Hacettepe Üniversitesi Eğitim Fakültesi Dergisi

Hacettepe University Journal of Education

e-ISSN: 2536-4758

Development of a Scoring Rubric to Assess Training in an Immersive Experience Environment*

Esma ÇUKURBAŞI ÇALIŞIR, Eralp ALTUN***, Yasin ÖZARSLAN****

Article Information

Received: 26.05.2025

Accepted: 07.07.2025

Online First: 15.07.2025

Published: 15.07.2025

ABSTRACT

This research focuses on the development of a scoring rubric (SR) designed to assess occupational health and safety (OHS) training delivered through immersive experience environments. To this end, an immersive training module was implemented in the field of electrical safety, supported by immersive technologies. The participants were 30 students enrolled in the electrical program of a public university located in the Aegean region. The primary goal was to enhance participants' awareness of occupational safety and to improve their practical skills in managing potentially hazardous situations. For performance assessment, each participant was independently evaluated by three raters using the developed rubric. The collected data were analyzed through Intraclass Correlation Coefficient (ICC) and Generalizability Theory (G-Theory). The ICC analysis yielded a coefficient of 0.936, indicating a strong level of inter-rater consistency. G-Theory results supported the high reliability of the evaluations. Additionally, expert evaluations contributed to the validation process of the rubric. Overall, the findings indicate that the SR is a valid and reliable tool for evaluating learning performance in immersive OHS training contexts.

Keywords: Immersive experience, scoring rubric, occupational health and safety

doi: 10.16986/HUJE.2025.537 Article Type: Research Article

Citation Information: Çukurbaşı Çalışır, E., Altun, E., & Özarslan, Y. (2025). Development of a scoring rubric to assess training in an immersive experience environment. *Hacettepe University Journal of Education, 40*(2), 70-84. https://doi.org/10.16986/HUJE.2025.537

1. INTRODUCTION

Immersive experience technologies offer users the opportunity to step outside the limitations of the physical world and engage with digital environments through lifelike interactions (Slater & Sanchez-Vives, 2016; Stefan et al., 2024). When applied to occupational health and safety (OHS) training, these technologies create safe yet realistic scenarios in which individuals can confront hazardous situations and learn how to manage them effectively. This type of training not only fosters active involvement but also strengthens learners' ability to apply theoretical knowledge in practice (Dede, 2009).

In contemporary professional settings, OHS plays a critical role in safeguarding both the physical and psychological well-being of workers. The purpose of OHS training is to equip individuals with the awareness and competencies needed to identify workplace hazards, implement preventive strategies, and reduce the likelihood of accidents (Hale & Borys, 2013; Ricci et al., 2016). Conventional training models tend to rely heavily on theoretical instruction within classroom settings, which can limit engagement and applicability. However, recent technological advancements have opened new pathways for making training more engaging and effective. In this regard, immersive experience environments present a promising alternative that addresses the shortcomings of traditional methods and transforms the learning process into a more dynamic and impactful experience.

Immersive experience environments provide users with a strong sense of presence, allowing them to feel as though they have stepped into an entirely different reality. These environments overcome the constraints of the physical world by enabling individuals to explore complex and potentially dangerous scenarios in a secure setting (Lawson et al., 2019; Wang et al., 2018). Their application in educational contexts—particularly in high-risk professions—has been shown to offer significant

* This article is derived from the doctoral dissertation conducted by the first author as part of her Ph.D. studies at Ege University. This research was carried out by obtaining the necessary ethical permissions from the Ege University Social and Human Sciences Scientific Research and Publication Ethics Committee (08.02.2024 – 01/09-2279 protocol number).

** Lecturer, Hatay Mustafa Kemal University, Hatay-TÜRKİYE. e-mail: esma.cukurbasicalisir@mku.edu.tr (ORCID: 0000-0002-4951-0728)

*** Prof. Dr., Ege University, Faculty of Education, Department of Computer Education and Instructional Technology, Division of Computer Education and Instructional Technology, İzmir-TÜRKİYE. e-mail: eralp.altun@ege.edu.tr (ORCID: 0000-0002-4309-7493)

**** Prof. Dr., Yaşar University, Faculty of Arts and Sciences, Department of Science Culture, İzmir-TÜRKİYE. e-mail: yasin.ozarslan@yasar.edu.tr (ORCID: 0000-0003-0831-6985)

pedagogical advantages. Immersive experiences help bridge the gap between theoretical instruction and practical application, enhancing learners' preparedness for real-life tasks. By fostering deeper engagement and stronger motivation, such technologies contribute to more meaningful and participatory learning processes (Blair et al., 2021; Dede, 2009).

Previous studies have demonstrated that immersive experiences can significantly improve performance in OHS training by refining users' skills and decision-making abilities (Choi et al., 2021; Ryan et al., 2022). The safe simulation of dangerous situations allows learners to gain confidence in their responses to real-world risks (Babalola et al., 2023). In this regard, immersive technologies not only reshape general learning environments but also offer transformative potential in vocational education. Especially in fields where safety and accuracy are critical, immersive environments make it possible to design learning experiences that are simultaneously safer, more effective, and highly practical.

In the context of electrical OHS training, immersive technologies play a vital role in promoting safety and operational competence. Electrical tasks often involve high-risk activities, where mistakes can result in severe injury or even death (Babalola et al., 2023). Immersive environments give trainees the opportunity to interact with electrical systems and hazardous components in a risk-free digital setting, improving both their theoretical understanding and hands-on proficiency (Stefan et al., 2024). This mode of training is especially valuable in situations that demand repeated practice—such as troubleshooting electrical faults—where learners benefit from multiple safe iterations (Alnagrat et al., 2022; Renganayagalu et al., 2021). Therefore, in domains like electrical OHS, it becomes essential not only to deliver high-quality immersive instruction but also to assess learning outcomes and skill acquisition through reliable evaluation tools.

Immersive experiences serve as powerful platforms for experiential learning, particularly when it comes to mastering complex tasks that demand both cognitive engagement and motor coordination (Magi et al., 2023; Ryan et al., 2022). Within such environments, the assessment process plays a crucial role in evaluating participants' cognitive abilities, psychomotor skills, and decision-making capacity. At this point, the SR emerges as an essential instrument for conducting structured, consistent, and transparent evaluations. SRs provide learners with insights into their strengths and areas requiring improvement, while also enabling educators to assess achievement levels in line with predetermined standards. Especially in high-risk domains such as electrical OHS training, the detailed feedback offered by SRs enhances learners' readiness for real-world applications and supports the delivery of higher-quality instruction. As a result, evaluation tools like SRs have become indispensable for maintaining and improving the overall quality of educational environments.

Traditional methods of assessment often focus narrowly on theoretical knowledge conveyed in classroom settings (Saher et al., 2022). While such approaches may reveal what learners know in principle, they fall short of capturing whether this knowledge can be translated into action. For instance, a student might perform well on a written test about how to safely close an electrical circuit, yet be unable to carry out the procedure in a practical scenario. This gap underscores the importance of performance-based evaluation tools that assess not just cognitive understanding but also behavioral competence and professional attitudes. By using an SR, it becomes possible to observe directly whether the learner can successfully apply the intended skill in practice. This multidimensional approach to assessment deepens feedback and makes the learning process more effective and relevant.

SRs function by rating participants' performance on specific tasks according to predefined criteria and clearly articulated achievement levels. Typically, these levels are categorized as beginner (1), acceptable (2), and successful (3). This tiered structure ensures that assessments are more systematic, fair, and objective. However, the utility of any SR hinges on the clarity, relevance, and alignment of its criteria with instructional goals. Well-designed criteria highlight the most essential aspects of a task and provide educators with a roadmap for performance evaluation. In fields such as electrical OHS training, where tasks can be intricate and high-stakes, well-formulated criteria allow for more precise judgment of learners' competencies. This, in turn, facilitates the identification of both strengths and areas for improvement, ultimately elevating the effectiveness of the training program. Moreover, scoring criteria help clarify expectations for both educators and learners, promoting transparency in what is being evaluated, the standards of performance, and how learners can progress (Rahayu, 2017; Stanley, 2021).

1.1. Statement of the Problem

A growing body of research has examined how immersive experiences contribute to learning within OHS education. These studies often emphasize the pedagogical value of immersive technologies, particularly in enhancing learners' engagement and understanding. For instance, Baxter and Hainey (2023) noted that participants viewed immersive environments as helpful in improving the comprehension and recall of complex subjects. Similarly, Azis and Cantafio (2023) highlighted the potential of interactive virtual reality (VR) tools to support the teaching of advanced scientific and technical concepts in OHS contexts. However, most of this research tends to concentrate on general learning outcomes and does not delve deeply into how performance is measured—especially through structured tools such as scoring rubrics (SRs).

Although immersive experiences have been widely recognized for their positive effects on motivation and knowledge retention, there remains a notable lack of standardized tools for evaluating their educational impact. In the absence of validated and structured assessment methods like SRs, it becomes difficult to assess the full value of these experiences or to ensure the consistency of evaluations across learners and contexts.

The integration of immersive technologies into OHS training represents a significant opportunity to link innovative instructional design with systematic performance evaluation. While existing literature supports the general effectiveness of immersive environments in education, further inquiry is needed to explore how these environments affect measurable learning outcomes (Jayadurga & Rathika, 2023; Jiang et al., 2018; Shevchuk et al., 2023). Specifically, there is limited research on how SRs function within immersive training settings. To address this gap, the current study focuses on developing a tailored SR aimed at assessing learner performance in immersive OHS education. It is anticipated that such a tool will contribute to enhancing the quality of training, promoting safety awareness, and reducing risks in real-world work environments.

1.2. Purpose of the Study

This study aims to develop a SR that evaluates students' performance in an immersive technology-supported OHS experience and to present evidence of its validity and reliability. Accordingly, a structured and objective measurement tool compatible with immersive experiences was designed to evaluate students' skills in handling hazardous situations in their professional lives, their safety awareness and their adherence to procedures. The validity and reliability of the developed SR were analyzed and its effectiveness in evaluating educational performance was demonstrated.

2. METHODOLOGY

2.1. Research Design

This research is designed as a developmental study, aiming to construct a scoring rubric (SR) for evaluating educational performance in immersive experience environments. Developmental research involves the systematic design, development, and evaluation of instructional products, tools, or procedures that meet identified needs (Fraenkel, Wallen, & Hyun, 2012). In this study, the developmental process focused on creating a valid and reliable tool that could capture observable learning outcomes aligned with performance-based assessment principles. The SR development was guided by expert opinions, and a systematic and iterative approach was followed to ensure content relevance, clarity, and usability.

2.2. Participants

The study sample consisted of 30 first-year students from the Electrical Department of a vocational school affiliated with a public university in the Aegean region. A purposive sampling strategy was adopted to select participants, as this approach allows researchers to focus on cases that are especially informative and aligned with the purpose of the research (Creswell, 2012). Accordingly, the participant group was purposively selected from students enrolled in the Electrical Department of the vocational school affiliated with the same public university. This decision was based not only on the university's ability to provide the necessary physical and technical infrastructure for the implementation of the study, but also on the curricular alignment between the immersive experience scenario and the program's practice-oriented content in areas such as electrical safety, PLC diagnostics, and circuit operations.

Table 1.
Demographic Characteristics of the Study Group

Demographic Characteristics		N	%
Gender	Male	30	100.00
Age	18-20	27	90.00
	21-31	3	10.00
Type of High School Graduated	Anatolian High School	19	63.33
	Vocational High School	11	36.67
Computer Ownership	Yes	22	73.33
	No	8	26.67
Computer Usage (Daily)	Less than 1 hour	7	23.33
	1-2 hours	10	33.33
	3-5 hours	11	36.67
	6-7 hours	2	6.67
Internet Usage (Daily)	1-2 hours	3	10.00
	3-5 hours	12	40.00
	6-7 hours	10	33.33
	8 hours and above	5	16.67
Use of Immersive Technologies	Yes	7	23.33
	No	23	76.67
Age	18-20	27	90.00
	21-31	3	10.00
Type of High School Graduated	Anatolian High School	19	63.33
	Vocational High School	11	36.67

As shown in Table 1, all participants in the study were male. This distribution reflects a common trend in vocational schools, where male students are more likely to enroll in electrical programs. While the majority of the group (n = 27) was between 18 and 20 years old, three participants were aged between 21 and 31. Regarding educational background, 19 students had graduated from Anatolian High Schools, which typically require entrance exams, whereas 11 had completed their education at Vocational High Schools. In terms of access to technology, 22 participants reported owning a personal computer, while 8 did not. When asked about daily computer usage, 11 students reported using a computer for 3 to 5 hours, 10 indicated 1 to 2 hours, 7 stated less than an hour, and 2 used it for 6 to 7 hours per day. Internet use followed a similar pattern: 12 participants accessed the internet for 3 to 5 hours daily, 10 for 6 to 7 hours, 5 for more than 8 hours, and 3 for 1 to 2 hours. When asked about their familiarity with immersive technologies, 23 students indicated they had no prior experience, whereas 7 reported having used such technologies. All of those who had previous exposure stated they had used VR goggles, and two of them had also used a handheld controller. Participants' self-assessed levels of prior knowledge were gathered through the demographic information form and are detailed in Table 2.

Table 2.
Participants' Perceived Prior Knowledge Levels

Perceived Prior Knowledge Level		N	%
I recognize occupational health and safety equipment.	Low	11	55.00
	Medium	5	25.00
	High	4	20.00
I always use occupational health and safety equipment in my professional practices.	Low	7	35.00
	Medium	8	40.00
	High	5	25.00
I am aware of the accident risks I may encounter in my professional life.	Low	5	25.00
	Medium	10	50.00
	High	5	25.00
Before working on faults related to energy, I open and close the energy at the main panel and power distribution panel to eliminate the risk of electric shock.	Low	3	15.00
	Medium	12	60.00
	High	5	25.00
I always use warning signs in my professional life.	Low	6	30.00
	Medium	9	45.00
	High	5	25.00
I can make connections of circuit elements with appropriate technical specifications by adhering to occupational health and safety measures.	Low	9	45.00
	Medium	7	35.00
	High	4	20.00
I can identify faults in the inputs and outputs of a PLC.	Low	15	75.00
	Medium	4	20.00
	High	1	5.00
I take measurements using the appropriate measuring instrument according to the usage technique.	Low	10	50.00
	Medium	6	30.00
	High	4	20.00

As illustrated in Table 2, participants evaluated their prior knowledge across a set of key occupational safety and health (OHS) practices, using three self-assessed levels: low, medium, and high. When asked whether they could recognize OHS-related equipment, more than half (55%) indicated a low level of knowledge, while 25% rated themselves at a medium level and 20% reported high familiarity. For the statement regarding consistent use of safety equipment in professional practice, 35% selected "low," 40% "medium," and 25% "high." In response to whether they are aware of accident risks they may face in their future profession, 25% of students rated their awareness as low, half reported a medium level, and another 25% claimed a high level of awareness. Regarding safe operation of electrical panels—specifically turning power on and off at main and distribution points to prevent electric shock—only 15% identified their knowledge as low, whereas 60% and 25% rated it as medium and high, respectively. Use of warning signs in professional contexts revealed slightly more variation: 30% assessed their practice at a low level, 45% at medium, and 25% at high. When asked about their ability to correctly connect circuit elements according to technical specifications and safety protocols, 45% indicated low proficiency, 35% medium, and 20% high. The statement related to diagnosing faults in PLC inputs and outputs yielded the lowest levels of confidence, with 75% of students identifying as having low knowledge, 20% as medium, and only 5% as high. Finally, for the task of taking measurements with proper instruments and techniques, 50% reported low competency, 30% medium, and 20% high. Overall, these responses suggest that participants' prior knowledge concerning essential OHS practices is generally concentrated at the low to medium levels, with relatively few students expressing high confidence across the evaluated areas.

2.3. Expert Group of the Study

Throughout the development and evaluation phases of the SR, expert input was gathered to strengthen the validity of its criteria. The selected experts had professional backgrounds in both the electrical field and OHS, ensuring their feedback was grounded in practical and pedagogical relevance. They reviewed the alignment between the SR and the intended learning goals, assessed

the clarity of each criterion, and evaluated the distinctiveness of the performance levels. Their insights were instrumental in refining the tool to ensure both its validity and real-world applicability. In addition to reviewing the SR itself, experts also evaluated whether the scenarios embedded in the immersive environment aligned with the objectives of the training. They participated in the implementation process as both observers and raters, contributing directly to the application of the SR. These expert contributions played a key role in reinforcing the rubric's credibility and supporting the overall reliability of the evaluation framework.

2.4. Data Collection and SR Development Process

Within the immersive environment, participants were asked to complete specific tasks based on a pre-designed scenario. To assess their performance, an SR consisting of three achievement levels was created. SRs are widely accepted tools in performance-based evaluation, as they help reduce potential rater bias and improve scoring consistency across evaluators (Goodrich, 1997). Moreover, SRs support the recognition of individual differences among learners and facilitate structured feedback by simplifying the assessment process (Iltar & Karataş, 2022). As a teacher-guided tool, the SR developed in this study was intended to evaluate students' performance across a range of applied skills and competencies.

The initial phase of the rubric design involved defining performance criteria that were aligned with the goals and content of the immersive experience. Performance levels were structured into three categories: beginner (1), acceptable (2), and successful (3). To ensure the relevance and clarity of each criterion and level, expert opinions were collected from four professionals with expertise in both the electrical domain and OHS. These experts reviewed the draft SR using a structured feedback form that included dimensions such as content relevance, usability, clarity of items, level appropriateness, and alignment with expected learning outcomes.

The expert review form included three response options—"appropriate," "needs revision," and "not appropriate"—for each criterion and level descriptor. Additionally, a comment section allowed experts to provide qualitative feedback. The responses were later analyzed to determine the degree of agreement among the experts. Table 3 presents the percentage of agreement for each criterion, offering insight into the perceived appropriateness and clarity of the SR components.

Table 3.
Experts' Agreement Percentages According to Criteria

Criteria	Expert-1	Expert-2	Expert-3	Expert-4
1	100.00	100.00	100.00	100.00
2	100.00	100.00	100.00	100.00
3	100.00	100.00	100.00	100.00
4	100.00	50.00	50.00	100.00
5	100.00	50.00	100.00	100.00
6	100.00	100.00	100.00	100.00
7	100.00	100.00	100.00	100.00
8	100.00	100.00	100.00	100.00
9	100.00	100.00	100.00	100.00
10	100.00	100.00	100.00	100.00
11	100.00	100.00	100.00	100.00

According to Table 8, Expert-1 and Expert-4 demonstrated 100% agreement across all criteria. Expert-2 indicated that two criteria needed revision, resulting in a 50% agreement rate for those items. Likewise, Expert-3 reported a 50% agreement for only one criterion. These findings suggest that the majority of the criteria were deemed appropriate by the experts, although minor revisions were required for a few items. The criteria with lower agreement levels were revised in accordance with the experts' qualitative feedback, and the final version of the scoring rubric was shaped accordingly.

In addition to field-specific expert input, feedback was also obtained from specialists in measurement and evaluation. Their contributions helped refine the overall structure of the rubric and further supported its content validity. Following this iterative development process, the scoring rubric was finalized with 5 learning outcomes, 11 criteria, and 3 performance levels. Within this framework, students could receive a minimum score of 0 and a maximum of 33 points.

2.4.1. Immersive experience study

As part of the OHS training program for electrical students at vocational schools, this study incorporated an immersive experience designed to simulate real-world tasks in a controlled virtual environment. Participants interacted with a scenario-based simulation using VR headsets, allowing them to engage in realistic tasks within a digitally constructed space. The immersive environment was intended to reinforce adherence to OHS procedures while also supporting the development of technical skills such as operating PLC panels, identifying faults, and performing system corrections. Throughout the experience, students actively applied OHS practices using VR goggles and handheld controllers, navigating tasks that mirrored real-life workplace conditions. The design of the environment aimed to balance safety with realism, enabling learners to encounter

critical situations without physical risk. Figure 1 displays visual examples of participants interacting with the immersive environment during the training.

Figure 1. Participants Engaged in an Immersive Experience

The immersive environment was modeled after a realistic factory layout and was divided into four main areas: an equipment room, a production line system, a PLC control panel room, and a fuse panel room (see Figure 2). The space was designed to support a wide range of user interactions via VR controllers, enabling actions such as touching, dragging, lifting, dropping, zooming in and out, rotating, and pushing. Additional functions—including grasping, removing, placing, and repositioning objects—were carried out by pressing and releasing buttons on the handheld controllers. Items could be moved by holding the button down, while fuses were operated by vertical movements of the controller to simulate opening and closing. Contextual information such as visual warnings and directional guidance was also integrated into the environment to support learners during task execution.

Figure 2. Images from the Immersive Experience Environment

According to the scenario, the general framework of the study is as follows:

> *SCENARIO:*
> *In the immersive environment, users are first required to apply OHS measures before moving on to control the PLC panel, diagnose potential malfunctions, and perform necessary corrective actions. Throughout the scenario, the use of personal protective equipment is mandatory, and strict adherence to OHS protocols is expected. Given the nature of the tasks—which involve simulated exposure to potentially dangerous voltage levels—the scenario is designed to emphasize risk awareness and safe working practices. Learners are challenged to manage hazards such as electric*

shock, fire, and safety threats caused by inattentiveness or human error. The experience aims to reinforce procedural discipline while allowing participants to engage with complex systems in a safe and controlled digital setting.

During the immersive experience, participants were evaluated by three different raters using the SR according to the following plan.

LEVEL: Associate and Bachelor's Degree
DEPARTMENT: Electrical Associate Degree Program, Electrical-Electronics Engineering Bachelor's Degree Program
TARGET AUDIENCE: Undergraduate students enrolled in electrical programs
COURSE: Occupational Health and Safety
OBJECTIVE: This immersive training scenario was developed to provide students with hands-on experience in recognizing and managing potential hazards and risks associated with OHS in professional settings. The aim is to equip learners with practical skills in prevention and control, while promoting the development of safe work habits.
EXPECTED LEARNING OUTCOMES:

- Identifies possible accident risks likely to be encountered in the workplace
- Recognizes relevant OHS tools, equipment, and materials
- Uses OHS equipment correctly and consistently
- Connects circuit components in accordance with technical standards and OHS procedures
- Detects and diagnoses faults in PLC input and output terminals while maintaining adherence to OHS protocols

2.4.2. Implementation and Scoring Process

The implementation and scoring process of the SR involved the following key stages:

1. Determination of the Passing Score (Angoff Method):
    - The minimum passing score was established using the Angoff method (Cizek & Bunch, 2007).
    - Three subject matter experts estimated the likelihood of a borderline student receiving a score of 1 (fail), 2 (adequate), or 3 (successful) for each of the 11 SR criteria.
    - These estimates were multiplied by their corresponding score levels and averaged to calculate the expected score per item.
    - The sum of these expected scores yielded a total of 23.3.
    - Based on this total:
        - The 70% success threshold was set at 23.1 points.
        - The 50% success threshold was set at 16.5 points.
    - In accordance with the institutional policy that considers 50% as the minimum passing level in vocational programs, 16.5 was adopted as the passing score.
2. Application of the SR During the Immersive Experience:
    - The finalized rubric was applied while participants engaged with the immersive training scenario.
    - Each of the 30 participants was evaluated independently by three expert raters.
3. Scoring and Aggregation:
    - For each criterion, the scores assigned by the three raters were averaged to produce a single criterion-level score per participant.
    - These averaged scores were summed to calculate each participant's total SR score.
    - The total possible score ranged between 0 and 33.
4. Analysis of Reliability and Rater Consistency:
    - Inter-rater reliability was assessed using:
        - Intraclass Correlation Coefficient (ICC)
        - Generalizability Theory (G-Theory)
    - The results demonstrated strong consistency across raters and confirmed that the SR could be used reliably to assess performance in immersive environments.

2.5. Data Analysis

In the immersive experience, the performance of 30 students was evaluated by three independent raters using the SR. To assess the level of agreement among raters, Intraclass Correlation Coefficient (ICC) analysis was carried out. Additionally,

Generalizability Theory (G-Theory) was applied to identify and quantify the contribution of potential sources of error in the scoring process.

All analyses were conducted using Python. Data preparation was completed with the help of Pandas and NumPy libraries. Once the validity and reliability processes were completed, the passing threshold for the SR was determined using the Angoff method—a widely used standard-setting technique (Angoff, 1971; Buckendahl et al., 2002). The calculation of passing scores employed SciPy and NumPy, while descriptive analysis was used to examine supporting documents collected from participants.

The ICC analysis, used to assess the degree of consistency among raters, produces values ranging from 0 to 1—where values approaching 1 suggest high reliability, and those near 0 indicate poor agreement (Koo & Li, 2016; Shrout & Fleiss, 1979). The analysis was performed using the Pingouin library following data preparation.

G-Theory was implemented to further analyze reliability by identifying specific sources of measurement error. It provides detailed estimates of how much variance can be attributed to different factors—such as individual raters—and offers insight into the robustness of the overall measurement system (Brennan, 2021; Shavelson & Webb, 1981; Merrifield, 1974). This analysis was conducted using the gStudy package.

2.6. Validity

To assess the validity of the developed SR, content validity evidence was gathered through expert evaluation. During this process, input was obtained from four professionals with expertise in both the electrical field and OHS. The experts assessed the appropriateness of the SR's criteria and performance levels using a three-point scale: “appropriate,” “needs correction,” and “not appropriate.” Based on their feedback, several refinements were made to improve the clarity and alignment of specific items.

There was a high level of agreement among the experts regarding the relevance and adequacy of the criteria. Expert-1 and Expert-4 showed full agreement, approving all items without suggesting revisions. Expert-2 flagged two criteria as needing modification, resulting in 50% agreement for those items, while Expert-3 identified one item requiring revision. These responses indicate that most criteria were deemed suitable, though a few required adjustments. The revisions made in response to this feedback helped strengthen the content validity of the SR.

The performance levels were defined in three categories: beginner (1), acceptable (2), and successful (3). These levels were developed in alignment with the perspectives of the experts. The high level of consensus among reviewers and the improvements made following their suggestions provide strong evidence that the SR is both well-structured and valid for evaluating learner performance in immersive training environments.

3. FINDINGS

3.1. Calculating the Passing Score of SR with the Angoff Method

The Angoff method was used to determine the passing score for the scoring rubric (SR). The Angoff method is a widely used standard-setting technique that estimates item difficulty levels based on expert judgments and determines an overall passing score from these estimates (Cizek & Bunch, 2007). To implement the Angoff method, data were first collected from three subject matter experts. These experts estimated the probability that students would receive a score of 1 (fail), 2 (adequate), or 3 (successful) for each criterion in the SR. For each item, these probabilities were multiplied by their respective score values, and the resulting values were averaged to calculate the expected score for each criterion. Table 4 presents the average scores for each item, based on expert estimates, along with the total cumulative score derived using the Angoff method.

Table 4.
Average Scores for Each Item Based on Expert Estimates Using the Angoff Method and Their Total

Item	Expert 1	Expert 2	Expert 3	$\bar{X}$
Item 1	2.1	2.05	2.0	2.05
Item 2	2.1	2.05	2.0	2.05
Item 3	2.05	2.1	2.1	2.08
Item 4	2.0	2.05	2.0	2.02
Item 5	2.4	2.35	2.3	2.35
Item 6	2.15	2.2	2.2	2.18
Item 7	2.05	2.1	2.05	2.07
Item 8	2.1	2.1	2.05	2.08
Item 9	2.3	2.25	2.2	2.25
Item 10	2.1	2.05	2.0	2.05
Item 11	2.2	2.1	2.05	2.12

Table 4 presents the average scores estimated by three experts for each SR item, offering insight into how each item was perceived in terms of expected student performance. The expert evaluations demonstrated a high degree of consistency, with score differences generally ranging between 0.1 and 0.2 points. Such minimal variation suggests strong agreement among the reviewers and no significant discrepancies in judgment.

Based on these estimates, the average score for each item was calculated, resulting in a total cumulative score of 23.3. This score was used to establish benchmark thresholds according to the Angoff method: 23.1 points corresponded to a 70% success level, while 16.5 points represented a 50% threshold. As the host institution applies a minimum passing score of 50% in vocational programs, the cut-off score for this study was set at 16.5 points.

3.2. Examining the Consistency Among Raters: ICC

The level of consistency among the three raters was assessed through ICC analysis, using a two-way random effects model with the ICC(2,1) coefficient. Prior to this, Pearson correlation coefficients were calculated to evaluate the degree of agreement between raters and to assess the reliability of their scoring behavior. High correlation values suggest a strong level of consistency, indicating that the scoring process was both objective and dependable (Doğan & Yosmaoğlu, 2015; Kocakülah, 2022). These statistical measures are essential for identifying any inconsistencies in scoring, such as potential rater bias or inattentiveness. Table 5 presents the Pearson correlation values calculated between each pair of raters.

Table 5.
Pearson Correlation Coefficients Among Raters

Raters	Inter-rater Correlation
Rater 1 and Rater 2	0,961
Rater 1 and Rater 3	0,954
Rater 2 and Rater 3	0,996

As shown in Table 5, the Pearson correlation coefficients between raters were calculated as 0.961 between Rater 1 and Rater 2, 0.954 between Rater 1 and Rater 3, and 0.996 between Rater 2 and Rater 3. These high correlation values indicate a strong level of agreement, suggesting that the scoring across raters was highly consistent.

To further confirm inter-rater reliability, the ICC(2,1) value was computed. The ICC provides a statistical estimate of how closely aligned raters are in their evaluations. The resulting ICC(2,1) value was 0.936, which exceeds the commonly accepted threshold of 0.90. This outcome reflects a high degree of reliability among the raters and supports the consistency of the evaluation process.

3.3. Analysis of Rater Error Based on G-Theory

To examine the sources of error in rater evaluations, G-Theory analysis was conducted. As a comprehensive framework for assessing measurement reliability, G-Theory enables researchers to identify specific contributors to measurement error by examining variance components. In this study, the analysis aimed to uncover the extent to which inconsistencies stemmed from participants, raters, or residual factors. An ANOVA procedure was applied to calculate the variance associated with each of these sources. The resulting statistics—sum of squares (sum_sq), mean square (Mean Sq), degrees of freedom (df), F-values, and p-values—were used to interpret the reliability of the scoring process. Detailed results from this analysis are presented in Table 6.

Table 6.
ANOVA Results of G-Theory Analysis

Source of Variance	Sum Sq	df	Mean Sq	F	p
Participants	103.582	29	3.572	11.018	<0.001
Raters	0.016	2	0.008	0.024	0.976
Error (Residual)	281.384	868	0.324		

An examination of Table 6 reveals that the variance attributable to participants (Sum Sq = 103.582, df = 29, Mean Sq = 3.572) was substantial and statistically significant ($F = 11.018$, $p < .001$), indicating meaningful differences in performance across students. In contrast, the variance associated with raters (Sum Sq = 0.016, df = 2, Mean Sq = 0.008) was minimal and not statistically significant ($F = 0.024$, $p = 0.976$), suggesting that scoring across raters was consistent and unbiased. This result indicates that the influence of raters on the measurement process was negligible and that inter-rater reliability was successfully maintained.

The residual error (Sum Sq = 281.384, df = 868, Mean Sq = 0.324) accounted for the largest portion of the total variance. This relatively high error term likely reflects the natural variability in participants' individual performance, as well as small, random inconsistencies inherent in the measurement process. Such variation is typical in applied educational settings—especially in

performance-based assessments—and does not necessarily indicate flaws in the evaluation tool. Nevertheless, to improve measurement precision, future implementations may benefit from minimizing the impact of external factors such as fatigue, environmental distractions, or task complexity.

To complement these findings and provide an overall estimate of score reliability, the generalizability coefficient was calculated. As a result of the G-Theory analysis, the generalizability coefficient (G) was found to be 0.971. This high coefficient confirms that the assessment system was robust and dependable, and that the scoring results can be considered both reliable and generalizable, even in the presence of some residual error variance.

4. RESULTS, DISCUSSION AND RECOMMENDATIONS

This study set out to develop an SR for evaluating the effectiveness of immersive experiences in OHS training and to assess the validity and reliability of the tool. Conducted with vocational school students in the field of electrical education, the study demonstrated that the SR yielded consistent and dependable evaluation results. Structured around three performance levels—beginner, acceptable, and successful—the SR was developed in consultation with field experts and was designed to provide objective assessments within immersive training contexts. The findings revealed a high degree of agreement among raters. The ICC analysis confirmed strong consistency in scoring, underscoring the objectivity and reliability of the evaluation process (Koo & Li, 2016). These results align with findings from previous research, where high ICC values are associated with reliable measurement tools and valid evaluations of educational performance (Kim & Kwak, 2022).

To further explore the consistency of ratings and identify possible sources of measurement error, a G-Theory analysis was conducted. As Brennan (2021) emphasizes, G-Theory plays a critical role in enhancing objectivity and minimizing error in educational assessments. In this study, the low variance among raters and the strong inter-rater agreement supported the reliability of the evaluation process. The analysis showed that most of the variance originated from differences in participant performance, while the residual error variance—though substantial—suggested that further refinement of the evaluation process may be beneficial. This relatively high error term may reflect a combination of natural variation in student performance and measurement-related inconsistencies. While such variance is not unusual in applied settings, it does point to potential influences from uncontrolled external factors or possible areas for refining the scoring process.

Improving reliability in future applications of the SR could involve several strategies. Clarifying evaluation criteria, offering training sessions for raters, and conducting periodic calibration activities are all useful for minimizing inconsistencies. When multiple raters assess the same participants, discrepancies can be monitored and corrected more effectively. Providing structured feedback to raters and supporting their ongoing development can also help reduce scoring errors. In addition, reviewing the relevance and clarity of the SR at regular intervals is essential to maintain its accuracy. Including a larger pool of raters may help to balance out individual biases, and re-evaluating participant performances over time can assist in detecting measurement variability. Finally, integrating digital tools into the assessment workflow can support greater objectivity and consistency in scoring.

The improvement strategies outlined above can be applied to enhance both the reliability and consistency of the evaluation process. Following the implementation of these measures, the reliability analysis yielded results that further support the strength of the system. The G-Theory coefficient was calculated as 0.971, indicating excellent reliability and strong generalizability across both raters and participants (Brennan, 2021; Shavelson & Webb, 1981; Merrifield, 1974). The observed variance between participants and the stability across raters reinforce the robustness and applicability of the SR.

Although the correlation coefficients between raters were high, it is important to acknowledge that small variations in absolute scoring levels are natural in performance-based assessments. While raters generally followed similar scoring patterns, individual differences in rating tendencies may still occur due to subjective interpretations of criteria. Despite this, both ICC and Pearson correlation results demonstrated strong consistency among raters. Furthermore, the G-Theory analysis found that the rater variance was low and not statistically significant ($p = 0.976$). These findings confirm that the overall evaluation process was consistent and reliable across raters.

The observed range of performance across participants suggests that the SR successfully captured a broad spectrum of skill levels, allowing for fairer assessment among learners with varying competencies. However, the presence of items that may be perceived as too easy or too difficult could create imbalances in overall scoring. To address this, it is recommended that all assessment tools undergo periodic review and refinement to ensure fairness and accuracy. The diversity in participant performance also emphasizes the need to consider individual differences when designing and applying assessment frameworks.

The potential for variation in rater scoring behavior highlights the ongoing need for calibration and standardization in assessment practices. Organizing regular training sessions and calibration exercises can help ensure that raters interpret the SR consistently. Clearly defined criteria reduce the likelihood of differing interpretations and strengthen inter-rater reliability. Future studies could explore the use of larger rater groups or comparative evaluations across different learner populations to further improve scoring reliability in immersive learning environments.

The Angoff method was employed in this study to determine the passing score for the SR. This approach is widely recognized for establishing performance standards based on expert judgment and is particularly suited for high-stakes educational assessments. One of its core strengths lies in its systematic and replicable nature, which helps ensure fairness and consistency in score interpretation. As emphasized by Thomas et al. (2021), a sound standard-setting process should be objective, transparent, and resistant to arbitrary variation.

Unlike fixed cutoff thresholds—such as the commonly used 60%—the Angoff method allows for more tailored and context-sensitive standards. For example, Kamal et al. (2018) reported that a 64.5% threshold was established using this method in a medical education context. Other studies have also shown that the Angoff method often results in higher and more rigorous benchmarks (Yim & Shin, 2020; Yousef et al., 2017). In this study, a 70% threshold was adopted based on expert-derived estimates, confirming the SR's alignment with a defensible and valid standard-setting process (Cizek & Bunch, 2007).

The method has also gained credibility due to its association with strong inter-rater reliability. Previous research reports reliability coefficients ranging from 0.78 to 0.85 (Shah et al., 2014). In the present study, inter-rater consistency was particularly high, with an ICC value of 0.936, indicating strong agreement among evaluators. This consistency is largely attributable to the clarity and objectivity of the criteria and performance descriptors defined within the SR.

Overall, this study contributes valuable insights regarding the validity and reliability of SR in the context of immersive OHS training. The findings support the use of SR not only as a robust evaluation tool in immersive settings but also as a mechanism for enhancing instructional quality. By offering structured and transparent performance measurement, SR enables educators to better align teaching strategies with intended learning outcomes. Prior research highlights SR as an effective means of assessing educational performance (Blair et al., 2021; Doğan & Yosmaoğlu, 2015; Kim & Kwak, 2022), particularly in environments requiring the integration of multiple skill domains (Udeozor et al., 2023). Immersive learning applications—especially in practice-intensive fields like engineering and medicine—benefit significantly from SR-based evaluations. For instance, Bulut and Sönmez (2020) found that VR-based environments can enhance students' applied skills, and when these experiences are assessed using SR, they allow for a more objective and targeted analysis of learner performance. Ultimately, the use of SR in immersive education strengthens the link between instructional design and meaningful learning outcomes.

This study centered on the use of SR in immersive OHS training environments. Future research could investigate the applicability and effectiveness of the developed rubric across various vocational education domains such as healthcare, construction, and aviation. Such studies would help assess the generalizability and adaptability of the SR in different instructional settings. Comparative analyses between SR and other evaluation methods could also highlight its specific advantages and limitations.

Moreover, the long-term effects of SR on learners' retention, skill development, and professional growth present a promising area of exploration. Integrating SR with emerging technologies may lead to the development of automated assessment systems, enhancing scoring efficiency and consistency. Particularly, AI-supported evaluation tools could improve the performance and responsiveness of SR applications. In addition, investigating the use of SR in workplace settings—especially in high-risk industries—could offer valuable insight into its practical relevance beyond educational contexts.

To ensure that SR structures used in immersive environments become more effective and widely adopted in the future, both pedagogical and technological innovations are required. With the advancement of AI-supported assessment systems, it will become possible to enrich SR applications with automated data analysis and enable real-time adaptations based on learner performance. Additionally, the development of interdisciplinary rubric libraries and their open-access dissemination could enhance the transparency, equity, and quality of assessment practices. Lastly, it is recommended that future research expands beyond academic settings and explores the practical implementation of SRs in professional development and workplace training contexts. Such efforts would contribute to broader and more sustainable integration of structured performance assessment in diverse learning environments.

An important consideration for future studies is the role of presence in immersive environments. The sense of presence has been shown to influence learner engagement and performance. According to Geriş and Tunga (2020), interface design elements such as user perspective, spatial layout, and video resolution significantly impact perceived presence. Since this perceptual factor may influence how participants engage with tasks and how their performance is evaluated, it should be taken into account in future SR development and validation efforts. Understanding the relationship between presence and assessment outcomes is essential for advancing the credibility of SR in immersive learning environments.

In summary, assessment plays a pivotal role not only in measuring learning outcomes but also in guiding educational processes and supporting learners' development. A well-designed evaluation system enables instructors to align instruction with intended learning goals and provides learners with meaningful feedback on their progress. Especially in high-risk training scenarios, accurate and objective assessment improves both performance and awareness of safety practices. Within this context, the developed SR represents a significant step toward integrating innovative assessment approaches into immersive education.

As a result, the developed SR is supported by strong statistical evidence in terms of both content validity and measurement reliability. The high levels of expert agreement indicate that the criteria adequately represent the intended learning outcomes. The passing score determined using the Angoff method was calculated in accordance with both academic standards and the contextual requirements of the implementation environment. Furthermore, the analyses conducted using ICC and Generalizability Theory revealed a high degree of inter-rater consistency, providing strong support for the reliability and robustness of the evaluation framework. These findings suggest that the SR is not only valid and reliable within the scope of this study but also holds potential for use as a systematic assessment tool in other immersive experience-based training contexts with similar content.

Limitations

While the findings of this study provide strong evidence regarding the validity and reliability of the developed scoring rubric in immersive OHS training, certain limitations must be acknowledged. First, the sample consisted of only 30 students from a single vocational school, which may limit the generalizability of the results. Additionally, although inter-rater reliability was high, the presence of natural variability in student performance and broader implementation are needed. Future studies should consider replicating this research with larger and more diverse samples across different institutions and domains to enhance external validity.

Another limitation of this study is the absence of longitudinal validation to assess the sustained effectiveness of the scoring rubric over time. Future research may consider repeated administrations of the SR to examine its stability and long-term reliability in various instructional settings. Additionally, while the current SR was structured as a 3-level analytic rubric, alternative formats such as holistic or expanded multi-level rubrics may provide different insights into student performance. Comparative studies involving diverse SR structures could contribute to a deeper understanding of how rubric design influences assessment outcomes how rubric design influences assessment outcomes, learner engagement, and instructional decision-making.

Research and Publication Ethics Statement

This research was carried out by obtaining the necessary permissions from the Ege University Social and Human Sciences Scientific Research and Publication Ethics Committee and the Rectorate of Manisa Celal Bayar University. Ethical principles and rules were taken into consideration in the collection, analysis, and reporting of data. The authors declare that all information in this study has been obtained and presented in accordance with academic rules and ethical conduct.

Contribution Rates of Authors to the Article

The authors declare that each author made an important contribution to every stage of the study. The three authors worked together during the analysis and reporting of the data.

Statement of Interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Acknowledgement

This study was conducted at the Extended Reality Laboratory (XR LAB, https://xrlab.mcbu.edu.tr) of Manisa Celal Bayar University. We would like to express my gratitude to the XR Lab team for their support throughout this research. We also appreciate the participation of the students and staff who contributed to this study.

5. REFERENCES

Alnagrat, A. J. A., Ismail, R. C., & Idrus, S. Z. S. (2022). The effectiveness of virtual reality technologies to enhance learning and training experience: during the covid-19 pandemic and beyond. *Journal of Creative Industry and Sustainable Culture,* 1, 27-47. https://doi.org/10.32890/jcisc2022.1.2

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.), 508–600. American Council on Education.

Azis, I. R. & Cantafio, G. (2023). The role of virtual reality in science and technology education. *Journal of Training, Education, Science and Technology,* 13-18. https://doi.org/10.51629/jtest.v1i1.170

Babalola, A., Manu, P., Cheung, C., Yunusa-Kaltungo, A., & Bartolo, P. (2023). Applications of immersive technologies for occupational safety and health training and education: A systematic review. *Safety Science*, *166*, 106214. https://doi.org/10.1016/j.ssci.2023.106214

Baxter, G. & Hainey, T. (2023). Using immersive technologies to enhance the student learning experience. *Interactive Technology and Smart Education,* 21(3), 403-425. https://doi.org/10.1108/itse-05-2023-0078

Blair, C., Walsh, C., & Best, P. (2021). Immersive 360° videos in health and social care education: a scoping review. *BMC Medical Education,* 21(1). https://doi.org/10.1186/s12909-021-03013-y

Brennan, R. L. (2021). Generalizability theory. In *The history of educational measurement*, 206-231. Routledge.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement*, 39(3), 253-263. https://doi.org/10.1111/j.1745-3984.2002.tb01177.x

Bulut, A. & Sönmez, O. (2020). Diş hekimliği preklinik eğitimi için sanal gerçeklik ortamında diş modellerinin oluşturulması: Pilot çalışma. *Turkish Journal of Clinics and Laboratory,* 11(2), 43-49. https://doi.org/10.18663/tjcl.676506

Choi, J., Thompson, C. E., Choi, J., Waddill, C., & Choi, S. (2021). Effectiveness of immersive virtual reality in nursing education. *Nurse Educator*, 47(3), E57-E61. https://doi.org/10.1097/nne.0000000000001117

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: a guide to establishing and evaluating performance standards on tests*. Sage Publications.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, 323(5910), 66-69. https://doi.org/10.1126/science.1167311

Doğan, C. D. and Yosmaoğlu, H. B. (2015). The effect of the analytical rubrics on the objectivity in physiotherapy practical examination. *Turkiye Klinikleri Journal of Sports Sciences,* 7(1), 9-15. https://doi.org/10.5336/sportsci.2014-39517

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). New York: McGraw-Hill.

Geriş, A., & Tunga, Y. (2020). *Sanal gerçeklik ortamlarında bulunma hissi*. *Manisa Celal Bayar Üniversitesi Sosyal Bilimler Dergisi*, *18*(4), 261–282. https://doi.org/10.18026/cbayarsos.818457

Gittinger, F. P., Lemos, M., Neumann, J. L., Förster, J., Dohmen, D., Berke, B., … & Jonas, S. (2022). Interrater reliability in the assessment of physiotherapy students. *BMC Medical Education,* 22(1). https://doi.org/10.1186/s12909-022-03231-y

Goodrich, H. (1997). Understanding Rubrics: The dictionary may define" rubric," but these models provide more clarity. *Educational leadership,* 54(4), 14-17.

Hale, A. R., & Borys, D. (2013). Working to rule, or working safely? Part 1: A state of the art review. *Safety Science*, 55, 207-221. https://doi.org/10.1016/j.ssci.2012.05.011

Humphry, S. & Heldsinger, S. (2020). A two-stage method for obtaining reliable teacher assessments of writing. *Frontiers in Education*, 5. https://doi.org/10.3389/feduc.2020.00006

Iltar, L., & Karataş, A. G. (2022). Türkçenin yabancı dil olarak öğretiminde anlatmaya/göstermeye dayalı metinler için yazma becerisi dereceli puanlama anahtarı. *Okuma Yazma Eğitimi Araştırmaları,* 10(2), 194-213. https://doi.org/10.35233/oyea.1177730

Jayadurga, R. & Rathika, M. (2023). Significance and impact of artificial intelligence and immersive technologies in the field of education. *International Journal of Recent Technology and Engineering*, 12(2), 66-71. https://doi.org/10.35940/ijrte.b7802.0712223

Jiang, Y., Clarke-Midura, J., Baker, R. S., Paquette, L., & Keller, B. (2018). How Immersive Virtual Environments Foster Self-Regulated Learning. In R. Zheng (Ed.), *Digital Technologies and Instructional Design for Personalized Learning*, 28-54. *IGI Global Scientific Publishing.* https://doi.org/10.4018/978-1-5225-3940-7.ch002

Kamal, D., ElAraby, S., Kamel, M., & Hosny, S. (2018). Evaluation of two applied methods for standard setting in undergraduate medical programme at the faculty of medicine, suez canal university. *Education in Medicine Journal,* 10(2), 15-25. https://doi.org/10.21315/eimj2018.10.2.3

Kim, C. & Kwak, E. (2022). An exploration of a reflective evaluation tool for the teaching competency of pre-service physical education teachers in korea. *Sustainability*, 14(13), 8195. https://doi.org/10.3390/su14138195

Kocakülah, A. (2022). Development and use of a rubric to assess undergraduates' problem solutions in physics. *Participatory Educational Research,* 9(3), 362-382. https://doi.org/10.17275/per.22.71.9.3

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Lawson, G., Shaw, E., Roper, T., Nilsson, T., Bajorunaite, L., & Batool, A. (2019). Immersive virtual worlds: Multi-sensory virtual environments for health and safety training. *arXiv preprint arXiv:1910.04697*. https://doi.org/10.48550/arXiv.1910.04697

Magi, C. E., Bambi, S., Iovino, P., El Aoufy, K., Amato, C., Balestri, C., Rasero, L., & Longobucco, Y. (2023). Virtual reality and augmented reality training in disaster medicine courses for students in nursing: a scoping review of adoptable tools. *Behavioral Sciences*, 13(7), 616. https://doi.org/10.3390/bs13070616

Merrifield, P. R. (1974). Book Reviews: Cronbach, Lee J., Gleser, Goldine C., Nanda, Harinder, and Rajaratnam, Nageswari. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. *American Educational Research Journal*, *11*(1), 54-56. https://doi.org/10.3102/00028312011001054

Rahayu, E. Y. (2017). Raters' bias, background and perception in awarding score of writing performance. *Journal of English Education,* 6(2), 69. https://doi.org/10.24127/pj.v6i2.1022

Renganayagalu, S. K., Mallam, S., & Nazir, S. (2021). Effectiveness of vr head mounted displays in professional training: a systematic review. *Technology, Knowledge and Learning*, 26(4), 999-1041. https://doi.org/10.1007/s10758-020-09489-9

Ricci, F., Chiesi, A., Bisio, C., Panari, C., & Pelosi, A. (2016). Effectiveness of occupational health and safety training: A systematic review with meta-analysis. *Journal of workplace learning*, *28*(6), 355-377. https://doi.org/10.1108/JWL-11-2015-0087

Ryan, G., Callaghan, S., Rafferty, A., Higgins, M., Mangina, E., & McAuliffe, F. (2022). Learning outcomes of immersive technologies in health care student education: systematic review of the literature. *Journal of Medical Internet Research,* 24(2), e30082. https://doi.org/10.2196/30082

Saher, A. S., Ali, A. M. J., Amani, D., & Najwan, F. (2022). Traditional Versus Authentic Assessments in Higher Education. *Pegem Journal of Education and Instruction*, *12*(1), 283-291. https://doi.org/10.47750/pegegog.12.01.29

Shah, C., Parmar, D., & Parmar, R. (2014). Study of standard setting in constructed response type written examination. *International Journal of Medical Science and Public Health,* 3(9), 1046. https://doi.org/10.5455/ijmsph.2014.170620142

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology,* 34(2), 133-166. https://doi.org/10.1111/j.2044-8317.1981.tb00625.x

Shevchuk, I., Filippova, L., Krasnova, A., & Bazyl, O. (2023). Virtual Pedagogy: Scenarios for Future Learning with VR and AR Technologies. *Futurity Education*, *3*(4), 95–117. https://doi.org/10.57125/FED.2023.12.25.06

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428. https://psycnet.apa.org/doi/10.1037/0033-2909.86.2.420

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. Frontiers in Robotics and AI, 3, 74. https://doi.org/10.3389/frobt.2016.00074

Smith, J. D., Dishion, T. J., Brown, K., Ramos, K., Knoble, N. B., Shaw, D. S., … & Wilson, M. N. (2015). An experimental study of procedures to enhance ratings of fidelity to an evidence-based family intervention. *Prevention Science,* 17(1), 62-70. https://doi.org/10.1007/s11121-015-0589-0

Stanley, T. (2021). *Using rubrics for performance-based assessment: A practical guide to evaluating student work*. Routledge. https://doi.org/10.4324/9781003239390

Stefan, H., Mortimer, M., Horan, B., & McMillan, S. (2024). How effective is virtual reality for electrical safety training? Evaluating trainees' reactions, learning, and training duration. *Journal of Safety Research*, *90*, 48-61. https://doi.org/10.1016/j.jsr.2024.06.002

Thomas, D., Abdalla, A., McKeirnan, K., & Khalifa, S. (2021). Standard-setting for continuing education assessment of select new competencies. *Journal of Continuing Education in the Health Professions,* 42(1), e96-e98. https://doi.org/10.1097/ceh.0000000000000397

Udeozor, C., Chan, P., Abegão, F. R., & Glassey, J. (2023). Game-based assessment framework for virtual reality, augmented reality and digital game-based learning. *International Journal of Educational Technology in Higher Education,* 20(1). https://doi.org/10.1186/s41239-023-00405-6

Wang, P., Wu, P., Wang, D., Chi, H., & Wang, X. (2018). A critical review of the use of virtual reality in construction engineering education and training. *International Journal of Environmental Research and Public Health*, 15(6), 1204. https://doi.org/10.3390/ijerph15061204

Yim, M. & Shin, S. (2020). Using the angoff method to set a standard on mock exams for the korean nursing licensing examination. *Journal of Educational Evaluation for Health Professions,* 17, 14. https://doi.org/10.3352/jeehp.2020.17.14

Yousef, M., Alshawwa, L., Tekian, A., & Park, Y. (2017). Challenging the arbitrary cutoff score of 60%: standard setting evidence from preclinical operative dentistry course. *Medical Teacher,* 39, 75-79. https://doi.org/10.1080/0142159x.2016.1254752