# User Evaluation during Development

## Lessons Learned from Ten Years of Studies of Virtual Environments[1]

Mary C. Whitton, Effective Virtual Environments Research Group
University of North Carolina at Chapel Hill

In 1997 Fred Brooks came back from a sabbatical spent in London (UK) with an idea for an extension of a user study performed at University College London in Mel Slater's laboratory (Slater, Usoh, & Steed, 1995). In 1998 we performed that study to evaluate how locomotion technique influences sense of presence (Usoh, 1999). That work was the first in a thread of research studies examining how a variety of technologies influence the effectiveness of virtual environments.

This paper reports lessons the Effective Virtual Environments (EVE) research team learned while doing a dozen or more studies and lessons learned by another, cross-disciplinary team in the Distributed nanoManipulator project. In that project we designed, implemented, and evaluated a tool for distributed scientific collaboration (D. H. Sonnenwald, R. Berquist, K. Maglaughlin, E. Kupstas Soo, and M. Whitton, 2001), (Hudson, Helser, Sonnenwald, & Whitton, 2003), and (D. H. Sonnenwald, K. Maglaughlin, M. Whitton, 2004). The lessons are presented as freely of the context of a specific study as possible, but examples from particular studies and references to published works are included. The lessons are not intended to exhaustively cover all the issues that arise in designing and executing a user study. A useful primer in experimental design is *Doing Psychology Experiments* (Martin, 2007).

### 1. Know the question you're trying to answer

Too often people perform an evaluation study without having first determined what question they are trying to answer and what hypothesis they will test to answer the question. In the extreme, this can result in nothing being learned and wasted effort.

Different questions and hypotheses are appropriate at different stages of project development (Gabbard, Hix, & Swan, 1999). Are you doing an early stage requirements study? A design review before implementation starts? A functionality and usability review of small to medium-sized work products? A functionality and usability evaluation of an integrated system? Or are you comparing your method to other established methods of doing the same thing either to establish the validity of your method or to demonstrate the superiority of your method? Additional references: (Hix et al., 1999) and (Helms, Arthur, Hix, & Hartson, 2006).

One way to focus yourself on clearly defining your question and hypotheses is, before you go any further, to write the abstract of your paper/report of this study as if you had already completed the study and everything went perfectly.

---

[1] While my presentation "User Evaluation During Development" is organized by the types of evaluation done in the process of developing a product or technique, the lessons in these notes are organized around and focus on lessons learned in designing, planning, and executing user studies..

## 2. Designing your Experiment

**Reuse experimental procedures and measures.** Re-use, or minimally modify, experimental methods, measures, and analysis protocols from published works. The methods have already been vetted by publication reviewers, and it makes it easier to compare work across studies.

### 2.1. Basic Design

#### 2.1.1. How many independent variables and conditions?

It is temping to study everything at once, but adding conditions has an impact on the statistics you run, on how many participants you must have, on how long a participant session is, and on how long it will take you to complete the evaluation project.

**Logistics--how long does it take?** When possible, strive for within-subjects designs that expose all participants to all conditions. In our locomotion study reported in 2005 (M. C. Whitton et al., 2005) participants were able to rate and rank all five virtual locomotion interface techniques because each had experienced all five. In a subsequent study, the length of sessions dictated that each participant experience only one of the five conditions. That limited our ability to use participant comments to make sharp distinctions among the conditions.

The formative evaluation of the Distributed nanoManipulator (see course slides, Sections 1 and 4) involved participants in only two of the four possible conditions in a full 2 X 2 study design (D. Sonnenwald, Whitton, & Maglaughlin, 2003). Had we had participants who did both labs face-to-face and both labs distributed, we could have eliminated any difference in difficulty of the two laboratory tasks as a confounding factor. Did it have an influence on the difference of scores between the first and second sessions? We're unable to tell from the data we have. However, including the other two conditions in our work would have required twice as many pairs of participants and another 6-8 months.

| | Face-to-face first | Distributed first |
|---|---|---|
| Face-to-face second | | X |
| Distributed second | X | |

The evaluation of the distributed nanoManipulator used only two of the four possible condition pairs.

**Between-Subjects or Within-Subjects (repeated measures) or Mixed?** Many considerations go into the decision of whether to run a between-subjects or a within-subjects design. Factors to look for include learning effects—doing better on later trials because you've learned something; order effects—it is easier (or harder) to do a task after you've completed some other condition. The statistics become somewhat more complex with mixed designs: some variables are between subjects, and some are within. An example of a mixed design is our Locomotion study #5 (Loco5): the locomotion condition was between-subjects (due to the length of time the study took, all participants couldn't do all conditions) and was within-

subjects for the task performance scores of exposure to gunfire and counts of jets and explosions.

### 2.1.2. Participants, Multiple sessions, Compensation

The traditional participant pool of undergraduate Psychology students may or may not be available to you. Our computer science students have had access to them when taking the Experimental Design course in the Psychology Department and, occasionally, when their doctoral committee has included a professor from Psychology. Don't assume you can use this pool. Also, be aware that tapping this pool early in the semester will bring you a different type of participant than later on in the semester. Early in the semester you'll attract participants who are organized and motivated to complete their requirement: later the population may be dominated by less-motivated students.

**Professionals as Study Participants.** With funding from NIH we performed a multi-year, multi-faceted project to develop and evaluate of an instrument to enable scientific collaboration over the internet. The product is called the Distributed nanoManipulator (Dist nM). The question was whether "good" scientific collaboration could be done when the collaborators were not located together in the same laboratory.

As conceived, study participants were to have been the system's target users—graduate research assistants, post-docs, and working scientists. We quickly realized we were unlikely to find forty of that population willing to participate in a study requiring eight hours over three different days. Our decision to use undergraduate students broadened the participant pool, but constrained the sophistication of the science lab tasks (D. Sonnenwald et al., 2003).

**Active military as study participants.** Using military personnel as study participants may require review by the military human-subjects protection organization. This includes ROTC cadets as they are considered active duty military personnel.

**It is hard to get people to come back.** Require only one session with each participant if at all possible. It is often difficult to get volunteer or minimally-compensated ($7-$10/hour) participants to return to the lab for the multiple sessions that studying, for instance, training retention requires. Expect to offer larger incentives for multi-session studies. To encourage participants to complete the study, you can withhold most of the payment until after the final session.

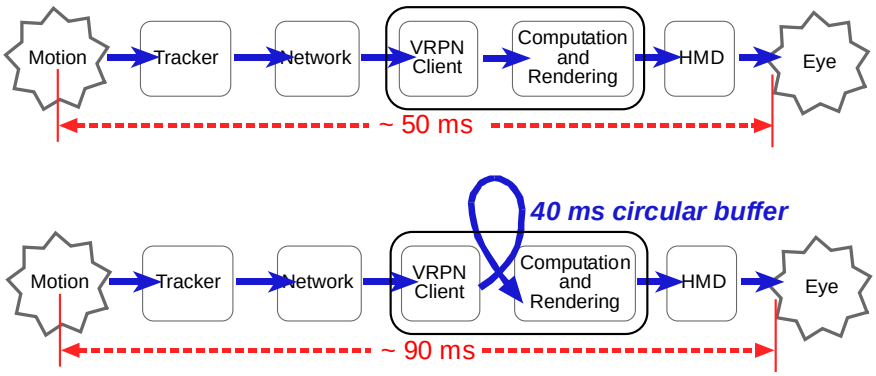
## 2.2. Conditions: Choose Wisely

### 2.2.1. Independent Variables

**Pragmatism 1.** Select the levels of the independent variable carefully, balancing the number of conditions and the number of research questions with the reality of study design complexity and number of participants required. For reasons of expediency, a study on the value of passive haptics for training a maze task (Insko, 2001) did not include a condition exposing participants simultaneously to passive haptics and synthetic audio and visual cues indicating collisions of the user's hands with the maze. Had he included this additional condition, he could have examined the questions of whether using all cues together would result in even better real-maze performance than with passive haptics alone, and whether

training with the audio tones, clearly absent in the real world, would, in fact, mistrain and lead to poorer performance. This is the perennial "training wheels" question for all simulation-based training.

**Pragmatism 2.** In 2002 we ran a user study while demonstrating our virtual environment system to guests in the Emerging Technologies venue at SIGGRAPH. Our independent variable was the amount of end-to-end latency in the system (M. Meehan, S. Razzaque, M. Whitton, F. Brooks, 2002). Pragmatism helped us choose both our high- and low-latency values. Although we could achieve 40 ms latency with our best hardware, we chose 50 ms as the low-latency condition in case we had an equipment failure and had to continue with less capable hardware. A goal for the exhibition was that every participant have a very good VE experience, so we selected a high latency value, 90 ms, that is 10% less than 100 ms, a generally accepted upper bound for interactive systems.




Block Diagram of the end-to-end latency in the system demonstrated at SIGGRAPH 2002.

### 2.2.3. Dependent Variables

**Compare to a Gold Standard.** It is helpful to include an experimental condition that compares the component or technique being evaluated to a similar, familiar component. The wide-FOV HMD used in (Arthur, 2000) is radically different from the commercial Virtual Research V8: It has six liquid crystal display panels per eye, it weighs twice as much as the V8, and brightness and contrast vary across the display panels. The data gathered for participants using the V8, in a condition where their natural eyesight was field-of-view restricted, gave us confidence that the data we collected for wide-FOV HMD users were reasonable.

**Baseline and delta-from-baseline.** A series of studies led by Meehan and Insko (M. Meehan, B. Insko, M.C. Whitton, and F.P. Brooks, 2002) , (M. Meehan, S. Razzaque, M. Whitton, F. Brooks, 2002), and (Meehan, Razzaque, Insko, Whitton, & Brooks, 2005), used the physiological measures of rise in heart rate, skin temperature, and skin conductivity (sweat) as dependent variables. In this series of studies we were looking for stress responses when users experienced the virtual-cliff environment that we call the PIT. In each study, baseline values for the physiological measures were gathered in the normally-floored,

low-stress ante-room of the environment and gathered again in the high-stress PIT room. Heart rate was the most successful for us.

Because baseline heart-rate varies widely between people, the standard deviation of the sampled values was large, making statistical significance difficult to achieve. The alternative was to make the dependent variable the *difference in heart rate between the two rooms.* This delta-heart-rate variable varies much less across the participant population.

**Dependent variable's ability to discriminate among conditions.** Always pilot test your study to ensure that the measured values of the dependent variables will discriminate among conditions. Insufficient pilot testing before a study turned a summer's work into an overly elaborate pilot test for a later study. The pilot test showed our task was too easy: Participants in every condition performed well; they were able to move through the virtual environment and successfully hide behind barriers. In the next iteration of the study (described in my course slides for Section 5—Summative Evaluation) we added a distracter cognitive task, reasoning that if people were counting explosions and jets flying overhead, there would be fewer mental resources available for moving through the environment and hiding behind pillars to avoid being shot at. This strategy was generally successful, but if we were doing it again we would make the task even more difficult. If a relevant taxonomy of tasks is available, e.g., Bloom's taxonomy of cognitive tasks, consult it when choosing tasks.

**Eliminate Potential Confounders.** Eliminate, or at least mitigate, confounders—conditions of the experiment that are uncontrolled and may influence the outcomes—by carefully considering all aspects of your stimuli and evaluation set-up. The virtual environment for the Zimmons' visual search task was very simple—a room with a table and a picture frame—and was easy to develop. However, development of the *stimulus models* and images was complex and time-consuming because lighting, brightness, and colors had to be matched in order that inadvertent differences in them would not confound the results (Zimmons, 2004).

**Study designs usually demand compromises.** In the Distributed nanoManipulator evaluation (Sonnenwald 2003), the metrics for the quality of science would, ideally, have been long-term measures such as number and quality of papers and grants that result from the work. The study required short-term measures that were plausibly related to scientific quality: participants each wrote a report of their laboratory work and results. The lab reports were graded using a rubric developed iteratively by three graders scoring a subset of the lab reports. One grader, using the rubric, completed the grading.

### 2.3. Experimental Tasks

#### 2.3.1. Ecological Validity

Ecological validity is how well your study conditions mimic the environment in which the item under test will really be used. Study designers should consult with their target users and collaborators and engage them in defining experimental tasks. This will insure both that the task is a reasonable one and that it is as much like a real use situation as possible. Ecological validity is easy to achieve for an algorithm that will be used in a hospital image analysis center; it is much more difficult to achieve in a virtual environment, particularly for tasks that would normally be done outdoors.

**Training Transfer Studies.** Training-transfer studies are particularly difficult because they require a "real" condition. The laboratory environment imposes space and other limitations on the "real" condition, resulting in low ecological validity. There is not yet enough literature to enable us to define how generalizable laboratory training-transfer study results are to real-world training.

**Evaluate using real collaborator-supplied data and use cases.** First, this makes the evaluation fair, and secondly, subject to the limits of ecological validity in a lab setting, shields you from the accusation of working on a "toy problem." A third and major advantage is that if you are using your collaborator's data, they will be much more engaged in the process.

### 2.4. Analysis

Designing the analysis procedures for studies is always difficult because many of us doing studies are not experts in sophisticated statistical methods. Available statistics courses tend to be totally applied or totally mathematically-oriented, neither of which is satisfying for a technologist who needs to know how to use the tests, but is also capable of understanding the mathematics! Field and Hole (Field & Hole, 2003) is a good introduction that blends material on the application of statistics with some of the mathematics.

**What is the proper statistical test?** Do not expect all statistical analyses to be as simple as t-tests and ANOVAs. The experimental design may dictate more sophisticated techniques than those learned in a first statistics or experimental design course. In a recent locomotion study, the complexity increased unexpectedly when we found that the exposure-to-gunfire data did not meet the normality criteria required for use of parametric techniques. Beginning statistics courses don't teach you about setting up tests for mixed-model, non-parametric data.

**What can go wrong will.** We unexpectedly added complexity to the data analysis of the locomotion study reported on (Whitton 2005) because the path segments the participants walked were not all the same length. The consequence was that the data from the different segments could not be naively combined in a repeated-measures analysis.

**Use on-campus statistics consulting services.** We are fortunate to have an on-campus consulting service whose mission is to assist in developing the analysis component of studies. Make use of such a service if it is available. Expert advice while you are developing the study helps ensure that you will be able to answer your research questions with the hypotheses, study design, and analysis you have planned.

**Do you think you need a new measurement tool?** Avoid developing a completely new measurement tool if you can. Developing new measurement tools and validating them is complex. Seek outside expertise. A center offering consulting on statistics will often also help with measurement tool development. We did develop a new questionnaire as part of the Distributed nanoManipulator project and used it in our summative study (D. H. Sonnenwald, Maglaughlin, & Whitton, 2001) and (D. Sonnenwald et al., 2003).

Colleagues in psychology departments frequently know of already existing standard tests for things such as baseline 3D spatial reasoning ability that you may be interested in as baseline characteristics of your participant population.

## 3. Ethics Committees—protecting human subjects

The protection of the health and rights of human subjects participating in controlled studies is the job of what is called the Institutional Review Board in the United States and the Ethics Committee in the European Union. Persons proposing to run studies apply to the IRB for approval to do the work. In the United States, any research done with human subjects is not publishable unless the study has IRB approval.

**Design the study before you write the application.** Preparing an application for Ethics Committee approval of a study has a reputation for being painful. My observation is that this is because people begin filling out the form *before* they have designed their study. People use the application form, with its systematic series of questions about study purpose, hypotheses, methods, materials, consent forms, etc. as their experiment design template. While that is one strategy for designing a study, it gives the ethics committee a bad reputation: Just because designing a study is hard work, there is no reason to blame the ethics board for it. Design your study; then write the application.

**Get to know your Ethics Committee.** We have developed and maintain a good working relationship with the IRB at the University of North Carolina. The UNC IRB has, over the years, become familiar with our work and the precautions we take to ensure participant safety. This good relationship worked to our advantage when we sought permission to run the Latency study at SIGGRAPH 2002 (M. Meehan, S. Razzaque, M. Whitton, F. Brooks, 2002). Although the study locale was quite unusual, getting IRB approval for the exhibition-based study was straightforward.

## 4. Planning and Piloting

### 4.1. It will be harder than you think

Don't underestimate the space, equipment, programming, modeling, logistical, time, and management resources required to design, implement, and execute studies, particularly if you are striving for ecological validity. Just the paper design of the four virtual scenes for the Loco5 study took well over 80 hours. The layouts were constrained by analysis requirements, available building blocks for passive haptics, cable management issues, the need to be able to switch from one physical (passive haptics) environment to another in three to four minutes, and the need for them to be of comparable difficulty.

Large, multifaceted studies are resource-intensive. For the Distributed nanoManipulator study, two rooms, each with two computers, a force-feedback Phantom device, four cameras, two video recorders, two audio recorders, and wireless telephones, were tied up for eight months. Seven people shared the study execution and observation duties; on the order of 400 person hours to simply gather the data. The forty participants were each paid $100. Including system development and the study, an average of four graduate students

worked on the project each semester for four years and three to five faculty members were involved over the life of the project (D. Sonnenwald et al., 2003).

**Supporting Equipment:** Evaluating early stage prototype devices may require additional specialized equipment. Arthur's studies (Arthur, 2000) comparing performance across head-mounted displays with different fields-of-view required not only access to a DARPA-funded, Kaiser Electro-Optics-developed, experimental wide-field-of-view HMD, but also required a large graphics system –at the time an SGI Reality Monster with 12 separate graphics pipelines and video outputs—to drive the 12 display tiles in the HMD. The department's (then prototype) HiBall wide-area tracker (3rdTech, 2006) enabled Arthur to design the maze-walking task so that participants really walked in the lab.

**Make session control easy on the experimenter.** For computer-based studies, devise a control application that enables relatively naïve computer users (including the project PIs) to oversee and conduct study sessions. This makes it easier on the study lead and allows sharing the task of running subjects.

### 4.2. Try it all out first: Will you be able to answer your question?

**Debug the process; test everything: meet and greet through payment and goodbye.** Always run pilot studies. Besides bringing procedural problems to light, running a pilot study all the way from greeting participants through data analysis enables a statistical power analysis to determine if the experiment is likely to differentiate among the conditions without an untenable number of subjects.

**Can the users do the task?** Pilot the task. Participants must be able to learn the interfaces and complete the task. Some participants were never able to successfully use the neural-network-based walking-in-place (WIP) interface (Usoh 1999) and the accelerometer based WIP (Whitton, 2005). Feasel's LL-CM WIP (Feasel, Whitton, & Wendt, 2008) works sufficiently well (i.e., no complaints in the post-session interviews) that we feel that with it, for the first time, we can fairly compare WIP to other means of virtual locomotion.

**Train to competence in all conditions.** Train in a setting with complexity comparable to the experimental scenario. Our training scene for the Joystick (JS) and Walking-in-Place (WIP) conditions in Locomotion study #5 was, unfortunately, less cluttered than the test scenes; it did not force the participants to maneuver through spaces as tight as those in the test scenes. Both JS and WIP users showed improved performance on the exposure measure (lower is better) over the first 6 of the 12 trials; for the final 6 trials, their performance approximated that of the other (real-walking) conditions.

**Time the task.** You want to learn during piloting how long the experimental sessions will be and, from that, judge if participant fatigue is going to be an issue.

## 5. Execution

**Go slowly and carefully. Don't waste your efforts.** Small errors can render months of work worthless. Simple errors include lost video tapes, bad batteries in an audio recorder, and paper notes mistakenly thrown out. Be careful and take your time.

**Record observations and/or record sessions.** Log experimenter observations and reports from sensors in the system—e.g., keystrokes, tracker data. The logs can help explain outlier data points and support the exclusion of those data from the statistical analysis. In a passive haptics study (Insko, 2001), observations caught a consistent, but unexpected, wrong-turn behavior when, after training in the virtual environment, blindfolded subjects walked the real maze.

The observation that participants consistently tipped their heads to locate sound sources in 3D helped explain why our (unpublished) results comparing sound localization performance in 2D (Microsoft DirectSound) and 3D (AuSIM Gold Series) sound-generation conditions differed from those reported in the literature. We found no significant performance differences attributable to sound rendering method. Our participants could freely walk about and move their heads. In previous studies, participants, who were seated with their heads held stationary, performed better on the localization task with the stimuli presented in 3D sound (Wenzel, Wightman, & Foster, 1988).

## 6. Reporting

### 6.1. Assumptions

Devising an evaluation study often requires assumptions. Burns (Burns, Razzaque, Whitton, & Brooks, 2007) used a single up-staircase method in a psychophysics study to determine the position-discrepancy detection threshold between the location a person's real hand (felt) and the location of the avatar hand. In a later study Burns used multiple, interleaved, adaptive staircases to determine the velocity-discrepancy detection threshold. Because the outputs of the two studies were not strictly comparable, Burns had to make some major, but plausible, assumptions in order to complete development of his technique. The lesson is the importance of reporting and justifying all assumptions. If the results seem implausible, revisit the assumptions.

### 6.2. Null Statistical results

Not all is lost if your quantitative results are not statistically significant. Null results do not mean the work is valueless, but never claim that lack of statistical significance of differences implies that the conditions are the same. There are two ways to emphasize the practical significance of any differences in measured values.

Field and Hole (2003) suggest that authors always report effect size as part of their statistical results. Reporting effect size allows readers to judge for themselves if differences matter practically.

Statistical techniques for equivalence testing, testing the hypothesis that sample populations do not differ, are available. An important application is in studies verifying the efficacy of generic compared to brand name drugs. Wellek (Welleck, 2002) is a comprehensive study of equivalence testing written for statisticians.

**Triangulation.** Multifaceted studies enable data *triangulation* (Berg, 1989). Triangulation, common in the social sciences, is the use of multiple research methodologies to study the same phenomena. The theory is that using multiple methodologies overcomes any biases

inherent in the individual methods and, consequently, enables the researcher to draw conclusions from the aggregate data more confidently than from a single measure or method. In the Distributed nanoManipulator study, the null statistical results were plausibly explained by the interview data that showed participants found positive and negative elements for both face-to-face and distributed collaboration conditions; they developed workarounds when they encountered problems. We were trying to find out if there were problems with scientific collaboration systems that would suggest that development stop. Looking at the whole of our data, we are comfortable saying that we found no showstoppers and development should continue.

## 7. Post-Experiment Debrief

We learn by doing, but we forget if we don't write it down. This is particularly true in an environment such as a graduate school research team with constantly changing members. The fact that a key individual from the early 2000s still works in the area and regularly reads EVE group email has helped us a number of times. The electronic tools are there, so keeping notes is logistically easy; it is the will that is weak. While the experience of running subjects is fresh, before the data analysis is done, sit together as a team, and record what went right and wrong. Do it again when the analysis is finished and you know whether you are able to answer the question that you started with.

**Acknowledgements.** The work reported here was largely funded by the Office of Naval Research (VIRTE Project), the NIH National Institute for Biomedical Imaging and Bioengineering, and SAIC. Additional support was provided by the Link Foundation and NC Space Grant.

Many of these same lessons are included in a book chapter "Evaluating VE Component Technologies" (Whitton & Brooks, 2008).

**References**


Arthur, K. W. (2000). *Effects of field of view on performance with head-mounted displays (CS Tech Rpt. # TR00-019).* Ph.D Dissertation, University of North Carolina, Technical Report #00-019.

Berg, B. L. (1989). *Qualitative Research Methods for the Social Sciences*. Boston: Allyn and Bacon.

Burns, E., Razzaque, S., Whitton, M., & Brooks, F. (2007). MACBETH: Management of Avatar Conflict by Employment of a Technique Hybrid. *International Journal of Virtual Reality, 6*(2), 11-20.

Feasel, J., Whitton, M. C., & Wendt, J. D. (2008). LLCM-WIP: Low Latency, Continuous - Motion Walking-In-Place. *Proceedings of IEEE Symposium on 3D User Interfaces, 97-104*.

Field, A., & Hole, G. (2003). *How to design and report experiments*: Sage Publishing.

Gabbard, J., Hix, D., & Swan, E. (1999). User-centered design and evaluation of virtual environments. *IEEE Computer Graphics and Applications*, 51-59.

Helms, J. W., Arthur, J. D., Hix, D., & Hartson, H. R. (2006). A field study of the Wheel--a usability engineering process model. *The Journal of Systems and Software, 79*, 841-858.

Hix, D., Swan, J., Gabbard, J., McGee, M., Durbin, J., & King, T. (1999). User-Centered Design and Evaluation of a Real-Time Battlefield Visualization Virtual Environment. *IEEE Virtual Reality 1999*, 96-103.

Hudson, T., Helser, A., Sonnenwald, D. H., & Whitton, M. (2003). Managing Collaboration in the Distributed nanoManipulator. *Proceedings of IEEE on Virtual Reality 2003*.

Insko, B. (2001). *Passive Haptics Significantly Enhances Virtual Environments (CS Technical Report #01-017).* Unpublished Ph.D Dissertation, University of North Carolina at Chapel Hill

Martin, D. W. (2007). *Doing Psychology Experiments, 7th Ed.* Belmont, CA: Wadsworth Publishing.

Meehan, M., B. Insko, M.C. Whitton, and F.P. Brooks. (2002). Physiological Measures of Presence in Stressful Virtual Environments. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2002), 21*(3), 645-652.

Meehan, M., S. Razzaque, M. Whitton, F. Brooks. (2002). Effects of Latency on Presence in Stressful Virtual Environments. *IEEE Virtual Reality 2003, 1*, 141-148.

Meehan, M. S., Razzaque, S., Insko, B., Whitton, M. C., & Brooks, F. P. (2005). Review of Four Studies on the Use of Physilogical Reaction as a Measure of Presence in Stressful Virtual Environments. *Applied Physiological and Biofeedback., 30(3)*, 239-258.

Slater, M., Usoh, M., & Steed, A. (1995). Taking Steps: The Influence of a Walking Technique on Presence in Virtual Reality. *ACM Transactions on Computer-Human Interaction, 2*(3), 201-219.

Sonnenwald, D., Whitton, M., & Maglaughlin, K. (2003). Evaluating a Scientific Collaboratory: Results of a Controlled Experiment. *ACM Transactions on Computer Human Interaction, 10*(2), 151-176.

Sonnenwald, D. H., K. Maglaughlin, M. Whitton. (2004). Designing to Support Situational Awareness Across Distances: An Example from a Scientific Collaboratory. *Information Processing & Management, 40*(6), 989-1011.

Sonnenwald, D. H., Maglaughlin, K., & Whitton, M. (2001). Using innovation diffusion theory to guide collaboration technology evaluation: Work in progress *IEEE 10th Workshop on Enabling Technologies: Infrastrucrure for Collabrative Enterterprises.*, Video Paper.

Sonnenwald, D. H., R. Berquist, K. Maglaughlin, E. Kupstas Soon, and M. Whitton. (2001). Designing to Support Scientific Research Across Distances: the nanoManipulator Environment. In D. S. E. Churchill, and A. Munro (Ed.), *Collaborative Virtual Environments* (pp. 202-224). London: Springer Verlag.

Usoh, M., K. Arthur, M.C. Whitton, A. Steed, M. Slater and F.P. Brooks. (1999). Walking>Virtual Walking>Flying, in Virtual Environments. *Proceedings of ACM SIGGRAPH 1999 (Computer Graphics Annual Conference Series 1999)*, 359-364.

Welleck, S. (2002). *Testing statistical hypotheses of equivalence*. Boca Raton, FL: Chapmen & Hall/CRC Press.

Whitton, M., & Brooks, F. (2008). Evaluating VE Component Technologies. In D. Nicholson, J. Cohn & D. Schmorrow (Eds.), *Virtual Environments for Training and Education: Developments for the Military and Beyond* (Vol. 2, pp. 240-261). Westport, CN: Praeger Security International.

Whitton, M. C., Cohn, J., Feasel, J., Zimmons, P., Razzaque, S., Poulton, S., et al. (2005). Comparing VE Locomotion Interfaces. *Proceedings of IEEE Virtual Reality 2005*, 123-130.

Zimmons, P. (2004). *The Influence of Lighting Quality on Presence and Task Performance in Virtual Environments (CS Tech Rpt. # TR04-017).* The University of North Carolina at Chapel Hill.