Knowledge Analysis Automatic Evaluation in Virtual Reality Immersive Experiences

José Vieira[1], Rui Nóbrega[1,2], Vasco Pereira[1], António Coelho[1,2], Alexandre Jacinto[3] and Carla Morais[4,5]
[1]FEUP, Faculdade de Engenharia da Universidade do Porto
[2]INESC TEC, Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência
[3]ESAD, Escola Superior de Artes e Design, Matosinhos
[4]FCUP, Faculdade de Ciências da Universidade do Porto
[5]CIQUP, Centro de Investigação em Química da Universidade do Porto
Porto, Portugal
Email: evenilink@gmail.com, {ruinobrega, vpm, acoelho}@fe.up.pt, alexandrejacinto@esad.pt, cmorais@fc.up.pt

***Abstract*—Museums and exhibitions usually attempt to evaluate visitors' obtained knowledge through the use of traditional evaluation methods such as questionnaires. These are intrusive and may not provide correct results, especially due to the fact that visitors are usually not interested in being evaluated and may consider such questionnaires as intelligence tests. This paper proposes methods of design and creation of automatic evaluation techniques that make use of Virtual Reality (VR) in order to evaluate users' obtained knowledge after playing through a VR museum game experience. This Analysis System is non-intrusive (its methodology does not impact users' immersion and engagement), valid (can draw conclusions regarding users' obtained knowledge), and replicable (designed techniques can be used in a variety of experiences). Results indicate that the designed assessment techniques can be used to automatically evaluate the knowledge obtained by users throughout the experience, as well as some considerations to keep in mind when designing game experiences with these techniques.**

***Index Terms*—automatic analysis system, interaction interfaces, virtual reality, non-intrusive evaluation**

I. Introduction

Communication between science and scientists with the general population is utterly important. When someone wishes to learn about a specific topic, either because it is part of their education, work, or something they enjoy doing in their free time, one of the possibilities is to visit museums. But in recent years, the number of visitors has been declining [1], mostly because the population prefers to spend their leisure time doing other activities, such as enjoying the latest technological gadget. Museums exhibitions may also not be engaging: going from exhibition to exhibition, reading information regarding the displayed artifacts may not captivate the visitor's attention, and therefore the knowledge acquired by the visitor may be lower than expected. When users are required to memorize knowledge without an actual engaging context, they tend to forget it after a while [4]. In a science museum, a specific process can be taught to their visitors, either by making them read about it or watch a video, but since they are not required to use that knowledge for anything other than memorizing it, that knowledge soon fades away since it is not stimulated.

When the visit ends, museums may attempt to grade and evaluate the museum's experience and the acquired knowledge from a visitor, asking them to fill in questionnaires. These questionnaires are usually inconvenient and intrusive, since visitors do not feel the need to be evaluated, and most of the times they do not want to, especially if their visit was only out of curiosity.

To address this, one can look at the rise of popularity in virtual reality (VR) experiences [5], especially in the entertainment industry, with users being more aware of the technology than ever. Experiences developed for this medium allow the player to be immersed and engaged while they are interacting in the virtual world [6]. Games have also proven to be an efficient way of learning [7] since they provide an interactable environment for users to explore, succeed and fail without real-life consequences, empowering users to try anything they want [4]. Since they provide a more interactable experience, they can better engage and motivate the player than traditional methods, such as reading or watching documentaries [3].

By joining virtual reality with educational games, it is possible to address the lack of engagement that a museum may suffer from, providing users with a rich experience that focuses on the content knowledge that the museum has to share. If users can learn inside a VR experience, they can also be evaluated using that same medium, without the need for outside evaluation systems [8] [3]. A system that evaluates what the user learned and what they missed throughout the game could be used to conclude the amount of information provided to the user. This should be a non-intrusive evaluation system.

A new method to evaluate users when they interact with educational content is necessary, a method that does not remove the engaging aspect of a learning experience, that is non-intrusive and can actually make users learn in a situational context, rather than just through memorization. This work is part of the project iSea [14] for developing non-obtrusive, valid and replicable methods to evaluate audience attitudes about science communication projects.

II. Related Work

Many science museums try to attract their audience with the promise of interactive experiences inside the museum

[1]. There have been many experiences in museums that use VR to support educational experiences. They vary from basic experiences such as viewing artifacts or virtual tours [10] to rich and immersive interactive games, which can target single-player [11] or multiplayer [12]. These experiences are focused on trying to teach users about a specific topic, but not necessarily in evaluating the knowledge users acquired or the exhibit experience itself. The British Museum held a "Virtual Reality Weekend" event [13] in 2015, allowing users to explore a scene during the Bronze Age, using the Samsung Gear VR. In this experience, users could walk around in the landscape using a touchpad on the Head Mounted Display (HDM) and look around by moving their heads. Users could interact with certain objects, and in order to give clues as to which objects were interactable, those objects glow blue to highlight the fact that they were there for the user to know more about them. Users could select them by looking at them and by tapping the touchpad, changing to a closer view in which they were able to rotate the selected object while hearing a description. That description was the message that the museum wanted their visitors to learn, so the audio description had to be engaging and significant in order to avoid boring the user. Since the experience was not linear, users could walk around and interact with objects at their own pace, keeping immersion correlated to its main objective: providing knowledge. The evaluation from that weekend alone was great, with most visitors saying that the experience was good and that it provided a great opportunity to learn more about the Brozen Age.

The French National Museum of Natural History has a permanent exhibition with a catalog of VR experiences that change based on the museum's events. One of those experiences is called *Journey Into the Heart of Evolution*[1], in which participants can interact with a network of hierarchical species, by manipulating that network in 3D space, selecting the species that they wish to learn about, and details of that species are then presented to the user. It also has a mini-game regarding relationships between different species and a model viewer for each one.

Some of these experience also have systems to evaluate the user and their obtained knowledge. Garcia-Cardona et al. [3] developed an application that offered an immersive experience and evaluated users while they were playing the experience itself. The application allowed users to visit a portion of an ancient Aztec city, in which they had to explore the environment while answering questions inside the virtual world. Users wondered around the environment, guided through audio cues and interactive visual feedback (objects being highlighted), encountering several pop-ups referring to specific objects and/or scenes related to the Aztec city, which would present images or text information. Users could also find pop-ups presenting questions about the newly obtained information from the image/text seen before. To increase the user's motivation to complete all the questions available and explore everything the application had to offer, the authors implemented a scoring system as a positive feedback loop, in which each question answered had audios cues to inform the user if they answered correctly or not. Since the evaluation was actually inside the experience, answering questions would still be done in an immersive environment, so users would still be engaged even when under evaluation. Around 88% of users that went through this experience answered that the experience itself was more engaging than being provided the same information on a physical paper.

Allison et al. [8] wanted to teach students about gorilla interactions and the place each one occupied in the dominance hierarchy. They designed an experience in which students take on the role of a juvenile gorilla and must interact with other gorillas. If they approached an older gorilla in a threatening way or just stared continuously at them, the older gorilla would start to intimidate the user by, for example, beating his chest. If users insisted on continuing with the same behavior, they would leave the gorilla's interacting zone and move to a new zone, which is a metaphor for the species removal and reintroduction in a different gorilla group. In the beginning, users would completely ignore the older gorilla's warnings, resulting in the gorilla charging at them, and with users (as the young gorilla) being reintroduced in a new environment, but they quickly understood the warning messages from more dominant gorillas, keeping their distance from the stronger ones, which was ultimately the goal of the experiment. If users finished the experience by being constantly reintroduced into a new zone, then users failed in learning the fundamental information that the experience wanted to portray, but if they completed it by not being reintroduced after a long time, then they understood with success how to interact with the portrayed species, learning concepts of gorilla interaction and dominance hierarchies in an interesting and fun way, as stated by them after the experience ended.

III. Analysis System

In order to evaluate users through gameplay, the Analysis System (AS) must be able to be aware of their actions so that it can associate them with a specific conclusion. The AS must be able to detect what the user sees, what they do and how long they take to do it. The AS uses their overall interactions to evaluate what users learned or paid attention to.

The first metric to take into account is the user's gaze direction, which can give the AS information regarding what the user is looking at. Many applications that require the gaze direction for gameplay reasons usually display it as a white dot/circle in the middle of the screen, but in order to maintain the non-intrusive nature of the system, any information regarding the gaze should not be displayed, since the gaze is only used by the AS in order to detect the object users are currently focused at. By making use of this functionality, the AS can observe where the user is looking, and if they are looking at a specific point of interest (POI) that the system considers to be

[1]Journey Into the Heart of Evolution, 2017 (released year): https://www.mnhn.fr/en/explore/virtual-reality/journey-into-the-heart-of-evolution, Last accessed: 20/12/2019.

important or contains precious information that they can learn from, certain conclusions become possible.

Only using the gaze direction is not enough, as further analysis is necessary when drawing conclusions. If users just look at a POI momentarily, it is wrong to expect they learned the POI's intricacies, as there was not enough time for the user to fully analyze it. In order to improve on this, each POI should be focused by the player for a specific amount of time (acknowledge time), time that should be enough to carefully consider the importance of specific POIs located in the game's environment. Each POI requires a certain level of attention that is dependent on each one, based on their learning complexity. For instance, assuming there is a screen that displays important information that the user can learn from, in order for the system to understand if users learned what the screen portrays is to set its acknowledge time as the screen's reading time. If the user looks for that required time to the screen, then the system can assume that the user gave the screen enough attention as to understand and learn what was written there. This should enable the AS to more carefully conclude about the knowledge obtained by users when using their gaze direction since it requires a certain level of attention that is dependent on each POI. The moment users' gaze direction intersects a POI, the focus time starts counting towards the POI's acknowledge time, stopping counting when they stop looking at the POI.

It is possible to improve on this concept by associating different levels of attention to each object. Instead of specifying only a single acknowledge time in which users that looked for long enough are considered to have browsed the information and users that stayed under that acknowledge time are considered to not have browsed, by specifying, for instance, 2 acknowledge times, this restrictiveness can be mitigated (see Fig. 1). If users looked long enough to go over the first acknowledge time, the AS knows the user is somehow interested in the POI and what that POI has to offer in terms of information, increasing the probability of users actually learning about it, establishing a linear relationship between the amount of time looking after reaching the first acknowledge time and the probability of users learning the portrayed information. By continuing to focus on the object, the probability continues to increase, until it reaches the second acknowledge time, which is assumed that users should definitely have learned what the object has to offer. It is worth pointing out that this extra acknowledge time was not implemented, and thus not tested, but it is, nevertheless, and important suggestion for this metric.

Another important metric is interaction. Certain interactions may have an underlying objective: when the player interacts with an object and depending on the design of the experience, that can be an indication of awareness towards understanding what the experience portrays. The system should be able to detect when certain interactions take place, and if such interactions are important, it can conclude if users understood the knowledge that the experience wanted to present. After being taught how a specific interface works, when users use

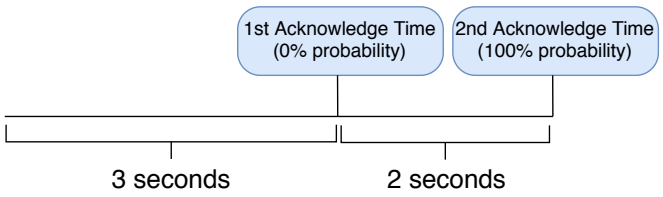

Fig. 1. Acknowledge time with 2 thresholds. The probability of the user learn goes increases between the 3rd and 5th second.

that same interface, they usually do so with a purpose in mind. For instance, the user is taught that by pressing a specific button, that button triggers an action that changes the game state. If, throughout the game experience, users press that button when that change to the game state is positive regarding a specific problem they are trying to solve, the AS can conclude that users understood when to press the button and use such functionality in the right moment.

One more possible metric to evaluate users is by measuring time. This metric can be used to understand how the player performed under certain situations, such as the time they took to complete a specific action. For example, if users take too long to execute a certain action, that can either suggest they did not understood how to use the required interface to perform it or that they lacked the ability to execute the required action. But if they performed the action in a short amount of time, that suggests they knew how to use the interface and had the ability to perform it. This metric can also be used to evaluate users' decision making: if they take an unusually big amount of time to decide, that can suggest users were careful when making their decision, taking the required time to measure all the different possibilities and their impact on the game world, as opposed to when the decision time is very short, suggesting that users' mind was set on a specific choice and they had no doubts about what they decided.

IV. Implementation

The implemented solution is a VR experience with an incorporated AS that is able to attribute meaning to player's actions within the game. This AS operates according to the evaluation metrics described in the previous section. When users are immersed in the experience, the evaluation system works in a non-intrusive way, analyzing player's actions, classifying said actions, and in the end exporting this information to a file that can be read in order to verify what the user learned or not regarding the played experience.

In order to make this experience feel as a complete game experience, a lot of systems that react to one another had to be implemented, including the interfaces users interact with [9]. Some of the interactions with the interface are used to evaluate users' actions, and draw conclusions about their obtained knowledge or information. The main setting for the story takes place at Azores deep sea. In the story the Azores' government has to make a decision about where to invest in order to enrich the local economy. On one hand, there has been some pressure from technological companies to invest

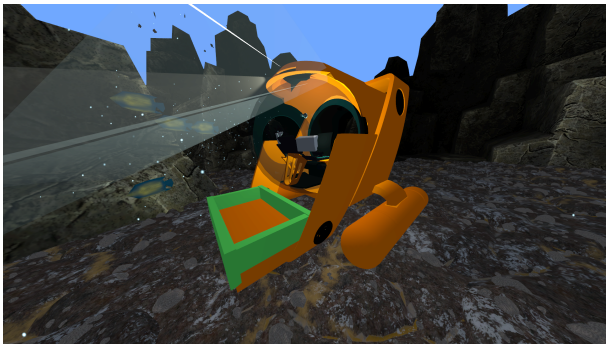

Fig. 2. Submarine in Azorean sea.

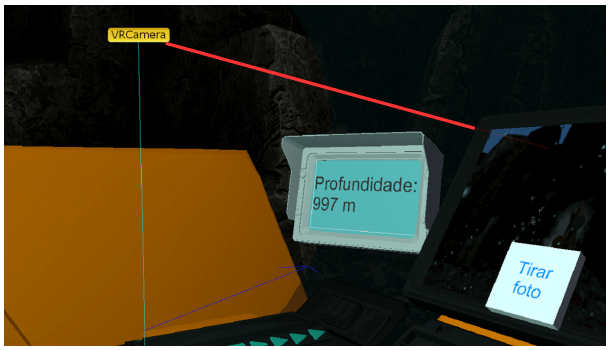

Fig. 3. Using a simple raycast (red line), the POI is not detected, since it is not at the center of users' gaze, even if they are using their peripheral vision.

Fig. 4. Using a sphere cast (yellow line, ending with a white sphere), the POI is detected, since it casts a sphere with a large enough radius to detect it.

in deep-sea mining, in order to use the mined minerals in building computers and smartphones. On the other hand, local communities are worried about the possible consequences of such activities on the Azores deep sea and its ecosystem, which itself contributes to economic growth providing an attractive area for tourism and fishing activities. The user's mission is to give their opinion about whether the government should approve deep-sea mining or stay away from it. They should embark on a submarine mission and evaluate mining possibilities. An outside view of the submarine can be seen in Figure 2. The setting and the issues users have to deal with, the knowledge, the experience and it´s design, are of major importance and have a direct impact on the evaluation methods used.

A. *Look-at Evaluation*

The look-at evaluation consists of determining if the user looked long enough at any specific objects that are valuable to the AS, using the user's gaze direction.

In order for the AS to be able to scan the environment for points of interest (POI), a component was created in order to detect this. This component can be attached to the user's virtual camera, so as to mimic the player's vision. This component can, right from the start of the experience, be constantly shooting a raycast into the scene to look for interesting POI that are important to the AS. But only using a simple raycast proved to be problematic, as this would generate a lot of detection problems.

For example, the player could be using its peripheral vision in order to look at a specific POI. Using a simple raycast, the component would not be able to detect if the player was indeed looking at the POI, since the raycast would not intersect the POI in any way, because it is only shot directly at the center of the screen (based on the users' gaze direction), missing the POI (see Fig. 3). This problem can be minimized using a sphere cast. By using an appropriate sphere radius, it is possible to cast a sphere into the environment that will cover a larger percentage of the user's field of view, allowing the system to more easily detect where the player is looking, even when they are using their peripheral vision (see Fig. 4). The radius does not need to be big, as current generation HMDs only allow field of views with angles up to 110º, and since HMD's lens allow relatively narrow focus points, users would not exactly be using the edge of their peripheral vision since that part of the environment would be too blurred for them to truly notice or understand what was presented there without actually rotating their head towards that point.

The radius value that is used throughout the whole experience is 10 centimeters. This value was selected based on preliminary tests since the results regarding the object that the sphere cast was detecting versus what users were actually looking at were favorable with this value. It is worth pointing out that this value may need tweaking when applying this evaluation technique to other VR experiences. Other games may need to adapt this value based on the scale of their own virtual world, as bigger scale also means that the sphere cast radius could be bigger, since big objects would automatically occupy a bigger portion of the screen, not requiring users to look directly at the center of the object.

In order for an object to be considered for evaluation as a look-at POI in the eyes of the AS, that object should be marked as such, and its acknowledge time specified. This is required because different objects may require different look-at times to be acknowledged by the system. Basically, each POI needs to have its own acknowledge time. This is very important, as this is the required time that a user must look at an object in order for the object to be considered acknowledged by the user, meaning that the user has paid sufficient attention to

it that they will most likely understand what the object is or represents. In other words, users should look at the object for a specific amount of time in order for the object to be considered acknowledged by them, as different objects require different levels of attention, since they have different complexities. If the user is asked to read something in a virtual piece of paper in order to understand a specific problem, the acknowledge time of that piece of paper may be its reading time, allowing the user to read it all. For a computer screen that only indicates, for example, the current depth in meters, a simple glance at the screen of about 1 second may be enough, and when that POI is acknowledged, that information is saved by the AS in the form of a table, with 1 of the columns stating if the screen was acknowledged (see Look-at in Table I). Only when the user has looked at the object for its own specific amount of time is the object acknowledged.

If users start looking at an object but they look away from it without going over the acknowledge time, the time starts from zero the next time they look at it. This prevents non-intentional focusing, as users could just be looking around and checking the environment, with the possibility of the sphere cast intersecting a specific POI, accumulating time that was not actually used to focus the object, but in just generally exploring and looking around.

It is worth noting that a look-at evaluation could possibly benefit from the use of eye tracking technologies such as the Tobii Eye Tracker. Eye tracking technologies direct near-infrared light to the user's eyes, creating reflections that are tracked by an infrared camera. After some calculations, it is possible to detect where the user is looking, even if they are using their own peripheral vision to look at certain objects. Instead of trying to understand and iteratively adjust the best possible value for the sphere cast radius (as to simulate the player's attention zone), eye tracking could greatly improve the efficiency of such methods.

B. Interact Evaluation

The interact evaluation consists of determining if the user interacted with specific objects that are valuable to the AS and learned something from their interaction. If a certain information is only provided through a certain interaction, that is, in order for users to have access to a specific information they are required to interact with a certain interface, then, if they interact with that interface, they learn the information locked under it.

For instance, in this experience, users can ask for advice regarding the decision problem they are facing (explained at the beginning of the section). Such advice is only provided if they press the corresponding virtual button that triggers the dialog audio that provides that advice. At the start of the experience, they were taught how to interact and press virtual buttons, so that they know how to trigger them. If such advice is only available to users if they press the corresponding advice button, and if the AS wants to evaluate if they learned some of the points stated in the advice, they must verify if users actually interacted with the advice button. If the dialog audio is simple and short, then users would have learned that information, and the corresponding table is generated, stating which advice they heard (see Interact in Table I.

TABLE I
DIFFERENT EVALUATION EXAMPLES

Look-at	**Deepness Association**	**Noticed Depth Screen**	
		Yes/no	
Interact	**Mining Exploration**	**Advice Heard**	
		In Favor / Against	
Time	**Mining Exploration**	**Decision**	**Decision Time**
		Favor / Against	seconds

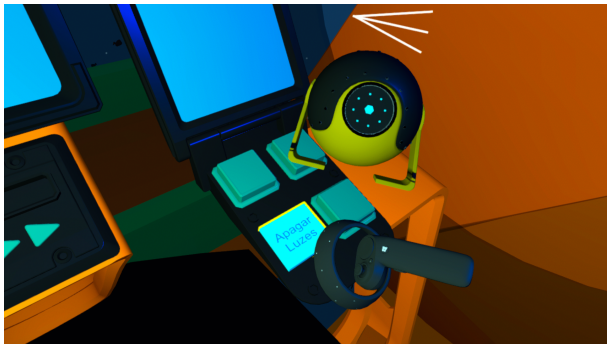

Fig. 5. Analysis System evaluates if users pressed the light switch button.

Another example is when the submarine starts descending and the sunlight starts to fade away, eventually getting completely dark, so the AS wants to analyze if users understand if they know they need to press the light switch virtual button in order to light up the scene (see Fig. 5). This moment starts when it gets dark enough to justify turning on the lights. If the user takes more than 3 seconds to turn it on, a dialog hint plays informing the player that it is getting too dark, and if 7 more seconds pass without the user pressing the light switch, the lights get turned on automatically. Since this moment's particular objective is to understand if users know that they need to turn on the lights in dark places, the AS only considers the interaction with the button to turn on the lights important after it gets dark, since previous or future interactions with the button do not count towards verifying if the player actually understands they have to press the button. This introduces situational context to the evaluation, in which a certain evaluation may only happen in a given context and only at a specific moment during the experience in order to make sense. This way the system needs to have the ability to evaluate at specific times only, depending on the experience's design. If the user interacts with the light switch within the time frame of this moment, the AS will create the corresponding evaluation stating that they indeed turned. If they did that before or after the hint was provided, stating they understood that light changes based on depth. If not, the moment's evaluation reflects their failure.

Fig. 6. Users must press one of the decision button to make their decision. The Analysis System records the time they take.

C. Time Evaluation

The time evaluation consists of counting the amount of time a user took to accomplish a specific task.

It is possible to easily record the starting time of a specific task, and when the user finishes that task by means of gameplay, the starting time is subtracted from the finish time, offering an accurate time duration of the embraced task.

Every moment or event from the game can be timed for evaluation. This value can be used to understand if users had doubts regarding what to do or what to decide. If users are faced with a specific decision and are struggling to effectively decide, this might give insights that the user was carefully considering all the nuances of the task at hand, with the objective to make the best decision possible. If users took a surprisingly short amount of time to decide, it might be possible to conclude the users did not consider everything there was to measure or that they simply were certain regarding a particular choice. This evaluation gives some insights into their decision making.

One particular example is when users need to actually state their opinion regarding whether the government should approve deep-sea mining or stay away from it, by pressing the corresponding decision button (see Fig. 6). From the moment the decision is possible, time starts counting, and based on that time, it is possible to see if users struggled to decide or were very firm in their decision. Table I references this moment in the Time cell, where ”seconds” represents a possible time value.

V. Evaluation

35 participants attended the user studies. 71% of users were male and 26% were female, while 1 user preferred to not state their gender. Participants' age has an average of 26.83, with a median of 23, ranging from 16 to 53. In terms of virtual reality experience, 91% of users had at least one previous opportunity to try VR, 57% were intermediate or above users, and 9% never tried VR before.

A. Protocol

Users play through the developed VR experience from start to finish. They face all the challenges and decisions while the AS is constantly tracking users' behaviour. There are a total of 5 evaluation moments. The first one evaluates if users know the temperature the submarine should not exceed, information that is displayed on a post-it, above the temperature screen as seen in Fig. 6. The second moment evaluates if players can visually describe 2 ecosystems that are displayed as images on a computer screen. The third moment evaluates if players acknowledge some of the information regarding two distinct ecosystems, displayed as bullet points on brochures available inside the virtual submersible. The fourth moment evaluates if players know how deep one of the ecosystems is located, by looking at the current depth screen when accomplishing the proposed tasks at that same ecosystem. The fifth and final moment evaluates if players know some of the reasons to be in favor or against deep-sea mining, by detecting if certain virtual buttons are pressed, playing an audio track with the corresponding advice dialog.

When users finish the experience, a set of questions are orally asked to the participants by an assistant, and depending on their answer, those questions are marked as correct, incorrect, or left blank (indicating that users did not know or did not notice the aspect of the question in the VR experience). The AS evaluated users regarding their obtained knowledge, providing conclusions as to what users learned and did not learn. The questions asked to participants were regarding the same evaluated knowledge, so as to match the answers provided by them with the evaluation from the AS. For instance, the AS evaluated the user with regards to the depth they were in (explained in section IV-A), generating Table I first section, which stated if participants looked at the screen that displayed their depth. One of the questions asked to participants was how deep they were, and if they answered correctly, in order to get a positive evaluation out of the AS, the AS also had to conclude that the user did indeed look at the screen.

B. Results

Table II displays the respective moment, aligned with the AS and user accordance and discordance percentage, so, for instance, when looking at ”Maximum Temperature Detection”, 85% of times the AS correctly concluded that the user knows or did not know the maximum temperature value (AS stated they know and they answered correctly, or the AS stated they did not know and the user answered incorrectly or did not answer at all), failing 15% of times (AS stating they know and they answered incorrectly, or the AS stating they did not know and they answered correctly). Each of the 5 moments have their own accordance and discordance percentage, displaying the success rate of the AS in each one.

Taking a look at Table III, it is possible to see the total amount of times the AS and the user were in accordance or discordance, pertaining the results of all the moments merged together. Table IV displays the results from a precision and recall test, including its accuracy, followed by the calculation of the F1 score.

C. Analysis

The AS had very high accuracy on some of the moments, and the majority of them were correctly evaluated, except the "Deepness Association" moment, which had an accordance rate of 40%.

Starting with the "Maximum Temperature Detection" moment, which had a accordance rate of 85% and discordance of 15% according to Table II, it is possible to see that this moment was the second highest rated. This high success rate in this specific moment may suggest that the moment itself was overall well designed, and that no major interference from other interactions took place. This also supports, at the first sight, the usage of look-at evaluations.

Moving to the "Image Recognition" moment, Table II shows that this moment had a 60% accordance rate. Although this moment still has a positive accordance rate, the drop in accordance is probably based on the less than ideal acknowledge time. The images on the computer screen had an acknowledge time of 3.2 seconds, which was probably not the best value, since this made the AS predict more times than not that the player should indeed know how to visually describe those images, since they acknowledged the computer screen. For some users, 3.2 seconds were enough, but some required more time in order to absorb the information displayed on the images in order be able to visually describe them. This suggests that the acknowledge time is a very important factor in the look-at evaluation, and a lot of consideration should be put into figuring out the best time for a specific moment.

The next moment, "Brochure Recognition", had an accordance rate of 73%. Since users had to verbally answer with at least one bullet point on the brochure that they could remember, and since those bullet points were brief and short, the acknowledge time for each brochure was 3 seconds. There are 2 factors that may contribute to the fact that this moment does not have an accordance rate similar to the last moment (98%): acknowledge time and object positioning. The acknowledge time design problem was addressed in the above moment analysis, but this moment may also suffer a new design problem, which is its own positioning. Both brochures are placed near the center of the submarine's frontal view, one slightly to the left and the other to the right, leaving some space between them so that players can see where they are going when moving the submarine (see Fig. 7 for a visually description of the problem). When this happens, the sphere cast that exists in order to detect POIs may detect one of the brochures, since users can have their head slightly rotated to one side as they move the submarine with the brochures still in display, starting counting towards the acknowledge time. When that time is surpassed, the AS receives the evaluation that the user did indeed look at the brochure for at least the intended amount of time, even if that was not their intention.

"Deepness Association" moment is the moment with the lowest accordance rate, and by watching the participants' recorded footage, the the strongest hypothesis is that this moment may suffer from asset design issue, also described above. When players wanted to drive the submarine using the thrusters control, the depth screen was too close to the movement thruster (interface that allows the player to move the submarine), so when users looked to the thruster in order to grab it, the sphere cast radius attached to the player's camera was big enough to detect the depth screen, sending an evaluation to the AS stating that players looked at it for the required amount of time. Memory problems may have also occurred. Some users, were visibly frustrated when trying to answer the question, since they stated they saw the value on the depth screen, but could not remember the answer, contributing to the decay of accordance.

The last evaluation moment, "Advice Rephrasing", is the one with the highest accordance rate. The advice users heard was brief and very concise, in order to make it easy to

TABLE II
ACCORDANCE AND DISCORDANCE IN EACH MOMENT.

	Accordance	**Discordance**
Maximum Temperature Detection	85%	15%
Image Recognition	60%	40%
Brochure Recognition	73%	27%
Deepness Association	40%	60%
Advice Rephrasing	98%	2%

TABLE III
ACCORDANCE AND DISCORDANCE BETWEEN THE AS AND USER.

	User knows	**User does not know**
AS says user knows	**121** (True Positive)	**48** (False Positive)
AS says user does not know	**11** (False Negative)	**41** (True Negative)

TABLE IV
PRECISION, RECALL, ACCURACY AND F1 SCORE CALCULATION.

Metric	**Score**
Precision	0.72
Recall	0.92
Accuracy	0.73
F1	0.80

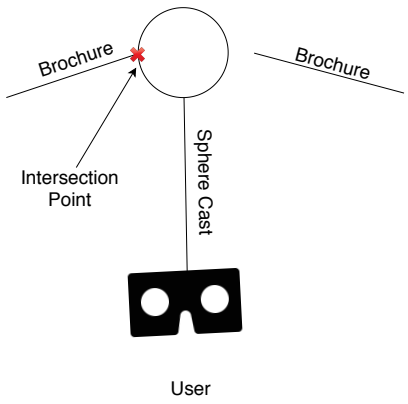

Fig. 7. Brochure being detected by the sphere cast while users are slightly rotated when moving the submarine.

understand. Since this moment used an interact evaluation, as opposed to a look-at one, it was easier for the AS to detect when an advice button was pressed, which would result in the AS concluding that the user should know the reasons explained when the button was pressed, and since this moment did not suffer from any apparent design problem, its success rate is very high.

In order to better understand the efficiency of the AS as a whole, its precision and recall were calculated, based on the information in Table III, with results displayed in Table IV. The calculated value for the recall was 0.92 (between 0 and 1), meaning that the system can find and correctly classify almost all the relevant information the user knows at the end of the experience, providing a good recall result. For precision, 0.72 was the calculated value, meaning that 72% of the results that the system states as relevant are actually relevant which, in this case, is when the AS states that the user learned and they actually did, and although less than the recall value, this value still offers a satisfying result. The system's accuracy was also calculated (0.73), although not as relevant, since this metric can sometimes be deceiving. F1 score was also calculated, giving a value of 0.80, suggesting that the AS is precise and robust, offering a good balance between precision and recall.

These results suggest that this AS and the implemented evaluation techniques, as part of the many components that constitute this VR experience as a museum application, can be used with relative success to determine if users effectively learned what the application is teaching to the user.

VI. Conclusions

When creating experiences with emphasis on learning, intrusive methods are still used in order to evaluate if users actually learned. When these experiences are used in the context of a museum, validating the user's knowledge acquired during the experience, has proven troublesome, since most users are not interested in answering further questions This work offers solutions to solve this issue, by proposing an Analysis System with evaluation techniques that is able to provide valuable insights about users obtained knowledge. User studies were performed with the objective to evaluate the quality and efficiency of the AS. These tests, which were performed by 35 users, validated the feasibility of the AS, which was able to accurately evaluate what was the user's knowledge when the experience ended most of the time.

Some parts of the systems were found that can be improved upon. Since one of the moments only had 40% of success due to the depth screen positioning, that screen should be placed in a different area or at least moved away from the movement thruster. As previously mentioned, this reinforces the need to design the experience in tandem with the Analysis System. This system may also be improved by using 2 acknowledge times as opposed to 1, as it was first planned. This is, of course, subject to further study.

With the implementation of this AS we hope that the base for user evaluation methods regarding scientific communication was established and can be further improved. These evaluation methods deserve continuous study and development due to their high importance. By designing and implementing new methods, the AS would improve on its robustness and flexibility, offering more evaluation variety. Applying eye tracking technologies to this system would be interesting, since it provides a more accurate way to process the user's gaze direction, hopefully removing some of the design restrictions mentioned during the analysis.

VII. Acknowledgments

This work is financed by FCT (Fundação para a Ciência e a Tecnologia), by national funds through the UT AUSTIN Portugal Program within project UTAP-EXPL/CD/0106/2017.

References

[1] Department for Culture Media and Sport. Sponsored Museums Performance Indicators 2015/16 - Statistical Release. (January), 2017.

[2] Sebastian Garcia-Cardona, Feng Tian, and Simant Prakoonwit. Tenochtitlan - An Interactive Virtual Reality Environment That Encourages Museum Exhibit Engagement. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10345 LNCS, pages 20–28, 2017.

[3] Sebastian Garcia-Cardona, Feng Tian, and Simant Prakoonwit. Tenochtitlan - An Interactive Virtual Reality Environment That Encourages Museum Exhibit Engagement. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10345 LNCS, pages 20–28, 2017.

[4] S. A. Barab, M. Gresalfi, and A. Ingram-Goble. Transformational Play: Using Games to Position Person, Content, and Context. Educational Researcher, 39(7):525–536, 2010.

[5] Cristiano Carvalheiro, Rui Nóbrega, Hugo da Silva, Rui Rodrigues. User redirection and direct haptics in virtual environments, Proceedings of the 24th ACM international conference on Multimedia, ACM, pages 1146-1155, 2016.

[6] Jason Jerald. The VR Book: Human-Centered Design for Virtual Reality. ACM; Claypool, New York, NY, USA, 2016.

[7] Thomas M. Connolly, Elizabeth A. Boyle, Ewan MacArthur, Thomas Hainey, and James M. Boyle. A systematic literature review of empirical evidence on computer games and serious games, 2012.

[8] Don Allison, Brian Wills, Doug Bowman, Jean Wineman, and Larry F. Hodges. The Virtual Reality Gorilla Exhibit. IEEE Computer Graphics and Applications, 17(6):30–38, 1997.

[9] Rui Nóbrega, Diogo Cabral, Giulio Jacucci, António Coelho. NARI: Natural Augmented Reality Interface, GRAPP 2015 - Proceedings of the 10th International Conference on Computer Graphics Theory and Applications, SciTePress, pages 1–9, 2015.

[10] B. Bonis, J. Stamos, S. Vosinakis, I. Andreou, and T. Panayiotopoulos. A platform for virtual museums with personalized content. Multimedia Tools and Applications, 42(2):139–159, 2009.

[11] Fabio Bruno, Loris Barbieri, Antonio Lagudi, Marco Cozza, Alessandro Cozza, Raffaele Peluso, and Maurizio Muzzupappa. Virtual dives into the underwater archaeological treasures of South Italy. Virtual Reality, 22(2):91–102, 2018.

[12] L. Li and J. Zhou. Virtual Reality Technology Based Developmental Designs of Multiplayer-interaction-supporting Exhibits of Science Museums: Taking the Exhibit of Virtual Experience on an Aircraft Carrier in China Science and Technology Museum as an Example. In Proceedings - VRCAI 2016: 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry, volume 1, pages 409–412, 2016.

[13] Juno Rae and Lizzie Edwards. Virtual reality at the British Museum: What is the value of virtual reality environments for learning by children and young people, schools, and families? MW2016: Museums and the Web 2016, pages 1–9, 2016.

[14] António Coelho, Carla Morais, Lucy Atkinson, Alexandre Jacinto, Rui Nóbrega, M. Varzim, , João Paiva, Luciano Moreira, Teresa Aguiar, Ana Teixeira, José Vieira and Diogo Coelho. I SEA – Virtual reality to evaluate audience attitudes about science communication. UT Austin Portugal Program' 2019 Annual Conference, Braga, Portugal, 2019.