Journal on Interactive Systems, 2021, 12:1, doi: 10.5753/jis.2021.2063

Virtual look around: comparing presence, cybersickness and usability for virtual tours across different devices

Jean Felipe Patikowski Cheiran [Federal University of Pampa and Institute of Informatics, UFRGS | jeancheiran@unipampa.edu.br]
Adriel Rodrigues [Federal University of Pampa | adriel.rodrigues07@hotmail.com]
Marcelo Soares Pimenta [Institute of Informatics, UFRGS | mpimenta@inf.ufrgs.br]

Abstract

Virtual Reality has become readily available in the last few years through different devices, from desktop computers to head-mounted displays (HMD). Also, virtual tours became popular with 360º panoramic photographs and video clips on online social media, so people could visit remote locations without being exposed to crowded transportation or long travels. Also, virtual tours demonstrate considerable potential as a form of escapism and even for remote teaching. Since we lack studies that evaluate the User Experience (UX) in virtual tours on different devices, this article aims to compare aspects of the User Experience (regarding sense of presence, cybersickness, and usability) in a virtual tour website developed in WebXR across different devices. To achieve our objective, we developed a virtual tour based on 360º pictures using WebXR API and React 360 framework and conducted an experiment with 41 undergraduate students using four different devices: a laptop computer, a smartphone, a Google Cardboard headset, and a Samsung Gear VR HMD. We evaluated users' perceptions by adapting and translating the Suitability Evaluation Questionnaire (SEQ) and users' performance by measuring the time to fulfill a set of tasks. The main findings from this study include that (i) the overall self-reported experience using Google Cardboard is worse than using other devices, (ii) the users' performance is quite similar between the platforms, (iii) there is evidence of unexpected cybersickness symptoms in tests with the smartphone, and (iv) the development of a plausible hypothesis concerning low usability having an effect upon the sense of presence. Additional contributions of our research are the adaptation, translation into Portuguese, psychometric analysis, and revised scoring procedures of the SEQ.

Keywords: *UX, User Experience, Suitability Evaluation Questionnaire, Virtual Tour, WebXR*

1 Introduction

Even though Virtual Reality (VR) had been created more than 40 years ago[1] according to Costa and Ribeiro (2009), it has evolved and it has become more accessible. Parisi (2015) and Jerald (2016) state that VR aims to convince the users that they are somewhere else using the illusion of presence and immersion to change their physiological and psychological condition. For Jerald (2016), the immersion is a flexible aspect of VR because it is more linked to the technology that leads people to realize and to interpret sensory stimuli in a wide, coherent, vivid, and interactive way. Therefore, one VR environment in combination with different devices could create different levels of the sense of "being there".

One example of VR application is the "virtual tour" that, according to Lee et al. (2013) and Osman et al. (2009), allows users to navigate within a simulated environment that contains virtual reality elements, so offering the opportunity of looking around places far away in space and time. One of the first virtual tours was an installation in a British Museum in 1994 as presented by Boland and Johnson (1996) and Pujol (2004): the representation of Dudley Castle (England) as it had been in the year 1550. When this kind of application is available on the web, it usually simulates places through 360º pictures or videos as it is done by many universities worldwide as described by Osman et al. (2009).

Nowadays, the Covid-19 outbreak and the physical distancing policies caused an increase of seven times on searches for "virtual tour" terms on search engines. Many websites have become a hub for these tour experiences as stated by Bloom (2020) and CatracaLivre (2020). This scenario just makes clear, according to Tarcia (2020), the escapism needs of isolated people and the search for didactic resources for remote teaching since Klippel et al. (2019) say that virtual tours are a good alternative (in many aspects) to replace the experience of real trips.

Considering the increasing search for virtual tours, the compatibility between VR API (Application Programming Interfaces) and web browsers, and the number of available visualization devices, it is necessary to assess aspects of User Experience (UX) of virtual tours in different platforms to assure that users are able to make a good decision based on cost-benefit relations and usage context.

Most scientific papers on virtual tours do not directly address the comparison of the platforms as they focus either on isolated usability evaluation, like Osman et al. (2009) and Oprean et al. (2018), or application development, like Sathe et al. (2017), Butcher and Ritsos (2017), and Ye et al. (2017). Moreover, the most significant studies on this subject performed by Lee et al. (2013) and Klippel et al. (2019) address, respectively, the comparison of different visualization modes using tablets in a tour and the comparison of real field trips and virtual ones, but neither contrasts different devices. Finally, none of the related studies include a discussion of differences in the evaluation results if the researchers choose a holistic approach (evaluating the experience as a whole) or a multidimensional analysis (evaluating separately the distinct experience dimensions, e.g., presence, usability, flow, etc.).

[1] 30 years ago in the original reference, but we updated this number since the source was published in 2009.

Considering the gaps in the mentioned papers, we put the following research questions regarding the tested devices (two non-immersive devices and two VR immersive devices): (1) Is there any significant difference in the experience reported by users of virtual tours among different devices? (2) Is there any difference in performing a holistic or a multidimensional UX analysis? (3) Is there any significant difference in the users' performance among different devices?

So, this research aims to compare the User Experience (regarding sense of presence, cybersickness, and usability) in a virtual tour website developed in WebXR and used on four different devices: a laptop computer, a smartphone, a Google Cardboard platform, and a Samsung Gear VR HMD.

To perform this comparison: (1) we developed a virtual tour website for Federal University of Pampa (UNIPAMPA) using 360º pictures; (2) we performed user tests with 41 participants using the four devices; (3) we collected users' opinion regarding the virtual tour experience by a standardized questionnaire that had been translated and adapted; (4) we recorded performance data on participants; and (5) we analyzed the results through inferential statistics. The Suitability Evaluation Questionnaire (SEQ) of Gil-Gomez et al. (2013) was chosen for the translation and adaptation process because of its multidimensional perspective, the small amount of items, and other features presented in Subsection 3.1.

Finally, this paper is organized as follows: the Section 2 describes the main studies related to our scope that were found on scientific databases; the Section 3 includes the apparatus and the methods applied in this research, the experimental procedures, and the description of the developed virtual tour; the Section 4 presents the detailed results of our study, and the answers to the research questions; and the Section 5 shows the key contributions of this research, the faced limitations, and the suggestions to further work.

2 Related Work

Few works have been developed focused on virtual tours in the last years. We highlight the studies found in the scientific databases in this field that address different perspectives.

Cho et al. (2002) performed an academic essay on the effects and implications of a virtual web tour in tourism marketing. Based on theory and evidence from the scientific literature about tourists' experience, the authors claim that it is necessary to keep the good experience, the interaction and the sharpness at high levels to (1) lead the users to a player status instead of a spectator, (2) allow more effective information search, (3) create proper and evaluative envisioning of a destination, (4) allow the users to evaluate their expectations regarding a real destination, and (5) cause satisfaction.

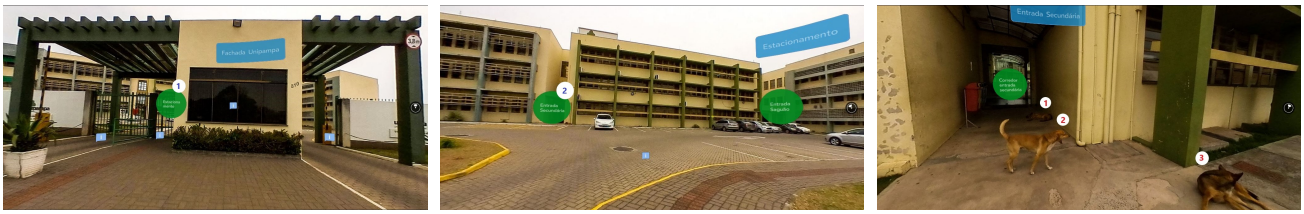
Osman et al. (2009) developed and evaluated the usability of a virtual tour of four Malaysian places. Two experiments were performed: the first aimed to find usability issues and to receive feedback of 10 participants through an interview carried out after some task had been performed, and the second aimed to measure the 5 participants' satisfaction with regard to movement speed, image quality, sounds, trip attractiveness, terminology, text descriptions, and navigability. The ad hoc questionnaire developed by the authors for the second study is not available in the paper.

Osman and Wahab (2011) also evaluated the feasibility of using virtual tours with children in kindergarten. 12 children experienced two different panoramas (a playground and a zoo) using a Flash program. This application was projected on the wall and the children used either mouse or keyboard to interact. The environments had animations and sounds. The authors observed kids' reactions during the interaction, and they identified that children with previous experience with computers preferred to use the mouse while inexperienced kids chose to use the keyboard more often. Furthermore, the general experience with the panoramas was well-accepted and positive concerning the children, even though cognitive benefits had not been assessed.

Lee et al. (2013) reported the development and the evaluation with users of an Augmented Reality (AR) virtual tour in the Antarctic. The miniaturized versions of some Antarctic regions were mapped into a 90.000 squared meters space in a park, and users could visualize these environments in AR through tablets. The tour was composed of 3D virtual models to represent Antarctic elements in AR, a map that allowed top view interaction, and some pictures, videos and panoramas that could be seen without AR. The authors collected data from 50 participants through an ad hoc questionnaire on usability and an adapted version of the GEQ (Game Experience Questionnaire). The analysis of the subjective measures considered the recommended scoring procedures for the GEQ in a subscale approach: competence, immersion, flow, tension annoyance, challenge, negative affect, and positive affect. Differences between the use of the application in an open environment (the park) and a close environment (a booth) were not noticed, but the authors observed that younger users reported a more negative experience even so they felt more confidant in fulfilling tasks. Finally, most users reported that panoramic pictures were the favorite feature, and the open environment users spent more time touring in AR.

Sundar et al. (2017) carried out a detailed analysis on immersive journalism and how it affects our perceptions and mental processes. The authors exposed 129 participants to two stories from the New York Times (they differ on emotional intensity) in three mediums: text read on desktop computers, 360º video also watched on desktop computers, and VR accessed by using a Cardboard VR headset and a smartphone. They measured dispositions and outcomes through a broad range of questionnaires (some of them shortened or adapted by the authors), including the Interpersonal Reactivity Index, Arrival and Departure telepresence questionnaires, and the Reality Judgement and Presence Questionnaire. The story recall was also assessed. The sense of presence was approached in a multidimensional perspective (sense of being there, interaction, and realism). While the 360º video experience and the VR experience presented higher scores than the text experience for the sense of presence scales, the recalling was slightly better for participants that had read the stories. Also, the source credibility, the empathetic link, and the sharing intention were all significantly higher among participants that experienced 360º video and VR.

Oprean et al. (2018) evaluated the differences in a virtual field trip in a settlement using three device configurations:



(a) UNIPAMPA's facade **(b)** UNIPAMPA's main parking lot **(c)** UNIPAMPA's secondary entrance

Figure 1. Sequence of scenes in the easy task of the virtual tour.

an HTC Vive HMD with joysticks on a Unity application, a Google Cardboard platform also on a Unity application, and a website developed using WebVR. The number of participants (Architecture students) was not reported, and the data collection approach was an informal interview. The main problems that the authors identified were related to image quality, lack of georeferencing, lack of camera settings in the VR environment, lack of sounds, user controls on the Cardboard version of the tour, and lack of user control in the videos. No comparison between the devices was conducted.

Klippel et al. (2019) carried out a study that contrasted conventional field trips with immersive virtual field trips. Moreover, the authors proposed a taxonomy for the area. 37 students took part in the experiment in a between-subject perspective. They visited an outcrop either physically or virtually by using an HTC Vive device. The users of the virtual trip were guided through 14 scenes with 360º images, and a series of ad hoc questionnaires was applied to assess technological satisfaction, learning experience, orientation abilities, and sense of presence. The subjective measures were analyzed in a multidimensional perspective with scores for each factor. The results pointed out that immersive virtual trips present advantages regarding satisfaction, learning, and grades when compared to conventional field trips.

Sathe et al. (2017), Butcher and Ritsos (2017), and Ye et al. (2017) developed different web VR applications: a shopping website, a prototype of an app for data visualization, and a system for visiting control system facilities that also allows observing experiments. The three systems were developed using WebVR and other technologies, and all of them run on web browsers. Even though the overall software architecture is detailed by Sathe et al. (2017) and by Ye et al. (2017), user testing is not presented by any author[2].

3 Methodology and Case Study

3.1 Virtual Tour Case Study

We adopted the WebXR API to develop the virtual tour because it assures that the same virtual environment can be viewed on different devices through a web browser. The React 360 framework, created by the Facebook team, was also used as a productivity tool since it is a popular framework for developing immersive and semi-immersive web scenes.

We also created a scene mesh with 90 scenes to represent the UNIPAMPA in the virtual tour. The navigation through scenes was done by teleportation when the user selected a navigational element. Each scene is composed of a 360º picture taken with a Samsung Gear 360 camera, a scene title to aid users orientation, one or more navigation elements to reach adjacent scenes, and (occasionally) one or more information texts to better explain the details about the place that the user is seeing. The pictures had originally been captured in a spherical shape and so they were converted to a panoramic shape (the only one supported in React 360 framework) with 5472 x 2736 pixels size and 96 dpi resolution. We compressed the scene images in JPEG files with 90% or higher quality, resulting in files with about 1MB storage size.

Four tasks regarding the virtual tour were performed by users in experiment sessions. Each task was performed in 10 minutes or less. The four tasks had different complexity levels: easy, medium, hard, and long trip. The complexity of each task was directly related to the course size and to the difficulty in locating the right elements in each scene.

The easy task demanded to navigate through two scenes from the first scene (in the shortest path) and to count the number of dogs in the scene. The medium task required the user to navigate through at least six scenes and to count the number of students in the UNIPAMPA's library. The hard task demanded to navigate through at least 17 scenes and to identify four distinct geometric shapes located in the university's main staircase. And finally, the long trip required the participant to travel through more than 18 scenes at UNIPAMPA and to read aloud the information text of an item located in the last scene. For each performed task, the participant wore a different device. The full sequence for the **Easy Task: how many dogs can be found at the Secondary Entrance** is presented in Figure 1 and described in details next:

1. At the beginning of the virtual tour, the participant sees the first scene (Figure 1a).
2. Next, the participant selects the navigation element (1) in Figure 1a and they are teleported to the second scene (Figure 1b).
3. Next, the participant selects the navigation element (2) in Figure 1b and they are teleported to the third scene (Figure 1c).
4. In Figure 1c, the participant is able to count the number of dogs in the scene (three) and the task is finished.

To avoid creating a bias, we did a rotation of devices and tasks for participants by using two configuration models based on Latin Squares design as recommended by Zaiontz (2018) (Tables 1 and 2). We point out that after the 4th participant in Table 1 and after the 16th participant in Table 2, the pattern is restarted.

[2]The paper of Ye et al. (2017) includes a section titled "a case study" but it just contains the planning for a future experiment.

Table 1. Latin Square Model for participants and devices.

Part.	Device Sequence			
1	Gear VR	Computer	Cardboard	Smartphone
2	Computer	Smartphone	Gear VR	Cardboard
3	Cardboard	Gear VR	Smartphone	Computer
4	Smartphone	Cardboard	Computer	Gear VR
the sequence is restarted for the next participants				

Table 2. Latin Square Model for participants and tasks.

Participants	Task Sequence			
1 to 4	Easy	Medium	Hard	Long trip
5 to 8	Medium	Long trip	Easy	Hard
9 to 12	Hard	Easy	Long trip	Medium
13 to 16	Long trip	Hard	Medium	Easy
the sequence is restarted for the next participants				

After the participant had accomplished a task, the experimenter handed the adapted SEQ questionnaire to them and wrote down the time to conclude that trip. If the user did not finish a task within the time limit (10 minutes), the experimenter would interrupt them to deliver the adapted SEQ questionnaire and to note the limit time down.

The complete flow of activities during a test session is summarized next:

1. **Greetings**: the experimenters present themselves and thank the availability of the guest.
2. **Experiment presentation**: the experimenters describe and present the experiment and the devices.
3. **Screening**: the guest fills out the Screening Form (SF) and they might be prevented from carrying on with the experiment if there is any risk.
4. **Terms and Profile**: the guest screened that accepts to take part in the experiment fills out the Informed Consent Form (ICF) and the Demographic Survey (DS).
5. **1st task**: the participant receives instructions about the first device and the first task, and they try to accomplish the task. Next, the participant fills out the adapted SEQ while the experimenter write down the elapsed time.
6. **Other tasks**: the previous step is repeated to the remaining three devices and three tasks.
7. **Acknowledgments**: the experimenter thanks the participant and the test session is finished. The participant might also experience other VR apps on the Samsung Gear VR device.

We also performed a search in scientific databases to identify standardized questionnaires related to virtual experience measurement since these tools would support the gather of participants' opinions on the experience and the interaction with the virtual tour. The Suitability Evaluation Questionnaire (SEQ) of Gil-Gomez et al. (2013) is noteworthy for its multidimensional perspective (user satisfaction, sense of presence, perceived success, perceived control, realism, comprehensibility of instructions, cybersickness symptoms, and general discomfort), for the small number of items, for being based on another standardized instrument, and for having a prior psychometric assessment. The questionnaire was based on the Short Feedback Questionnaire (SFQ) of Kizony et al. (2005), which in turn was based on the Witmer and Singer's Presence Questionnaire. Although both SEQ and SFQ have been created to evaluate the VR experience in the context of rehabilitation systems, the latter is much simpler than the former and it does not contain specific items to assess the user's perception of progress in rehabilitation. Also, we understand that suitability represents the degree of appropriateness of a system designed for a particular domain, and it covers a subset of the UX construct according to the original SEQ study. Thus, we use in this paper "suitability" as a synonym for "User Experience" regarding the UX factors assessed through the questionnaire. The items of the original SEQ are presented in Table 3. The details on the translation, adaptation and assessment process can be found in Subsection 4.3.

We applied two additional questionnaires in our experiment: a Demographic Survey (DS), and a Screening Form (SF) to assure the participants safety while and after using VR devices. The latter questionnaire was crucial for identifying risks on using the Google Cardboard and the Samsung Gear VR, and it was based on instruction manuals of the main VR hardware that point out the user profiles more sensitive to intense side-effects from immersive VR experiences. The screening procedure aimed to forbid the following profiles from taking part in the experiment: pregnant people, people under the effects of psychoactive medication or other substances, people with psychiatric or neurological issues, people that feel ill, people with vision impairment, among others. We point out that the guests never informed the exact condition that could keep them from taking part in the experiment; they only informed that one or more items in the SF block their access to the experiment and thus they were immediately dismissed from the testing session after the acknowledgments.

3.2 Evaluated Devices

The chosen immersive devices are the Google Cardboard and the Samsung Gear VR. On the other hand, the non-immersive devices are a laptop computer and a smartphone.

In Figure 2a, we present the use of the Google Cardboard viewer. It was operated with an Asus Zenfone 3 Zoom smartphone (5.5" screen with 1080 x 1920 pixels and $\sim$401 ppi resolution) and a Dell WM126 wireless mouse that worked as a clicker.

In Figure 2b, we show the use of the Samsung Gear VR HMD. It was operated with a Samsung Galaxy S6 smartphone (5.1" screen with 1440 x 2560 pixels and $\sim$577 ppi resolution). The side touchpad was used for navigation.

In Figure 2c, we represent the use of the laptop computer Asus model K46CA with Windows 8.1 and 14" screen (1366 x 768 resolution). The same Dell wireless mouse was used to control navigation.

Finally, in Figure 2d, we show the use of the mentioned Asus Zenfone 3 Zoom smartphone. The navigation was made by touching the screen.

3.3 Ethical Considerations

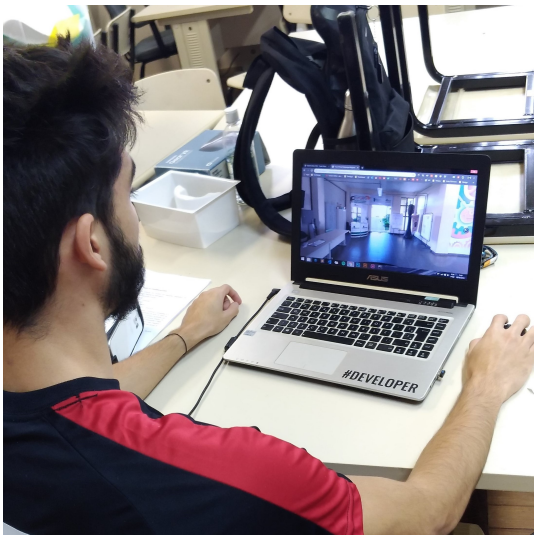
All participants were informed about the details of the experiment and about the risks regarding the side effects of using VR devices. The aforementioned SF was applied to every guest to reduce the hazard to participants.

Table 3. Original Suitability Evaluation Questionnaire from Gil-Gomez et al. (2013).

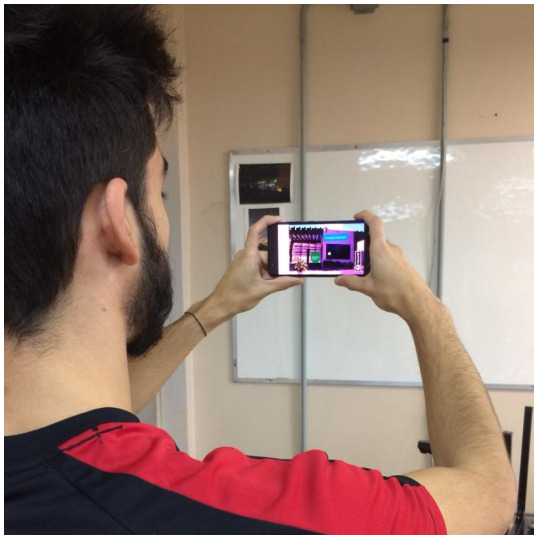
Question	**Answer**
Q1. How much did you enjoy your experience with the system? Q2. How much did you sense to be in the environment of the system? Q3. How successful were you in the system? Q4. To what extent were you able to control the system? Q5. How real is the virtual environment of the system? Q6. Is the information provided by the system clear? Q7. Did you feel discomfort during your experience with the system? Q8. Did you experience dizziness or nausea during your practice with the system? Q9. Did you experience eye discomfort during your practice with the system? Q10. Did you feel confused or disoriented during your experience with the system? *Q11. Do you think that this system will be helpful for your rehabilitation?*	*Not at all* (1) (2) (3) (4) (5) *Very much*
Q12. Did you find the task difficult? Q13. Did you find the devices of the system difficult to use?	*Very easy* (1) (2) (3) (4) (5) *Very difficult*
Q14. If you felt uncomfortable during the task, please indicate the reasons.	*Open response*

(a) Google Cardboard

(b) Samsung Gear VR



(c) Laptop computer



(d) Smartphone

Figure 2. Use representation of each device operated by participants.

The guests authorized to take part in the experiment filled out and signed an Informed Consent Form (ICF), and they were given copies of Confidentiality Agreement (CA) and ICF signed by the experimenters. Also, all the documents given to participants contained the phone number of the researchers and some emergency recommendations in case of presenting VR side effects after leaving the experimental installation.

All the protocols and experimental procedures have been designed in consonance with the recommendations of the UNIPAMPA's Ethics Committee on Research.

3.4 Software Tools

The virtual tour was developed based on the React 360 framework v 1.0.0 using the Sublime IDE.

The conversion from sphere shape to the panoramic shape of the 360º pictures taken with the Samsung Gear 360 camera was made through the Cyberlink ActionDirector software.

The following web browsers were adopted to experience the virtual tour in each device: Google Chrome for Windows 8.1 on the laptop, Oculus Browser on Gear VR, and Google Chrome for Android on the smartphone and the Cardboard.

We used Google Forms to create digital versions of the adapted SEQ questionnaire, the DS, and the survey of the adapted SEQ translation and back-translation process. The time for task accomplishment was also registered on a Google Drive spreadsheet.

Adapted SEQ and time data were imported on the RStudio (version 1.4.1106) for performing psychometric analysis, descriptive statistics, and inferential statistics. We used the R programming language (version 4.0.5) and the R packages `psych` (version 2.1.3), `GPArotation` (version 2014.11.1), and `irrCAC` (version 1.0).

4 Results

This section includes the demographic profile of the participants, the gathered data, the statistical methods applied, and the research question answers.

4.1 Participants

Fifty-one undergraduate students were invited to participate in this study by personal contact or by social media contact. Of these 51 guests, 41 took part in the experiment after the screening phase. Sixty-one percent of the participants (n = 25) were male.

Furthermore, the participants were distributed in undergraduate programs as following:

- 18 students of Software Engineering
- 10 students of Civil Engineering
- 7 students of Electric Engineering
- 3 students of Telecommunications Engineering
- 2 students of Agricultural Engineering
- 1 student of Computer Science

The average age of the participants was 23 years ($\bar{X} = 23.0$, $SD = 3.578$, Minimum $= 18$, Maximum $= 38$).

We also verified the sensitivity of participants to motion sickness in vehicles. Forty-four percent (n = 18) of the participants reported that they had already experienced sickness while being transported by car, bus, ship, or airplane. Among these people, only 1 participant reported that it happens often. This student was once again informed about the experiment risks regarding cybersickness and asked if he wanted to carry on with the test nevertheless.

4.2 SEQ adaptation, translation and psychometric analysis

The original SEQ developed by Gil-Gomez et al. (2013) aims to evaluate usability, suitability, and safety aspects of VR experiences for rehabilitation software. To use SEQ for evaluating a virtual tour, we have adapted it. The adaptation process just demanded the removal of one question of the original SEQ which is exclusively related to rehabilitation systems: Q11. *Do you think that this system will be helpful for your rehabilitation?* The remaining items compose the broad and multidimensional evaluation tools presented in Subsection 3.1 and they have gone through a translation process in which the results can be seen in Table 4.

The SEQ translation (except for Q11 from the original questionnaire) into Brazilian Portuguese was made by the authors of this paper (one of them proficient in English). Next, the translation quality was assessed in a translation and back-translation process based on recommendations of Coster and Mancini (2015). The evaluators had no prior knowledge of the SEQ and also had no access to the complete instrument during the assessment. It contributed to avoid a bias in the evaluation process.

Three independent evaluators, researchers on Human-Computer Interaction (HCI) and proficient in English, assessed the semantic equivalence between the original SEQ items and the translated items. Based on the answers of a digital form that had been sent to the evaluators, we achieved an agreement percentage of 89.7% and an AC_1 Gwet's (2008) agreement coefficient of .886 ($p < .001$) that is considered **very good** according to the Altman's benchmark scale described by Gwet (2016). Also, there was a majority of answers **Yes** to the question "Are the two items below semantically equivalent?" for each pair of sentences consisting of both the original SEQ item in English and the same item translated into Brazilian Portuguese.

Next, an independent professional translator performed the back-translation of the Brazilian Portuguese SEQ back into English.

The last three independent evaluators, experienced in HCI research and proficient in English, assessed the semantic equivalence between the original SEQ items and the back-translated items. Once again, we collected the responses through a digital form that we had sent to the evaluators. There was a majority of responses **Yes** to the question "Are the two items below semantically equivalent?" for each pair of sentences consisting of the original SEQ item and the back-translated item both in English, except for Q6 that received two out of three **No** responses. Thus, we achieved an agreement percentage of 59% and an AC1 Gwet's agreement coef-

Table 4. Suitability Evaluation Questionnaire adapted and translated into Brazilian Portuguese.

Question	Answer
Q1. Quanto você gostou da sua experiência com o sistema? Q2. Quanto você sentiu estar no ambiente do sistema? Q3. Quão bem-sucedido você foi no sistema? Q4. Até que ponto você conseguiu controlar o sistema? Q5. Quão real é o ambiente virtual do sistema? Q6. As informações fornecidas pelo sistema são claras? Q7. Você sentiu desconforto durante a sua experiência com o sistema? Q8. Você sentiu tontura ou náusea durante a prática com o sistema? Q9. Você sentiu desconforto ocular durante a prática com o sistema? Q10. Você se sentiu confuso ou desorientado durante sua experiência com o sistema?	*Nem um pouco* (1) (2) (3) (4) (5) *Muito*
Q12. Você achou a tarefa difícil? Q13. Você achou os dispositivos do sistema difíceis de usar?	*Muito fácil* (1) (2) (3) (4) (5) *Muito difícil*
Q14. Se você se sentiu desconfortável durante a tarefa, indique as razões.	*Resposta aberta*

ficient of .364 [3] (p = .098) that is considered **fair** considering the Altman's benchmark.

The comments of evaluators about Q6 were sent to another expert translator for inspection after being anonymized. The analysis included checking the verbal tense and the redundancy since the back-translated version is "Was the information provided by the system clear enough?" After receiving the expert's feedback, we concluded that the disagreement is fair, but our Portuguese translation of Q6 is reliable and closer to the original SEQ version than the back-translated version.

Ultimately, we performed an exploratory psychometric analysis of the adapted and translated SEQ using the answers of all participants of our experiment. So, we raised validity evidence based on the internal structure of the questionnaire.

We carried out an Exploratory Factor Analysis (EFA) procedure to identify the general organization of the adapted SEQ structure and its factors. This technique is based on the analysis of covariation of the observable variables according to Nunnally and Bernstein (1994), and Bandalos and Finney (2010). Firstly, we confirmed the adequacy of our sample to EFA procedures through the Kaiser-Meyer-Olkin's test (p = .812) and Bartlett's sphericity test ($p < .001$). Considering the ordinal nature of our data and the expectation of correlations between the factors, we chose the extraction method of Principal Axis Factoring with Oblimin Direct oblique rotation and Kaiser's normalization for EFA as recommended by Costello and Osborne (2005).

All the commonalities were higher than .3 and the total explained variance was 53% using three factors found through the scree plot analysis of the eigenvalues Figure 3. Table 5 presents the distribution of the adapted SEQ questions in each factor (items with an * had their scores inverted, and bold numbers represent the factor which those items are significantly linked into).

We can observe in Table 5 that for every question just one loading factor is either higher than .32 or lower than -.32 in one factor, thus fitting the Costello and Osborne's (2005) threshold that represents about 10% of overlapped variance. However, we notice that Q6 is slightly above the minimum

[3] This coefficient is different from the one in our original article because we redid the calculations with new software and included Q6 ratings even though they had been analyzed separately. Nonetheless, the results in our previous work remain valid and the conclusions unchanged.

Table 5. Standardized loadings matrix for the adapted SEQ in the Exploratory Factor Analysis.

	Factors		
	1	**2**	**3**
Q1	**-.738**	.024	.152
Q2	**-.796**	-.077	.032
Q3	-.203	-.088	**.559**
Q4	-.100	-.132	**.766**
Q5	**-.839**	.124	-.074
Q6	**-.325**	.224	.285
Q7*	-.106	**.759**	-.074
Q8*	.140	**.791**	.034
Q9*	-.184	**.590**	.133
Q10*	.052	.297	**.546**
Q12*	.137	.097	**.561**
Q13*	-.073	-.002	**.588**

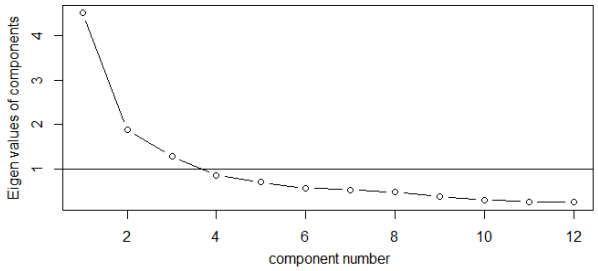
loading proposed by Costello and Osborne (2005). Once it is not strongly loaded to any factor (i.e., factor loading higher than .5 or lower than -.5 in one factor) and perspicuity is also covered by tasks' difficulty (Q12), we suggest as future study the identification of the cause of such effect and the possibility of the removal of this question in a process of continuous improvement.

Furthermore, the Cronbach's alpha coefficient (α), which is used for estimating the reliability of the instrument through the items intercorrelation according to Nunnally and Bernstein (1994) and Hutz et al. (2015), indicated a **good** internal consistency for the adapted Brazilian Portuguese SEQ (α = .844). This measure surpasses the acceptable coefficient found in the original SEQ (α = .700) and indicates a more cohesive internal structure after the adaptation process and within the experiment context (virtual tours).

The original SEQ scoring procedure consists of adding up the items' scores to achieve a global score that ranges from 13 (poor suitability) to 65 (excellent suitability). The items Q7, Q8, Q9, Q10, Q12, and Q13 (Table 3) need to be reversed before adding since they are all negative items.

According to Avila et al. (2015), the methodological procedures to evaluate the factor structure of the adapted SEQ in light of Classical Test Theory (CTT) and the original scoring strategy allow us to classify the questionnaire as a reflective measurement model. Considering that the technique of simple summation is the most commonly used for this type of model, we decided to keep it in our analyses. However, the

Figure 3. Scree plot of eigenvalues for identifying the optimal number of factors.



multidimensional structure of the SEQ requires additional investigation into the scoring procedures: should we report one total score or the factor's scores separately? Furthermore, is this imperative to the interpretation of results?

We used Haberman's approach described by Reise et al. (2013) to evaluate if the test total score (TOT_X) is a better predictor of the collection of true factor scores (SUB_{true}) than the individual factor scores (SUB_X). The strategy involves computing "the proportional reduction in mean squared error based on total scores (PRMSE_{TOT}) and compare it with PRMSE_S [the proportional reduction in mean squared error based on subscale scores estimated through the Cronbach's α for each factor]."

First, we computed the Cronbach's α (PRMSE_S) for each factor: .812 (presence), .762 (cybersickness), and .761 (usability). The standard deviations for total score and factor scores are 6.870 (SEQ), 3.300 (presence), 2.002 (cybersickness), and 3.366 (usability). The factor intercorrelations are .394 (usability-cybersickness), .513 (usability-presence), and .322 (cybersickness-presence).

Following the Reise et al. (2013) algorithm, we achieved the following PRMSE_{TOT} values: .666 (presence), .426 (cybersickness), and .708 (usability). Then we compared these values to the original PRMSE_S values: .812 (presence), .762 (cybersickness), and .761 (usability). Since the latter are larger than the former, we can argue that the factors' total scores represent a better indicator of the factors' true scores and should be reported instead of the test score. Considering this piece of evidence, we have reasonable grounds for performing a comparison between the interpretation of results based on the total score (the original SEQ approach) or factor scores (the alternative approach).

4.3 Adapted SEQ analysis

Our sample contains 164 fully answered questionnaires without any missing data[4]. The descriptive statistics for each adapted SEQ item can be found in Table 7, including sample means ($\bar{X}$), standard deviations (SD), medians (Mdn), minimum and maximum values (Min. and Max.), skewness and kurtosis measures, and Shapiro-Wilk's normality tests' W and p[5].

[4]The answers of the adapted SEQ for each participant and each device is available at `https://figshare.com/s/f57da73f33726c634760`, except for Q14.

[5]A Shapiro-Wilk's test's $p < .05$ indicate that the data come from a non-normally distributed population.

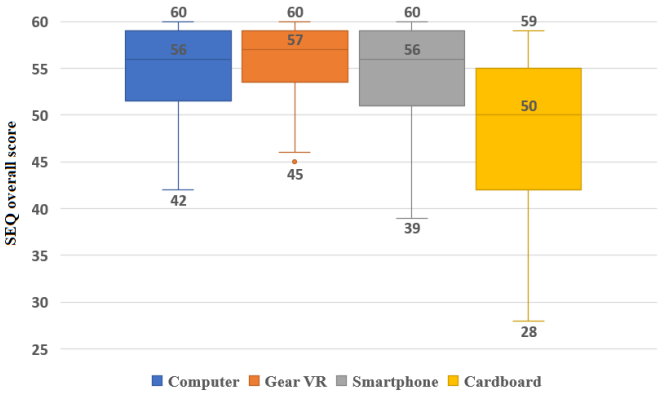


Figure 4. Box plot of total scores in different devices.

Table 6. Wilcoxon signed-rank test for the total scores for all pairs of devices in a within-subject perspective ($\alpha = .05$).

	N	Z	p	r
Computer - Gear VR	41	-1.238	.216	.137
Computer - Smartphone	41	-.260	.795	.029
Computer - Cardboard	41	-3.993	**<.001**	.441
Gear VR - Smartphone	41	-1.147	.251	.127
Gear VR - Cardboard	41	-4.947	**<.001**	.546
Smartphone - Cardboard	41	-3.984	**<.001**	.440

4.3.1 Analysis based on the total score

We computed the total score of the adapted SEQ (Table 4) of each participant by adding up the numerical value of every response. The sum was direct for items Q1, Q2, Q3, Q4, Q5, and Q6. Questions Q7, Q8, Q9, Q10, Q12, and Q13 have a negative tone and demanded being reversed (i.e., an answer with value 1 is mapped to value 5, an answer with value 2 is mapped to value 4, and so on). The total score for the adapted SEQ goes from 12 points (worst user experience) to 60 points (best user experience). The item Q14 is assessed separately from the others since it is an optional and open answer question that explains the discomfort faced by the users while performing the tasks during a virtual trip.

Figure 4 shows the box splot of the adapted SEQ scores in different devices.

We performed a Shapiro-Wilk's test on the adapted SEQ samples, and the result ($p < .01$) pointed out that all samples come from non-normal distributions.

Once non-parametric tests are more appropriate in this case of comparing non-normal samples, we carried out a Quade test ($p < .001$) to check an overall difference and multiple Wilcoxon signed-rank tests to analyze individual differences between scores of all devices. Table 6 presents the sample size for each pair in a within-subject perspective (N), the Z score (Z), the two-tailed probability value (p), and the correlation that represents the effect size (r).

We can notice that there is a significant difference only for the Computer - Cardboard, Gear VR - Cardboard, and Smartphone - Cardboard pairs ($p < .05$). By looking at the medians in Figure 4 , we can also remark that the Google Cardboard scores are lower than the others, hence revealing an experience significantly worse than other devices. Furthermore, it is possible to claim a clear contrast once the effect size for the observed differences is moderate ($r > .3$) or big ($r > .5$) according to Pallant (2016).

Table 7. Descriptive statistics on adapted SEQ items.

	$\bar{X}$	SD	**Mdn**	**Min.**	**Max.**	**Skewness**	**Kurtosis**	**Shapiro-Wilk's W**	**Shapiro-Wilk's *p***
Q1	4.372	.934	5	1	5	-1.511	1.869	.700	$< .001$
Q2	4.226	1.047	5	1	5	-1.253	.760	.744	$< .001$
Q3	4.22	.997	5	1	5	-1.257	.934	.760	$< .001$
Q4	4.482	.818	5	1	5	-1.682	2.583	.670	$< .001$
Q5	4.274	1.005	5	1	5	-1.177	.399	.731	$< .001$
Q6	4.299	1.131	5	1	5	-1.460	.925	.666	$< .001$
Q7	1.433	.908	1	1	5	2.180	4.130	.544	$< .001$
Q8	1.207	.622	1	1	5	3.490	13.170	.383	$< .001$
Q9	1.366	.872	1	1	5	2.699	6.954	.482	$< .001$
Q10	1.604	.976	1	1	5	1.639	1.966	.667	$< .001$
Q12	1.762	.99	1	1	5	1.126	.408	.757	$< .001$
Q13	1.463	.916	1	1	5	2.297	5.136	.568	$< .001$

4.3.2 Analysis based on subscale scores

We also computed the factor scores of the adapted SEQ of each participant according to the EFA output structure. The presence factor was computed by adding up the raw scores of items Q1, Q2, Q5, and Q6, ranging from 4 (worst sense of presence) to 20 (best sense of presence). The cybersickness factor was computed by adding up the raw scores of Q7, Q8, and Q9, ranging from 3 (no cybersickness symptoms) to 15 (intense cybersickness symptoms). Finally, the usability factor was computed by adding up the raw scores of items Q3 and Q4 and the reversed scores of Q10, Q12, and Q13, ranging from 5 (worst usability) to 25 (best usability).

We want to call attention to the way that the cybersickness factor score is calculated. While computing the adapted SEQ total score demanded the inversion of items Q7, Q8, and Q9, calculating the cybersickness score did not. It happens because there is a relationship between cybersickness items and other factors items when calculating the total score: since higher SEQ scores represent better user experience, the responses "(5) Very much" should be reversed when they come from negative items like "Q7. Did you feel discomfort during your experience with the system?" in order to contribute positively to the final score. On the other hand, the cybersickness factor score does not depend on other factors when computed alone. Thus, the lower the score, the better the user experience for this particular case. Figure 5 presents the box splot of the factor scores in different devices.

Shapiro-Wilk's tests on each factor score allowed us to identify that all samples come from non-normal distributions ($p < .01$) Therefore, we performed Quade tests (presence-factor $p < .001$, cybersickness-factor $p < .001$, and usability-factor $p = .013$) and multiple Wilcoxon signed-rank tests to analyze the differences among devices. Table 8 shows for every factor the sample size of each pair in a within-subject perspective (N), the Z score (Z), the two-tailed probability value (p), and the effect size (r).

We confirm a significant difference in all pairs containing the Cardboard device ($p < .05$). For all factors (presence, cybersickness, and usability), the mean scores of Google Cardboard are lower than the others and the effect size for these differences is moderate ($r > .3$) or big ($r > .5$), except for the usability factor of pair Computer-Cardboard and Smartphone-Cardboard.

Nevertheless, the pairs Computer - Gear VR and Gear VR - Smartphone also presented significant differences in the presence factor mean scores ($p < .05$) with moderate effect size ($r > .3$). Also, the pair Computer - Gear VR achieved a significant difference in cybersickness factor mean scores ($p < .05$) with a low effect size ($r \leq .3$).

4.3.3 Analysis of the open response question (Q14)

As question 14 (Table 4) demanded a full written answer, we chose to copy[6] the comments from all participants in Table 9.

We can notice in Table 9 that just **computer** use did not cause vision stress, dizziness, or headache. Of course, this result was already expected since the computer is a stable, immobile object with a big screen that relies on mouse interaction. The long loading times might be caused, indeed, by the rendering time of a scene on a big screen since all devices used the same Internet wireless connection. Moreover, the need for changing the direction of the field of view after the users had been positioned backward (Cardboard 6th item in Table 9) is related to all devices and is caused by the same usability problem: sometimes, the user is positioned facing a different, predefined direction when they enter a new scene.

Regarding the **Gear VR** device, we observe that one participant reported a dizziness effect. It is a common problem in immersive devices, and it is also a cybersickness symptom addressed by many authors such as Kennedy et al. (1993) and LaViola Jr. (2000). The temporary frame rate slowdown might be caused by overheating once cellphones usually deal with this problem through thermal throttling and other similar approaches that degrade performance. Besides, the blurred lens happened because of the skin's warmth around the HMD region and the cold, wet weather in the facilities.

Smartphone use resulted in an unexpected statement of discomfort. Even though we did not expect any cybersickness symptom once the smartphone is a non-immersive device, the report on headache and eye pain is intimately related to oculomotor issues as presented by Kennedy et al. (1993). We believe, in this case, that the look-around movements that participants had performed while focusing on a small screen to accomplish every task could have caused such sickness symptoms. The second reported issue in Table 9 has been aforementioned in the paragraph about the computer device, and it is a usability noise shared by all devices.

[6] We translated the comments from Brazilian Portuguese into English.

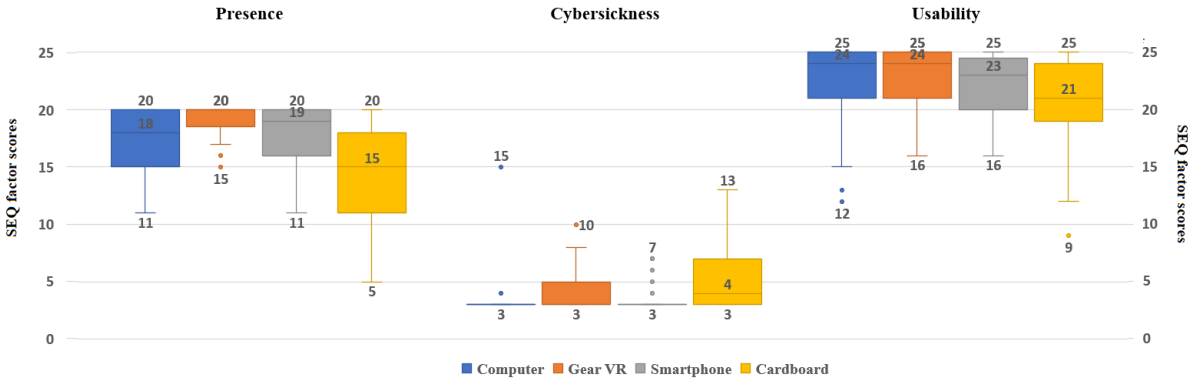


Figure 5. Box plot of factor scores in different devices.

Table 8. Wilcoxon signed-rank test for the subscale scores for all pairs of devices in a within-subject perspective ($\alpha = .05$).

		Presence factor			**Cybersickness factor**			**Usability factor**		
	N	Z	p	r	Z	p	r	Z	p	r
Computer - Gear VR	41	-3.834	<**.001**	.423	-2.700	<**.007**	.298	-.424	.672	.047
Computer - Smartphone	41	-.502	.615	.055	-1.794	.073	.198	-.313	.755	.035
Computer - Cardboard	41	-4.360	<**.001**	.482	-3.712	<**.001**	.410	-2.346	**.019**	.259
Gear VR - Smartphone	41	-2.880	**.004**	.318	-1.888	.059	.208	-.845	.398	.093
Gear VR - Cardboard	41	-5.051	<**.001**	.558	-3.069	**.002**	.339	-3.254	**.001**	.359
Smartphone - Cardboard	41	-4.121	<**.001**	.455	-3.571	<**.001**	.394	-2.361	**.018**	.261

Table 9. Answers from all participants for question Q14 of the adapted SEQ after using all devices.

Computer
1. Just after I'd clicked to access other areas, the system took a considerable amount of time to respond. The use itself didn't cause discomfort, but it did as I needed to wait for new scenes to appear.
2. Sometimes, while navigating, the next scene was loaded turned to where I'd been, so it was necessary that I rotate my view through 180ºto go on.
Gear VR
1. At some moments, the information labels, the scene title, and the navigation buttons were blurred, thus causing discomfort.
2. A little bit uncomfortable cause the slight system slowdowns, but it didn't really disturb either the experience or the task performance.
3. Dizziness.
Smartphone
1. I guess that being too much time doing the task caused me a bit of headaches and eye pain.
2. I felt uncomfortable regarding the delay in showing the images, and I also felt confused regarding being turned to one direction and, after I clicked a button and entered the next scene, getting this direction reversed (I needed to turn myself back to the path I'd been following).
Cardboard
1. Because of the visualization through the device, the images were quite blurred at some moments, and it forced me to make more effort to see, thus causing discomfort.
2. Low quality, I couldn't see a thing.
3. I felt uncomfortable regarding the visual quality of the images in which the green buttons were nearly impossible to read. I was able to accomplish the task just because I'd recalled from the previous task that it was written "secondary entrance" in that same button.
4. Because of the low-quality visualization, it made me feel uncomfortable due to forcing my vision a lot.
5. This device made me feel a little dizzy because of the low resolution.
6. Poor image. In some scenes, I entered out of order. For instance: I'd entered some scene facing forward and, in the next scene, I was facing backward.
7. Low resolution caused me discomfort.
8. The images are blurred, and the text unreadable.

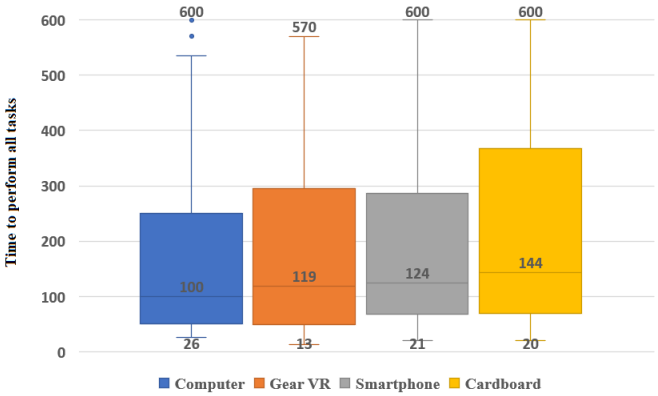


Figure 6. Box plot of time data (in seconds).

Finally, the **Cardboard** device piled up problems of the blurred lens, user's dizziness, and poor image quality. Although we consider that the smartphone used together with the Google Cardboard has a decent resolution (Subsection 3.2), we are not sure if the main cause of discomfort is the smartphone resolution, the lens quality of the cheaper viewer, or even the rendering mechanism on the web browser.

4.4 Time Analysis

We present an overall view of the elapsed time for completing tasks in each device in Figure 6 [7]. It is worth mentioning that the time limit for each task was 10 minutes.

Again, a Shapiro-Wilk's test applied over time samples revealed non-normal distributions ($p < .01$), but we were unable to identify significant differences through a Quade test ($p = .912$). We can notice that the Samsung Gear VR device was the only one in which participants accomplished all the tasks without exceeding the time limit (it does not reach the top of the box plot chart in Figure 6) but, nonetheless, the difference is not substantial.

Since the tasks have different levels of complexity and therefore unequal times for accomplishment, we performed a fairer analysis task by task. An overview of the test results can be seen in Figure 7.

We applied Kruskal-Wallis tests (easy task $p = .256$, medium task $p = .054$, hard task $p = .758$, and long trip $p = .105$) to assess differences among devices in each task type. We adopted this approach because all samples represent temporal data with similar non-normal distribution and because the comparison between devices in each task is performed in a between-subject perspective (one single participant appears just once in one of two compared samples). Even though the tests had not revealed significant dissimilarity, we chose to present the overall tendencies via multiple Mann-Whitney U tests[8]. Table 10 presents the size of each compared sample (N-N), the Z score (Z), the two-tailed probability value (p), and the effect size (r).

Next, we looked into the discomfort reported by participants in adapted SEQ Q14 (Table 9) to help us interpreting such results since the differences in Table 10 were not supported by previous Kruskal-Wallis tests.

Regarding the medium task and long trip, participants using the Google Cardboard device declared discomfort related to readability problems (items 2, 6, and 8 in Table 9) while participants using Samsung Gear VR and computer devices did not report any problems. In this case, we suppose the existence of rendering issues on the Android Chrome web browser for VR mode (where the screen image is split in two separate images - one for each eye) resulting in lower visual quality. However, it remains unclear why participants were not equally affected in easy and hard tasks.

Unfortunately, there were no discomfort records concerning the easy task that could help us to understand better the difference between Gear VR and the smartphone.

4.5 Discussion

(1) Is there any significant difference in the experience reported by users of virtual tours among different devices? Yes. The Google Cardboard platform presented a significantly worse User Experience than the computer, smartphone, and Samsung Gear VR. While the main measures for this conclusion have been the adapted SEQ total score (by measuring self-reported user satisfaction, sense of presence, perceived success, perceived control, realism, comprehensibility of instructions, cybersickness symptoms, and general discomfort) and the adapted SEQ factor scores (presence, cybersickness, and usability), the discomfort reported by participants in each device corroborates the score analysis. We also observed that the key negative factor in Cardboard was the readability problems caused, most likely, by rendering issues on the web browser used and by the low budget lens of the platform. We are not sure about the impact of the screen resolution that is slightly lower in the used with the Cardboard smartphone.

We were not able to identify any substantial difference among the other devices (computer, smartphone, and Gear VR) by the adapted SEQ overall score, but the scores of the presence factor alone indicate that the sense of being there was higher when users experienced the virtual tour through the Samsung Gear VR. This result was expected since the Gear VR is the only immersive, high-quality device under testing. Sundar et al. (2017) also identified higher presence scores in VR experiences using a different questionnaire despite using a Cardboard as VR equipment. Since the experiments of Sundar and colleagues (2017) involved watching VR content passively, the participants probably did not face many usability issues related to control like blurred navigation widgets and hard-to-read instructions.

Furthermore, the scores of the cybersickness factor show that symptoms are more intense for Gear VR users in comparison with desktop computer users. This was also anticipated because desktop computers are more stable and thus unlikely to induce cybersickness symptoms, especially when compared to immersive VR devices.

(2) Is there any difference in performing a holistic or a multidimensional UX analysis? Yes. The multidimensional analysis based on factor scores revealed differences among devices that otherwise would go unnoticed.

[7] Data with elapsed time for each participant and each device is available at `https://figshare.com/s/b2fe9fd05d5373333f90`.

[8] We also decided to keep Table 10 to maintain consistency with the original paper.

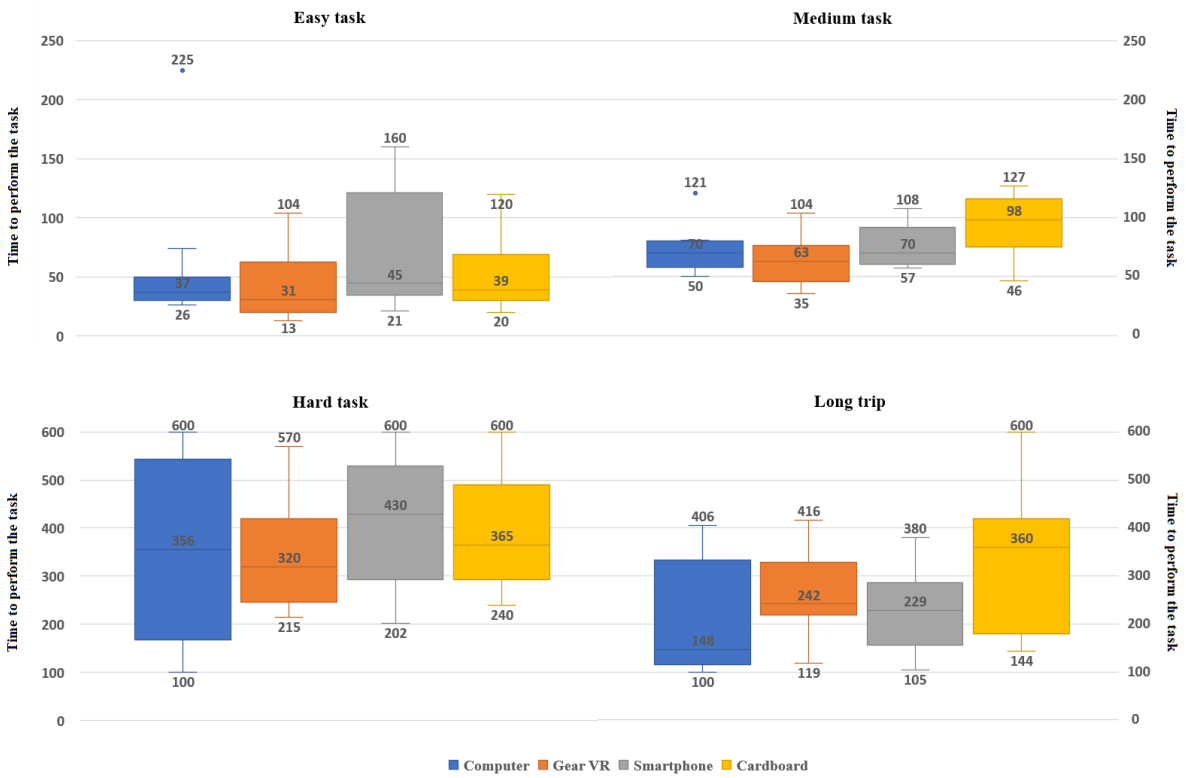


Figure 7. Box plot of time data per task (in seconds).

Table 10. Mann-Whitney U test for the elapsed time for all pairs of devices in each task and a between-subject perspective.

	Computer - Gear VR				Computer - Smartphone				Computer - Cardboard			
	N-N	*Z*	*p*	*r*	**N-N**	*Z*	*p*	*r*	**N-N**	*Z*	*p*	*r*
Easy task	11-10	-.638	.523	.139	11-10	-1.165	.244	.254	11-10	-.567	.571	.124
Medium task	10-10	-1.211	.226	.051	10-11	-.601	.548	.131	10-10	-1.633	.103	.365
Hard task	10-11	-.035	.972	.008	10-10	-.681	.496	.152	10-10	-.341	.733	.076
Long trip	10-10	-1.512	.131	.338	10-10	-.756	.450	.169	10-11	-2.290	**.022**	.500

	Gear VR - Smartphone				Gear VR - Cardboard				Smartphone - Cardboard			
	N-N	*Z*	*p*	*r*	**N-N**	*Z*	*p*	*r*	**N-N**	*Z*	*p*	*r*
Easy task	10-10	-2.005	**.045**	.448	10-10	-1.216	.224	.272	10-10	-.794	.427	.178
Medium task	10-11	-1.551	.121	.338	10-10	-2.534	**.011**	.567	11-10	-1.340	.180	.292
Hard task	11-10	-1.021	.307	.223	11-10	-.740	.459	.161	10-10	-.454	.650	.102
Long trip	10-10	-.832	.406	.186	10-11	-.775	.438	.169	10-11	-1.585	.113	.346

Some of these differences are obvious, like the higher sense of presence for Samsung Gear VR users and the weaker cybersickness symptoms for desktop computer users when compared to immersive devices. On the other hand, there is some new insight about factor correlations. It is important to notice that even if presence and usability factors are strongly correlated ($r = .513$), the lower sense of presence does not seem to be enough for impacting the overall usability as we can observe in pairs Computer - Gear VR and Gear VR - Cardboard. This corroborates the results of Chow (2016) that achieved the same .51 correlation coefficient between "perceived ease of use" and "presence" factors, but no strong direct effect based on Structural Equation Modeling analysis.

Since there are two cases of a different sense of presence and indistinguishable usability (Gear VR compared to non-immersive devices) and there is no case of the opposite, we hypothesize that the interaction noises and obstacles faced by users (namely usability issues) affect the sense of being there, but not the contrary. Another piece of evidence supporting this hypothesis is the consistently lower presence scores for the Google Cardboard, although it provides an immersive VR experience. Sundar et al. (2017) got higher presence scores from Google Cardboard users in comparison with desktop computer users (what is contradictory to our findings), but their stimuli contained just 360º videos with almost no interaction and thus no room for substantial usability issues.

(3) Is there any significant difference in the users' performance among different devices? We did not observe any significant difference in that in our case study. Even though the elapsed time analysis to accomplish tasks presented dissimilarity for three cases in tasks with distinct complexity levels, the overall test showed no statistical significance and a detailed analysis revealed that the issues faced by participants (causing a lower performance) had not been constant. Thus, we cannot claim that the readability problems consistently affected users of the Cardboard platform since the decreasing performance in two out of twelve comparisons (Table 10) does not impact the overall performance with the device.

Besides, we did not identify any performance difference among the distinct interaction equipment for selection: mouse on the computer, trackpad on the Gear VR, touchscreen on the smartphone, and clicker on the Cardboard.

5 Conclusion

As the seeking for web virtual tours grows and the compatibility between web Virtual Reality and visualization devices increases, it's important to check the differences in the users' experiences and performance on distinct platforms, since it leads the users to a better choice based on cost-benefit and context of use.

Thus, we developed a virtual tour in UNIPAMPA using WebXR and React 360 technologies, and carried out a case study with 41 participants. Four distinct platforms (a laptop computer with a mouse, a smartphone, a Google Cardboard viewer with a mouse that was used as clicker, and a Samsung Gear VR HMD) were alternately used by all participants while performing tasks with different levels of complexity.

We conclude with this research that the adapted SEQ scores point out a significantly worse overall experience on the Google Cardboard viewer when compared to the others in our virtual tour context. Moreover, the sense of presence was higher for Samsung Gear VR users and the cybersickness symptoms were weaker for desktop computer users as expected. Finally, differences in the elapsed time to accomplish tasks are indistinguishable among the platforms.

The key contributions of our case study are (1) the identification of worse user experience on Google Cardboard in virtual tours that are mainly related to visualization problems, (2) the observation of similar users' time performance among the platforms, (3) the discovery of cybersickness symptoms while using the smartphone for the virtual trip in a non-immersive context, and (4) the possibility of low usability exerting a direct effect upon the sense of presence (the lower the usability, the lower the presence).

An important secondary contribution of our paper is the translation, adaptation, and psychometric analysis of a standardized questionnaire to assess UX aspects related to presence, cybersickness, and usability in virtual tours. The adapted SEQ presents enough evidence for being used as a standardized instrument in a multidimensional perspective by measuring user satisfaction, sense of presence, perceived success, perceived control, realism, comprehensibility of instructions, cybersickness symptoms, and general discomfort.

Based on the factor structure of the adapted SEQ, we also identified that computing factor scores is a more reliable strategy for interpret results than computing a total score. The scoring procedure analysis as presented by Reise et al. (2013) is, to the best of our knowledge, innovative for User Experience standardized questionnaires. It is also important to mention that this approach is just feasible for multidimensional questionnaires.

We would like to draw attention to some limitations on this research: (1) the impossibility of using 360º images in higher quality due to server-side network bandwidth capacity, (2) the invitation to participate in the study that was exclusively sent to undergraduate students of UNIPAMPA which had already visited the tour environment physically, (3) the population was not controlled for representativeness regarding gender and occupation, (4) the slightly lower resolution of the Asus Zenfone 3 could have worsened readability problems on Google Cardboard even though the lens quality or the web browser rendering engine is the central issue, (5) the lack of a standardized instrument widely validated for virtual tours, and (6) the lack of access to high-end HMD like Oculus Rift and HTC Vive.

Finally, we suggest as future work (1) the creation of a set of guidelines for web virtual tour development, (2) the inclusion of new platforms regarding both visualization and interaction in the comparison, (3) the increase and diversification of the participant sample, and (4) the detailed evaluation of the psychometric quality of the adapted SEQ instrument.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

This article is an extended version of the previous paper: Rodrigues, A., and Cheiran, J. F. P. (2020). Virtual look around: interaction quality evaluation for virtual tour in multiple platforms. In Proceedings of the 22nd Symposium on Virtual and Augmented Reality (SVR), (pp. 47-56). IEEE. DOI: 10.1109/SVR51698.2020.00023.

References

Avila, M. L., Stinson, J., Kiss, A., Brandão, L. R., Uleryk, E., and Feldman, B. M. (2015). A critical review of scoring options for clinical measurement tools. *BMC Research Notes*, 8(1):1–11.

Bandalos, D. L. and Finney, S. J. (2010). Factor analysis. exploratory and confirmatory. In Hancock, G. R. and Mueller, R. O., editors, *The reviewer's guide to quantitative methods in the social science*, pages 93–114. Routledge, New York.

Bloom, L. B. (2020). Ranked: The world's 15 best virtual tours to take during coronavirus. Available at https://www.forbes.com/sites/laurabegleybloom/2020-/04/27/ranked-worlds-15-best-virtual-tours-coronavirus.

Boland, P. and Johnson, C. (1996). Archaeology as computer visualization: virtual tours of dudley castle c. 1550. In *Imaging the Past: Electronic Imaging and Computer Graphics in Museums and Archaeology*, pages 227–234. British Museum Press.

Butcher, P. W. and Ritsos, P. D. (2017). Building immersive data visualizations for the web. In *2017 International Conference on Cyberworlds (CW)*, pages 142–145. IEEE, IEEE Conference Publications.

CatracaLivre, R. (2020). Passeios virtuais mostram as belezas dos destinos brasileiros. Available at https://catracalivre.com.br/viagem-livre/passeios-virtuais-mostram-as-belezas-dos-destinos-brasileiros.

Cho, Y.-H., Wang, Y., and Fesenmaier, D. R. (2002). Searching for experiences: The web-based virtual tour in tourism marketing. *Journal of Travel & Tourism Marketing*, 12:1–17.

Chow, M. (2016). Determinants of presence in 3D virtual worlds: A structural equation modelling analysis. *Australasian Journal of Educational Technology*, 32(1):1–18.

Costa, R. M. and Ribeiro, M. W. S. (2009). *Aplicações de Realidade Virtual e Aumentada*. Editora SBC – Sociedade Brasileira de Computação, Porto Alegre, RS, Brazil.

Costello, A. B. and Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7):1 – 9.

Coster, W. and Mancini, M. (2015). Recommendations for translation and cross-cultural adaptation of instruments for occupational therapy research and practice. *Revista De Terapia Ocupacional Da Universidade De São Paulo*, 26(1):50–57.

Gil-Gomez, J.-A., Pilar, M.-H., Albiol, S., Carmen, A. V., Hermenegildo, G.-G., and José-Antonio, L. Q. (2013). Seq: Suitability evaluation questionnaire for virtual rehabilitation systems. application in a virtual rehabilitation system for balance rehabilitation. In *Int. Conf. on Pervasive Computing Technologies for Healthcare and Workshops*, pages 335–338.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Gwet, K. L. (2016). *Handbook of Inter-Rater Reliability*. Advanced Analytics, LLC, Gaithersburg, USA, 4 edition.

Hutz, C. S., Bandeira, D. R., and Trentini, C. M., editors (2015). *Psicometria*. Artmed, Porto Alegre, RS, Brazil.

Jerald, J. (2016). *The VR Book: Human-Centered Design for Virtual Reality*. ACM Books, Morgan & Claypool, Williston, VT, USA.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3):203–220.

Kizony, R., Raz, L., Katz, N., Weingarden, H., and Weiss, P. L. T. (2005). Video-capture virtual reality system for patients with paraplegic spinal cord injury. *Journal of Rehabilitation Research & Development*, 42(5):595–609.

Klippel, A., Zhao, J., Jackson, K. L., Femina, P. L., Stubbs, C., Wetzel, R., Blair, J., Wallgrün, J. O., and Oprean, D. (2019). Transforming earth science education through immersive experiences: Delivering on a long held promise. *Journal of Educational Computing Research*, 57(7):1745–1771.

LaViola Jr., J. J. (2000). A discussion of cybersickness in virtual environments. *SIGCHI Bull.*, 32(1):47–56.

Lee, G. A., Dünser, A., Nassani, A., and Billinghurst, M. (2013). Antarcticar: An outdoor ar experience of a virtual tour to antarctica. In *Int. Symp. on Mixed and Augmented Reality*, pages 28–37. IEEE.

Nunnally, J. and Bernstein, I. (1994). *Psychometric Theory*. MacGraw-Hill, New York, NY, USA, 3rd edition.

Oprean, D., Wallgrün, J. O., Klippel, A., Zhao, J., Duarte, J., and Verniz, D. (2018). Developing and evaluating vr field trips. In Fogliaroni, P., Ballatore, A., and Clementini, E., editors, *Int. Conf. on Spatial Information Theory (COSIT 2017)*, pages 105–110. Springer International Publishing.

Osman, A. and Wahab, N. A. (2011). Virtual excursions for tiny fingers: A shared experience. In *International Conference on Information Science and Applications*, pages 1–5. IEEE.

Osman, A., Wahab, N. A., and Ismail, M. H. (2009). Development and evaluation of an interactive 360º virtual tour for tourist destinations. *Journal of Information Technology Impact*, 9:173–182.

Pallant, J. (2016). *SPSS Survival Manual*. McGraw-Hill Education, Berkshire, England, 6 edition.

Parisi, T. (2015). *Learning Virtual Reality*. O'Reilly Media, Sebastopol, CA, USA.

Pujol, L. (2004). Archaeology, museums and virtual reality.

Digithum, 6:none.

Reise, S. P., Bonifay, W. E., and Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95(2):129–140.

Sathe, V., Gupta, P., Kaushik, K., Bhat, S., and Deshpande, S. (2017). Virtual reality websites (vr web). In *International Conference of Electronics, Communication and Aerospace Technology*, volume 1, pages 647–652. IEEE.

Sundar, S. S., Kang, J., and Oprean, D. (2017). Being there in the midst of the story: How immersive journalism affects our perceptions and cognitions. *Cyberpsychology, Behavior, and Social Networking*, 20(11):672–682.

Tarcia, L. (2020). Projeto minas faz ciência 360. Available at https://minasfazciencia.com.br/infantil/2020/04/13/projeto-minas-faz-ciencia-360.

Ye, Q., Hu, W., and Zhou, H. (2017). Implementation of webvr-based laboratory for control engineering education based on ncslab framework. *2017 36th Chinese Control Conference (CCC)*, pages 7880–7885.

Zaiontz, C. (2018). Latin squares design. Available at http://www.real-statistics.com/design-of-experiments/latin-squares-design.