A critical analysis of human-subject experiments in Virtual Reality and 3D User Interfaces

Carlos Andujar   Pere Brunet

April 7, 2015

Abstract

This paper is about the major peculiarities and difficulties we encounter when trying to validate research results in fields such as virtual reality (VR) and 3D user interfaces (3DUI). We review the steps in the empirical method and discuss a number of challenges when conducting human-subject experiments. These challenges include the number of independent variables to control to get useful findings, the within-subjects or between-subjects dilemma, hard-to-collect data, experimenter effects, ethical issues, and the lack of background in the community for proper statistical analysis and interpretation of the results. We show that experiments involving human-subjects hinder the adoption of traditional experimental principles (comparison, repeatability, reproducibility, justification and explanation) and propose some ideas to improve the reliability of findings in VR and 3DUI disciplines.

1   Introduction

A substantial part of the knowledge in science and engineering fields comes in the form of findings supported by experimental research. In contrast to correlational studies, experimentation allows researchers to make strong claims about causality, and is thus one of the key methodologies that sustain the development of science and engineering. Unlike pure observation, experimentation is, in essence, controlled experience, and as such it involves actions to manipulate, in a controlled manner, some of the variables involved in the phenomena under investigation.

Experimentation has been used for years in both science and engineering fields, although addressing different objects and for different purposes. In science, the goal of experimentation is to *understand* a natural phenomenon, whereas in engineering the goal is *test* a man-made artifact. We can thus distinguish multiple types of experiments (Table 1) according to the subjects under investigation (natural objects, man-made artifacts and humans). Notice that we have included Virtual Reality (VR) experiments in a special category, along with Human-Computer Interaction (HCI), 3D user interfaces (3DUI)[5] and disciplines alike. VR has a dual nature in the sense that it shares goals and methodologies from both science and engineering. We are interested both in understanding some aspects of human-related phenomena (e.g. human perception, human performance) as well as to test VR-related artifacts (e.g. software,

1

Subject	Disciplines
Natural objects (excluding humans)	Natural sciences
Artifacts	Engineering
Humans	Social sciences
Artifacts' use by humans	Human factors, ergonomics, HCI, VR...

Table 1: Types of experiments according to the entities under investigation.

hardware, interaction techniques), although the relative weights of these two goals vary across subfields. This dual nature implies that experimentation in VR has to deal with the inherent complexities of two worlds: that of the technical artifacts being tested, and that of the humans that test these artifacts. As we shall see, the resulting complexity often limits severely the generalizability of findings from VR experiments.

The key role of human-subject experiments in VR and 3DUI fields is evidenced by the large number of papers thoroughly reporting such experiments in major conferences and journals. It is therefore natural to ask ourselves questions such as: are we doing experimentation in VR the right way? Are the findings of published papers on VR supported by solid evidence? If so, are these findings really useful? How should experiments be documented? In this paper we treat some of these questions by discussing the major peculiarities and challenges of human-subject experiments in VR and 3DUI fields. We show that experiments involving human-subjects hinder the adoption of traditional experimental principles (comparison, repeatability, reproducibility) and propose some ideas to improve the reliability of findings in VR and 3DUI disciplines. We argue that there is a lack of background on experimental research in a large part of our community, and that fundamental ideas of the experimental method should be included in the curricula of computer science students like in all other science curricula. The conclusions are partially based on some discussions of the Dagstuhl Seminar 13241 *Virtual Realities*, June 9th-14th 2013, Dagstuhl.

2 Experimentation in related disciplines

VR teams often bring together professionals from multiple disciplines, including formal sciences (e.g. mathematicians, statisticians), social sciences (e.g. psychologists) and engineering (e.g. computer scientists). However, the concept of experimentation, its methodology, and the aspects that distinguish good experiments from bad experiments vary substantially across these disciplines.

From a scientific point of view, experiments should obey the classical principles of comparison, repeatability, reproducibility, justification and explanation [1]. *Comparison* refers to the possibility of knowing similar research already done in the past, to avoid repeating uninteresting experiments and to get insights on worthy ones. *Repeatability* is the property of an experiment that yields consistent results when repeated in the experimenter's own setup at different times and in different places. *Reproducibility* is the possibility for independent researchers to verify the results of a given experiment by repeating it with similar conditions. Achieving reproducibility is much harder than repeatability because it requires researchers to document thoroughly the experimental setup and to expose all variables that might bias the experiment. *Justification* and
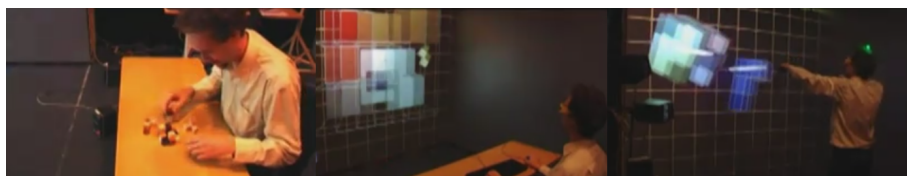
Figure 1: Different interfaces for a puzzle-solving problem.

explanation refer to the possibility of drawing well justified conclusions based on all the collected data.

Experimental research in natural and social sciences most often follow these principles. Psychology experiments, for example, adopt these principles through rigorous protocols. In contrast, experimental methodologies in computer science have not yet reached the level of maturity of other scientific disciplines. Despite its wide-spread use, experimentation is still one of the less well-understood methodologies in computer science, and computer scientists do not seem to agree on the role and relevance of experiments in their field [1]. Whether classic experimentation principles should also apply to computer science experiments is still under discussion [1].

3   Sample problems from VR

Let us introduce a couple of problems that will help us to illustrate the need for running controlled experiments in VR research. These examples will be used through the text to exemplify major difficulties encountered in VR experiments. The first example is the evaluation of presence in VR systems, that is, to which extent users feel and behave as if physically present in a virtual world. Users of immersive systems might forget about the real environment and the virtual environment can become the dominant reality. The evaluation of presence is very important in many VR applications, from phobia therapies to psychological experiments through pain relief for patients with serious injuries. Many presence evaluation studies report surprising findings when analyzing the human behavior when presented a stress situation in a virtual environment [10, 11]. Nowadays, the common practice is to evaluate presence by observing users' behavior and measuring their physiological response (as captured by specialized devices such as heart rate monitors and galvanic skin sensors) in a controlled experiment. The use of unconventional measurement devices might hinder participant recruitment and data collection.

The second example is about the comparison of 3D UIs in terms of their usability. Figure 1 shows three different user interfaces to solve a 3D puzzle [14], using either a physical puzzle, a classic keyboard-and-mouse interface, or a Wii controller. For this task, users must be able to select pieces, manipulate them, and explore the puzzle from different viewpoints. The key problem thus is to determine which UI is better in terms of usability. The only way to know the answer is through a controlled experiment comparing these techniques. Despite some models have been proposed to predict human performance for some tasks (the well known Fitts' law [9] is the most notable example), typical spatial tasks are just too complex to be predicted by such simple models.

4 Major challenges in VR experiments

The typical steps of experimental methods are shown below (the list has been adapted from [12] to refer explicitly to experiments involving human subjects):

1. Formulate a hypothesis and make it testable

2. Design an experiment

3. Get approval by ethics committee

4. Recruit participants

5. Conduct the experiment and collect data

6. Pay participants

7. Analyze the data

8. Accept or refute the hypothesis

9. Explain the results

10. If worthy, communicate your findings

The list above will serve to guide the discussion about some major challenges in human-subject experiments for VR and 3DUI not found in other types of experiments.

4.1 Hypothesis formulation

A first step is to formulate a hypothesis and make it testable. Using the 3D puzzle problem as an example, a general hypothesis might be formulated as follows:

Using the Wii controller will make people more effective when doing manipulation tasks.

A possible testable version of it could be formulated as:

We measured the time it takes for users to solve a particular 3D puzzle, using either Wii or mouse; we hypothesize that users will be faster using the Wii.

Here we find a first major problem: to make the hypothesis testable, we had to choose a particular task which we take as representative (besides fixing some other important variables). Unfortunately, many problems have a task space so large and heterogeneous that it can be really difficult to find a small set of representative tasks. Here we wrote *task* for the 3D puzzle example, but we could have written 3D model, image, movie, stimulus and whatever other elements are fixed to make the hypothesis testable. As a result, a large body of findings in VR and 3DUI only apply to very specific conditions, and it is unclear if the results can be generalized to other contexts. As an example, many interaction techniques for 3D selection have been tested with a single

scene attempting to mimic a *typical* real-world scene, but no single 3D scene is really representative of the huge variety of scenarios we encounter in real 3D applications. Some other selection techniques have been tested with synthetic, well-controlled scenes (e.g. varying object size, position and density) and thus provide more insights on which factors influence user performance, but even in these cases many other variables need to be fixed arbitrarily to keep the experiment complexity within reasonable limits.

4.2   Experiment design

A key aspect of experiment design is to decide which variables will be varied systematically (independent variables or factors) and which variables will be measured (dependent variables). Consider for example the 3D puzzle task. The completion time for a 3D puzzle (dependent variable) might be influenced by a number of variables including the interaction device (Wii, mouse), the viewing conditions (stereoscopic, mono), the mapping between user actions and application tasks, the size of the pieces, and the quality of the graphics [4]. Complex VR setups are incredibly hard to specify, and it is hard to define a set of 'normal conditions' for the ceteris paribus assumption. In an ideal world, VR systems would have zero end-to-end latency, ghosting-free stereo separation, no accommodation-convergence mismatch, accurate, encumbrance-free full-body tracking, natural locomotion and many other features beyond the state-of-the-art. Current VR systems are thus suboptimal to a certain degree, and the best hardware configuration is unclear, making the prediction of user behavior even more difficult.

These facts would suggest trying to control a large number of variables, but as the number of independent variables (and the corresponding levels at which they occur) increases, the number of conditions for a complete factorial design increases rapidly. Therefore, despite the large number of variables that might bias the outcome of a VR experiment, considering more than five or six independent variables is often unpractical, prompting investigators to seek alternative designs that require fewer conditions [6].

Experiment reproducibility is certainly a hard task in VR. The use of non-conventional, rapidly evolving displays, tracking, haptics, props and other specialized devices hinders experiment reproducibility. In some extreme cases, the VR hardware being tested is so sophisticated that, in practice, the experiment cannot be reproduced by independent researchers. Furthermore, VR hardware platforms are hardly ever selected according to an analysis of requirements; instead, hardware selection is influenced by already purchased hardware, prior expertise, available resources and compatibility factors, leaving little opportunities to run the experiment on multiple hardware configurations [16]. From the point of view of software development, VR systems are often 'complex, chaotic and difficult' [16], and VR developer communities are separated by tools that hinder code reuse and shared implementations.

Another important aspect of experiment design concerns the assignment of independent variables. Independent variables can vary in two ways: within-subjects (each participant sees all levels) and between-subjects (each participant sees only one level). The decision on which design is better to confirm or refute a given hypothesis is often controversial. Within-subject designs are more time consuming for the participants, and require the experimenter to coun-

terbalance for carry-over effects (the order of presentation might influence the participant behavior) such as fatigue and learning effects. Between-subject designs, on the other hand, are not free from limitations, as more participants need to be recruited and we are likely to loose statistical power during Null Hypothesis Significance Testing (NHST), and thus have less chances to proof our hypothesis. Within-subject designs are usually preferred in HCI experiments because having fewer participants overall simplify recruiting, scheduling, briefing, demonstrating and practicing [8], and because it is easier to detect differences across levels because each subject's behavior under one condition is compared to that subject's behavior under the other condition.

4.3   Ethics

Many usability guides address in depth all the ethical issues related with user studies [2, 7]. Most organizations require experimenters to get the approval by an ethics committee before running the experiment. After the approval, it is often a hard task to recruit participants and get their informed consent, in particular when participants should be chosen from a specific target user group (such as physicians for VR medical application).

Experiments involving immersive VR systems often need a detailed informed consent. Researchers should never deceive participants about aspects that would affect their willingness to participate, such as risks (VR users might bump into walls, trip over cables), discomfort (many 3D interaction techniques for spatial tasks are physically demanding) and unpleasant experiences (some VR systems cause motion sickness). Frustration handling is also important when measuring user performance. In case of failure to complete a task, experimenters should make it clear that the responsible is the technology. Usability tests should not be perceived as tests of the participant's abilities [7].

4.4   Running the experiment and collecting data

Wingrave and LaViola summarize multiple VR issues that can impact scheduled appointments and impact experiment repeatability [16]. VR systems are still not robust; some specialized VR hardware such as data gloves are fragile and might stop working, some projectors might go out of alignment, cables tangle, and cluster rendering is less robust than single-machine rendering.

Data collection is a further issue in typical VR applications. At the lowest level, a classic WIMP interface requires tracking a single pointing device whose position can be described by a stream of (x,y) values. In contrast, VR interfaces often requires 6-DoF tracking of multiple body parts. Although some dependent variables can be measured easily (e.g. task completion times, error counts, scores), some VR experiments need to measure hard-to-collect data such as user heart rate variability and galvanic skin response.

Pilot experiments are often required prior to running the main VR experiment because many issues cannot be identified until participants experience the VR system. Besides software and hardware affordances, the multisensory nature of VR experiences hinders the prediction of human factors such as fatigue, stress, and nausea, that should be detected in pilot experiments [16].

Experimenters should be careful to avoid manipulating the experiment. In addition to the well known placebo effect, there are other experimenter issues

that often hinder data collection. The Hawthorne effect occurs when increased attention from superiors or colleagues increases user performance. The performance of a participant might change if somebody else, e.g. the previous participant, is observing. Observer-expectancy effect occurs when the researcher unconsciously manipulates the experiment, using e.g. body language. Experiments should be double-blind, but researchers in non-life critical fields often disregard these issues.

4.5   Data analysis and explanation

Proper data analysis is absolutely required to accept or refute the hypothesis and to provide statistical evidence of the findings. The usual statistical procedure is to run Null-Hypothesis Significance Testing (NHST) to compute a p-value that expresses the conditional probability of getting an outcome as extreme as or even more extreme than what we observed, assuming the null hypothesis holds. The p-value p is compared against a previously-defined significance level α (often 0.05). If $p < \alpha$, we reject the null hypothesis H_0 and embrace H_1 (thus we get a result). Otherwise, we fail to reject H_0 and conclude that there is no evidence in the data to confirm H_1. Since NHST is well-known to be biased by sample size (i.e. with a sufficiently large sample we can detect statistically significant differences even when such differences are really tiny), a measure of the effect size should be included in the experiment report.

Unfortunately, a part of the VR community seems to lack enough background on experimental design and statistical analysis to conduct the user studies required to evaluate their own research. This is evidenced by the large number of submitted and even published papers with serious evaluation errors related with the statistical analysis of the results (see next section). Unlike other disciplines like medicine, where journal editors often send manuscripts to expert statistical reviewers to check the statistical analysis, this is no common practice in major VR journals.

Experimenters are expected to provide a plausible explanation of the observed data and to draw well justified conclusions based on the experiment findings. The explanation should rely on widely accepted guidelines and theories. Due to the multidisciplinary nature of VR experiments, plausible explanations might require acquaintance with theories from a variety of fields (e.g. perceptual, cognitive, social and cultural psychology, control theory, ergonomics, biomechanics, proxemics...) beyond the experimenter expertise. These experiments would greatly benefit from feedback provided by other members of the VR community, but papers with poorly explained outcomes have little options to be accepted for publication and this reduces the opportunities to get such a feedback.

4.6   Reporting the experiment

Documenting an experiment is a substantial task within experimental research, and one of the issues most frequently criticized by reviewers. A typical report would include detailed information about the hypotheses being tested, the experiment design, the procedure, apparatus, participants, and statistical analysis.

The way data analyses are reported in VR should not differ much from other disciplines, but in practice the VR community does not seem to demand

the same level of rigor found in other disciplines. The American Psychological Association provides well-established protocols and guidelines for reporting experiments on social and behavioral sciences [3]. Unfortunately, no such protocol exists for VR papers and alike; although some authors do follow to some extent some of these guidelines, a large body of VR literature still reports data analysis in a non-standard, often incomplete manner.

Besides data analysis, the description of the experiment setup itself is another widely criticized aspect. We have already argued about the extraordinary complexity of VR experiments, which have to deal with (complex, immature, imperfect, fragile, heterogeneous) technical artifacts, as well as with many aspects of human phenomena (perception, cognition, human factors...). As a consequence, the number and variety of variables potentially influencing the outcome of a VR experiment is often beyond reasonable complexity limits. Experimenters are thus faced with a hard dilemma: if too few details about the experiment setup are reported, the experiment will be far from being reproducible; if too many details are included, this might result in an overly long paper, a large portion of it would be of interest only for those researchers aiming to reproduce the experiment. In such cases, experiment reproducibility must be seen as a goal rather than a strict requirement to be enforced during reviewing.

Concerning the information about the participants, it has been shown that humans vary widely in their ability to perform tasks in VR environments [16]. Besides age, gender and education, demographics and cognitive aptitudes can also account for variance [15]. For example, a new interaction technique might prove to outperform competing techniques on young gamers but not on older adults. Unfortunately, there is no de-facto standard for reporting such information about participants, making results difficult to generalize.

5   Common errors in submitted papers

Precise criteria for deciding which VR papers should include results from a human-subject experiment and which should not is out of the scope of this paper, although obviously we can find examples of papers where such a validation does not apply or it is unnecessary (e.g. rendering papers with strongly convincing visual results), as well as examples where validation through a controlled experiment should be enforced (most papers on new interaction techniques fall into this category). We now summarize some common experimentation-related errors found in submitted (and sometimes published) VR papers.

A first error is the lack of inferential analysis; some manuscripts include graphs and descriptive statistics but draw conclusions directly from the raw comparison of summary statistics (e.g. simple mean comparison). Although in some fields hypothesis testing is culturally unnecessary provided that plots convincingly show enough power [13], most experiments have outcomes with subtle effect sizes and do require proper statistical analysis.

A second common error is improper hypothesis formulation. Before describing an experiment, it is important to state clearly its purpose in terms of one or more hypotheses. Continuing with the 3D puzzle example, consider the following hypothesis:

We measure the time it takes for subjects to solve a particular 3D puzzle

under Wii and mouse conditions. We hypothesize that users will be slower with the Wii controller due to the lack of a physical support for free space interaction.

What is wrong with the hypothesis above is that it includes both a testable hypothesis (Wii slower) and one of the many plausible explanations for it; NHST could confirm the first part, but not the second (unless we manage to vary systematically the physical support condition).

Now consider this hypothesis:

The stereo condition (mono/stereo) has no influence on user performance.

The problem here is that NHST can be used to detect significant differences between conditions (i.e. confirm the alternative hypothesis) but not to prove the truth of the null hypothesis. We could, however, test if the difference between means is below some effect size threshold [13].

Another common error is to choose the wrong statistical test to analyze the data. It is not uncommon to find manuscripts analyzing within-subjects data with a classic (independent-samples) t-test or ANOVA instead of using the repeated-measures version of these tests. Even when the right test is chosen, the ANOVA table is often miss-constructed. For example, consider a between-subjects version of the 3D puzzle example. With n subjects and m trials per subject we get $n \times m$ measures. These m trials, however, need to be aggregated per participant before running the ANOVA. Running the ANOVA with $n \times m$ rows violates the assumption of independence and ruins the analysis. Experimenters should always report the parameters of the sampling distribution so that reviewers could double-check for these errors. For example, when running an ANOVA test, the two parameters of the F-distribution must be reported [12].

Many manuscripts also forget to check the assumptions underlying the chosen statistical test, or at least forget to report it. A last and more serious mistake concerns the interpretation of p-values. Failure to reject the null hypothesis does not imply the null hypothesis is true, although this fundamental mistake can be found in many manuscripts containing no power analysis at all. For example, the following outcome report is misleading:

We found no significant effect of interaction technique on completion time. Therefore we conclude both techniques perform equally well.

6 Conclusions

In the last decades some computer science disciplines such as VR are experiencing a shift of focus from *implementing the technology* to *using the technology*, and validation through experimentation is becoming critical. In this paper we have discussed some major issues of such validation experiments: lack of background on experimentation, psychology and psychophysics, time-consuming and resource-consuming nature of human-subject experiments, and the difficulties to fulfill all requirements (double-blind experiments, informed consent, representative users, representative datasets/models/tasks, reproducibility).

The user performance during a typical computer-related task depends on

a number of domain-specific factors as well as hardware-related factors. Considering all these factors simultaneously as independent variables in controlled experiments is clearly not practical. This fact limits the validity of the findings reported in the VR literature to a specific domain and a particular setup. The lack of protocols and de-facto standard datasets for testing purposes (more common in other scientific communities) along with the plethora of VR hardware configurations makes it difficult to make fair comparisons. Furthermore, many techniques are still proposed and evaluated in isolation, whereas in the real world user tasks are mixed with other tasks. These are issues that must still be addressed.

Although guaranteeing reproducibility of VR experiments is often unfeasible or impractical, researchers should take some steps to facilitate fair comparisons. One way of doing this is by sharing code and data so that other researchers can verify the results by repeating the experiment with the same or similar conditions. In this sense the concepts of reproducible research and executable papers (which are being promoted by some publishers) is quite promising. Reproducible research refers to the idea that academic papers should include not only text, tables and figures, but also a complete environment including code, models and data, that can be used to reproduce the results and to create new work based on the research. Nevertheless, non-standard VR hardware configurations still limit reproducibility under identical conditions. Experimenters should also be careful to avoid manipulating experiments, although sophisticated VR equipment make double-blind experiments hard to implement.

Researchers, and in particular journal editors and reviewers, should enforce correct data analysis and experiment reporting. Papers with serious experimental flaws should not be accepted for publication, as they contain unreliable findings and can mislead novice researchers who might follow mistaken experimental procedures.

Finally, our recommendation for VR professionals, researchers and students is to ensure a strong education on the principles of sound experimentation with human subjects. In research projects, at least one of the research team members should be an expert on experiment design and statistical data analysis.

References

[1] Carlos Andujar, Viola Schiaffonati, Fabio A. Schreiber, Letizia Tanca, Matti Tedre, Kees van Hee, and Jan van Leeuwen. The role and relevance of experimentation in informatics. In *Workshop on the Role and Relevance of*

Experimentation in Informatics, 8th European Computer Science Summit (ECSS 2012) of Informatics Europe, 2012.

[2] American Psychological Association. Ethical principles of psychologist and code of conduct, 2010.

[3] American Psychological Association. *Publication manual of the American psychological association, 6th edition.* American Psychological Association, 2013.

[4] Doug A. Bowman, Joseph L. Gabbard, and Deborah Hix. A survey of usability evaluation in virtual environments: classification and comparison of methods. *Presence: Teleoper. Virtual Environ.*, 11(4):404–424, 2002.

[5] Doug A. Bowman, Ernest Kruijff, Joseph J. LaViola, and Ivan Poupyrev. *3D User Interfaces: Theory and Practice.* Addison Wesley, 2004.

[6] Linda M. Collins, John J. Dziak, and Runze Li. Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs. *Psychological methods*, 14(3):202, 2009.

[7] European Telecommunications Standards Institute (ETSI). Usability evaluation for the design of telecommunication systems, services and terminals. ETSI Guide EG 201472. Sophia Antipolis, 2000.

[8] I. Scott MacKenzie. *Human-computer interaction: An empirical research perspective.* Morgan Kaufmann, 2013.

[9] I. Scott MacKenzie, Tatu Kauppinen, and Miika Silfverberg. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '01, pages 9–16, New York, NY, USA, 2001. ACM.

[10] Michael Meehan, Brent Insko, Mary Whitton, and Frederick P. Brooks. Physiological measures of presence in virtual environments. In *Proceedings of 4th International Workshop on Presence*, pages 21–23, 2001.

[11] Mel Slater, Angus Antley, Adam Davison, David Swapp, Christoph Guger, Chris Barker, Nancy Pistrang, and Maria V. Sanchez-Vives. A virtual reprise of the Stanley Milgram obedience experiments. *PLoS ONE*, 1(1), 2006.

[12] J. Edward Swan, Stephen R. Ellis, and Bernard D. Adelstein. Conducting human-subject experiments with virtual and augmented reality. In *Virtual Reality Conference, 2006.* IEEE, 2006.

[13] J. Edward Swan and Joseph L. Gabbard. Quantitative and qualitative methods for human-subject experiments in augmented reality. In *Tutorial presented at the International Symposium on Mixed and Augmented Reality (ISMAR 2012).* IEEE, 2012.

[14] Tuukka M. Takala, Roberto Pugliese, Paivi Rauhamaa, and Tapio Takala. Reality-based user interface system (RUIS). In *2011 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 141–142. IEEE, 2011.

[15] Chadwick A. Wingrave and Doug A. Bowman. Baseline factors for ray-casting selection. In *Proceedings of Virtual Reality International*, 2005.

[16] Chadwick A. Wingrave and Joseph J. LaViola. Reflecting on the design and implementation issues of virtual environments. *Presence: Teleoperators and Virtual Environments*, 19(2):179–195, 2010.