Hindawi
Advances in Human-Computer Interaction
Volume 2022, Article ID 7527457, 11 pages
https://doi.org/10.1155/2022/7527457

Research Article

A Framework of the Training Module for Untrained Observers in Usability Evaluation Motivated by COVID-19: Enhancing the Validity of Usability Evaluation for Children's Educational Games

Marwan Alshar'e,[1] **Ali Albadi**,[2] **Malik Mustafa**,[3] **Noman Tahir**,[1] **and Marya Al Amri**[1]

[1]*Faculty of Computing Sciences, Gulf College, Seeb, Oman*
[2]*Faculty of Computing Sciences and Centre for Postgraduate Studies and Research, Gulf College, Seeb, Oman*
[3]*Center for Foundation Studies, Gulf College, Seeb, Oman*

Correspondence should be addressed to Marwan Alshar'e; marwan@gulfcollege.edu.om

Received 22 December 2021; Revised 23 January 2022; Accepted 10 February 2022; Published 9 March 2022

Academic Editor: Christos Troussas

The usability evaluation of educational games is an important task, especially for children. By applying Jakob Nielsen's ten heuristics, most of the HCI designs can be evaluated, but when educational games are involved, where the user being observed is a child between the ages of six and eight, many questions arise. Is the observer trained well enough to observe the child's reactions to the game with regard to its memorability, learnability, ease of use, and enjoyment? Will it be necessary for the observer to have a training session exploring the game before evaluating a child? Our research suggests that a training module designed to train an untrained facilitator (observer) in how to evaluate four usability dimensions (learnability, memorability, ease of use, and enjoyment) would be very useful. The usability evaluation data was collected by observing users playing generic educational games, using the Mann–Whitney U test, which was conducted by two groups of observers, one trained and one untrained. This was then reviewed, and a distinct difference was found between the results of evaluations in the two groups, thus validating the importance of training for an observer.

1. Introduction

The slogan "user friendly" appeared popular during the 1980s, but since the 1990s, the focus of usability engineering has relied heavily on the elaboration of usability evaluation methods. The usability engineering books by [1, 2] set the basis of encompassing the concept of human-computer interaction (HCI). The first decade of the twenty-first century developments regarding usability analysis had software-flavored tactics such as user interface implementation through software tools, standards, and "look and feel" aspects. This move enhanced the awareness of the need to work on evaluating usability through the user interface as a medium. Nielsen defined the usability through five elements (i.e., ease of learning, efficiency of interaction, ease of remembering, frequency and seriousness of errors, and user satisfaction). According to [3], the usability is the degree of efficiency, effectiveness, and user satisfaction obtained from a product used in a particular environment. Similarly, there is a definition of usability from the International Organization for Standardization (ISO), which is "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [4].

Traditionally, there are certain usability evaluation methods such as behavioral analysis, heuristic evaluation,

cognitive walking, interviews, and questionnaires. Besides the traditional methods, a variety of other interface evaluation methods are in use nowadays. However, few of them have considered the importance of the role and relevant education of the evaluator. Reference [5] worked on the evaluator effect and revealed that there was a considerable variation in the results reported by different evaluators using the same application in similar conditions. This casts doubt on the effectiveness of these methods.

The heuristic evaluation method, conducted by experts, is one of the preferred methods for assessing the usability of games. Most of the modular heuristic evaluations are conducted using Neilson's proposals that focus on software [6, 7]. Recent research by [8], in which a systematic review on heuristics was conducted, have reported that the heuristics and usability of educational games still offer potential for exploration in specific evaluation and validation of the process.

Reference [9] presents a set of heuristics resembling traditional heuristics while emphasizing the context of their use. Many of the recent studies have tried to provide platforms to enhance the outcomes of usability evaluations, but none of the studies so far has attempted to explore the competence level of evaluators in a specific cultural context. Reference [10] reveals that the inappropriateness of feedback on errors and the inadequate interaction with educational games are still little explored. Therefore, the evaluation of game-based learning software needs to be accurate in observation, which makes the role of observers a vital one.

Regarding the usability evaluation method, Cognitive Usability Evaluation (CUE-E) is a new dimension in addition to traditional heuristics [11]. There is continuous optimism about increased access to the network-based technologies for encouraging young learners to pursue their interest in technology-based learning [12, 13]. The development and improvement in games design are very much needed in order to create a learner-centric approach in development which facilitates learning. Feedback on learning through the observations provides the input for improvement. In this scenario, the role of the observers is very important in the context of their competencies of observation and their ability to provide reliable and valid input. A framework of heuristics reflecting the specific learning role support for educators is proposed, which is very important for them [14]. Many of the previous studies on the role of the observer have been focused mainly on the perspective of learning [15–17]. In studying the role of the observers, the active observation is scripted and promoted as potentially serving the learning experience by providing stimulation. To offer learning experiences, observers dwell in learning intentions by maintaining distance and detachment as these depict their part into evaluations [18–21].

Keeping the evidence from literature in view, the current study infers that the role of the observers under the preview of their competence is still unexplored. Cultural context and symbolism in educational games have a great potential for catching the attention of users. While the effectiveness of an evaluation method may not result in similar outcomes in varying cultural contexts and learning perceptions, the competence of the observers in evaluating the usability is very important in order to ensure the reliability of the input in usability evaluations.

Therefore, the current study aims to propose a training module conducted remotely for the observers. The research aims to improve the role of the observers based on learning from the divergence in observations of the learners in a real-life situation. The research oversees the observer's role in user system interaction.

2. Literature Review

2.1. Usability Testing Methods. Moderated testing is done using phone, video, or interviews with the users in any HCI design. Lab and guerrilla testing methods are also concerned with moderated testing. In this usability testing platform, phone, interviews, and video testing can all be conducted remotely. Lab and guerrilla testing must be carried out in person. Moderated testing conducted remotely has a high success rate in collaborative usability testing of virtual reality systems [22, 23].

2.1.1. Lab Usability Testing: In Person vs. Remote. A user's ability to complete the task or set of tasks within the time frame can be tested by using lab usability testing. This testing can be in person or remote. An empirical comparison between the lab and the remote usability testing of websites was conducted, where 8 participants were tested through the lab and 38 people were tested remotely [24]. The average subjective ratings of usability given by the users are shown in Figure 1.

2.2. Affordance Testing. "Affordance" refers to the features of an object, software program, website, or any other application that provides a default clue about how to use it. Bower conducted an affordance analysis on an online educational site that offers additional affordance testing including "link-ability," "highlight-ability," and "permission-ability." The study concluded that interactions between affordance and operations performed make a significant impression on learning experiences. The study further proposed numerous levels of awareness for the online educational program [25, 26].

2.2.1. Learnability Testing. The feature of "learnability" helps the users to familiarize themselves quickly with the tasks allowed by the provided interface. Evaluation of learnability for learning management systems is of great importance. A study was conducted at King Abdul Aziz University, Saudi Arabia, in order to evaluate the usability and learnability of its LMS "Blackboard." The investigation concluded that LMS is reliable and is well designed but still lacks the ability to guide distance education learners. It also found that Blackboard violates some of the basic usability guidelines [27].

2.2.2. Ease of Use Testing. A study was conducted for usability testing of a school website developed to provide

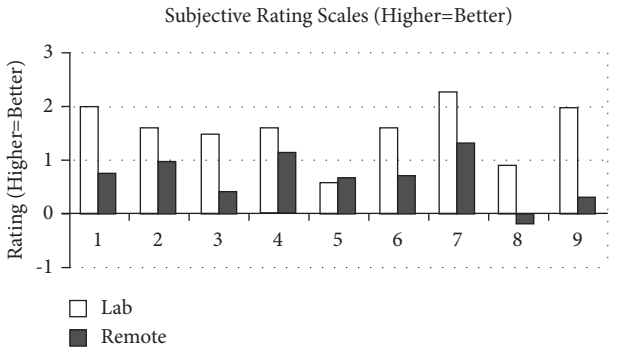

Figure 1: Average subjective ratings given on nine rating scales for both lab and remote tests. Source: [24]. Scale: −3 to +3 where higher ratings are better. Averages: lab = 1.6, remote = 0.7, $r = 0.49$.

relevant information to parents and visitors at Kennesaw State University. A qualitative approach was adopted to conduct testing using observation and think-aloud methods. There were different tasks given to be performed by users with a rating system of zero to ten. In some of the more challenging tasks, it was found that the website was not easy to use at a certain level. It was concluded that Dunwoody School's website contains beneficial resources for school community members but that testing did not result in enhancing its "ease of use" factor [28].

2.3. Mobile Usability Testing. Millions of apps are in use but, at the same time, the number of applications fulfilling the HCI and usability standards is very small. In fact, it is important to know what usability evaluation methods have been applied to different mobile applications and where we stand.

2.3.1. Mobile Usability Evaluation Models. The first model was introduced in 2002 for the evaluation of a set of usability dimensions. These measures included navigation, input rate and menu visualization, presentation, error prevention, navigation, and contents and architecture of information design, but this study was not able to specify how each dimension is explicitly connected to each usability dimension [29]. In 2006, a study purely related to highlighting the challenges faced in evaluating the usability dimensions was conducted. This discussion was based on one hundred and eighty publications published in core human-computer interaction journals [30].

In [31], a model for usability measurements of mobile applications has been introduced. This model was designed after a review of hundreds of empirical papers. This model proposed guidelines for researchers to adopt the way of usability evaluation of mobile applications in general, but this research still lacks guidance on which usability dimension should be chosen for specific types of mobile application.

The mGQM model was built in 2011 (and revised in 2017) based on ISO 9241-11 usability measurements [32]. This model is comprehensive enough to measure effectiveness, efficiency, and satisfaction but also lacks guidance on setting the dimensions of specific mobile applications [33].

The PACMAD model opens the discussions, where the authors argue that mobile applications require a specific model and there is an extension required in existing models such as Nielsen or ISO models to measure usability dimensions [34].

2.3.2. Usability Evaluation for Children's Mobile Learning Applications. It is important to understand the user preferences when designing the behavior of any system, whether it is online or mobile-based [35]. A study conducted in 2016 was based on evaluating the quality attributes of mobile learning applications for children. It reviewed the top four usability quality attributes which are efficiency, effectiveness, learnability, and user satisfaction. The purpose of this research is to explore the current literature on the subject matter as well as creating a simulation for further studies that may improve the usability and design for mobile learning apps for children [36].

The literature explains succinctly how usability and HCI are taught. In [38], the authors explained how the use of a case study performed by the students relates to the life cycle of usability. Furthermore, [39] presents an assessment developed by the students to explain the application of heuristic evaluation. Reference [40] finalizes a method that enables students to apply certain techniques in addressing usability conducted through the testing and analysis of results. In [41], the authors have conducted a usability study involving students who used a set of web pages to answer questions about the usability of these pages. An investigation of the usability of "Blackboard" at King Abdul Aziz University, Saudi Arabia, concludes that LMS is reliable and is well designed but still lacks the ability to guide distance education learners and that it also violates some basic usability guidelines [27]. In [36], the authors carried out a study reviewing the top four usability quality attributes which are efficiency, effectiveness, learnability, and user satisfaction. The study recommends conducting further studies to improve the usability and design of mobile learning apps for children.

The literature review provides systematic information about the usability evaluations of four dimensions which are learnability, memorability, ease of use, and enjoyment for educational online systems as well as mobile applications. It also refers to the fact that usability evaluations are normally conducted by researchers who are directly involved in data gathering and validates the fact that in-person evaluation results are more efficient than remotely done evaluations.

3. Training Module Framework

This study proposes a training module framework to train a novel observer to conduct usability evaluations of mobile app-based educational games for the age group of 6–8 years. This section elaborates on the phases of the suggested training module, as well as conducting a pilot study to validate the module.

3.1. Training Framework Proposition. Our training framework specifies a fast-track comprehensive descriptive training design to train a novel observer who has no background in information technology or HCI and usability. The training is very specific for children's educational mobile app games for the 6–8-year age group in order to assess the learnability, memorability, enjoyment, and ease of use. The proposed framework is described in Figure 2.

3.2. Phase 1: Training Need Assessment. Whatever research has been conducted so far where the observation of children is involved to evaluate the learnability, memorability, ease of use, and enjoyment for any educational game, it was not believed that there might occur circumstances where researchers would not be able to conduct direct observation themselves all the time, which is currently the situation because of COVID-19. Section 2.1.1 of the literature review also justifies the validity of in-person evaluation being better than remote observations, specifically when children are involved. In this situation of COVID-19, where classes and most of the educational procedures have gone online, it is felt to be essential to design a training module to train any untrained observer that might be one of the parents of the children, their guardian, or a person who can directly observe a child in order to evaluate the mentioned dimensions to test the usability of educational games.

3.3. Phase 2: Training Requirements Analysis

3.3.1. Environment Analysis. In the current circumstances of COVID-19, it would seem difficult for a researcher to conduct usability evaluations of any educational game directly, especially when children are to be observed. One of the main barriers is to follow a comprehensive ethical procedure to reach someone's child, which may include a pre-COVID-19 negative test certificate obtained within an appropriate number of hours [42] and the confirmation of not infecting the child under any circumstances. Even then, parents are extremely protective towards their children.

3.3.2. Selection of the Observer. The training reflects absolutely the knowledge and understanding of usability testing of the said dimensions. The potential observer might be a novel computer user and may not have any idea about how to use the IT tools, and even with an efficient computer expert there is no guarantee that they will be a good usability tester.

3.3.3. Game Selection. With insufficient availability of educational games in other languages, a good ranking English generic educational mobile application for children in the 6–8-year age group is appropriate. It should also be ensured that the subjects of the selected game should be generic, such as English, Maths, or Science which are common across the globe.

3.3.4. Trainer's Selection. A certified usability testing training from ISTQB (International Software Testing Qualifications Board) or a B.S. in HCI and usability engineering with sufficient industrial experience of mobile app testing and evaluation procedures is considered to be appropriate training for the proposed training module.

3.3.5. Training Material. Training material is generic, where the trainer is required to cover the following learning outcomes:

(1) To provide an overview of usability evaluation for learnability, memorability, ease of use, and level of enjoyment [43]

(2) To evaluate the usability of an educational game for learnability, memorability, ease of use, and level of enjoyment [44]

It is the choice of the trainer to develop the training material to cover the above learning outcomes.

3.4. Phase 3: Training Module Delivery

3.4.1. Training Mode. In the circumstances of COVID-19 referred to above, the best way to deliver the training is by adopting an online training strategy. Any of the recommended online teaching tools can be utilized for this purpose. The tools adopted must be made available to the trainees, and it should be ensured that they either know how to use them or are provided with an understanding of their usage.

3.4.2. Tools. The trainer is free to select the tools that can be utilized for all training purposes. Trainers can prepare the handouts as well as utilizing any multimedia tools that facilitate the training sessions.

3.4.3. Training Sessions. Training sessions can be designed as per the trainer's delivery plan, which should be sufficient to train the trainees. However, the recommended number of training sessions should be at least four, where each of the dimensions should be covered within an hour [45]. The learning outcomes should be covered to a satisfactory extent.

3.4.4. Training Activities and Feedback. The arrangements of the training session and the expected procedures and outcomes are explained in Table 1.

3.5. A Pilot Run of a Training Session. A group of 20 observers including parents and teachers were selected for this research purpose. The selection was made randomly by inviting school teachers and parents from Muscat, a city of Sultanate of Oman, to the training session. It was ensured that none of the selected trainees were aware of usability evaluation procedures and that they understood English very well. For research purposes, "Math Games" by GunjanApps Studios was selected because of the number of downloads it had received and the star ranking given to the mobile app-based game. The selected game is the highest ranking, having received 4.4 stars and been downloaded

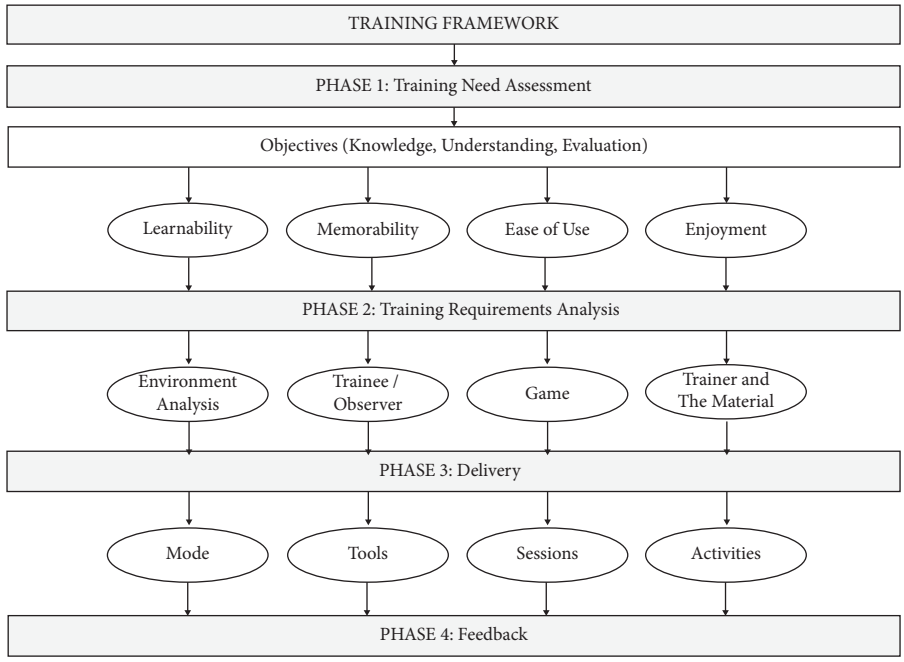

Figure 2: Proposed training framework.

153,000 times. The selected users were children in the 6–8 age group studying in an English medium School in Oman. For this research purpose, an academician with 5 years' experience of teaching HCI and usability was selected. Online training was conducted where the introduction, knowledge, and understanding of learnability, memorability, enjoyment, and ease of use were delivered for 1 hour each, with a break of 15 minutes between each topic. The trainer utilized MS PowerPoint to prepare the presentation as training material, and the observers were requested by the trainer to join online through MS Teams at a specified time. He developed the training handouts by utilizing the book "Student Usability in Educational Software and Games" [46] in order to cover the dimension of "learnability." To cover the "memorability" and "ease of use" dimensions, he designed the handouts from the book "Usability and User Experience Studies" [47]. In order to design the handouts for the "enjoyment" dimension, "The Mobile Learning Voyage - From Small Ripples to Massive Open Waters" [48] was reviewed by the trainer. A live question-and-answer session was delivered after the training to gather feedback from the trainees and make sure they had gained a thorough understanding of the subject.

4. Methodology

The current research used a descriptive design based on the experiment. The purpose of experimental design is to find out how evaluators allocated to different groups provide data based on their observations. To describe a phenomenon systematically, the researchers used field data to investigate and compare the behavior of data related to different variables. Hence, analytical research established the existence of observed facts in order to describe the phenomenon, which was not otherwise possible without applying this method [49]. A total of 40 observers were selected for two groups through two-stage sampling. Firstly, two groups were identified. Secondly, observers for each targeted group were selected. The first group were 20 observers trained for the proposed training framework given in Section 3 as explained in Section 3.5. The second group were the parents and private mentors of students, from which 20 observers were selected with informed consent. Each of the observers was assigned to record observation data on one learner only. Hence, a sample of 40 observers participated in the investigation in order to provide their observations of learners' experiences on the game's usability. To compare the difference in variables between groups, the Mann–Whitney U test [50], as cited in [51, 52], was used to determine whether the distribution of variables was the same for the two groups and that the samples were likely to have been derived from the same population.

4.1. Measurement Scale. To measure the responses of trainers regarding usability dimensions (i.e., learnability, memorability, ease of use, enjoyment), an instrument was developed by adapting the measurements of the different items from previous research. A self-administered structured questionnaire scaling the items on the 5-point Likert scale was used to collect the data. Table 2 shows the measurement of the dimensions of the items for the four criteria under observation, which is directly related to the learners' experience of the selected dimensions of usability of the game under investigation.

4.2. Training Module Evaluation

4.2.1. Procedure of Experiment. The procedures of the experiment consist of four steps as follows:

Table 1: Training activities.

Session duration	Training activities	Expected outcomes
4 hours	1 hour session for each selected dimension of usability evaluation design: Explanation and demonstration on understanding and assessing the cognitive aspects of dimensions (i.e., learnability, memorability, enjoyment, ease of use)	Observer's ability to observe and report the behavior of users on any given dimension
1 hour	Informative feedback from attendees of the session on current training using a scale-based adaptive instrument	Evaluation of trainees' understanding in observing the dimensions of the proposed usability evaluation design

(1) Explanation of the purpose of data collection to the observers

(2) Instructions on how to execute data collection (instrument)

(3) Selection of the game and explanation to the observers of how to choose learners

(4) Determination of the duration and time frame of the observation

(5) Collection of the data for the completed tasks and analysis of this data

4.2.2. Results and Analysis. To evaluate the difference between trained and untrained observers regarding the reporting for learnability, the data was tested using the Mann–Whitney U test. As shown in Table 3, there is not enough evidence to support the assumption of similarity between observer groups for selected dimensions of usability design (i.e., learnability, memorability, enjoyment, and difficulty of use).

The Mann–Whitney test indicated a statistically significant difference in learnability: Group 1 (Mdn = 2.75, $n = 20$), Group 2 (Mdn = 3.00, $n = 20$), $U = 79.500$, $p = 0.001$, $r = 0.75$; memorability: Group 1 (Mdn = 3,00, $n = 20$), Group 2 (Mdn = 3.442, $n = 20$), $U = 112.00$, $p = 0.017$, $r = 0.55$; enjoyability: Group 1 (Mdn = 3.00, $n = 20$), Group 2 (Mdn = 3.40, $n = 20$), $U = 76.00$, $p = 0.001$, $r = 0.77$; and difficulty of use: Group 1 (Mdn = 2.88, $n = 20$), Group 2 (Mdn = 3.20, $n = 20$), $U = 121.500$, $p = 0.033$, $r = 0.48$. Therefore, there are significant differences in the medians of the different factors between the groups of observers. As shown in Table 4 and Table 5, the null hypothesis for similarity is rejected for all four dimensions of usability design.

4.3. Descriptive Analysis. A descriptive analysis executed in this study evaluates the perceptions of the respondents on the five study variables, as regards the importance levels of the variables. Hence, the success of the model for untrained observers in usability evaluation could be determined. Accordingly, the means, the standard deviations, the minimum level, and the maximum level of the feedback on the research variables obtained from the respondents can be viewed in Table 6. There are five levels of agreement that can be selected by the respondents for the items and variables, and in order to ease interpretation, the levels were split into three categories of low level, high level, and moderate level. Specifically, low level is concluded when the mean scores are lower than 2.33, high level is concluded when the obtained mean scores are higher than 3.67, and moderate level is concluded when the mean scores are between 2.33 and 3.67.

4.4. Structure Model. The evaluation of the measurement model or the outer model involves the measurement of convergent validity and discriminant validity. Convergent validity determines the ability to differ instruments in measuring the exact construct [57] and it also demonstrates to what extent the instruments are in agreement with one another. Convergent validity also includes reliability of construct measurement using composite reliability and internal consistency. Meanwhile, the internal consistency can be measured using Cronbach's Alpha coefficient where the alpha value should be greater than 0.7 to be interpreted as having reliability. Additionally, Table 7 shows composite reliability (CR) of greater than 0.7 for all constructs.

Further, all constructs in this study had their scales' internal consistency (ICR) verified. Additionally, convergent validity can be determined through the use of the factor loadings of the model's construct items. Accordingly, [58] stated that items with loadings of 0.70 or higher should be retained as they denote convergent validity. Five constructs (learnability (EF), ease of use, memorability, enjoyability, and educating untrained observers in usability evaluation) showed loading value of greater than 0.7 (factor loading) and those fulfilling the requirement of threshold value were analyzed further.

5. Discussion

The outcomes of the research have revealed that there is a significant difference in the observations on the usability dimensions of the game between the two groups of observers. The research successfully answered the research question posed in the Introduction. The study expected that the observations by the trained observers, who are considered to be familiar with usability and experienced in the HCI environment and its issues, will vary from those by the group of untrained observers, who had no background in technical and cognitive issues regarding HCI. This was indeed revealed by the statistical analysis of ranks and medians, which showed a significant difference in the observation data between the two groups relating to selected dimensions of the usability model of the selected game. The results are consistent with previous research reported in the

Table 2: Dimension criteria.

Dimension	Measurement items
Learnability [53]	(i) Selection of menu to start the game (ii) The command on navigational controls from one activity to another (iii) The level of completion of the task given to the user (iv) The level of completion of the task within the time frame given to the user (v) The number of errors committed while performing a task (vi) Fixation and scan path duration
Ease of Use [54]	(i) The amount of trouble it takes to move from one activity to another (ii) Time spent on understanding a process (game step) (iii) Time spent on watching a process (game step) (iv) Time spent on understanding (how and when) to pause the game (v) Time spent on understanding (how and when) to move from one level to another
Memorability [55]	(i) The time spent in accomplishing the first level of the game in the first attempt (ii) The time spent in accomplishing the second level of the game in the second attempt (iii) The difference in timings between the first and second attempts
Enjoyment [56]	(i) The level of concentration of the user when playing the game (ii) The level of immersion in the game that is experienced by the user (iii) The ease of distracting the user from the game (iv) The strength of feeling and sense of control over the game experienced by the user (v) The strength of motivation of the user to play again

Table 3: Ranks.

	Group	N	Mean rank	Sum of ranks
Learnability	Trained observers	20	14.48	289.50
	Untrained observers	20	26.53	530.50
	Total	40		
Memorability	Trained observers	20	16.10	322.00
	Untrained observers	20	24.90	498.00
	Total	40		
Enjoyability	Trained observers	20	14.30	286.00
	Untrained observers	20	26.70	534.00
	Total	40		
Difficulty of use	Trained observers	20	16.58	331.50
	Untrained observers	20	24.43	488.50
	Total	40		

Table 4: Mann–Whitney U test statistics.

	Test statistics[a]			
	Learnability	Memorability	Enjoyment	Difficulty of use
Mann–Whitney U	79.500	112.000	76.000	121.500
Wilcoxon W	289.500	322.000	286.000	331.500
Z	−3.353	−2.478	−3.456	−2.149
Asymp. sig. (2-tailed)	0.001	0.013	0.001	0.032
Exact sig. [2*(1-tailed sig.)]	0.001[b]	0.017[b]	0.001[b]	0.033[b]

a: grouping variable: group. b: not corrected for ties.

Table 5: Summary of assumptions validation.

Hypothesis test summary (independent-samples Mann–Whitney U test)		
Supposition	Sig	Decision
The distribution of learnability is the same for both groups	0.001*	Rejected
The distribution of memorability is also same for both groups	0.017*	Rejected
The distribution of enjoyment is the same across categories of group	0.001*	Rejected
The distribution of difficulty of use is the same across categories of group	0.033*	Rejected

Asymptotic significance is displayed. The significance level is 0.05. *, $p < 0.05$.

Table 6: Descriptive statistics for and coefficients of correlations variables ($N = 20$).

Construct	M[a]	SD[b]	1	2[c]	3	4
Learnability	3.74	3.643	0.75	0.55**	0.43**	0.547**
Ease of Use	3.82	3.849		0.47**	0.42**	0.781**
Memorability	3.74	3.767			0.38**	0.354**
Enjoyment	3.88	3.829				0.374**
Educating untrained observers in usability evaluation	3.78	3.871				

[a]Median; [b]standard deviation; **significant at 0.01 Level.

Table 7: Result of construct assessment.

Constructs	Items	Factor loading	Mean ± SD	CR	Cronbach's α	AVE
Learnability	**Lea 1**	0.787	3.815 ± 0.926	0.928	0.815	0.521
	Lea 2	0.745	3.825 ± 0.936			
	Lea 3	0.741	3.845 ± 0.956			
	Lea 4	0.786	3.816 ± 0.921			
	Lea 5	0.787	3.817 ± 0.936			
	Lea 6	0.785	3.715 ± 0.826			
Ease of Use	**EoU 1**	0.783	3.972 ± 0.971	0.814	0.981	0.576
	EoU 2	0.882	3.712 ± 0.825			
	EoU 3	0.784	3.713 ± 0.974			
	EoU 4	0.742	3.714 ± 0.975			
	EoU 5	0.881	3.854 ± 0.923			
Memorability	**Mem 1**	0.787	3.716 ± 0.923	0.821	0.912	0.531
	Mem 2	0.787	3.717 ± 0.932			
	Mem 3	0.786	3.815 ± 0.822			
Enjoyment	**Enj 1**	0.787	3.716 ± 0.923	0.834	0.971	0.586
	Enj 2	0.787	3.717 ± 0.932			
	Enj 3	0.786	3.815 ± 0.822			
	Enj 4	0.785	3.612 ± 0.975			
	Enj 5	0.784	3.775 ± 0.921			
Educating untrained observers in usability evaluation	**EUOUE 1**	0.874	3.647 ± 1.072	0.834	0.991	0.556
	EUOUE 2	0.758	3.624 ± 1.024			

AVE: average variance extracted, SD: standard deviation, Lea= learnability; EoU= ease of use; Mem= memorability, Enj= enjoyment; EUOUE= usability evaluation.

literature review which concludes that the context and ability of observers may produce different results.

The outcomes of the research give merit to the proposal for the training of observers. The uniformity of approach and common understanding of the aims of observation for observers will increase the validity of observations on the usability of games. Observers indicate usability and provide precise and constructive feedback that clarifies the practical aspects of games. An observer's ability to estimate and record the data, to be alert, and to control emotions are key qualities for a good observer. The observers' training will give them a broader reach, enabling them to understand and appreciate the need for the observation of the cognitive skills displayed by the user. Additionally, understanding the importance of neutrality and noninterfering behavior would become a part of knowledge enhancement related to observation.

6. Implications

The present research provides evidence on understanding issue-based empirical data related to actual conditions. The study guides the readers to consider the contextual environment of the learners, for whom learning games are designed and who, indeed, benefit. The results provide a clear guideline on evaluating the usability of games and how evaluation carried out by heterogeneous groups of observers may produce misleading results. Theoretically, the research provides an insight into understanding the importance of the cognitive learning aspects of learners and the competence of observers in the evaluation of usability. The research also guides the directions of future research into linguistics, exemplification, and esthetic utility of games.

7. Research Finding

The present study is significant to the domain of educational games evaluation as it sheds light on the subject in the midst of COVID-19 pandemic. This study addressed the issue of reaching the target audiences in the assessment of a certain systems amidst the pandemic, by bringing forth an innovative approach. Accordingly, the needed theoretical knowledge in dealing with the problems is offered by this study. A new model comprising various variables with strict coordination and other variables was proposed.

This study looks into the important considerations in remote evaluation of systems in situations where it is not

possible or feasible for evaluators or practitioners to reach their target audiences, such as during COVID-19 pandemic situation.

Usability evaluation of educational games is regarded as an important task to several bodies including game developers, departments of research and innovation, and, in the context of this study, the ministry of education. Accordingly, the usability evaluation can be applied in evaluating the potential of education games and the children's acceptance of these games. Accordingly, the usability evaluation in Oman was clarified succinctly in this study. Moreover, the factors impacting usability evaluation of educational games for children specific to the context of Oman during COVID-19 pandemic were identified. The measurements and conceptual framework illuminating the relationship between trainers were developed, where learnability (EF), ease of use, enjoyment, and memorability were the independent variables, while the construct of "educating untrained observers in usability evaluation" was the dependent variable.

7.1. Research Contributions. Several issues significant to the situation at hand were discussed in this study. For this purpose, this study established and examined a framework of the training module for untrained observers in usability evaluation during COVID-19 pandemic, to improve the validity of usability evaluation for children's educational games. In view of that, this study had made several contributions as discussed below:

(i) This study helps authorities in their remote execution of usability evaluation as needs arise

(ii) This study becomes a valued input to the scrutiny on the impact factors affecting usability evaluation of educational games

(iii) This study enriches the knowledge on the benefits of and prerequisites for the improvement of usability evaluation of educational games in Omani context

(iv) This study presents a model that describes the impact of some factors on usability evaluation of educational games in Oman

More importantly, the research model proposed was applied in the examination of variables analysis of usability evaluation for educating untrained observers in usability evaluation in Oman.

8. Recommendations

Based on the findings, the researchers recommend the training module (framework) detailed in Section 3 in this paper for untrained observers who want to observe the children for usability evaluation of an educational game to produce more valid and accurate results.

The major considerations in developing an observer's deployment would cover the extent of knowledge needed by an independent observer.

9. Limitations of the Study

Because of the unavailability of games in a local language (Arabic) interface, we decided not to investigate the differences in communicative attributes of games. Due to time constraints and limited access due to COVID-19, a large enough sample size was not assured.

Another limitation in assessing the differences between observers involved the issue of exposure of the learners to the game. The learners were not classified as being first-time users or already experienced in using games. The difference in experience itself may have impacted on the comparison between the two groups.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research leading to these results has received funding from the Ministry of Higher Education, Research and Innovation (MoHERI) of the Sultanate of Oman under the Block Funding Program (block funding agreement no. TRC/BFP/GULF/01/2019; project code: BFP/RGP/ICT/19/292).

Supplementary Materials

The supplementary materials include the questions that composed the questionnaires used in this study. . (*Supplementary Materials*)

References

[1] J. Nielsen, *Usability Engineering*, Morgan Kaufmann, Burlington, MA, USA, 1994.

[2] D. J. Mayhew, "The usability engineering lifecycle," *CHI '99 extended abstracts on Human factors in computing systems-CHI '99*, p. 147, 1999.

[3] N. Bevan and M. Macleod, "Usability measurement in context," *Behaviour & Information Technology*, vol. 13, no. 1-2, pp. 132–145, 1994.

[4] M. Hertzum, "Usability testing: too early? Too much talking? Too many problems?" *J. Usability Stud.*vol. 11, no. 3, pp. 83–88, 2016.

[5] M. Hertzum, "A usability test is not an interview," *Interactions*, vol. 23, no. 2, pp. 82–84, 2016.

[6] M. I. Alsharâ™e, R. Sulaiman, M. R. Mokhtar, and A. MohdZin, "Design and implementation of the TPM user authentication model," *Journal of Computer Science*, vol. 10, no. 11, pp. 2299–2314, 2014.

[7] R. Yanez-Gomez, J. L. Font, D. Cascado-Caballero, and J.-L. Sevillano, "Heuristic usability evaluation on games: a

modular approach," *Multimedia Tools and Applications*, vol. 78, no. 4, pp. 4937–4964, 2019.

[8] E. A. O. Vieira, A. C. d. Silveira, and R. X. Martins, "Heuristic evaluation on usability of educational games: a systematic review," *Informatics in Education*, vol. 18, no. 2, pp. 427–442, 2019.

[9] L. Goosen and S. A. Ajibola, "Information systems architecture and technology security aspects relating to the usability attributes and evaluation methods of mobile commerce websites," *Advances in Intelligent Systems and Computing*, Springer Verlag, vol. 1050, pp. 328–337, 2020.

[10] A. Granić, J. Nakić, and N. Marangunić, "Scenario-based group usability testing as a mixed methods approach to the evaluation of three-dimensional virtual learning environments," *Journal of Educational Computing Research*, vol. 58, no. 3, pp. 616–639, 2020.

[11] M. S. Pfaff, A. Anganes, O. Eris, A. Prior, M. Ward, and J. Nebeker, "Cognitive usability evaluation of electronic health record systems (CUE-E)," *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, vol. 8, no. 1, pp. 13–17, 2019.

[12] R. L. Maata and S. Pagcaliwagan, "A proposed human computer interaction (HCI) model through the use of computerized speech laboratory (CSL)," *International Journal of Research in Engineering and Technology*, vol. 3, no. 12, pp. 20–26, 2016.

[13] M. Mustafa, S. Alzubi, and M. Alshare, "The moderating effect of demographic factors acceptance virtual reality learning in developing countries in the Middle East," *Communications in Computer and Information Science*, vol. 1244, pp. 12–23, 2020.

[14] D. Nacu, C. K. Martin, and N. Pinkard, "Designing for 21st century learning online: a heuristic method to enable educator learning support roles," *Educational Technology Research & Development*, vol. 66, no. 4, pp. 1029–1049, 2018.

[15] G. Funke, E. Greenlee, M. Carter, A. Dukes, R. Brown, and L. Menke, "Which eye tracker is right for your research? Performance evaluation of several cost variant eye trackers," *Proceedings of the Human Factors and Ergonomics Society-Annual Meeting*, vol. 60, no. 1, pp. 1240–1244, 2016.

[16] D.-H. Kim, T. J. Y. Kim, X. Wang et al., "Smart machining process using machine learning: a review and perspective on machining industry," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 5, no. 4, pp. 555–568, 2018.

[17] M. Alshar'e and M. Mustafa, "Evaluation of autistic children's education in Oman: the role of eLearning as a major aid to fill the gap," *Elementary Education Online*, vol. 20, no. 5, 2021.

[18] E. Thorup, P. Nyström, P. Nyström, G. Gredebäck, S. Bölte, and T. Falck-Ytter, "Reduced alternating gaze during social interaction in infancy is associated with elevated symptoms of autism in toddlerhood," *Journal of Abnormal Child Psychology*, vol. 46, no. 7, pp. 1547–1561, 2018.

[19] A. B. Landman, L. Redden, P. Neri et al., "Using a medical simulation center as an electronic health record usability laboratory," *Journal of the American Medical Informatics Association*, vol. 21, no. 3, pp. 558–563, 2014.

[20] O. Rantatalo, D. Sjöberg, and S. Karp, "Supporting roles in live simulations: how observers and confederates can facilitate learning," *Journal of Vocational Education and Training*, vol. 71, no. 3, pp. 482–499, 2019.

[21] R. A. Jones and S. A. Bogle, "An investigation of the use of Facebook groups as a Learning Management System to improve undergraduate performance," *Proceedings of the World Congress on Engineering and Computer Science*, vol. 1, 2017.

[22] M. Alshar'E, A. M. Zin, R. Sulaiman, and M. R. Mokhtar, "Evaluation of the TPM user authentication model for trusted computers," *Journal of Theoretical and Applied Information Technology*, vol. 81, no. 2, 2015.

[23] K. Chalil Madathil and J. S. Greenstein, "An investigation of the efficacy of collaborative virtual reality systems for moderated remote usability testing," *Applied Ergonomics*, vol. 65, pp. 501–514, 2017.

[24] T. Tullis, S. Fleischman, M. McNulty, C. Cianchette, and M. Bergel: An Empirical Comparison of Lab and Remote Usability Testing of Web Sites," 2002.

[25] M. A. Masethe, H. D. Masethe, and S. A. Odunaike, "Scoping review of learning theories in the 21 st century," *Proceedings of the World Congress on Engineering and Computer Science*, vol. 1, pp. 25–27, 2017.

[26] M. Bower, "Affordance analysis-matching learning tasks with learning technologies," *Educational Media International*, vol. 45, no. 1, pp. 3–15, 2008.

[27] K. Al-Omar, "Evaluating the usability and learnability of the "blackboard" LMS using SUS and data mining," in *Proceedings of the 2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 386–390, Erode, India, February 2018.

[28] U. Kokil and S. Scott, "Usability testing of a school website using qualitative approach," in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications {(VISIGRAPP} 2017) - Volume 2: HUCAPP*, pp. 55–64, Porto, Portugal, February 2017.

[29] R. Agarwal and V. Venkatesh, "Assessing a firm's web presence: a heuristic evaluation procedure for the measurement of usability," *Information Systems Research*, vol. 13, no. 2, pp. 168–186, 2002.

[30] K. Hornbæk, "Current practice in measuring usability: challenges to usability studies and research," *International Journal of Human-Computer Studies*, vol. 64, no. 2, pp. 79–102, 2006.

[31] C. K. Coursaris and D. Kim, "A meta-analytical review of empirical mobile usability studies," *J. Usability Stud. Arch.*vol. 6, pp. 117–171, 2011.

[32] F. Zahra, A. Hussain, and H. Mohd, "Usability evaluation of mobile applications; where do we stand?" *AIP Conference Proceedings*, vol. 21891, p. 020056, 2017.

[33] A. Hussain: Metric Based Evaluation of Mobile Devices: Mobile Goal Question Metric (mGQM)," Salford Bus. Sch. Univ. Salford, Salford, UK Submitt. Partial Fulfilment Requir. Degree Dr. Philos, January, 2012.

[34] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model," *Journal of Interaction Science*, vol. 1, no. 1, p. 1, 2013.

[35] A. S. Halibas, R. L. Maata, and R. O. Sibayan, "Usability evaluation of selected Oman-based eCommerce websites," *Asian Journal of Business Management*, vol. 4, no. 4.

[36] E. O. C. Mkpojiogu, A. Hussain, and F. a. Hassan, "A systematic review of usability quality attributes for the evaluation of mobile learning applications for children," *AIP Conference Proceedings*, vol. 2016, p. 020092, 2018.

[37] J. Moizer, J. Lean, E. Dell'Aquila et al., "An approach to evaluating the user experience of serious games," *Computers & Education*, vol. 136, pp. 141–151, 2019.

[38] J. M. Carroll, M. B. Rosson, G. Convertino, and C. H. Ganoe, "Awareness and teamwork in computer-supported collaborations," *Interacting with Computers*, vol. 18, no. 1, pp. 21–46, 2006.

[39] A. Alshehri, M. Rutter, and S. Smith, "Assessing the relative importance of an E-learning system's usability design characteristics based on students' preferences," *European Journal of Educational Research*, vol. 8, no. 3, pp. 839–855, 2019.

[40] P. Conn, T. Gotfrid, Q. Zhao et al., "Understanding the motivations of final-year computing undergraduates for considering accessibility," *ACM Transactions on Computing Education*, vol. 20, no. 2, pp. 1–22, 2020.

[41] M. Bond, O. Zawacki-Richter, and M. Nichols, "Revisiting five decades of educational technology research: a content and authorship analysis of the British Journal of Educational Technology," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 12–63, 2019.

[42] A. Woodruff, "COVID-19 follow up testing," *Journal of Infection*, vol. 81, no. 4, pp. 647–679, Oct. 2020.

[43] M. Unsld, *Measuring Learnability in Human-Computer Interaction*, Ulm University, Ulm, Germany, 2018.

[44] A. M. Saleh, R. Ismail, N. Fabil, N. M. Norwawi, and F. A. Wahid, "Measuring usability: importance attributes for mobile applications," *Advanced Science Letters*, vol. 23, no. 5, pp. 4738–4741, 2017.

[45] T. Vanbellingen, S. J. Filius, T. Nyffeler, and E. E. H. van Wegen, "Usability of videogame-based dexterity training in the early rehabilitation phase of stroke patients: a pilot study," *Frontiers in Neurology*, vol. 8, p. 654, 2017.

[46] C. Gonzalez, *Student Usability in Educational Software and Games: Improving Experiences*, IGI Global, Hershey, PA, USA, 2013.

[47] A. Marcus, "Usability and user experience studies," in *Design, User Experience, and Usability: Novel User Experiences*-Springer, Berlin, Germany, 2016.

[48] T. H. Brown and H. J. van der Merwe, *The Mobile Learning Voyage - from Small Ripples to Massive Open Waters*, Springer International Publishing, Berlin, Germany, 2015.

[49] W. Fox and M. S. Bayat, *A Guide to Managing Research*, Juta and Company Ltd., Cape Town, South Africa, 2008.

[50] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[51] N. Nachar, "The mann-whitney U: a test for assessing whether two independent samples come from the same distribution," *Tutorials in Quantitative Methods for Psychology*, vol. 4, no. 1, pp. 13–20, 2008.

[52] S. M. Karadimitriou, E. Marshall, and C. Knox, *Mann-Whitney U Test*, Sheffield Hallam University, Sheffield, UK, 2018.

[53] S. H. Bibi, R. M. Munaf, N. Z. Bawany, A. Shamim, and Z. Saleem, "Usability evaluation of islamic learning mobile applications," *Elkawnie*, vol. 6, no. 1, p. 1, 2020.

[54] A. Hussain, E. O. C. Mkpojiogu, J. a. Musa, and S. Mortada, "A user experience evaluation of Amazon Kindle mobile application," *AIP Conference Proceedings*, vol. 1891, p. 020060, 2017.

[55] B. Saket, A. Endert, and J. Stasko, "Beyond usability and performance," *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization-BELIV '16*, pp. 133–142, 2016.

[56] L. López-Faican and J. Jaen, "EmoFindAR: evaluation of a mobile multiplayer augmented reality game for primary school children," *Computer Education*, vol. 149, p. 103814, 2020.

[57] L. G. Portney and M. P. Watkins, "Statistical measures of reliability," in *Foundations of Clinical Research: Applications to Practice*, Prentice-Hall, Hoboken, NY, USA, 2nd edition, 2000.

[58] D. Barclay, C. Higgins, and R. Thompson: The Partial Least Squares (PLS) Approach to Casual Modeling: Personal Computer Adoption Ans Use as an Illustration. 1995.