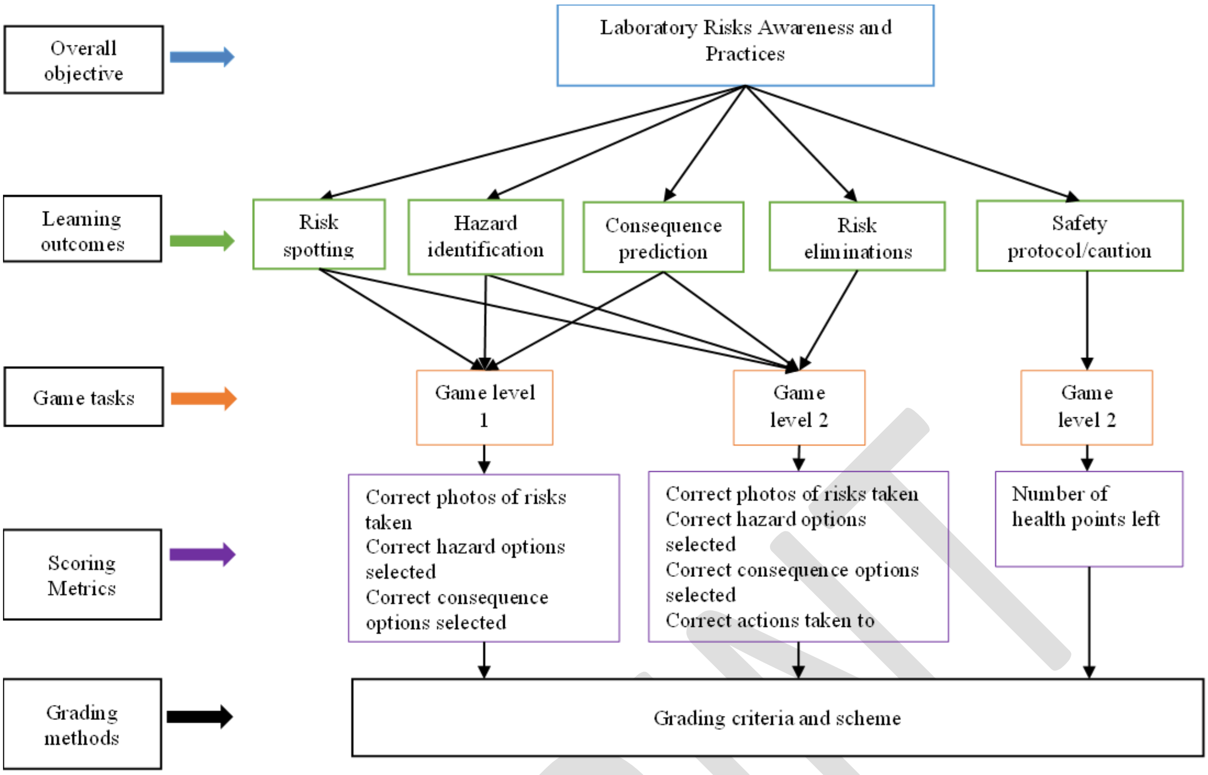

Overall objective
Laboratory Risks Awareness and Practices
Learning outcomes
Risk spotting
Hazard identification
Consequence prediction
Risk eliminations
Safety protocol/caution
Game tasks
Game level 1
Game level 2
Game level 2
Scoring Metrics
Correct photos of risks taken
Correct hazard options selected
Correct consequence options selected
Correct photos of risks taken
Correct hazard options selected
Correct consequence options selected
Correct actions taken to
Number of health points left
Grading methods
Grading criteria and scheme

Figure 6: Assessment design for a VR laboratory health and safety game. (Health point is described below)

Overall objective: This VR game aimed to provide an interactive and authentic learning context for students to learn about the identification and mitigation of laboratory risks in order to improve their knowledge and skills in laboratory safety practices.

Learning outcomes: The acquisition and demonstration of laboratory risk awareness and practices, the focus of this game, is broken down into measurable and achievable ILOs. These ILOs informed the design of tasks that measure the laboratory risk competencies of students. The intention was to measure higher-order cognitive process dimensions. Five ILOs were assessed as outlined below. The cognitive process level and knowledge dimensions following Bloom's taxonomy are respectively indicated too:

1. Students should be able to *distinguish hazardous conditions from non-hazardous* conditions in the laboratory (Risk spotting) –Analyse; Factual

2. Students should be able to *evaluate the hazard types* associated with each condition identified (hazard identification/classification) –Evaluate; Conceptual
3. Students should be able to *infer the consequences* of laboratory hazards (consequence prediction) –Understand; Conceptual
4. Students should be able to *apply measures to minimise or eliminate* laboratory safety risks (risk elimination) –Apply; Procedural
5. Students should be able to *safely execute* laboratory tasks to minimise risks to themselves (safety protocols) –Apply; Procedural

Game tasks: The VR game at the time of this study had two levels that teach and assess laboratory risk awareness and practices. In the first level of the game, students had to search and correctly spot five safety risks in the chemical laboratory. There were always more than five risks in the laboratory and these were randomly presented to students. To complete the task, students were required to take pictures of the identified risky scenario using the virtual tablet provided. Examples of the safety risks presented were a flammable chemical product in a non-explosion-proof fridge, and a laboratory technician performing hazardous experiments in a fume hood where the sash was fully open. After spotting the risks, students were presented with MCQs such as: “Which hazard types are present in this safety risk?”, and “What are the possible consequences of this risk?”. Answering these questions correctly would infer knowledge of hazard types and consequences. In the second level of the game, students were presented with similar tasks as in the first level. Additionally, students were required to eliminate the risks identified. This could mean moving the flammable products from a standard to an ignition-free fridge, or dressing up a laboratory technician in appropriate PPE. Students could choose to skip this risk-elimination step if they could not find a solution, however, this would affect their scores. The game tasks were completed when all five safety risks are spotted in the first level and when all safety risks are (correctly or incorrectly) minimised or skipped in the second level.

Scoring metrics: The scoring metrics selected to assess the ILOs were seamlessly woven into the game tasks in such a way that the gameplay activities of students provided evidence of their competencies. For risk spotting, ILO1, the percentage of correct photos taken was the scoring metric utilised. For ILO2, hazard identification, students were expected to evaluate the spotted hazard and determine what kind of hazard it possessed – chemical, physical, environmental, ergonomic or health hazard. Therefore, to assess this ILO, the percentage of correct hazard options selected from a list of possible options was the scoring metric used. In addition to identifying the hazard types, students were required to infer the potential consequences of such risks. Similar to ILO2, for ILO3, the percentage of correct consequences selected from the list of potential options was the scoring metric used. Furthermore, to mitigate potential risks in the laboratory, students were expected to make changes to the environment where needed. This involved moving objects, closing fume hoods, or dressing up technicians in the right PPE. For ILO4, students were assessed based on the percentage of correctly mitigated risks. For ILOs 1, 2, 3 and 4, incorrect answers and actions of students were also taken into account as shown in Table 1. This was done to account for guesswork from students and to evaluate the accuracy of actions. Lastly, as would happen in a real-world laboratory, interaction with dangerous chemicals could be unsafe without the right PPE. To simulate this effect in the VR game, a Health Point (HP) system was incorporated into the second level of the game. The HP of students at the beginning of the game was 5 (100%) but this value reduced with every inappropriate exposure to dangerous chemicals when appropriate PPE is not worn. Hence, to measure the observance of safety protocols in the laboratory environment, the percentage of HP left after the tasks were determined. The scoring of HP is important as it mimics real-life consequences of poor laboratory practices without causing any harm to students. This ILO was only assessed in the 2nd level of the game as it has been shown that the behaviours of students at the beginning of gameplay are exploratory and prone to mistakes compared to their behaviour subsequently (Udeozor *et al.*, 2021).

Grading methods: With all other components of the GBAF outlined, the grading criteria were developed. Given that this assessment was embedded during the design of the VR game, the formulas for calculating and grading the performance of students were also incorporated. Doing so made it possible to provide immediate feedback to students by presenting performance scores during gameplay. The scoring methods used to calculate performance on each measured learning outcome are presented in Table 1.

Table 1: Grading methods for each assessed learning outcome.

Learning outcomes	**Scoring methods**	**Description**
ILO 1	$R_{spot}=\frac{c}{c+i}$	*c* is the number of correct photos taken and *i* is the number of incorrect photos taken
ILO2	$R_{hazard}=\frac{h}{H+i}$	*h* is the number of correct hazards types identified, *H* is the total number of correct hazard options and *i* is the number of incorrect options selected
ILO3	$R_{conseq}=\frac{k}{K+i}$	*k* is the number of correct consequences chosen, *K* is the total number of correct options available and *i* is the number of incorrect options selected
ILO4	$R_{eliminate}=\frac{l}{L+i}$	*l* is the number of correctly eliminated risks, *L* is the total number of risks presented and *i* is the number of incorrect actions taken
ILO5	$R_{HPoint}=\frac{hp}{5}$	*hp* is the number of health points remaining after completing the level

Students were also graded for each level of the game completed to determine their proficiency levels. The following equations were used as scoring methods to calculate the performance of students on each game task /level:

Level 1: R_{spot} ×0.4 + R_{hazard} ×0.4 + $R_{consequence}$ ×0.2 …………………………………………….(1)

Level 2: R_{spot} ×0.3 + R_{hazard} ×0.3 +R_{conseq} ×0.2 + $R_{eliminate}$ ×0.1 + R_{HPoint} ×0.1 ……………...(2)

The coefficients in each scoring method/equation represent the weight attributed to the assessed outcomes. For Level 1 of the game, weightings of 40%, 40% and 20% were given to risk spotting, hazard identification and consequence prediction, respectively. For Level 2, risk spotting, hazard identification, consequence prediction, risk elimination and HP were weighted 30%, 30%, 20%, 10% and 10%, respectively. These weightings were applied specifically for this experiment taking into consideration the educational level of the participants and the level of knowledge expected of them at the time. The weightings could vary depending on the aim of the game, the knowledge level/expectations of the students and/or the goal of the assessment. These weightings were programmed in the back-end of the VR game during its development to allow for automatic grading of the gameplay actions of students. Potentially, these weightings could be easily altered by educators with the use of a graphical user interface (GUI). These will make it simpler for an educator to adapt a given game to different groups of students. In the case of the VR LaboSafe game, GUI was not implemented at the time of this study.

Finally, the grading scheme was drafted for each ILO and overall performance on each level of the game. The scoring of performance on the learning outcomes is considered a formative assessment as it provides students with information about their performance on each ILO. On the other hand, the grading of each level of the game acts as a summative assessment allocating ratings or scores to students given their competency levels. The grading scheme used for the VR game is shown in Table 2.

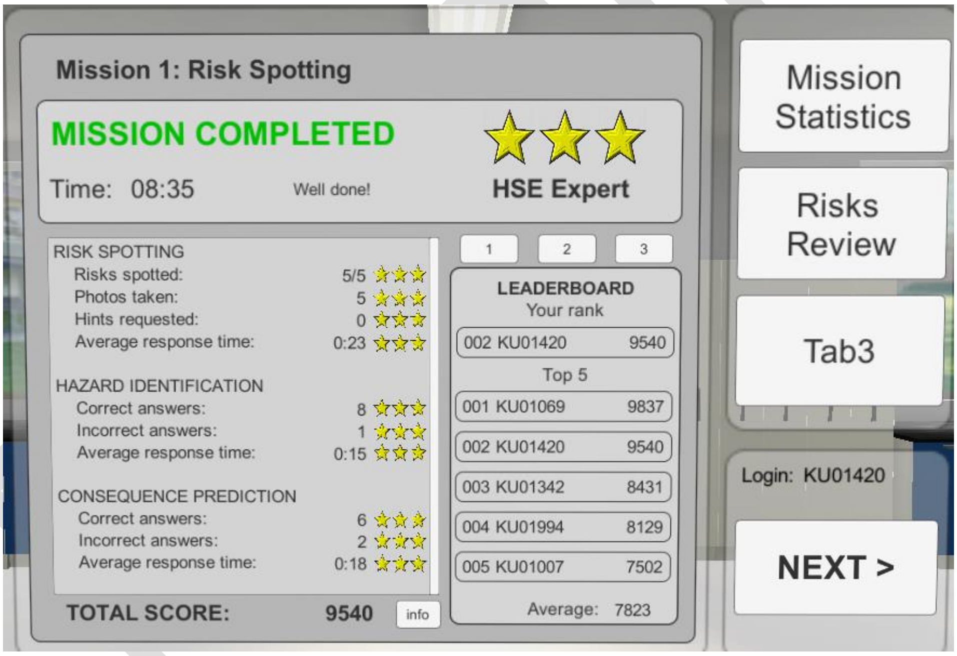
Table 2: Grading scheme for the VR laboratory H&S game

	Performance Rating		
	Novice	***Competent***	***Expert***
Risk spotting	R_{spot}= 0-40%	R_{spot}= 41-80%	R_{spot}= 81-100%
Hazard identification	R_{hazard}= 0-40%	R_{hazard}= 41-80%	R_{hazard}= 81-100%
Consequence Prediction	R_{conseq}= 0-40%	R_{conseq}= 41-80%	R_{conseq}=81-100%

Risk elimination	$R_{eliminate}$ = 0-40%	$R_{eliminate}$ = 41-80%	$R_{eliminate}$ = 81-100%
Safety protocols/caution	R_{HPoint} = 0-40%	R_{HPoint} = 41-80%	R_{HPoint}= 81-100%
Level 1	0-40%	41-80%	81-100%
Level 2	0-40%	41-80%	81-100%

4.2 Data Analysis

At the end of each level of gameplay, students were presented with summaries of their performance on scoreboards as shown in Figure 7. These show their performance and grades (Novice, Competent or Expert) on all measured ILOs as well as on each level of gameplay following the criteria outlined in Table 2.



Mission 1: Risk Spotting

MISSION COMPLETED ★★★

Time: 08:35 Well done! HSE Expert

RISK SPOTTING
- Risks spotted: 5/5 ★★★
- Photos taken: 5 ★★★
- Hints requested: 0 ★★★
- Average response time: 0:23 ★★★

HAZARD IDENTIFICATION
- Correct answers: 8 ★★★
- Incorrect answers: 1 ★★★
- Average response time: 0:15 ★★★

CONSEQUENCE PREDICTION
- Correct answers: 6 ★★★
- Incorrect answers: 2 ★★★
- Average response time: 0:18 ★★★

TOTAL SCORE: 9540 info

1 2 3

LEADERBOARD
Your rank

002	KU01420	9540

Top 5

001	KU01069	9837
002	KU01420	9540
003	KU01342	8431
004	KU01994	8129
005	KU01007	7502

Average: 7823

Login: KU01420

Figure 7: Scoreboard presented to students at the end of gameplay.

For a detailed insight into the performance of students, log files were collected and analysed. Since this assessment was designed and embedded into the game to measure specific learning outcomes, the data collected and stored were semi-structured and contained only relevant information. Data were collected for each student on each level of the game. To analyse the data, the raw .xml files were converted to readable table format (.csv) using Python. The outputs were tables containing rows of anonymised IDs of participants and columns of scores on all scoring metrics. In addition to the scoring metrics described in section 4.3.4 above, additional data that were considered relevant to the understanding learning process of students were collected and these included time on task and hints requested. Although relevant to understanding and enhancing learning in the game through immediate feedback, in the case of hints, these were not considered appropriate for assessment and grading as they can be affected by factors such as familiarity with VR devices, in the case of time on task, and learning style when it comes to hints used. At this time when much information about the VR experience of the students and their learning styles was unknown, incorporating these metrics into the assessment could unfairly affect the performance and grades of students. The speed of task completion was not considered releveant at this time. However, time could be considered relevant by an educator based on the measured outcomes, the level of knowledge and expertise of the students and the goal or purpose of the assessment. Using the grading methods in Table 1, the performance of students on all measured outcomes was computed as shown in Table 3.

4.3 Results and Discussion

From the results shown in Figure 8, the overall performance of students was better in Level 2 compared to Level 1. From the results, in Level 1, 67% and 23% of the students performed at novice and competent levels respectively, while in Level 2, 100% of the students performed at competent levels. This suggests that the VR game may have been effective for the acquisition of knowledge and skills on risk awareness and practices. This outcome is consistent with others that found digital games, VR and AR effective for improving the performance of engineering students (Bolkas *et al.*, 2022; Criollo-C *et al.*, 2021; Perini, Oliveira, *et al.*, 2018; Rossado Espinoza *et al.*, 2021; Urbina Coronado *et al.*, 2022).

The overall performance of students on the subject can be said to have generally improved for those students that spent more time in Level 1 (which had fewer tasks) compared to the time spent completing tasks in Level 2. This may indicate more extensive engagement with the game potentially leading to deeper knowledge and skills development. Similar observations were made for students that requested fewer hints in Level 2 than in Level 1 of the game.

Table 3: Performance of students on measured outcomes and levels of gameplay.

Level 1								Level 2								
ID	***Time on task (secs)***	***Hints***	***Risk-spot***	***Risk-hazard***	***Risk-consequences***	***Overall score***	***Overall Rating****	***Time on task (sec)***	***Hints***	***Risk-spot***	***Risk-hazard***	***Risk-consequences***	***Risk-eliminate***	***Risk-HPoint***	***Overall score***	***Overall Rating ****
CE1	1079	5	0.109	0.273	0.5	0.253	**N**	1147	2	0.556	0.167	0.444	1	0.2	0.481	**C**
CE2	1005	2	0.192	0.444	0.5	0.355	**N**	1109	2	0.625	0.429	0.625	1	0.2	0.624	**C**
CE3	1368	1	0.357	0.545	0.857	0.532	**C**	748	1	1	0	0	0.8	0.2	0.5	**C**
CE4	1549	5	0.063	0.375	0.875	0.350	**N**	830	3	0.217	0.5	0.5	1	0.2	0.457	**C**
CE5	337	1	0.455	0.5	0.5	0.482	**C**	677	1	0.5	0.571	0.75	0.8	0.2	0.621	**C**
CE6	1133	1	0.156	0.4	0.8	0.383	**N**	1211	2	0.278	0.5	0.75	0.8	0.2	0.511	**C**

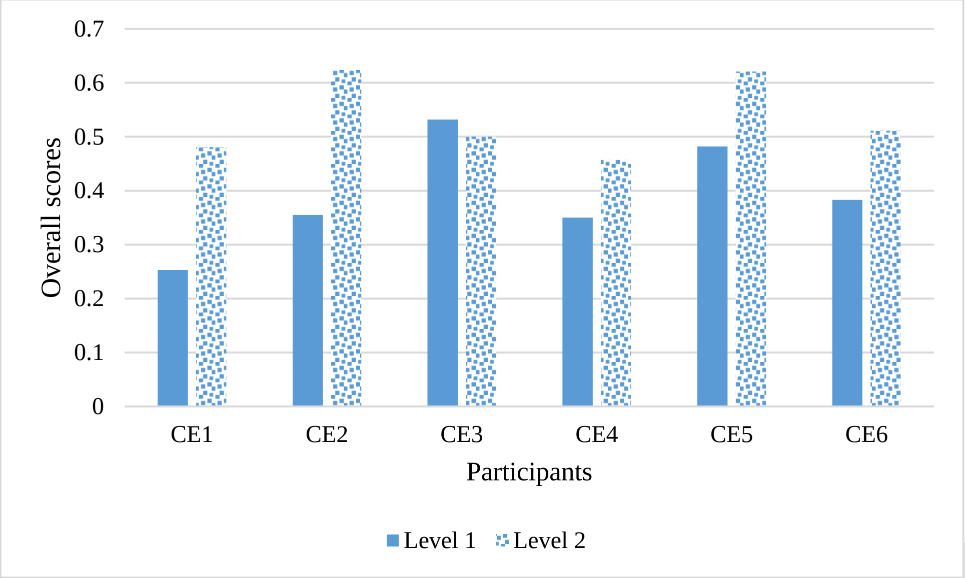


Figure 8: Overall performance of students in Level 1 and Level 2 of gameplay.

The highest improvement in the performance of students was seen in risk spotting with over 100% improvement in the scores of most of the students as highlighted in Table 3. However, for hazard identification and consequence prediction, 50% and 67% of the students performed worse in Level 2, respectively. This unexpectedly poor performance could be due to a lack of conceptual understanding of laboratory hazard types and their effects however, the small sample size of our study limits the conclusions that can be drawn based on these findings. DGs have been found to have the highest influence on procedural and factual knowledge compared to conceptual knowledge (Perini, Luglietti, *et al*., 2018; Perini, Oliveira, *et al*., 2018), which could be the reason for these outcomes. Nonetheless, the overall performance of students on the subject can be said to have generally improved for those students that spent more time in Level 1 (which had fewer tasks) compared to the time spent completing tasks in Level 2. Similar observations were made for students that requested fewer hints in Level 2 than in Level 1 of the game.

5 Conclusions and Limitations

With educators in mind, the current paper presents the Game-Based Assessment Framework (GBAF) grounded in the Evidence Centred Design (ECD) framework (Mislevy, Almond, *et al*., 2003) and the principles of Constructive Alignment (Biggs, 2003). This assessment framework offers a relatively simpler alternative to the ECD framework in response to the research question. Comprising five elements, the GBAF ensures adequate alignment between the learning outcomes, game tasks, scoring metrics, and grading methods for a valid and effective assessment of instructional objectives. It is educator-friendly in that it breaks down the assessment design process into steps that are familiar to educators. The GBAF is designed to facilitate the specification of assessments for VR, AR or digital games as demonstrated in its application in the VR LaboSafe game. The assessment design process with the GBAF is intended to be an easy, efficient, structured approach to designing assessments and measuring learning in immersive environments.

Compared to the ECD framework which is considered laborious and complex to use (Kim *et al*., 2016; Wallner & Kriglstein, 2012), the GBAF is less complex to use and requires no advanced mathematical/psychometrics skills. The GBAF applies similar principles and steps that educators apply to the design of conventional assessment tasks. In the case of the ECD and Stealth Assessment, designing assessments for measuring the proficiencies of students in laboratory health and safety awareness, a Bayesian Network would be created for each level of the VR LaboSafe game. Probability models will be developed to infer the proficiency levels of students based on their actions in the VR environment. These activities would often require the use of sophisticated machine learning software such as Netica (Shute & Rahimi, 2021). The initial values of the inferred competencies (prior probabilities) of students are then automatically updated as students progress through the VR application. This process is non-trivial and the demand on educators high given the skills required for it.

Nonetheless, the ECD framework is advantageous for large-scale testing as it automatically updates information about the proficiencies of students, requiring no additional grading inputs from educators. In the case of small-scale classroom assessment, applying the ECD to the design of assessments for classroom use would require extensive preparation time that offers little or no additional benefits to educators. In such cases, the GBAF would be a potentially better alternative to the ECD framework given that requires lesser preparation and implementation times, and no additional advanced skills from educators. This ease of use of the GBAF could enhance the adoption of immersive technologies-enabled authentic assessments in HE. The practical implication of the successful development and application of the GBAF is that it could potentially serve as a structured basis for researchers and educators interested in designing and implementing assessments of learning in immersive environments. This should also promote the adoption and use of immersive technologies for formal education. The presentation of the application of the GBAF to the design and implementation of assessment for a VR game shows the structured and simple steps educators can follow to design assessments for immersive learning. Starting with the overall objectives and ILOs to the grading methods to be used, the GBAF highlights relevant steps for assessment design and its application to a VR game presented in this paper offer a step by step approach that can be followed by anyone interested in assessments with immersive applications.

Although the purpose of this paper is not to determine the efficacy of VR for learning, the performance of students in the VR LaboSafe game was promising. The results showed that the overall performance of students in the assessed learning outcomes in Level 2 of the game was better than their performance in Level 1. This is particularly interesting given that for a majority of the students, it was their first time interacting with a VR device. Additionally, the performance of students in other metrics such as time on task and hints requested were seen to improve in Level 2 compared to Level 1. These findings are consistent with the results of other studies on VR for engineering education (Bolkas *et al*., 2022; Rossado Espinoza *et al*., 2021). VR applications are increasingly being explored for education due to the belief that they promote contextual and experiential learning that is considered beneficial for knowledge and skills acquisition (Radianti et al., 2020). The outcome of this research, although limited by the very small sample size, indicates that VR can lead to performance improvement when used for learning. To conclude, the GBAF introduced in this paper could provide a consistent and structured basis for designing assessments for immersive learning environments. Designed for educators, this framework can also be used by researchers, game designers and non-game experts. Developed at the conceptual level and not for one specific game environment, the GBAF can be applied to the design of assessments for mobile and computer games, as well as VR and AR games. One limitation of this paper is the lack of evaluation of the framework by educators. Future studies should consider carrying out studies to evaluate the usability of the GBAF, preferably with educator participants who have some experience using immersive technologies for teaching. Future works should also aim to test the robustness of the framework by applying it to the design of external assessment forms and for digital games and AR applications.

Declarations and ethics statement

Data availability statement: The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Funding: This project has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020 under Grant Agreement No 812716.

Acknowledgment: N/A

References

Alcañiz, M., Parra, E., & Chicchi Giglioli, I. A. (2018). Virtual Reality as an Emerging Methodology for Leadership Assessment and Training. *Frontiers in Psychology, 9*(SEP). https://doi.org/10.3389/fpsyg.2018.01658

Ali, L. (2018). The Design of Curriculum, Assessment and Evaluation in Higher Education with Constructive Alignment. *Journal of Education and E-Learning Research*, *5*(1), 72–78. https://doi.org/10.20448/journal.509.2018.51.72.78

Almond, R. G. (2015). Tips and Tricks for Building Bayesian Networks for Scoring Game-Based Assessments. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9505, pp. 250–263). Springer Verlag. https://doi.org/10.1007/978-3-319-28379-1_18

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2003). A four-process architecture for assessment delivery, with connections to assessment design. *University of Illinois, 1*(June), 147–171. http://www.cse.ucla.edu/products/reports/r616.pdf%5Cnhttp://www.education.umd.edu/EDMS/mislevy/papers/ProcessDesign.pdf%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.112.3158

Anderson, L. (2013). *A Taxonomy for Learning, Teaching and Assesing: a Revision of Bloom's Taxonomy.* Pearson.

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The Expanded Evidence-Centered Design (e-ECD) for Learning and Assessment Systems: A Framework for Incorporating Learning Goals and Processes Within Assessment Design. *Frontiers in Psychology*, *10*, 1–17. https://doi.org/10.3389/fpsyg.2019.00853

Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, *39*(2), 205–222. https://doi.org/10.1080/02602938.2013.819566

Behrens, J. T., Mislevy, R. J., Dicerbo, K. E., & Levy, R. (2012). Evidence Centered Design For Learning And Assessment In The Digital World. In *Technology-Based Assessments for 21st Century Skills* (pp. 13–53). http://www.cse.ucla.edu/products/reports/R778.pdf

Biggs, J. (1999). *Teaching for Quality Learning at University: What the Student Does*. OpenUniversity Press. https://books.google.com/books/about/Teaching_for_Quality_Learning_at_Univers.html?id=c3ElAQAAIAAJ

Biggs, J. (2003). Aligning Teaching for Constructing Learning. *The Higher Education Academy*, *1*(4). https://www.researchgate.net/publication/255583992

Biggs, J., & Tang, C. (2010). Applying constructive alignment to outcomes-based teaching and learning. *Training Material for "Quality Teaching for Learning in Higher Education" Workshop for Master Trainers, Ministry of Higher Education, Kuala Lumpur*, *53*(9), 23–25. https://teaching.yale-nus.edu.sg/wp-content/uploads/sites/25/2017/03/biggs.tang_.constructive.alignment.What-is-CA-biggs-tang.pdf

Bina, S., Mullins, J. K., & Petter, S. (2021). Examining game-based approaches in human resources recruitment and selection: A literature review and research agenda. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2020-Janua*, 1325–1334. https://doi.org/10.24251/hicss.2021.161

Bolkas, D., Chiampi, J. D., Fioti, J., & Gaffney, D. (2022). First Assessment Results of Surveying Engineering Labs in Immersive and Interactive Virtual Reality. *Journal of Surveying Engineering*, *148*(1), 04021028. https://doi.org/10.1061/(ASCE)SU.1943-5428.0000388

Chan, P., Van Gerven, T., Dubois, J.-L., & Bernaerts, K. (2021). Design and Development of a VR Serious Game for Chemical Laboratory Safety. In F. de Rosa, I. Marfisi Schottman, J. Baalsrud Hauge, F. Bellotti, P. Dondio, & M. Romero (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 13134 LNCS* (pp. 23–33). Springer. https://doi.org/10.1007/978-3-030-92182-8_3

Chittaro, L., & Buttussi, F. (2015). Assessing Knowledge Retention of an Immersive Serious Game vs. a Traditional Education Method in Aviation Safety. *IEEE Transactions on Visualization and Computer Graphics*, *21*(4), 529–538. https://doi.org/10.1109/TVCG.2015.2391853

Connolly, T., Stansfield, M., & Hainey, T. (2009). Towards the development of a games-based learning evaluation framework. In *Games-Based Learning Advancements for Multi-Sensory Human Computer Interfaces: Techniques and Effective Practices* (pp. 251–273). https://doi.org/10.4018/978-1-60566-360-9.ch015

Creswell, J. W. (2011). *Educational Research* (4th ed.). Pearson Education.

Criollo-C, S., Abad-Vásquez, D., Martic-Nieto, M., Velásquez-G, F. A., Pérez-Medina, J.-L., & Luján-Mora, S. (2021). Towards a New Learning Experience through a Mobile Application with Augmented Reality in Engineering Education. *Applied Sciences*, *11*(11), 4921. https://doi.org/10.3390/app11114921

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment*, *1*(1). https://doi.org/10.1145/950566.950595

Gorbanev, I., Agudelo-Londoño, S., González, R. A., Cortes, A., Pomares, A., Delgadillo, V., Yepes, F. J., & Muñoz, Ó. (2018). A systematic review of serious games in medical education: quality of evidence and pedagogical strategy. *Medical Education Online*, *23*(1), 1438718. https://doi.org/10.1080/10872981.2018.1438718

Jaffal, Y., & Wloka, D. (2015). Employing game analytics techniques in the psychometric measurement of game-based assessments with dynamic content. In *Journal of E-Learning and Knowledge Society* (Vol. 11, Issue 3, pp. 101–115). Italian e-Learning Association. https://doi.org/10.20368/1971-8829/1063

Kashive, N., Khanna, V. T., Kashive, K., & Barve, A. (2022). Gamifying Employer Branding: Attracting Critical Talent in Crisis Situations like COVID-19. *Journal of Promotion Management*, *28*(4), 487–514. https://doi.org/10.1080/10496491.2021.2008575

Kerr, D., & Chung, G. K. W. K. (2012). Identifying Key Features of Student Performance in Educational Video Games and Simulations through Cluster Analysis. *JEDM - Journal of Educational Data Mining*, *4*(1), 144–182. http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/25

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground. *International Journal of Testing*, *16*(2), 142–163. https://doi.org/10.1080/15305058.2015.1108322

Kumar, V. V., Carberry, D., Beenfeldt, C., Andersson, M. P., Mansouri, S. S., & Gallucci, F. (2021). Virtual reality in chemical and biochemical engineering education and training. *Education for Chemical Engineers*, *36*, 143–153. https://doi.org/10.1016/j.ece.2021.05.002

Loh, C. S. (2009). Researching and Developing Serious Games as Interactive Learning Instructions. *International Journal of Gaming and Computer-Mediated Simulations*, *1*(4), 1–19. https://doi.org/10.4018/jgcms.2009091501

Loh, C. S. (2012). INFORMATION TRAILS: IN-PROCESS ASSESSMENT OF GAME-BASED LEARNING. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in Game-Based Learning*. Springer.

McArthur, J. (2022). Rethinking authentic assessment: work, well-being, and society. *Higher Education*, 1–17. https://doi.org/10.1007/s10734-022-00822-y

Merrett, C. (2022). Using case studies and build projects as authentic assessments in cornerstone courses. In *International Journal of Mechanical Engineering Education* (Vol. 50, Issue 1). https://doi.org/10.1177/0306419020913286

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2020). DeepStealth: Game-Based Learning Stealth Assessment With Deep Neural Networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312–325. https://doi.org/10.1109/TLT.2019.2922356

Mislevy, R., Almond, R., & Lukas, J. F. (2003). *A Brief Introduction to Evidence-centered Design*.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Frezzo, D. C., & West, P. (2012). Three Things Game Designers Need to Know About Assessment. In *Assessment in Game-Based Learning* (pp. 59–81). Springer New York. https://doi.org/10.1007/978-1-4614-3546-4_5

Mislevy, R., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 3–62. https://doi.org/10.1207/s15366359mea0101_02

Perini, S., Luglietti, R., Margoudi, M., Oliveira, M., & Taisch, M. (2018). Learning and motivational effects of digital game-based learning (DGBL) for manufacturing education –The Life Cycle Assessment (LCA) game. *Computers in Industry*, *102*, 40–49. https://doi.org/10.1016/j.compind.2018.08.005

Perini, S., Oliveira, M., Margoudi, M., & Taisch, M. (2018). The Use of Digital Game Based Learning in Manufacturing Education – A Case Study. In P. Zaphiris & A. Ioannou (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10925 LNCS* (pp. 185–199). Springer. https://doi.org/10.1007/978-3-319-91152-6_15

Piaget, J. (1973). *To Understand is to invent: The future of education*. Grossman.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of Game-Based Learning. *Educational Psychologist*, *50*(4), 258–283. https://doi.org/10.1080/00461520.2015.1122533

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, *147*, 103778. https://doi.org/10.1016/j.compedu.2019.103778

Razak, A. A., Connolly, T., & Hainey, T. (2012). Teachers' Views on the Approach of Digital Games-Based Learning within the Curriculum for Excellence. *International Journal of Game-Based Learning*, *2*(1), 33–51. https://doi.org/10.4018/ijgbl.2012010103

Rossado Espinoza, V. P., Cardenas-Salas, D., Cabrera, A., & Coronel, L. (2021). Virtual Reality and BIM Methodology as Teaching- Learning Improvement Tools for Sanitary Engineering Courses. *International Journal of Emerging Technologies in Learning (IJET)*, *16*(06), 20. https://doi.org/10.3991/ijet.v16i06.13535

Shute, Rahimi, S., & Emihovich, B. (2017). Assessment for Learning in Immersive Environments. In D. Liu, C. Dede, R. Huang, & J. Richards (Eds.), *Virtual, Augmented, and Mixed Realities in Education.* (pp. 71–87). Springer. https://doi.org/10.1007/978-981-10-5490-7_5

Shute, V., Ke, F., & Wang, L. (2017). Assessment and Adaptation in Games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional Techniques to Facilitate Learning and Motivation of Serious Games* (pp. 59–78). Springer International Publishing. https://doi.org/10.1007/978-3-319-39298-1_4

Shute, V., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647. https://doi.org/10.1016/j.chb.2020.106647

Shute, V., & Ventura, M. (2013). Stealth Assessment: Measuring and Supporting Learning in Video Games. In *John D. and Catherine T. MacArthur Foundation Reports on Digital Media and Learning*. MIT Press. http://mitpress.mit.eduwww.macfound.org

Smith, G., Shute, V., & Muenzenberger, A. (2019). Designing and Validating a Stealth Assessment for Calculus Competencies. *Journal of Applied Testing Technology*, *20*(S1), 52–59. www.jattjournal.com

Squire, K. (2003). Video Games in Education. *International Journal of Intelligent Games & Simulation*, *2*(1). https://doi.org/10.4018/978-1-61520-781-7.ch020

Svinicki, M. D. (2004). Authentic assessment: Testing in reality. *New Directions for Teaching and Learning*, *100*, 23–29. https://doi.org/10.1002/tl.167

Sweet, S. J., & Rupp, A. A. (2012). Using the ECD Framework to Support Evidentiary Reasoning in the Context of a Simulation Study for Detecting Learner Differences in Epistemic Games. *Journal of Educational Data Mining*, *4*(1), 183–223.

www.epistemicgames.org].

Tsai, F. H., Tsai, C. C., & Lin, K. Y. (2015). The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment. *Computers and Education*, *81*, 259–269. https://doi.org/10.1016/j.compedu.2014.10.013

Udeozor, C., Russo Abegão, F., & Glassey, J. (2021). Exploring Log Data for Behaviour and Solution Pattern Analyses in a Serious Game. In U. Bakan & S. Berkeley (Eds.), *Gamification and Social Networks in Education*. MacroWorld Pub. Ltd. https://doi.org/10.15340/978-625-00-0106-6_10

Udeozor, C., Toyoda, R., Russo Abegão, F., & Glassey, J. (2022). Digital games in engineering education: systematic review and future trends. *European Journal of Engineering Education*, 1–19. https://doi.org/10.1080/03043797.2022.2093168

Urbina Coronado, P. D., Demeneghi, J. A. A., Ahuett-Garza, H., Orta Castañon, P., & Martínez, M. M. (2022). Representation of machines and mechanisms in augmented reality for educative use. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *16*(2), 643–656. https://doi.org/10.1007/s12008-022-00852-x

Villarroel, V., Boud, D., Bloxham, S., Bruna, D., & Bruna, C. (2019). Using principles of authentic assessment to redesign written examinations and tests. *Innovations in Education and Teaching International*, *57*(1), 1–12. https://doi.org/10.1080/14703297.2018.1564882

Wallner, G., & Kriglstein, S. (2012). A spatiotemporal visualization approach for the analysis of gameplay data. *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, 1115. https://doi.org/10.1145/2207676.2208558

Westera, W. (2019). Why and how serious games can become far more effective:

Accommodating productive learning experiences, learner motivation and the monitoring of learning gains. *Educational Technology and Society*, *22*(1), 59–69.

Westera, W., Prada, R., Mascarenhas, S., Santos, P. A., Dias, J., Guimarães, M., Georgiadis, K., Nyamsuren, E., Bahreini, K., Yumak, Z., Christyowidiasmoro, C., Dascalu, M., Gutu-Robu, G., & Ruseti, S. (2020). Artificial intelligence moving serious gaming: Presenting reusable game AI components. *Education and Information Technologies*, *25*(1), 351–380. https://doi.org/10.1007/s10639-019-09968-2