2017 IEEE 41st Annual Computer Software and Applications Conference

Quality Assessment of Interaction Techniques in Immersive Virtual Environments using Physiological Measures

Rafael Rieder
rieder@upf.br
Universidade de Passo Fundo
Rio Grande do Sul - Brazil

Christian Haag Kristensen
christian.kristensen@pucrs.br
Pontifícia Universidade Católica do Rio Grande do Sul
Rio Grande do Sul - Brazil

Márcio Sarroglia Pinho
pinho@pucrs.br
Pontifícia Universidade Católica do Rio Grande do Sul
Rio Grande do Sul - Brazil

***Abstract*—This paper presents a new methodology for quality assessment of interaction techniques in immersive virtual environments, based on the study of the relationships between physiological measures and usability metrics using multivariate data analysis. Our methodology 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. Besides, this study also shows that heart rate (HR) and skin conductance (SC) measures reflect the user's task performance during the interaction process. Despite these results, this work reveals that physiological measures still cannot be considered as substitutes of evaluation metrics for 3D interfaces, but may be useful in the interpretation and understanding process of them. Discussions also indicate the further studies are needed 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, 3D interaction techniques.

I. 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, such as prototypes, questionnaires and formative and summative tests, can be applied. These instruments are able to get relevant usability metrics also in 3DUIs, like variables to measure system performance, user task performance and user preferences. 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.

An alternative used to evaluate interfaces it is the use of the physiological measures. According to Malik et al. [19], 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 [15, 16, 17, 18] and to evaluate presence and user comfort in immersive virtual environments (VEs) [7, 8, 14, 20, 21]. 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 acquisition 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 [25]. The second problem concerns the reliability of subjective metrics, which may have influenced their results by external factors unrelated to the interaction process, such as user's physical and mental efforts, or cognitive mediation, such as omission or summarization of information.

Physiological measures, therefore, offer objective responses that are not controlled by the user, they are associated with factors such as fatigue and irritation, and provide data related to the behavior of the human organism. 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 questionnaire answers. Doing so, physiological responses may complement the current methods of assessment [5, 12, 22], allowing the understanding of the interaction process as a whole, and contributing to increasing the quality of the VR applications.

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 proposes 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

DOI 10.1109/COMPSAC.2017.196

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.

In order to validate these measures, this work uses physiological measures as usability metrics. To do so, we make comparisons between two manipulation ITs, between experiment stages, between process tasks and between performance user groups. Our idea is to identify if the user's physiological changes may contribute to improve 3D interface issues, and to detect situations that affect the user performance. In the same way, we intend also to indicate which IT generates a lower stress level and better suited to perform a particular interaction task.

II. Methods

A. 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.

The user's task is to get the books, turning them as necessary, and place them in transparent areas marked on the floor. In order to explore the VE and use the ITs resources, we used a Head Mounted Display (HMD) and a 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.

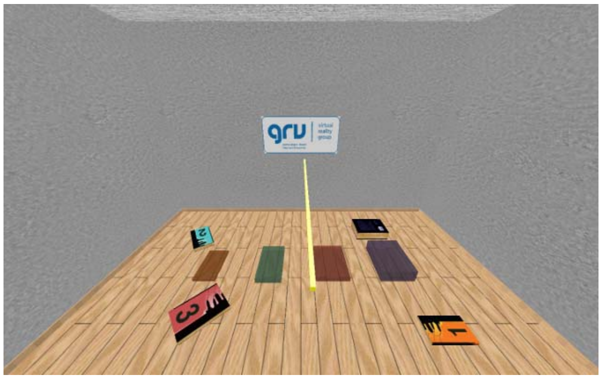

Figure 1. The virtual room application built to our experiment.

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 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.

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.

B. Physiological and Task Performance 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 heart rate variability (HRV) measures, generated by time domain and frequency domain methods from short-term 5-minutes recordings: mean SC, mean HR, standard deviation of the NN interval (SDNN), number of interval differences of successive NN intervals (NN50), proportion derived by NN50 (pNN50), mean total powers in very low frequency range (VLF), low frequency range (LF), high frequency range (HF), and ratio between LF and HF measures (LF/HF).

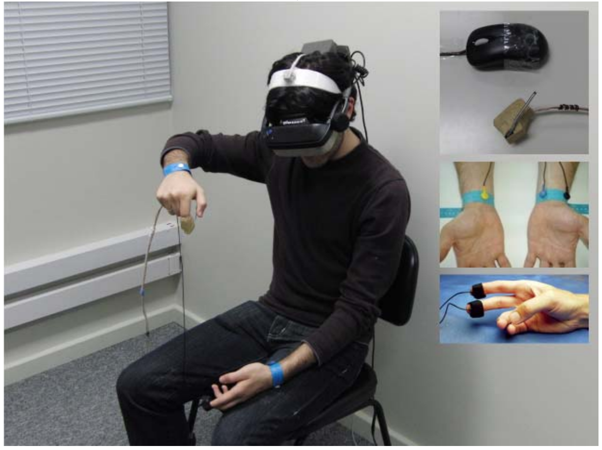

Figure 2. Subject wearing physiological and VR devices during the experiment.

In order to get the task performance during the experiment with ITs, the following measures were defined: total time, accuracy, collisions and grabs.

C. Questionnaires

Two questionnaires were created for this work. 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 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.

D. Test Protocol

The test protocol was established with nine stages, which can be executed in approximately 48 minutes, in the following order:

- Apply pre-test questionnaire (≈ 8 min);
- Prepare devices (≈ 3 min);
- Collect baseline data (≈ 2 min);
- Start training – first IT (≈ 3 min);
- Start experiment – first IT (≈ 10 min);
- Start training – second IT (≈ 3 min);
- Start experiment – second IT (≈ 10 min);
- Release devices (≈ 4 min);
- Apply post-test questionnaire (≈ 5 min).

"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.

"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 A and, finally, the user wears the HMD. "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.

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.

Again, the subject interacts with the same VE and IT in the "Start experiment – first IT" stage, and must perform all tasks. After this period, the subject rests 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".

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. 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.

E. Normalization and Exclusion Methods for Physiological 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 [10], 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 [24], 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 [13], 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.

$$\text{HRmax} = 208 - (0.7 \text{ x Age}) \quad (1)$$

$$\text{HRtarget} = [(\text{HRmax} - \text{HRrest}) \text{ x (Intensity level \%)}] + \text{Hrrest} \quad (2)$$

F. 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. [11], 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 (Descriptive statistics), test the normality (Kolmogorov-Smirnov), detect outliers of the data (stem and leaf and box-plots), and use techniques to verify correlations between variables (Pearson and Spearman coefficients). This approach allows identifying the consistency and distribution of the data, and avoiding the redundant variables;
- Apply multiple regression techniques (stepwise regression) 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. 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 ($p < 0.05$).

III. Results

Our evaluation included 54 healthy participants, 28 men and 26 women aged between 17 and 57 years old. The subjects were also distributed into two equal groups, 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.

A. Relationships between Physiological and Task Performance Measures

According to the Section F, it is necessary to apply a set of multivariate data analysis methods to assess the relationship between physiological and task performance measures. 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, detecting some abnormal HR measures in 22 subjects, which were 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 regressions, 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.

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 significant 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.

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**

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 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.

B. Relationships between Physiological and User Preferences Measures

We also applied a multivariate data analysis to verify the relationships between physiological and user preferences measures. In this study, we also used the data normalization and exclusion procedures, subdividing our dataset in two new sets as reported in Section A.

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.

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.

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*

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.

C. Comparison between interaction techniques

Considering the normalization procedures and the physiological measures presented in Section B, comparisons between ITs were performed with paired two-sample t-test and ANOVA. For this reason, we used the mean values of the physiological signals collected during the experiment.

As null hypothesis (H0), we supposed that there is no difference between the use of the techniques, and as alternative hypothesis (H1) we defined that the use of the ray-casting technique may cause more stress to the user than the use of the HOMER technique. In short, this means that the physiological stress using ray-casting tend be higher than HOMER.

Statistical tests didn't point to significant results for this comparison. As a consequence, H1 hypothesis was rejected. Therefore, for this case study there is no difference in using ray casting technique or HOMER technique.

D. Comparison between experiment stages

In order to highlight the physiological differences between experiment and resting stages, comparisons were performed with paired two-sample t-test and ANOVA. We used the mean values of the physiological signals collected during the baseline and experiment stages.

The first analysis aimed to identify if the interaction process changes levels of anxiety (measured by SC) or physical stress (measured by HR) of a user, considering the baseline data and the experiments data. As H0 hypothesis, we defined that there is no significant difference in anxiety or physical effort levels between baseline and experiment stages, and as H1 hypothesis, we defined that the baseline stage presents physiological values lower than the experiment stages. In short, the user is relaxed and calm before performing the interaction tasks in VE.

Comparing HR values between baseline and first experiment, strongly statistically significant results were observed ($p < 0.01$, $t = -3.97$, $F = 15.74$). Similar results also were obtained when comparing the baseline and the second experiment ($p < 0.01$, $t = -4.86$, $F = 23.65$), and the mean of the experiments ($p < 0.01$, $t = -5.32$, $F = 28.27$). So, we can say that the baseline reports fewer heartbeats per minute than during the experiments. This proves, in this study, the users pass from a resting state to a physical effort state, noted by experiences with the interaction process.

On the other hand, no significant results were observed in comparison with the same stages using SC values. As a result, H0 was accepted, pointing that there is no significant difference between the anxiety levels generated by the baseline and the experiment stages. This can be explained by the subjects' profile, since for most users it was the first VR experience.

The second analysis compared the changes in the levels of anxiety and in the physical stress between two experiment stages, regardless of the technique used. As H0 hypothesis, we supposed that there is no significant difference in anxiety or physical effort levels between the first and second experiment, and as H1 hypothesis, we supposed that the first experiment presents physiological values higher than the second experiment. In doing so, we hope this may be an indication that the user is more used to the interaction process in the second experiment.

Using HR values, the comparison between experiment stages showed no statistically significant results, indicating the physical effort required to interact in both experiments is similar, rejecting H1 hypothesis.

Using SC values, we observed strongly statistically significant results between the experiments ($p < 0.01$, $t = 3.16$, $F = 10.00$). So, we can say that the users are more anxious in the first experiment. This also suggests that the cognitive load required in a first is greater than the second experiment, because the user is learning to interact in the VE and adapting to the interaction process.

E. Comparison between interaction process tasks

In order to measure the relationship between physiological measures and interaction process tasks, we used paired two-sample t-test and ANOVA to compare selection and manipulation tasks performed by users. For this analysis, we used only normalized SC values. ITs were evaluated separately based on the SC mean values of each interaction task.

As H0 hypothesis, we supposed that there is no significant difference on the anxiety level between selection and manipulation tasks. As H1 hypothesis, we supposed that the manipulation task presents anxiety levels higher than the selection tasks.

As a result, statistical tests didn't point to any significant results for both ITs, rejecting H1 hypothesis.

In order to measure whether the occurrence of user events during the interaction tasks increases the levels of anxiety, we also compared the SC mean values between the interaction tasks (selection and manipulation) and the following user events: grab, release or collide.

As a result of this comparison, two distinct situations were observed between user's tasks and events. Firstly, considering only the HOMER technique, we noted a strongly statistically significant result between manipulation tasks and collision events ($p < 0.01$, $t = 3.13$, $F = 9.77$). In this case, manipulation tasks generated more anxiety than collision events. Secondly, considering only the ray-casting technique, there is a significant difference between collide events and selection tasks ($p < 0.05$, $t = -1.99$, $F = 3.96$). In conclusion, collide events generate less anxiety than selection tasks.

F. Comparison between performance user groups

In order to classify the subjects in rank groups, we used the sum of the four task performance measures (total time, accuracy, collisions and grabs), sorted from lowest to highest. The 10 lowest results were classified as the "Best Performance" (BP) group, while the highest 10 results as the "Worst Performance" (WP) group. We used the normalized data set to execute an evaluation based on performance among the rank groups, and according to the IT used. In summary, four rank groups were created: ray-casting BP (RCBP), ray-casting WP (RCWP), HOMER BP (HBP) and HOMER WP (HWP).

Comparisons were performed with independent two-sample t-test. In this case, we used Mean SC and Mean HR values as variables because they are indicators of the levels of anxiety and physical effort. NN50 measure also was used because it presented the most statistically significant results (Sections A and B).

We observed significant results using the SC values ($p = 0.02$, $t = 2.13$) and HR values ($p = 0.03$, $t = 2.21$) in order to compare RCBP and RCWP. These results showed the best performance users were more anxious and presented more physical effort than the worst performance users.

In addition, we also observed a strongly significant result using the NN50 measure ($p < 0.01$, $t = -7.17$) with unequal variances. As a consequence, we can say the RCBP group has adapted better to use the ray-casting technique than the RCWP group, although results of RCBP group indicate a higher number of heartbeats.

By contrast, the comparison between HBP and HWP groups showed only a statistically significant result for the NN50 measure ($p = 0.02$, $t = -2.89$) with unequal variances. As a result, we can also say the HBP group has adapted better to use the HOMER technique than the HWP group.

We also compare the performance between BP groups, considering the NN50 measure. This study presented a strong statistically significant result between ITs ($p = 0.01$, $t = -2.77$) with unequal variances. This showed that users of the RCBP group had a lower HRV than users of the HBP group. So, we can say the RCBP group has adapted better to the features offered by this technique during the interaction process, since it had a NN50 lower value, suggesting a more lilting rhythm and less heart rate oscillations [23].

IV. Discussion and Conclusions

A. Relationship between physiological measures and usability metrics

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 B, 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.

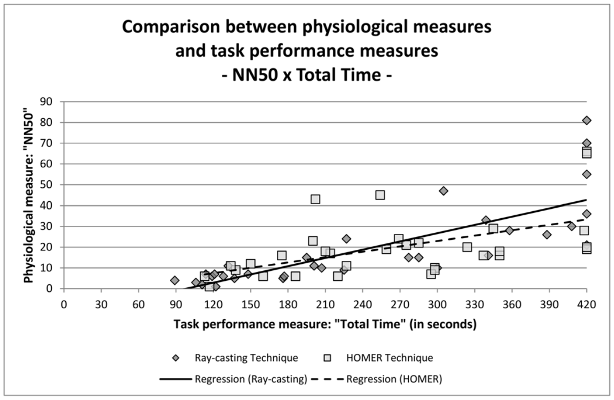

Figure 3. Correlations and trends between "Total time" and "NN50" using the two ITs.

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.

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.

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.

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.

B. Physiological measures as usability metric

In reference to the comparison between ITs, the result presented in the Section C could be repeated in a simple analysis with charts.

This study presents a balance between ITs: ray-casting is the best choice considering total time spent on the task; HOMER technique is the best considering the number of collisions and grabs executed in VE; and both techniques present a similar performance to the accuracy measure.

So, we can say that physiological measures could reveal the quality of a technique for an interaction task, and consequently, the initial evaluation process (preliminary tests) can be based on physiological responses, using their results for decision-making about the IT design. However, subsequent evaluations must consider the use of task performance metrics, as a way of studying in detail the IT during the interaction process.

Regarding the experiment stages, presented in Section D, physiological measures showed good results as a tool for distinguishing stages defined by our protocol. Firstly, comparisons using the Mean HR measure indicate that the experiments need a certain user physical effort to execute interaction tasks, and this physical effort presents no significant variations during the experiments, as shown in Figure 4.

SC measure shows the level of anxiety significantly reduces between the first and second experiences using the ITs. This situation proves that the user, over the interaction process, gets used to the devices and the techniques. Therefore, the SC measure can be used as a metric to indicate the user adaptation to the 3DUI. This trend is shown in Figure 5

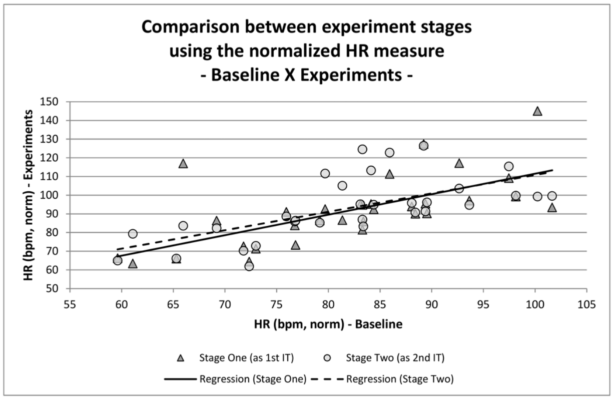

Figure 4. Correlation between baseline and experiment stages of the testing protocol, considering the HR measure.

On the other hand, changes in the user's level of anxiety have not clearly identified between baseline and experiment stages using the SC measure. One reason for this fact can be related to the profile of the evaluated group, composed by new VR users. In an effort to evaluate this feature, we recommend grouping subjects according to their VR experience and the inclusion of a new stage to collect baseline user data without the use of VR and physiological apparatus, in order to verify the device-related discomfort. Other suggestions are evaluations considering the periods of time when experiments were run (early, late, and night shift) or the user's physical or cognitive skills.

During the comparison between performance user groups (Section **F**), NN50 measure once again proved relevant, highlighting as a metric to evaluate user task performance, as already noted in Section **A**. This analysis demonstrates the need for a more detailed study of their behavior in assessing the quality of ITs for 3DUIs.

Regarding the comparison between physiological measures and interaction process tasks (Section **E**), we could not identify which moments required more physical effort from the user, or resulted in significant changes in his level of anxiety. As a result, we also could not found any relationship between physiological changes and application problems or failures reported by users in the post-test questionnaires. Probably, this situation was due to the VE properties, which were not stressful enough to cause irritation during the constant collisions between objects or to cause anxiety for completing the tasks within the time limit.

At last, we suggest a further study that considers VEs capable of providing high levels of anxiety and physical effort to the user. Likewise, we recommend decomposing the interaction process on more detailed parts and identifying studies about variations in physiological responses during frustration or satisfaction events. These future researches may contribute to a better assessment of 3DUIs.

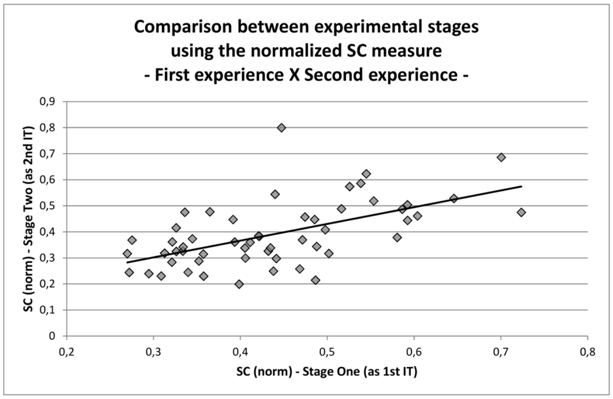

Figure 5. Correlation between experiment stages considering the SC measure.

V. Acknowledgments

Our research is partially funded by the National Institute of Science and Technology in Medicine Assisted by Scientific Computing(MACC) (Grant CNPq 181813/2010-6 and FAPERJ E-26/170.030/2008).

VI. 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. Guyton, A. C., Hall, J. E.: Textbook of Medical Physiology. Saunders, Philadelphia (2005)
11. Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E.: Multivariate Data Analysis. Prentice Hall, New Jersey (2005)
12. Hix, D., Hartson, H.: Developing User Interfaces: ensuring usability through product & process. John Wiley & Sons, New Jersey (1993)

13. 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)

14. 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)

15. 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)

16. 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)

17. 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)

18. 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)

19. 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)

20. 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)

21. 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)

22. Rosson, M., Carroll, J.: Usability Engineering: scenario-based development of human-computer interaction. Morgan Kaufmann, San Francisco (2001)

23. Stein, P. K., Bosner, M. S., Kleiger, R. E., Conger, B.M.: Heart Rate Variability: a measure of cardiac autonomic tone. American Heart Journal 127-5, 1376-1381 (1994)

24. 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)

25. Tullis, T., Albert, W.: Measuring the User Experience: collecting, analyzing, and presenting usability metrics. Morgan Kaufmann, San Francisco (2008)