Identifying Relationships between Physiological Measures and Evaluation Metrics for 3D Interaction Techniques

Rafael Rieder, Christian Haag Kristensen, Márcio Sarroglia Pinho

Identifying Relationships between Physiological Measures and Evaluation Metrics for 3D Interaction Techniques

Rafael Rieder[1,2], Christian Haag Kristensen[1], Márcio Sarroglia Pinho[1]

[1] Pontifical Catholic University of Rio Grande do Sul, Av. Ipiranga, 6681, Partenon
CEP 90619-900 Porto Alegre, RS, Brazil
[2] University of Passo Fundo, BR 285, Km 171, São José
CEP 99052-900 Passo Fundo, RS, Brazil
rafaelrieder@gmail.com, {christian.kristensen, marcio.pinho}@pucrs.br

Abstract. This project aims to present a methodology to study the relationships between physiological measures and evaluation metrics for 3D interaction techniques using methods for multivariate data analysis. Physiological responses, such as heart rate and skin conductance, offer objective data about the user stress during interaction. This could be useful, for instance, to evaluate qualitative aspects of interaction techniques without relying on solely subjective data. Moreover, these data could contribute to improve task performance analysis by measuring different responses to 3D interaction techniques. With this in mind, we propose a methodology that defines a testing protocol, a normalization procedure and statistical techniques, considering the use of physiological measures during the evaluation process. A case study comparison between two 3D interaction techniques (ray-casting and HOMER) shows promising results, pointing to heart rate variability, as measured by the NN50 parameter, as a potential index of task performance. Further studies are needed in order to establish guidelines for evaluation processes based on well-defined associations between human behaviors and human actions realized in 3D user interfaces.

Keywords: usability metrics, physiological measures, interaction techniques.

1 Introduction

In order to evaluate the characteristics of three-dimensional user interfaces (3DUI), like presence and immersion, methods and tools commonly used to evaluate two-dimensional user interfaces can be applied, such as prototypes, questionnaires and formative and summative tests. These instruments are able to get relevant usability metrics also in 3DUIs, like variables to measure system performance, user task performance and user preferences. The first two measures result in objective data for assessing, respectively, the computer or graphics system performance, and the quality of performance of specific tasks in the 3D application. The third measure results in subjective data for assessing the user satisfaction while using an interface. The

evaluation of these measures is important because they allow that different 3DUI elements, such as spatial perception, multimodal interaction and sensory stimulus are considered.

However, the adaptation of these tools to evaluate 3DUIs can lead to incomplete assessment of the particular characteristics of these applications, such as the use of non-conventional devices and 3D interaction techniques (ITs) [6]. These characteristics tend to influence the user performance and the user satisfaction, which requires a process to evaluate its various resources based on the user experience level.

A recent alternative used to evaluate interfaces it is the use of the physiological measures. According to Malik *et al* [22], the physiological monitoring provides information about the user's physiological balance, and its measures are associated with stress. Researches in the Virtual Reality (VR) area have been using this type of measurement to assess the user's physical and mental effort on the 2D games [17][18][19][20] and to evaluate presence and user comfort in immersive virtual environments (VEs) [7][8][16][23][24]. However, there are no studies about the relationship between physiological measures and metrics focused on the evaluation of the quality of ITs.

The use of physiological measures can still address other two classical problems in the evaluation of 3DUIs. The first concerns the collect of objective measures, which in some cases require modifications in the source code. In these situations, it is not always possible or desirable to alter an application, due to the complexity of the system [2] or the limited availability of development time [28]. The second problem concerns the reliability of subjective metrics, which may have influenced their results by external factors to the interaction process, such as user's physical and mental efforts, or user's cognitive mediation, such as omission or summarization of information.

Physiological measures, therefore, offer objective responses that are not controlled by the person, they are associated with factors such as fatigue and irritation, and provide data related to the organism's behavior. These are measures that can indicate, for example, the adaptation periods to a new device or new IT, because the user's stress level can be viewed along the timeline. Besides, they can aid in the comprehension of the performance results and answers of questionnaires. Doing so, physiological responses may complement the current methods of assessment [5][14][25], allowing the understanding of the interaction process as a whole, and contributing to increasing the quality of the VR applications.

In order to evaluate whether physiological measures may be or not a substitute for objective and subjective usability metrics, statistical methods of multivariate data analysis can be used. According to Hair *et al* [13], these methods allow analyzing simultaneously the influence of multiple measures on each subject or variable under investigation, regardless of the complexity or the context in which these variations occur. So, it is possible to verify if results obtained with physiological measures are able to indicate the same problems identified by traditional usability metrics.

Therefore, this work describes a methodology for assessing the quality of ITs in immersive VEs, comparing physiological measures and evaluation metrics for 3DUIs using multivariate data analysis. Our methodology contemplates the use of a testing protocol, data normalization and exclusion processes, and statistical methods for exploratory data analysis and regression analysis, in order to discover relationships

between variables that contribute to the evaluation process, complementing or assisting in the interpretation of results. In the same way, our methodology also determines the physiological measures able to indicate the same results expected by traditional usability measures, and these may eventually replace usual measures in projects in which the simplification of the testing stage is desirable. So, it is possible to reduce the dependency on subjective data, and to avoid changes to collect performance data in complex software.

This paper is organized as following: Section 2 presents the related work, whereas Section 3 describes the developed methodology. In Section 4, a case study is presented to evaluate the use of this approach. Section 5 shows a discussion about the results obtained using this methodology. Finally, Section 6 concludes the paper highlighting the potential of our approach.

2 Related Work

The use of physiological measures is a recent alternative in the evaluation of graphical interfaces. Latest researches apply this resource to measure presence and cybersickness in immersive VEs, and user's physical and cognitive effort in videogames.

Meehan *et al* [23][24] used physiological monitoring to measure presence in a stressful VE. These researches used heart rate (HR), skin conductance (SC) and skin temperature measures in four different experiments to compare participants' physiological reactions to a non-threatening virtual room and their reactions to a stressful virtual height situation. According to the authors, HR satisfied all the requirements for a reliable, valid, sensitive and objective measure of presence in a stressful VE. In addition, HR showed correlation with the well-established presence questionnaire. SC had some of the properties desired to measure presence, and skin temperature did not.

Slater *et al* [26] reported the difficulty to measure presence subjectively, and highlighted the need to evaluate alternatives that provided objective data to overcome, or, at least, supplement the use of questionnaires. With this in mind, the authors conducted an experiment to explore the relationship between physiological responses, breaks in presence and utterances by virtual characters towards the participants using a virtual bar scenario. The results showed that changes in HR and SC point to occurrence of breaks in presence during the interaction process. Changes in HR also indicated the moments when an avatar speaks to the subjects, whereas heart rate variability (HRV) parameters pointed to differ between participants with different social anxiety scores, classified by questionnaire.

Brogni *et al* [7][8] followed a similar approach to previous work. They studied the use of physiological responses to determine the impact of visual realism and the user stress level during the interaction in an urban VE. In this scenario, the sense of presence was subjected to the texture quality and the appearance of avatars. The results showed that HRV parameters indicated the reducing of level stress as time goes by. The authors also reported that these parameters were associated with the level of visual realism based on the texture strategy used in their experiments.

Kim *et al* [16] presented a study about the use of physiological measures to reduce cybersickness in immersive VEs. Their approach proposed a system based on 11 physiological signals, which detected the cybersickness and automatically reduced the user's field of view and slowed the travel velocity in the VE. The results indicated a significantly lower frequency of cybersickness when users use this system, compared to a situation in which it is not used. The authors also highlighted the gastric tachyarrhythmia as a physiological measure of a better outcome.

On the other hand, Lin *et al* [17][18] studied the use of physiological measures as a metric to evaluate the usability of video games. The experiments considered the use of HR, SC and blood pressure (BVP) measures to assess the user's performance and satisfaction while tasks were realized in three different difficulty game levels. The authors reported the SC measure as relevant in comparisons between game levels and performance user groups, indicating that a greater number of errors and difficulty consequently increases the SC. This work also studied the relationship between physiological measures and frustration events, which showed a range of more than 5% in SC in 70% of the frustration cases analyzed. Other measures did not showed significant results.

Complementary studies of Lin *et al* [19][20] added measures like pupillary response, eye tracking and HRV parameters to assess the user cognitive efforts in videogames. The results also showed the relationship between HRV parameters and game levels. Pupillary response and eye tracking presented promise results, but it was not consistent as SC and HRV measures.

From the literature review, it can be noticed that there are no studies exploring the relationship between physiological measures and usability metrics used to evaluate the quality of three-dimensional ITs in immersive VEs. The following section presents a methodology designed for this purpose.

3 Methodology

Firstly, the following sections present the platform for testing, the physiological measures, the task performance measures and the questionnaires used by our approach. Secondly, we present the steps to apply our methodology, which include the use of a test protocol, normalization and exclusion methods for physiological measures, and statistical analysis methods.

3.1 Platform for testing

In order to illustrate the use of our methodology, we built a virtual room with four numbered books, distributed on the floor inside the user's field of vision, as presented in Figure 1. Two well-known ITs also were implemented to select and manipulate objects: ray-casting [3] and HOMER [4]. These techniques were chosen because they are commonly used as parameters to evaluate new ITs. With ray-casting, the user selects and manipulates objects using a virtual light ray, with the ray's direction specified by the user's hand. With HOMER, the user selects an object using the

ray-casting technique, and manipulates it using a virtual hand, which instantly moves to the current object's position and attaches to it.

The user's task is to get the books, turning them as necessary, and place them in transparent areas marked on the floor. The books must be organized with their front cover visible to the user, and their spine toward to the left side of the virtual room. At the end of task, the books will be placed in ascending order, from left to right. The VE also provides two visual feedbacks to indicate when an object is ready for selection, and an aural feedback to inform collisions occurred into the virtual room.

This application was built using C++, OpenGL, GLUT, irrKlang[1], and the SmallVR[2] toolkit, which simplifies the development of VR applications by abstracting many implementation aspects such as device control and scene graph management, while maintaining the GLUT structure for the program.

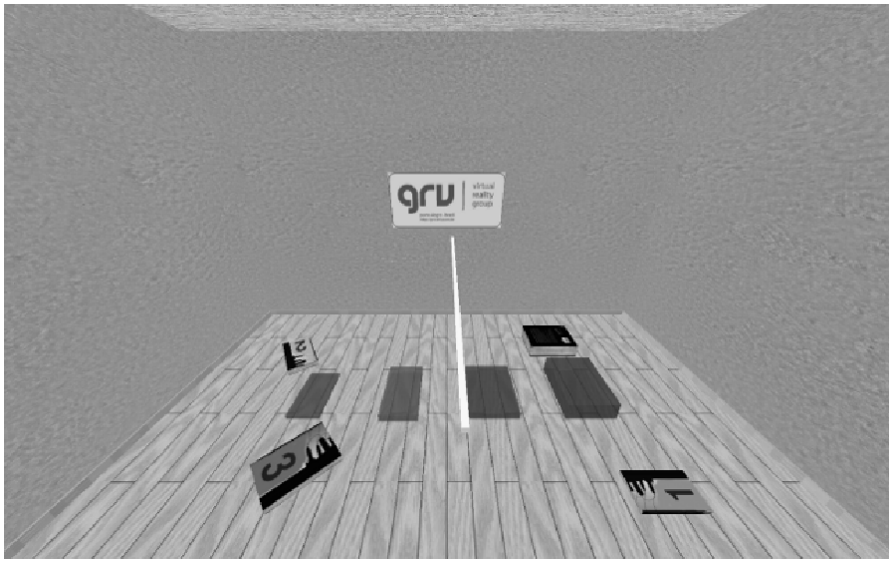

Fig. 1. The virtual room application built to our experiment.

In order to explore the VE and use the ITs resources, we used i-Glasses Head Mounted Display (HMD) and Polhemus FastTrak motion tracker with two tracking points enabled for interaction. The first tracking point was used to track the user's head movements, whereas second was used in the user's dominant hand to select and manipulate objects. Grab and release user's actions were confirmed using a push-button attached on the second tracking-point.

The physiological monitoring used an electrocardiogram (ECG) sensor and a SC sensor, on a non-invasive way. These equipments were connected to the Procomp Infiniti encoder, which captured and sent the physiological responses to the Biograph Infiniti software for data processing. Three electrodes connected to the ECG sensor

[1] http://www.ambiera.com/irrklang
[2] http://grv.inf.pucrs.br/projects

were fixed on the user's wrists with rubber wrist straps, whereas two electrodes connected to the SC sensor were strapped to two fingers of user's non-dominant hand using finger bands. All these solutions are manufactured by Thought Technology[3].

Figure 2 presents the device's default configuration, highlighting the push-button attached to one of the tracking-points, and positions of the ECG and SC sensors.

Fig. 2. Subject wearing physiological and VR devices during the experiment.

3.2 Physiological Measures

For this work we used the HR and SC physiological measures, collected by ECG and SC sensors, respectively. Our approach also includes the use of seven different HRV measures, generated by time domain and frequency domain methods from short-term 5-minutes recordings, according to Malik *et al* [22].

This way, the following physiological measures were selected:

- Mean SC, expressed in micro-Siemes (μS);
- Mean HR, expressed in beats per minute (bpm);
- Standard deviation of the NN[4] interval (SDNN), expressed in milliseconds (ms);

[3] http://www.thoughttechnology.com

[4] On an ECG recording, NN are the normal-to-normal intervals between adjacent QRS complexes.

– Number of interval differences of successive NN intervals greater than 50 ms (NN50);
– Proportion derived by dividing NN50 by the total number of NN intervals (pNN50);
– Mean total power in very low frequency range (VLF), expressed in milliseconds squared (ms^2);
– Mean total power in low frequency range (LF), expressed in ms^2;
– Mean total power in high frequency range (HF), expressed in ms^2;
– Ratio between the LF and HF measures (LF/ HF).

3.3 Task Performance Measures

In order to get the task performance during the experiment with ITs, the following measures were defined. These data were collected in real time by the own VE application:
– *Total time*: entire period of the interaction process, measured in seconds (s);
– *Accuracy*: distance between the center of an object and the center of its target position[5], measured in meters (m);
– *Collisions*: amount of collisions occurred with an object;
– *Grabs*: number of attempts to grab an object.

3.4 Questionnaires

Two questionnaires were created for this work. The pre-test questionnaire obtained demographic information and user's past experiences with VEs, whereas the post-test questionnaire evaluated the user preferences in relation to the interface, tasks, and ITs. A progressive scale of 1 to 7 was set to evaluate each question in both instruments, from a lower to a higher concept.

Pre-test questionnaire asked about the age, gender and education of each participant, and was composed of five questions. These questions were about the VR level of knowledge, the number of times there was presence in VE experiments, the number of times there were physiological measures, the sense of discomfort when interacting with new interfaces, and the sense of discomfort when interacting with graphical interfaces like games and VEs.

Post-test questionnaire was composed of seven questions. The aim of these questions were to evaluate the influence of wired devices in the user's performance, the level of irritation generated by the fact of requiring to the user pick up again the objects after each collision, the level of pressure generated by the time limit to complete the task, the level of satisfaction in using the ray-casting and the HOMER techniques, the level of confidence in performing the task correctly, and the level of satisfaction with the visual and aural feedbacks presented by the VE. Moreover, this questionnaire also provided areas to describe the occurrence of discomfort situations, and the moments of more irritation or more difficulty during the interaction process.

[5] The target position means the desired position of the object in the VE.

3.5 Test Protocol

In order to test our methodology, we designed a protocol based on related work presented in Section 2. Our approach supports steps to the simultaneous collection of task performance, subjective evaluation and physiological data during the experience.

The test protocol was established with nine stages, which can be executed in approximately 45 minutes, in the following order:

- Apply pre-test questionnaire;
- Prepare devices;
- Collect baseline data;
- Start training – first IT;
- Start experiment – first IT;
- Start training – second IT;
- Start experiment – second IT;
- Release devices;
- Apply post-test questionnaire.

"Apply pre-test questionnaire" stage contemplates the trainer and experiment presentations, the distribution of informed consent form to read, and the filling of pre-test questionnaire. After this it is provided to read the VE instructions, which explains how to execute tasks using the two ITs. The estimate duration to finish this stage is eight minutes. According to Kim *et al* [16], this period is also important to stabilize the physiological responses before next stages.

"Prepare devices" stage considers the arrangement of physiological sensors and VR equipments in the user's body. Firstly, the user is invited to turn off electronic devices, to remove watch and bracelets and to accommodate in a chair, sitting in a comfortably way. After this the trainer cleans the user's wrists with alcohol gel, and applies a conductive gel into the ECG sensors to reduce noise caused by electrical resistance of the skin. The physiological sensors are fixed as mentioned in Section 3.1 and, finally, the user wears the HMD. For this stage, we estimate approximately three minutes to complete it.

"Collect baseline data" is started after the user feels comfortable with the devices. The trainer requests the user to put his/her arms on his/her legs in a rest position. In this step, VR devices remain turned off, and it is asked for the user to keep his/her eyes open. Based on Slater *et al* [26] and Brogni *et al* [8] collecting physiologic data is done in two minutes and it will be the comparison base for the interaction process.

In the next stage, "Start training – first IT", the subject begins to interact with the VE to learn to use the devices and the first IT. Selection and manipulation tasks can be performed during three minutes.

Again, the subject interacts with the same VE and IT in the "Start experiment – first IT" stage, and must perform all tasks in seven minutes. After this period, the subject has three minutes to rest with all devices off.

The same procedures are applied to the next two stages for training and experiment of the "second IT".

It is important to highlight for an unbiased analysis, the use of two techniques must be balanced according the number of participants in an experiment. For this reason, our protocol divides the experiment between "first IT" and "second IT".

The above times, defined to complete the training and experiment stages, are based on Lin *et al* [17][18][19][20] and Brogni *et al* [7][8] works, whereas the interval to rest between stages is based on Kim *et al* [16].

During the "Release devices" stage, VR and physiological equipments are removed from the user. The trainer applies procedures to clean the user's wrist and devices, using dry wipes and dusters. The estimate time is four minutes to do it.

At last, "Apply post-test questionnaire" stage contemplates the subjective evaluation of test, which the user is invited to answer the post-test questionnaire. A brief period also is addressed for comments and thanks. For this stage, we estimate approximately five minutes, based on Slater *et al* [26] and Lin *et al* [20] works.

3.6 Normalization and Exclusion Methods for Physiological Measures

In order to statistically compare physiological responses and different usability variables, it is necessary to define data normalization and exclusion methods to the HR and SC measures.

For the SC data normalization, we adopted a scale of 0 to 1, which it attributed the minimum value of 0 to a lowest SC value, and the maximum value of 1 to a highest SC value. This procedure was applied to each user's SC signal, generated a new and normalized Mean SC measure. So, SC values became uniform and preserved the individual characteristics of each subject.

By contrast, for the HR measures we needed to apply a procedure to exclude some data, because the adopted way to collect this physiological response is susceptible to generation of noises in the HR signal. The procedure eliminated participants who presented HR values outside the normal range for a human. The exclusion criterion was executed in the following order:

- HR rest: based on the baseline data, subjects were excluded from the dataset when their mean HR was below 60 bpm or above 100 bpm. According to Guyton and Hall [12], typical healthy resting HR in adults is 60–100 bpm;
- HR max: subjects were excluded from the dataset when their mean HR during the experiment was above to the maximum HR, which it was estimated from the Tanaka formula [27], presented by the Equation 1;
- HR target: subjects were excluded from the dataset when their mean HR during the experiment was above to the target HR, which it was estimated from the Karvonen method [15], presented by the Equation 2. In order to use this method, we determined an intensity level of 50% to the interaction task, since the physical effort during the interaction process can be considered within a moderate activity zone [1], as a result of the subjects being seated and performed spatial movements using their arms and head during the test.
- HR min: subjects were excluded from the dataset when their mean HR during the experiment was below 60 bpm.

$$HR_{max} = 208 - (0{,}7 \text{ x Age}) \quad (\mathbf{1})$$

$$HR_{target} = [(HR_{max} - HR_{rest}) \text{ x (Intensity level \%)}] + HR_{rest} \quad (\mathbf{2})$$

3.7 Statistical Analysis Methods

In order to verify the relationships between different measures, we chose to use multivariate data analysis methods. According to Hair *et al* [13], these methods are able to investigate, simultaneously, multiple measures about each subject or object under study.

This work adopted the following analytical steps:

- Apply methods for exploratory data analysis to summarize, test the normality, detect outliers of the data, and use techniques to verify correlations between variables. This approach allows to identify the consistency and distribution of the data, and avoid the redundant variables;
- Apply multiple regression techniques to generate prediction models, considering methods to select relevant variables and its coefficients of determination. This approach allows discovering what measures are associated to the task performance and subjective responses.

For the exploratory data analysis, we defined the following tests:

- Summarize data: Descriptive statistics;
- Normalization: Kolmogorov-Smirnov test;
- Outliers' detection: stem and leaf and box-plots;
- Correlations: Pearson's coefficient to linear relationships, and Spearman's coefficient to nonlinear relationships.

The stepwise regression was chosen to create regression models and selects the predictor variables. In this process, each regression model may have one or more predictor variables and their coefficient of determination (r^2). These coefficients inform the power of these measures have to explain the variability of results. In this project, these results indicate whether a physiological measure can substitute or not a traditional usability measure. An analysis of variance (ANOVA) is also applied to test the significance of the regression model.

In order to generate regression models, our analysis selected only physiological measures with results statistically significant in the correlation tests ($p < 0,05$).

4 Case Study

In order to evaluate the effectiveness of physiological measures as indexes of quality to 3DUIs, this section aims to present a case study using our methodology, according to the definitions shown in Section 3.

Our evaluation included 54 healthy participants, 28 men and 26 women aged between 17 and 57 years old. Their tests were scheduled during a period of two weeks.

The subjects were also distributed into two equal groups (14 men and 13 women), in order to balance the use of ITs. The group "A" used as first IT the ray-casting technique, whereas the group "B" used as first IT the HOMER technique.

4.1 Relationships between Physiological and Task Performance Measures

According to the Section 3.7, it is necessary to apply a set of multivariate data analysis methods to assess the relationship between physiological and task performance measures, as presented in Sections 3.2 and 3.3. Thus, it is possible to identify which physiological responses are able to indicate task performance, or whether they can at least assist the interpretation of the results.

First of all, we used the testing protocol to collect the physiological, task performance and user preferences measures. After this we applied the normalization and exclusion procedures, in order to adjust our physiological datasets for the statistical comparisons. During this last stage, we detected some abnormal HR measures in 22 subjects, which needed to be discarded. Because of this situation, the original dataset had to be subdivided into two new groups: a dataset for SC measures, which included all the experiment participants (54 subjects), and another dataset for HR and HRV measures, which included only 32 subjects.

In the next stage, we applied the statistical methods for exploratory data analysis and multiple regression, looking for physiological measures able to indicate task performance.

Since our methodology was applied, two physiological measures (NN50 e HF) had a statistically significant relationship with two task performance measures ("Total time" and "Accuracy"). However, only one of these relationships indicated, on the regression model, strongly statistically significant results by both techniques ("Total time" x NN50, $p < 0,01$), as presented in the Table 1.

According to the results of the Table 1, the "accuracy" task performance measure only has statistically significance with the NN50 and HF physiological measures, for experiments using ray-casting technique.

Table 1. Regression models for physiological and task performance measures with strong correlation.

Interaction Techniques	Task Performance Measures	Physiological Measures	Regression	ANOVA	
			r^2 (%)	F-test	p(value)
HOMER	Total Time	NN50	28,83%	12,15	0,00**
	Accuracy	NN50	7,26%	2,35	0,13
	Accuracy	HF	2,35%	0,72	0,59
Ray-Casting	Total Time	NN50	61,98%	48,91	0,00**
	Accuracy	NN50	43,28%	22,89	0,00**
	Accuracy	HF	31,33%	13,69	0,00**

On the other hand, NN50 physiological measure may be considered as the variable with the most associated with the "Total time" measure, because results were strongly significants for both experiments, independently of two techniques ($p < 0,01$). Based on Table 1, the NN50 physiological measure is able to indicate the user task performance, for the "total time" measure, with a statistical power (r^2) of 61,98% to the experiments using ray-casting technique, and 28,83% to the experiments using HOMER technique.

We also generated a regression model using the NN50 and "Total time" means, in order to join the statistical power of the selected physiological response in a single

model, independently of two techniques. The result showed a coefficient of determination of 45,16% (ANOVA, $p < 0,01$, $F = 24,70$).

However, our results presented intermediary statistical power values. The coefficients of determination showed values far from an index close to perfect correlation ($r^2 = 100\%$), showing that the variance of NN50 measure cannot explain, alone and exactly, the variance of "Total time" measure. In other words, we can say that the NN50 physiological measure still cannot be used to replace the "Total time" measure during a task performance evaluation.

4.2 Relationships between Physiological and User Preferences Measures

Using the same approach presented in the Section 3.7, we also applied a multivariate data analysis to verify the relationships between physiological and user preferences measures, already presented in the Sections 3.2 and 3.4. In this study, we also used the data normalization and exclusion procedures, defined in the Section 3.6, which subdivided our dataset in two new sets (54 subjects for the SC measure, and 32 subjects for the HR and HRV measures).

In order to compare the questionnaire answers and physiological measures, we generated new physiological measure means from the values of experiences using the two ITs. Results can be visualized in the Table 2.

Table 2. Regression models for physiological and user preferences measures with strong correlation.

Questionnaires	User Preferences Measures	Physiological Measures	Regression	ANOVA	
			r^2 (%)	F-test	p(value)
Pre-test	Question 1	SC	11,44%	6,71	0,01*
	Question 1	NN50	15,26%	5,40	0,03*
	Question 2	NN50	13,82%	4,81	0,03*
	Question 2	LF	10,65%	3,57	0,07
	Question 2	HF	7,50%	2,43	0,13
	Question 4	SC	10,49%	6,09	0,02*
Post-test	Question 4	HR	12,83%	4,42	0,04*
	Question 4	NN50	17,02%	6,15	0,02*
	Question 4	HF	23,39%	9,16	0,01*
	Question 7	LF/HF	13,72%	4,77%	0,03*

Firstly, tests were applied to verify the relationships between physiological measures and pre-test questionnaire answers. In this study, NN50 and SC physiological measures presented statistically significant relationships with questions addressed the level of knowledge about VR (Question 1), the experience with non-conventional devices in VEs (Question 2) and the tendency to feel discomfort or irritation when interact with new interfaces (Question 4).

Based on these analysis, we may note that the regression models presented in the Table 2 showed significant results ($p < 0,05$) for SC and NN50 measures as predictors of assessment, but with low statistical power for the Question 1 (SC, $r^2 = 11,44\%$; NN50, $r^2 = 15,26\%$), Question 2 (NN50, $r^2 = 13,82\%$) and Question 4

(SC, $r^2 = 10,49\%$). In this way, we can say that these physiological measures – when solely used – still cannot be employed to indicate the user level of knowledge about VR, the user level of experience in VEs, and the user level of irritation during learning in new graphical interfaces.

Secondly, we applied the same tests to verify the relationships between physiological measures and post-test questionnaire answers. In this study, we did not compare the physiological measure means and the answers related to the Questions 4 and 5, because these questions aimed to evaluate the ITs, separately. In this case, the Question 4 responses were compared with physiological measures collected during the user experiences with the ray-casting technique, and Question 5 responses were compared with physiological data of the experiences using HOMER technique.

This study presented only one physiological measure (LF/HF) with statistically significant relationship ($p < 0,05$) for the evaluation about the quality of visual and aural feedbacks displayed during the interaction process (Question 7). However, only 13,72% of the variance of LF/HF can explain the variance of the Question 7 responses. So, we can say that the LF/HF is not able to indicate this item evaluation.

Comparisons between physiological measures and Questions 4 and 5, which aimed to evaluate the level of satisfaction in using the ITs, presented statistically significant relationship ($p < 0,05$) only between HR, NN50 and HF measures and the Question 4 answers (ray-casting technique evaluation), as shown in the Table 2. The generated regression models also showed determination coefficients of low explanatory power (HR, $r^2 = 12,83\%$; NN50, $r^2 = 17,02\%$; HF, $r^2 = 23,39\%$), impossible to indicate the level of satisfaction with the use of ray-casting technique through physiological measures.

The post-test questionnaire also evaluated aspects about physical discomfort, sense of irritation and difficulties to perform tasks during the interaction process. The subjects' responses showed that these sensations did not affect the user's performance during the test, pointing no significant results.

5 Discussion

Based on the previous section, physiological measures still cannot be considered as indexes of user task performance and user preferences measures for VEs. From the results, the physiological responses only indicate user behavior tendencies during the interaction process, which can help as an additional resource to understand usability metrics during the evaluation process.

Comparisons between physiological and task performance measures only highlight the NN50 measure, which presented statistically significant results for both evaluated techniques and statistical power near an acceptable level to the "Total time" measure. According to the Section 3.2, the NN50 attests the amount of interval differences of successive NN intervals greater than 50 ms, which indicates the level of stabilization of the heart rhythm. Figure 3 shows a correlation plot between "Total time" and "NN50" measures, considering both experiences using the two ITs. This figure also indicates the trend of the NN50 increases as the "Total time" increases too.

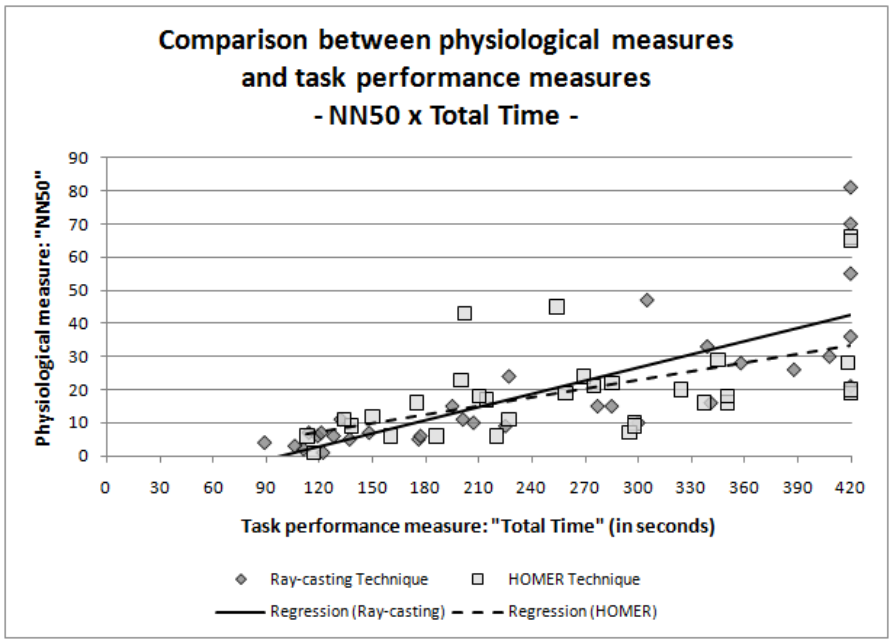

Fig. 3. Correlations and trends between "Total time" and "NN50" using the two ITs.

This result can be interpreted on two different points of views. Firstly, experiments completed in less time show more concentrated subjects, which also use more physical effort to perform the tasks, compared with those who spend more time. On the other hand, we can say that the subjects spend more time learning to use the 3DUI, which ends up leaving them more relaxed and with their HRs in levels around a baseline measurement. Anyway, it is a promising measure to be explored and evaluated again in future research.

Comparisons between physiological user preferences measures presented some relevant measures, such as SC, NN50 and LF/HF – but none of them showed significant statistical power. Probably, these low relationships can be associated with the first application of the pre- and post-test questionnaires, and their progressive scales.

With this in mind, we applied some tests to verify the questionnaire's reliability. Firstly, Cronbach's alpha [10] and split-half tests [11] were used, which presented the results shown in Table 3. According to Hair *et al* [13], the reliability coefficients indicate a good assessment tool when their index results in values above 0,7, which were not observed in this analysis. A probably reason for this result is the sample size used (54 subjects), which was below the literature recommendations, considering the scale dimension used by the instruments. As our scale ranged from 1 to 7, it would be necessary to apply our questionnaires to 70 subjects, at least, which guarantees obtaining more reliable results.

We also applied a factor analysis to the questionnaires to verify their relevance. According to Malhotra [21], it is recommendable to have the largest possible number

of factors to this analysis, which are generated based on the number of questions, satisfying at least 60% of the total variance. According to the Table 4, for the pre-test questionnaire, 63,54% of the total variance can be explained using two factors, considering eigenvalues greater than 1,00. For the post-test questionnaire, Table 5 presents that 67,60% of the total variance can be explained using three factors, but only two of them have eigenvalues greater than 1,00. These results are also corroborated by low values of the Kaiser-Meyer-Olkin (KMO) tests, which examine the appropriateness of factor analysis.

These results recommend the review of the progressive scale used in our questionnaires, before new evaluation. Certainly, these changes will result in a more refined analysis to discover the relationships between physiological measures and subjective data.

Table 3. Reliability tests for the pre- and post-test questionnaires.

Questionnaires	Cronbach's Alpha	Guttman Split-Half Coefficient	Items
Pre-test	0,50	0,10	5
Post-test	0,35	0,48	7

Table 4. Factor analysis of the Pre-test questionnaire (KMO measure = 0,51).

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	1,92	38,30	38,30
2	1,26	25,24	63,54
3	0,92	18,48	82,02
4	0,60	12,05	94,07
5	0,30	5,93	100,00

Table 5. Factor analysis of the Post-test questionnaire (KMO measure = 0,60).

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2,24	32,04	32,04
2	1,50	21,36	53,40
3	0,99	14,20	67,60
4	0,73	10,46	78,06
5	0,70	10,05	88,11
6	0,49	6,93	95,04
7	0,35	4,96	100,00

Moreover, our evaluation was partially hampered because some data were discarded during the normalization and exclusion procedures. Comparisons involving HR and HRV measures had a loss of almost 60% of data.

According to Combatalade [9], despite the precaution taken in relation to skin preparation, conductive gel application, electrode placement and user instructions, it is very difficult to save HR data absolutely clean and no noise. It forces the use of a

normalization process to the HR signal, especially to detect two types of artifacts: missed beats and extra beats.

Missed beats can occur when the signal is so distorted that the software is unable to identify the beat pattern and only picks up on the next good beat, which results in long inter beat interval value. On other hand, extra beats can occur when the program confuses a distortion in the signal for a beat and detect two or more beats, when there should be only one, resulting in short inter beat interval value at the end of analysis.

In order to reduce artifacts, a software solution to process HR signals and analyze HRV measures can be adopted. In this case, it is important to be assisted by a medical professional in order to ensure that the data cleaning does not interfere in future results.

It is important to mention that there are also real natural physiological events, similar to these artifacts, like premature trial or ventricular contractions. As a result of this, it is recommended to follow a specific protocol if there is presence of subjects with heart diseases and/or under medication.

Another suggestion to minimize the occurrence of noise in HR signal, it is the use of self-adjustable wrist straps, which prevents the electrodes become tightened or loosened on the user's wrists. It is possible to use a non-invasive device fixed on the user's chest, closer to the heart, which reports the HR to the ECG sensor without using wires.

At last, it is also recommended a collaborative effort between Computer Science and Medicine experts, in order to define guidelines allowing a better understanding about the user's behavior during the interaction process using physiological measures to do it.

6 Conclusions

This work presents a methodology for assessing the quality of ITs in 3DUIs, based on the use of physiological measures during the interaction process. A case study comparing two ITs was executed, in order to apply this new method through a well-defined testing protocol, two procedures to normalize and exclude some samples from physiological dataset, and the application of multivariate data analysis methods, which allow to highlight the relationships between physiological measures and usability metrics.

The use of our methodology pointed to promising results and interesting tendencies. We except our methodology to evolve and to be consolidate through new comparisons with other ITs, not only to evaluate selection, manipulation or navigation tasks in VEs, but also to consider another features of 3DUIs, like computer stereo vision, different degrees of freedom and multiple sensory stimuli.

Physiological measures still cannot be considered as substitutes of user task performance or user preferences. However, these measures can be used as a complementary resource for the interpretation of usability metrics commonly observed during the evaluation processes, highlighting the relationship between NN50 psychological measure and "Total time" task performance measure. We recommend a

more detailed study of this measure for future work, in order to contribute and qualify the ITs evaluation process.

Our research also presents a mode to validate the proposed questionnaires, as a means of establishing better relationships between physiological measures and subjective user preferences. This analysis pointed to the importance of scale adjustment for subjective evaluation before testing, based on the available sample size of subjects. This fact may have contributed to measures with significant differences, such as SC and NN50, have shown low statistical power in this work.

Regarding the comparisons between physiological measures and usability metrics, we highlighted the need to use multivariate data analysis techniques during the evaluation process. These statistical methods allowed the understanding and the detailed case study of relationships between different measures, encouraging the formulation of more refined conclusions and the indication of trends for future studies in this promising area.

Finally, we emphasize that our project is a first step to define a specific methodology to the 3DUI evaluation process, considering the use of physiological measures and appropriated statistic techniques to compare multiple datasets. The continuity of our research involves new case studies and a multidisciplinary effort between Computer Science and Medicine experts to create guidelines for clear and wide association of the user's behavior and their actions performed in 3DUIs using different interaction resources.

References

1. American College of Sports Medicine: ACSM's Advanced Exercise Physiology. Lippincott Williams & Wilkins, Philadelphia (2005)
2. Bisbal, J., Lawless, D., Wu, B., Grimson, J.: Legacy Information Systems: Issues and directions. IEEE Software 16–5, 103–111 (2002)
3. Bolt, R. A.: "Put-That-There": Voice and Gesture at the Graphics Interface. In: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, pp. 262–270. ACM, New York (1980)
4. Bowman, D. A., Hodges, L. F.: An Evaluation of Techniques for Grabbing and Manipulating Remote Objects in Immersive Virtual Environments. In: Proceedings of the 1997 Symposium on Interactive 3D graphics, pp. 35–38. ACM, New York (1997)
5. Bowman, D. A., Gabbard, J. L., Hix, D.: A Survey of Usability Evaluation in Virtual Environments: classification and comparison of methods. Presence: Teleoperators and Virtual Environments 11–4, 404–424 (2002)
6. Bowman, D. A., Kruijff, E., LaViola, J.J., Poupyrev, I.: 3D User Interfaces: theory and practice. Addison-Wesley, Boston (2004)
7. Brogni, A., Vinayagamoorthy, V., Steed, A., Slater, M.: Variations in Physiological Responses of Participants during Different Stages of an Immersive Virtual Environment Experiment. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology, pp. 376–382. ACM, New York (2006)
8. Brogni, A., Vinayagamoorthy, V., Steed, A., Slater, M.: Responses of Participants during an Immersive Virtual Environment Experience. The International Journal of Virtual Reality 6–2, 1–10 (2007)

9. Combatalade, D.: Basics of Heart Rate Variability Applied to Psychophysiology. Technical report MAR953-00, Thought Technology Ltd. (2010)
10. Cronbach, L. J.: Coefficient Alpha and the Internal Structure of Tests. Psychometrika 16–3, 297–334 (1951)
11. Guttman, L.: A Basis for Analyzing Test-Retest Reliability. Psychometrika 10–4, 255–282 (1945)
12. Guyton, A. C., Hall, J. E.: Textbook of Medical Physiology. Saunders, Philadelphia (2005)
13. Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E.: Multivariate Data Analysis. Prentice Hall, New Jersey (2005)
14. Hix, D., Hartson, H.: Developing User Interfaces: ensuring usability through product & process. John Wiley & Sons, New Jersey (1993)
15. Karvonen, M. J., Kentala, E., Mustala, O.: The Effects of Training on Heart Rate: a longitudinal study. Annales Medicinae Experimentalis et Biologiae Fenniae 35–3, 307–315 (1957)
16. Kim, Y. Y., Kim, E. N., Park, M. J., Park, K. S., Ko, H. D., Kim, H. T.: The Application of Biosignal Feedback for Reducing Cybersickness from Exposure to a Virtual Environment. Presence: Teleoperators and Virtual Environments 17–1, 1–16 (2008)
17. Lin, T., Omata, M., Hu, W., Imamiya, A.: Do physiological data relate to traditional usability indexes? In: Proceedings of the 17th Australia conference on Computer-Human Interaction, pp. 1–10. Computer-Human Interaction Special Interest Group of Australia, Narrabundah (2005)
18. Lin, T., Imamiya, A., Omata, M., Hu, W: An Empirical Study of Relationships Between Traditional Usability Indexes and Physiological Data. Australasian Journal of Information Systems 13–2, 105–117 (2006)
19. Lin, T., Imamiya, A.: Evaluating usability based on multimodal information: an empirical study. In: Proceedings of the 8th International Conference on Multimodal Interfaces, pp. 364–371. ACM, New York (2006)
20. Lin, T., Imamiya, A., Mao, X.: Using Multiple Data Sources to get Closer Insights into User Cost and Task Performance. Interacting with Computers 20–3, 364–374 (2008)
21. Malhotra, N. K.: Marketing Research: an applied orientation. Prentice Hall, New Jersey (2006)
22. Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., Schwartz, P. J.: Heart Rate Variability: Standards of measurement, physiological interpretation, and clinical use. European Heart Journal 17–3, 354–381 (1996)
23. Meehan, M.F., Insko, B., Whitton, M., Brooks Jr, F. P.: Physiological Measures of Presence in Stressful Virtual Environments. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 645–652. ACM, New York (2002)
24. Meehan, M.F., Razzaque, S., Insko, B., Whitton, M., Brooks Jr, F. P.: Review for Four Studies on the Use of Physiological Reaction as a Measure of Presence in Stressful Virtual Environments. Applied Psychophysiology and Biofeedback 30–3, 239–258 (2005)
25. Rosson, M., Carroll, J.: Usability Engineering: scenario-based development of human-computer interaction. Morgan Kaufmann, San Francisco (2001)
26. Slater, M., Guger, C., Edlinger, G., Leeb, R., Pfurtscheller, G., Antley, A., Garau, M., Brogni, A., Friedman, D.: Analysis of Physiological Responses to a Social Situation in an Immersive Virtual Environment. Presence: Teleoperators and Virtual Environments 15–5, 553–569 (2006)
27. Tanaka, H., Monahan, K. D., Seals, D. R.: Age-predicted Maximal Heart Rate Revisited. Journal of the American College of Cardiology 37–1, 153–156 (2001)
28. Tullis, T., Albert, W.: Measuring the User Experience: collecting, analyzing, and presenting usability metrics. Morgan Kaufmann, San Francisco (2008)