Applied Psychophysiology and Biofeedback, Vol. 30, No. 3, September 2005 (© 2005)
DOI: 10.1007/s10484-005-6381-3

Review of Four Studies on the Use of Physiological Reaction as a Measure of Presence in Stressful Virtual Environments

Michael Meehan,[1,4] **Sharif Razzaque,**[2] **Brent Insko,**[3] **Mary Whitton,**[2] **and Frederick P. Brooks Jr.**[2]

*A common measure of effectiveness of a virtual environment (VE) is the amount of presence it evokes in users. Presence is commonly defined as the sense of being there in a VE. There has been much debate about the best way to measure presence, and presence researchers need and have sought a measure that is reliable, valid, sensitive, and objective. We hypothesized that to the degree that a VE seems real, it would evoke physiological responses similar to those evoked by the corresponding real environment, and that greater presence would evoke a greater response. To examine this, we conducted four experiments, each of which built upon findings that physiological measures in general, and heart rate in particular, are reliable, valid, sensitive, and objective presence measures. The experiments compare participants' physiological reactions to a nonthreatening virtual room and their reactions to a stressful virtual height situation. We found that change in heart rate satisfied our requirements for a measure of presence, change in skin conductance did to a lesser extent, and that change in skin temperature did not. Moreover, the results showed that significant increases in heart rate measures of presence appeared with the inclusion of a passive haptic element in the VE, with increasing frame rate (*30 FPS $>$ 20 FPS $>$ 15 FPS*) and when end-to-end latency was reduced (50 ms $>$ 90 ms).*

KEY WORDS: presence; virtual environment; stress; heart rate; skin conductance; skin temperature; physiological measures.

INTRODUCTION

Presence and Virtual Environments

The effectiveness of a virtual environment (VE) might be defined in terms of enhanced task performance, more effective transfer of training, improved data comprehension and

[1]Stanford University, Crown Quadrangle, Stanford, California.
[2]Computer Science Department, University of North Carolina, Chapel Hill, North Carolina.
[3]3Dlabs, 9668 Madison Blvd., Madison, Alabama.
[4]Address all correspondence to Michael Meehan, Stanford University, Crown Quadrangle, 559 Nathan Abbott Way, Stanford, California 94305-8610; e-mail: mmeehan@stanford.edu.

increases in user preference for using the VE. A common metric of VE quality is the degree to which the VE creates in the user the subjective illusion of presence—a sense of being *in* the virtual environment, as opposed to the sense of *watching it* while being in the laboratory. For some applications such as phobia desensitization, presence is arguably the defining metric of VE quality (Hodges, 1994). Because presence is a subjective condition, it has most commonly been measured by self-reporting, either during the VE experience or immediately afterwards by questionnaires. There has been vigorous debate as to how to best measure presence (Dillon, Keogh, Freeman, & Davidoff, 2001; Ellis, 1996; Lombard & Ditton, 1997; Schubert, 2003; Sheridan, 1996; Slater, 1999; Slater, Usoh, & Steed, 1994; Slater & Steed, 2000, 2003; Witmer & Singer, 1998).

In order to study a VE's effectiveness in evoking presence, researchers need a well-designed and verified measure of the phenomena. This paper reports our evaluation of three physiological measures—heart rate, skin conductance, and skin temperature—as alternate operational measures for presence. Because the concept and idea of measuring presence are heavily debated, finding a measure that could find wide acceptance would be ideal. In that hope, we investigated the reliability, validity, sensitivity, and objectivity of each physiological measure.

Overview: Physiological Reaction as a Surrogate Measure of Presence

We conducted four experiments (in order): Effects of multiple exposures on presence (multiple exposures), effects of passive haptics on presence (passive haptics), the effects of frame rate on presence (frame rate), and effects of latency on presence (latency). The details of the user population and conditions of the four experiments are given in more detail below. However, it is instructive and interesting to note that the first three experiments, multiple exposures, passive haptics, and frame rate, were conducted at the University of North Carolina Computer Science Laboratory as within-participant studies. The latency experiment was conducted as part of the Emerging Technologies Exhibit at The Association of Computing Machinery's (ACM) Special Interest Group on Graphics (SIGGRAPH) 2002 Convention. In the latency study, we conducted a single trial for each participant and, therefore, conducted a between-groups study. Detailed discussion of some of the results from this research can also be found in Meehan (2001, Meehan, Insko, Whitton, & Brooks, 2002, Meehan, Razzaque, Whitton, & Brooks, 2003).

For each of the experiments, we had the dual goal of investigating and validating the use of physiological variables as measures of presence and for investigating how the modification of key VE parameters would impact the effectiveness of the VE. In order to validate the measures, we investigated whether the physiological measures were:

Reliable—produces repeatable results, both from trial to trial on the same participant and across participants,
Valid—measures subjective presence, or at least correlates with well-established subjective presence measures,
Sensitive—discriminates among multiple levels of presence, and
Objective—well shielded from both participant and experimenter bias.

We hypothesized that to the degree that a VE seems real, it will evoke physiological responses similar to those evoked by the corresponding real environment, and that greater

presence will evoke a greater response. If this hypothesis holds, these responses can serve as objective surrogate measures of subjective presence.

Overall, of the three physiological measures in our studies, Change in heart rate appeared to perform best. It consistently differentiated among conditions with more sensitivity and more statistical power than the other physiological measures, and more than most of the self-reported measures. It also best correlates with the reported measures.

Change in skin temperature was less sensitive, less powerful, and slower responding than change in heart rate, although its response curves were similar. It also correlated with self-reported measures. Our results and the literature on skin temperature reactions suggest that change in skin temperature would differentiate among conditions better if the exposures to the stimulus were at least 2 min (Slonim, 1974; D. R. McMurray, 1999, Director of Applied Physiology Lab, University of North Carolina, personal communication). Ours averaged 1.5 min in each experiment.

Change in skin conductance level yielded significant differentiation between conditions in some experiments but was not as consistent in doing so as change in heart rate. More investigation is needed to establish whether it can reliably differentiate among multiple levels of presence.

Because change in heart rate provided the best results of all the physiological measures used in our research, the remainder of this paper will focus primarily on this variable.

The Environment and Measures

In designing physiological measures for use in a virtual environment, one must consider at least three things. First, the virtual environment must evoke a physiological response. In our case, we use a virtual environment that evoked a height-response that is common for both height phobics and nonphobics (Abelson & Curtis, 1989; Andreassi, 1995). Second, the physiological response to the VE must be distinguishable from any response the user may have to the equipment or the laboratory. In our experiments, we realize that the participants may have some physiological reaction simply to wearing the head-mounted display and walking (see Fig. 1). In order to account for this, we define our measures based on a physiological baseline that is recorded while the users are walking while wearing the equipment. Third, the physiological reaction must be consistently measurable for the user population. In our case, we exclude any height phobics. Therefore, we need to ensure that the VE consistently elicits physiological reactions in the general population.

Our VR test environment was a derivative of the compelling VE reported by Usoh et al. (1999). Figure 2 shows the environment: a training room, quite ordinary, and an adjacent pit room, with an unguarded hole in the floor leading to a room 20 ft. below. On the upper level the pit room is bordered with a 2-ft wide walkway. The 18 ft $\times$ 32 ft, two-room virtual space fits entirely within the working space of our laboratory's wide-area ceiling tracker. Users, equipped with a head-tracked stereoscopic head-mounted display, practice walking about and picking up and placing objects in the training room. Then they are told to carry an object into the next room and place it at a designated spot. The door opens, and they walk through it to an unexpected hazard, a virtual drop of 20 ft. if they move off the walkway. Below the walkway is a furnished living room (Fig. 3).

Fig. 1. Participants wearing a head mounted display (HMD) and physiological monitoring equipment in the "pit room."

Users commonly report feeling frightened and some report vertigo. Some will not walk out on the ledge and ask to stop the experiment or demo at the doorway. Many walk the ledge, some hugging the wall for support. A few boldly walk out over the virtual hole, as if there were covered by a solid-glass floor. This environment, with its ability to elicit a fear reaction in users, enables investigation of physiological reaction as a measure of presence. If so strong a stress-inducing VE does not produce significant physiological reactions, a less stressful VE will not. This investigation is a first step. Follow-on research should investigate whether less stressful environments also elicit statistically significant physiological reactions.

Fig. 2. Side view of the virtual environment. Participants start in the training room and later enter the pit room.

Fig. 3. View of the 20-in. pit from the wooden ledge.

The Physiological Measures

As stated above, we investigated three physiological metrics that are commonly used to measure stress in real environments (Andreassi, 1995; Guyton, 1986; Weiderhold, Gervirtz, & Wiederhold, 1998): (1) change in heart rate (ΔHeart rate)—the heart beats faster in stress; (2) change in skin conductance (ΔSkin conductance level)—the skin of the palm sweats more and conductance rises in stress; and (3) change in skin temperature (ΔSkin temperature)—circulation slows in the extremities in stress, causing skin temperature to drop.

Each of these measures is constructed to increase when the physiological reaction to the pit room was greater:

$$\Delta\text{Heart Rate} = \text{mean HR}_{\text{Pit Room}} - \text{mean HR}_{\text{Training Room}}.$$

$$\Delta\text{Skin Conductance} = \text{mean SC}_{\text{Pit Room}} - \text{mean SC}_{\text{Training Room}}.$$

$$\Delta\text{Skin Temperature} = \text{mean ST}_{\text{Training Room}} - \text{mean ST}_{\text{Pit Room}}.$$

We measured heart rate with a convenient finger-mounted blood-pulse plethysmograph in the multiple exposures study but were not able to record usable signals for walking participants. In the three later studies, we switched to a more cumbersome chest-attached three-electrode electrocardiography (ECG) sensor—which consistently produced a usable signal. Skin conductivity and skin temperature were successfully measured on the fingers. Once connected, users reported forgetting about the physiological sensors—this suggests that did not cause "breaks" in presence during the experiments. Figure 1 shows a participant wearing the physiological monitoring equipment.

The Self-Report Measures

Reported Presence. In the first three studies (multiple exposures, passive haptics, and frame rate) we used the full University College London (UCL) questionnaire (Slater, Usoh, & Steed, 1995; Meehan, 2001; Usoh et al., 1999). The UCL questionnaire contains seven questions that measure presence (Reported presence) and three questions that measure behavioral presence (Reported Behavioral Presence)—does the user act as if in a similar real environment. Example questions include: "To what extent were there times during the experience when the virtual rooms you were in became the *reality* for you, and you almost forgot about the *real world* of the laboratory in which the whole experience was really taking place?" (Reported presence) and "To what extent was your reaction when looking down into the pit in the virtual reality the same as it would have been in a similar situation in real life?" (Reported Behavioral Presence). In the latency experiment (conducted at the ACM-SIGGRAPH convention) we used a single question from the Reported Behavioral Presence questionnaire (rate the amount of fear you experienced) to reduce the time required for participants to fill out the questionnaires, and five questions from the Reported presence measure.

Responses for each question are on a scale of 1–7 and these responses were coded as *high* (5, 6, or 7) or *low* (1–4) responses for the purposes of our analysis. There are also three questions that relate to ease of locomotion in the UCL questionnaire that were administered for consistency with earlier experiments, but these are not reported on it in this paper.

Methods and Procedures

Experimental Procedures

Each of the four studies investigated some relevant parameter of a VE along with examining the properties of the physiological measures as a correlate of presence. Table I summarizes all the questions studied. For the first three studies, we excluded participants who had previously experienced VEs more than three times. The experiments were also limited to participants who were ambulatory, could use stereopsis for depth perception, had

Table I. Questions Investigated in Each Study

	Multiple exposures	Passive haptics	Frame rate	Latency
Presence in VEs	Does presence decrease with exposures?	Passive haptics increase presence?	Higher frame rate increases presence?	Lower latency increases presence?
Reliability of measures	Are repeated measures highly correlated?	Regardless of condition, will the pit room evoke similar physiological reactions on every exposure?		Regardless of condition, will the pit room evoke similar physiological reactions in all users?
Validity	Do results correlate with reported measures?			
Sensitivity of measures	Do measures detect any effect?	Do measures distinguish between 2 conditions?	Do measures distinguish among 4 conditions?	Can measures differentiate between 50 and 90 ms latency?

no history of epilepsy or seizure, were not overly prone to motion sickness, were in their usual state of good physical fitness at the time of the experiment, and were comfortable with the equipment. In the latency experiment, we relaxed our usual practice of excluding individuals who had experienced VEs more than three times, had consumed alcohol or medications in the past 24 hr or were suffering from emotional stress or sleep deprivation because this was a less realistic restriction at SIGGRAPH.

Multiple Exposures. Ten participants (average age 24.4; $\sigma = 8.2$; seven female, three male) were trained to pick up books and move about in the training room—at which time a physiological baseline was taken. Participants then carried a virtual book from the training room and placed it on a virtual chair on the far side of the pit room. After that, they returned to the training room. The participants performed this task three times per day on 4 separate days. We investigated whether the presence-evoking power of a VE declines with multiple exposures. Heart rate was not successfully measured in this study due to problems with the sensor.

Passive Haptics. Fifty-two participants (average age 21.4; $\sigma = 4.3$; 16 female, 36 male) reported on 2 days. On day 1, a participant experienced the VE with the 1.5-in. wooden ledge. The 1.5-in. height was selected so that the edge-probing foot did not normally contact the real laboratory floor where the virtual pit was seen. On the other day, the participant experienced the VE without the ledge. Participants were counterbalanced as to the order of presentation of the ledge. Participants performed all exposures to the VE wearing only thin sock-like slippers (Fig. 4). The task was the same as in the multiple exposures study except participants were instructed to walk to the edge of the wooden platform, place their toes over the edge, and count to 10 before they proceeded to the chair on the far side of the room to drop the book. We investigated whether the 1.5-in. wooden ledge increased the presence-evoking power of the VE.

Frame Rate. Thirty-three participants (average age 22.3; $\sigma = 3.6$; 8 female, 25 male). Participants entered the VE four times on 1 day and were presented the same VE with a different frame rate each time. The four frame rates were 10, 15, 20, and 30 frames per second (FPS). Participants were counterbalanced as to the order of presentation of the four frame rates. Participants were trained to pick up and drop blocks in the training room and

Fig. 4. Participant in slippers with toes over 1.5-in. ledge.

then carried a red block to the pit room and dropped it on a red *X*-target on the floor of the living room, a procedural improvement that forced participants to look down into the pit. They then plucked from the air two other colored blocks floating in the pit room and dropped each on the correspondingly colored *X*s on the floor of the living room. The *X*-targets and the green and blue blocks are visible in Figs. 2 and 3. In this study, we investigated the effect of multiple frame rates on presence and hypothesized that the higher the frame rate, the greater the presence evoked.

Latency. This experiment was designed to accommodate as many participants as possible in a limited time. One hundred and ninety-five ACM-SIGGRAPH attendees participated in the full demo experience in just 4 days. One hundred and sixty-four individuals (average age 35; $\sigma = 10.9$; 32 female) passed the inclusion criteria and yielded full or partial data. All 164 had usable questionnaire data. Sixty-one participants had usable heart-rate data and 67 had usable skin-conductance data. These low yields were a consequence of our decisions not to interrupt any VE session to correct failures of the physiological monitoring equipment and of a higher rate of equipment failure than we have seen in our university lab. The later is likely attributable to greater wear-and-tear and less experienced experimenters attaching the sensors. We did not collect skin temperature data and used a task similar to that used in the frame rate experiment. We hypothesized that the lower the end-to-end latency (the time duration between a real-world head movement and the results of that movement appearing in the HMD), the higher the presence evoked. The lowest latency we could consistently achieve was 50 ms (roughly 25 ms was incurred by the HMD itself). We choose 90 ms for the high latency condition to ensure that every visitor had a good experience. In previous pilot versions of our system, some visitors complained when the end-to-end latency was above 120 ms. We decided to stay well below this unacceptable amount of latency. Details of how we increased the latency without changing the frame rate are described in (Meehan et al., 2003).

In all four studies, the amount of physical activity (walking, manipulating objects) was approximately balanced between the pit and training rooms. This lessened any difference between the two rooms in physiological reaction due to physical activity.

Statistics

To find the best statistical model for each measure, we used stepwise selection and elimination as described by Kleinbaum, Kupper, Muller, and Nizam (1998). As they suggest, to account better (statistically) for variation in the dependent variable (e.g. ΔHeart rate), we included all variables in the statistical models that were significant at the $p = .100$ level (strong trends).

The comparison of differences in physiological reaction between the pit room and the training room (Table II) was performed with a one-sample *t*-test. The correlations among measures were performed using the bivariate Pearson correlation. tests of the effects on presence of passive haptics and frame rate were performed with the univariate general linear model, using the repeated measure technique described in the SAS 6.0 manual (SAS, 1990). This technique allows one to investigate the effect of the condition while taking into account interparticipant variation, order effects, and the effects of factors that change from exposure to exposure such as loss of balance on the 1.5-in. ledge.

Table II. Means and Significance for One-Sample *t*-Test and Percentage of Times the Measure was >0

		All exposures				First exposure only (between-participants)			
Study	Variable	Mean	*p*-Values	% > 0	*N*	Mean	*p*-Values	% > 0	*N*
Multiple exposures	ΔSkin conductance	2.3 Δ μS	<.001	99%	112	2.9 Δ μS	.002	100%	9
	ΔSkin temperature	0.6 Δ °F	<.001	77%	94	1.2 Δ °F	.015	100%	7
Passive haptics	ΔHeart rate	6.3 Δ bpm	<.001	89%	92	6.2 Δ bpm	<.001	85%	46
	ΔSkin conductance	4.8 Δ μS	<.001	100%	100	4.7 Δ μS	<.001	100%	50
	ΔSkin temperature	1.1 Δ °F	<.001	90%	98	1.1 Δ °F	<.001	94%	49
Frame rate	ΔHeart rate	6. 3 Δ bpm	<.001	91%	132	8.1 Δ bpm	<.001	91%	33
	ΔSkin Conductance	2.0 Δ μS	<.001	87%	132	2.6 Δ μS	<.001	97%	33
	ΔSkin temperature	0.8 Δ °F	<.001	100%	132	1.0 Δ °F	<.001	100%	33
Latency (between-participants)	ΔHeart rate	8.6 Δ bpm	<.001	98%	61	8.6 Δ bpm	<.001	98%	61
	ΔSkin conductance	3.8 Δ μS	<.001	100%	67	3.8 Δ μS	<.001	100%	67

Note. The left side of the table shows the average of all exposures for all participants for each measure in each study. The right side shows the average of the first exposures for all participants for each measure for each study (latency study had only one exposure per person).

PHYSIOLOGICAL MEASURES OF PRESENCE

Reliability

Reliability is "the extent to which the same test applied on different occasions . . . yields the same result" (Sutherland, 1996). Specifically, we wanted to know whether the virtual environment would consistently evoke similar physiological reactions as the participant entered and remained in the pit room on several occasions. Inconsistency could manifest itself as either a systematic increase or decrease in reactions or in uncorrelated measures for repeated exposure to the same VE. In the multiple exposures study, the condition was the same each time, so this was our purest measure of reliability. We also hypothesized that in the passive haptics and frame rate studies, regardless of condition, that the pit room would also evoke similar physiological reactions on every exposure. We hypothesized that simply being exposed to the pit room would cause a greater physiological reaction than the difference between "high" and "low" presence conditions. Therefore, these three studies provide information on reliability. In the latency experiment, each user experienced the VE only one time. Therefore, we did not obtain data on exposure-to-exposure consistency, but did get interuser consistency information in this study.

As we hypothesized, the environment consistently evoked physiological reactions over multiple exposures to the pit room. There were significant physiological reactions to the pit room: heart rate and skin conductance were significantly higher and skin temperature was significantly lower in the pit room in all three studies. Heart rate was higher in the pit room for 90% of the exposures to the VE, skin conductance was higher for nearly 95%, and skin temperature was lower for 90%. Table II shows the mean difference, *t*-test, the percent of occurrences where the measure was above zero, and the total count for each physiological measure for each study. Table II also shows results both for all exposures taken together, which is the approach discussed primarily in this paper, and for analysis of the first exposure only, which we discuss in VE Effectiveness Results: Four Studies section.

We also wanted to examine whether the physiological reactions to the environment would diminish over multiple exposures. Because our hypotheses relied on presence in the VE evoking a stress reaction over multiple exposures (2–12 exposures), we examined whether physiological reactions to the VE would drop to zero or become unusably small due to habituation. In fact, ΔSkin temperature, Reported presence, Reported behavioral presence, and ΔHeart rate each decreased with multiple exposures in the three within-participant studies (though, not necessarily with statistical significance), and ΔSkin conductance decreased over multiple exposures in all but one study. None decreased to zero, though, even after 12 exposures to the VE. Table III shows the significant order effects.

Table III. There was a Significant Order Effect for Each Measure in At Least One Study

Order effects	ΔHeart rate (Δbpm)	ΔSkin conductance (ΔμS)	ΔSkin temperature (Δ°F)	Reported presence (Count "high")	Reported behavioral presence (Count "high")
Multiple exposures	NA		−0.9 (1st)	—	−0.7 (1st)
Passive haptics	—	—	—	−0.8 (1st)	−0.4 (1st)
Frame rate	−1.0 (Task)	−0.8 (1st)	−0.3 (1st)	—	−0.2 (Task)

Note. NA is "Not available." Full details given in (Meehan, 2001). The Latency experiment was interparticipant, single exposure, and therefore produced no data with respect to order effects; "(1st)" indicates a decrease after the first exposure only; "(Task)" indicates a decrease over tasks on the same day.

A decrease in physiological reaction over multiple exposures would not necessarily weaken validity, because the literature shows that habituation diminishes the stress reactions to real heights and other stressors (Abelson & Curtis, 1989; Andreassi, 1995). However, because the Reported presence measure, not just the physiological stress measures, decreased over multiple exposures, the decreases may not be due to habituation to the stressor; there may also be as Heeter (1992) hypothesized—a decrease in a VEs ability to evoke presence as novelty wears off.

Orienting Effect

In general, each measure decreased after the first exposure. Moreover, for each measure except ΔHeart rate, there was a significant decrease after the first exposure in at least one of the studies (see Table III). For physiological responses, this is called an *orienting effect*—a physiological reaction when one sees something novel (Andreassi, 1995). Though this term traditionally refers to physiological reactions, we will also use the term for the initial spike in the reported measures.

We attempted, with only partial success, to overcome the orienting effect by exposing participants to the environment once as part of their orientation to the experimental setup and prior to the data-gathering portion of the experiment. In the passive haptics and frame rate studies, participants entered the VE for approximately 2 min and were shown both rooms before the experiment started. These preexposures reduced, but did not eliminate, the orienting effects.

Validity

Validity is "the extent to which a test or experiment genuinely measures what it purports to measure" (Sutherland, 1996). Because presence is a subjective condition, researchers have developed various operational definitions of presence to allow it to be measured. Presence has most commonly been measured by self-report questionnaires. There has been vigorous debate in the community as to how to best do this and Sadowski (2002) presents a useful summary and extensive bibliography on this topic. Because the concept of presence has been best operationalized in questionnaires, we attempted to validate the physiological measures by investigating how well physiological reactions correlate with one or more of the questionnaire-based measures of presence. We investigated such correlations with two measures: Reported presence and Reported behavioral presence.

Reported Presence

Of the physiological measures, ΔHeart rate correlated best with the Reported presence. There was a significant correlation in the frame rate study ($r = .265$, $p < .005$) and no correlation ($r = .034$, $p = .743$) in the passive haptics study. In the multiple exposures study, where ΔHeart rate was not available, ΔSkin conductance had the highest correlation with Reported presence ($r = .245$, $p < .010$).

Reported Behavioral Presence

ΔHeart rate had the highest correlation, and a significant one, with Reported Behavioral Presence in the frame rate study ($r = .192$, $p < .050$), and there was no correlation

between the two in the passive haptics study. In the multiple exposures study, where ΔHeart rate was not measured, ΔSkin conductance had the highest correlation with reported behavioral presence ($r = .290$, $p < .005$). In the latency experiment, ΔSkin conductance and the single Reported behavioral presence question (Reported fear) correlated significantly ($r = .275$, $p = .024$).

The correlations between the physiological measures and the self-report questionnaire measures lends some support to their validity. The validity of ΔHeart rate appears to be better established by its correlation with the well-established reported measures. There was also some support for the validity of ΔSkin conductance from its correlation with reported measures.

Following Hypothesized Relationships

According to Singleton, the validation process includes "examining the theory underlying the concept being measured," and "the more evidence that supports the hypothesized relationships (between the measure and the underlying concept), the greater one's confidence that a particular operational definition is a valid measure of the concept" (Singleton, Straits, & Straits, 1993). We hypothesized that presence should increase with frame rate, with the inclusion of the 1.5-in. wooden ledge, and with lower latency, since each of these conditions provides increased sensory simulation fidelity. As presented in the next section, our physiological measures did increase with frame rate, with inclusion of the 1.5-in. wooden ledge, and with lower latency.

Sensitivity and Multilevel Sensitivity

Sensitivity is "the likelihood that an effect, if present, will be detected" (Lipsey, 1998). The fact that the physiological measures reliably distinguished between participants reaction in the pit room versus the training room in every study assured us of at least a minimal sensitivity. For example, heart rate increased an average across all conditions of 6.3 beats per min (bpm) in the pit room ($p < .001$) compared to the training room in both the passive haptics and frame rate studies and 8.6 bpm in the latency study. See Table II for a full account of sensitivity of physiological measures to the difference between the two rooms. A study on Acrophobic patients' heart rate reactions to climbing to the second story of a fire escape (with a hand rail), waiting 1 min, and looking down reported a 13.4 PBM increase on average (Emmelkamp & Felten, 1985). Our participants were nonphobic, and our height was virtual; so, we would expect, and did find, our participants' heart-rate reactions to be lower, but in the same direction.

Multilevel Sensitivity

For guiding VE technological development and for better understanding the psychological phenomena of VEs, we need a measure that reliably yields a higher value as a VE is improved along some "goodness" dimension, that is, is *sensitive* to *multiple* condition values. We distinguish this from sensitivity as described above and call this *multilevel sensitivity*. The passive haptics study provided us some evidence of the measures' ability

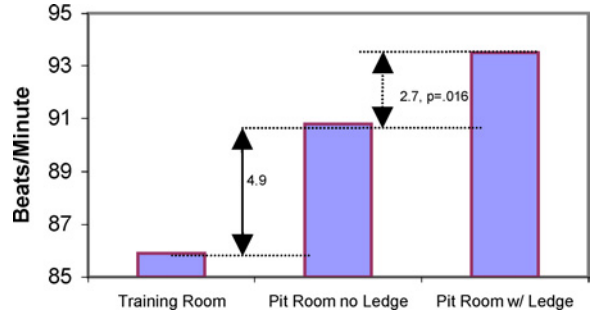

Fig. 5. ΔHeart rate in passive haptics study.

to discriminate between two "high presence" situations. We have informally observed that walking into the pit room causes a strong reaction in users, and this reaction seems greater in magnitude than the differences in reaction to the pit room between any two experimental conditions (e.g. with and without the 1.5-in. wooden ledge). Therefore, we expected the differences in reaction among the conditions to be less than the differences between the two rooms. For example, in passive haptics, we expected there to be a significant difference in the physiological measures between the two conditions (with and without the 1.5-in. wooden ledge), but expected it to be less than the difference between the training room and pit room in the "lower" presence condition (without the 1.5-in. wooden ledge). For ΔHeart rate, we found a significant difference between the two conditions of 2.7 bpm ($p < .050$), and it was less than the inter-room difference for the without-ledge condition: 4.9 bpm (see Fig. 5). Figure 6 shows that the differences among the conditions in the FR study are smaller in magnitude as compared to the differences between the two rooms.

In the passive haptics study, we investigated the multilevel sensitivity of the measures by testing whether presence was significantly higher with the 1.5-in. wooden ledge. Presence as measured by each of ΔHeart rate (2.7 bpm; $p < .050$), ΔSkin conductance (.8 μS; $p < .050$), and Reported Behavioral Presence (.5 more "high" responses; $p < .005$) was

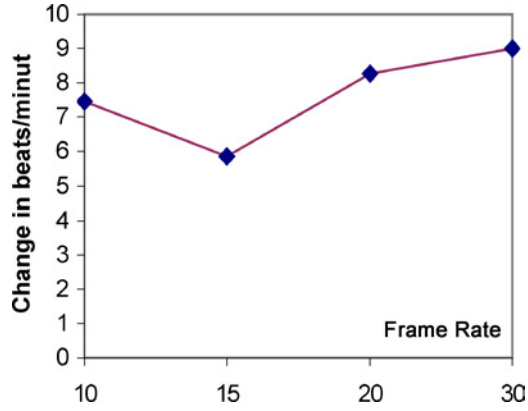

Fig. 6. ΔHeart rate, after correcting for Loss of Balance, at 10, 15, 20, and 30 frames per second.

significantly higher with the wooden ledge. Reported presence had a strong trend in the same direction (.5 more "high" responses; $p = .060$).

In the frame rate study, we investigated the multilevel sensitivity of the measures by testing whether presence increased significantly as graphic frame update rates increased. We hypothesized that physiological reactions would increase monotonically with frame rates of 10, 15, 20, and 30 FPS. They did not do exactly that (see Fig. 6). During the 10 FPS condition, there was an anomalous reaction for all of the physiological measures and for Reported Behavioral Presence. That is, at 10 FPS, participants had higher physiological reaction and reported more behavioral presence. We believe that this reaction at 10 FPS was due to discomfort, added lag, and reduced temporal fidelity while in the ostensibly dangerous situation of walking next to a 20-foot pit (Meehan, 2001).

We also observed that participants often lost their balance while trying to in. to the edge of the wooden platform at this low frame rate; their heart rate jumped an average of 3.5 bpm each time they lost their balance ($p < .050$). Controlling for these "Loss of Balance" incidents improved the significance of the statistical model for ΔHeart rate and brought the patterns of responses closer to the hypothesized monotonic increase in presence with frame rate—but did not completely account for the increased physiological reaction at 10 FPS. Loss of Balance was not significant in any other model.

Beyond 10 FPS, ΔHeart rate followed the hypothesis. After we statistically controlled for Loss of Balance, ΔHeart rate significantly increased between 15 FPS and 30 FPS (3.2 bpm; $p < .005$) and between 15 and 20 FPS (2.4 bpm; $p < .050$). There was also a nonsignificant increase between 20 and 30 FPS (.7 bpm; $p = .483$) and a nonsignificant decrease between 10 and 15 FPS (1.6 bpm; $p = .134$). Reported presence, and Reported behavioral presence also increased with frame rate from 15–20–30 FPS, but with less distinguishing power.

In the latency study, we investigated whether lower latency (50 ms as opposed to 90 ms) would affect presence. We recorded a difference of 3.1 beats per minute (bpm) between for ΔHeart rate the two conditions. In the 50 ms condition, heart rate in the pit room was, on average 10.1 bpm higher than in the training room ($n = 32$). In the 90 ms condition, this difference was 7.0 bpm ($n = 29$). See Fig. 7. The difference in reaction elicited by the two conditions is borderline significant with a p value of .050 (we define significance at the $p < .050$ level). We did not record a significant difference in ΔSkin conductance between conditions.

These findings support the multilevel sensitivity of ΔHeart rate.

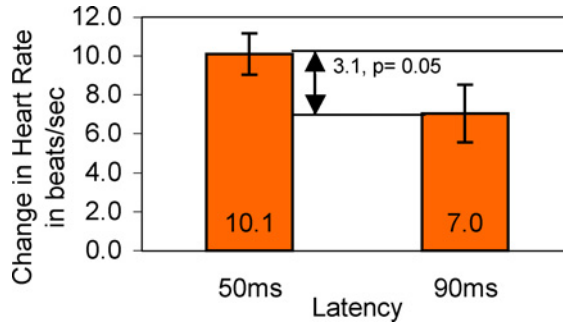

Fig. 7. Change in heart rate for 50 and 90 ms latency.

Objectivity

The measurement properties of reliability, validity, and multilevel sensitivity are established quantitatively. Objectivity can only be argued logically. We argue that physiological measures are inherently better shielded from both participant bias and experimenter bias than are either reported measures or measures based on behavior observations. Reported measures are liable to participant bias—the participant reporting what he or she believes the experimenter wants. Post-experiment questionnaires are also vulnerable to inaccurate recollection and to modification of impressions garnered early in a trial by impressions from later. Having participants report during the session, whether by voice report or by hand-held instrument, intrudes on the very presence illusion one is trying to measure. Behavioral observation measures, while not intrusive, can be subject to bias on the part of the experimenters who score such behaviors.

Physiological measures, on the other hand, are much harder for participants to affect, especially with no biofeedback. These measures are not liable to experimenter bias, if instructions given to the participants are properly limited and uniform. We read instructions from a script in the multiple exposures study. We improved our procedure in the later passive haptics and frame rate studies by playing instructions from a compact disk player located in the real laboratory and represented by a virtual radio in the VE. In the latency study, we played the instructions in headphones with 3D aural representation corresponding to a virtual loudspeaker in the VE.

Summary and Discussion

The data presented thus far support the idea that physiological variables can be used as reliable, valid, multilevel sensitive, and potentially objective measures of presence. Of the physiological measures, ΔHeart rate performed the best. There was also some support for ΔSkin conductance.ΔHeart rate significantly differentiated between the training room and the pit room, and although this reaction faded over multiple exposures, it never decreased to zero. It also correlated with the well-established reported measure, the UCL questionnaire. It distinguished between the presence and absence of passive haptics, among frame rates at and above 15 FPS, and between 50 and 90 ms latencies. Also, as we argued above, it is objective. In total, it satisfies all of the requirements for a reliable, valid, multilevel sensitive, and objective measure of presence. In addition, the measure produced significant results in our between-participant study (latency), providing evidence that, in addition to being a useful within-participants measure, it may be viable as a between-groups measure.ΔSkin conductance has some, but not all, of the properties we desire in a measure of presence. In particular, it did not differentiate among frame rates. We do not have a theory as to why. In addition, ΔSkin conductance did not perform well as a between-participant measure (in the latency study). We suspect this is in part because the magnitude of change in skin conductance due to stress is quite variable among people (Andreassi, 1995). In order to better utilize ΔSkin conductance in future studies, we may need to normalize the reactions among users. This is discussed further in Physiological Reactions Using a Between-Groups Design section.

Although, ΔHeart rate satisfied the requirements for a presence measure for our VE, which evokes a strong reaction, it may not for less stressful VEs. To determine whether

physiological reaction can more generally measure presence, a wider range of VEs must be tested, including less stressful, nonstressful, and relaxing environments. Investigation is currently under way to look at physiological reaction in relaxing 3D television environments (Dillon et al., 2001).

The height reaction elicited by our VE could be due to vertigo, fear, or other innate or learned response. The reactions are well known in the literature and manifest as increased heart rate and skin conductance and decreased skin temperature (Andreassi, 1995; Guyton, 1986). We hypothesized that the more present a user feels in our stressful environment, the more physiological reaction the user would exhibit. What causes this higher presence and higher physiological reaction? Is it due to a more realistic flow of visual information? Is it due to more coherence between the visual and haptic information? Is it due to the improved visual realism? All of these are likely to improve presence. We cannot, however, answer these questions definitively. However, we can say that we have empirically shown that physiological reaction and Reported presence are both higher when we present a "higher presence" VE. Whatever it is that causes the higher Reported presence and physiological reaction, it causes more as we improve the VE.

An additional desirable aspect of a measure is ease of use in the experimental setting. We did not record the time needed for each measure, but after running many participants we can anecdotally report with some confidence that use of the physiological monitoring and of the presence questionnaire each added roughly the same amount of time to the experiment. It took approximately 5 min per exposure to put on and take off the physiological sensors. It took about an extra minute at the beginning and end of each set of exposures to put on and take off the ECG sensor—it was left on between exposures on the same day (in the frame rate study). It took participants about 5 min to fill out the 16-item UCL Presence Questionnaire. It took some training for experimenters to learn the proper placement of the physiological equipment on the hands and chest of the participant—30 min would probably be sufficient.

Another aspect of "ease of use" is the amount of difficulty participants have with the measure and to what extent the measure, if concurrent with an experimental task, interferes with the task. No participants reported difficulties with the questionnaires. Only about 1 in 10 participants reported noticing the physiological monitoring equipment on the hands during the VE exposures. Our experiment, though, was designed to use only the right hand, keeping the sensor-laden left hand free from necessary activity. No participants reported noticing the ECG sensor once it was attached to the chest. In fact, many participants reported forgetting about the ECG electrodes when prompted to take them off at the end of the day. One participant actually left the building with the sensor still attached. There are groups investigating less cumbersome equipment, which would probably improve ease of use, including a physiological monitoring system that participants wear like a shirt (Cowings, Jensen, Bergner, & Toscano, 2001). Overall, questionnaires and physiological monitoring were both easy to use and nonintrusive.

PHYSIOLOGICAL REACTIONS USING A BETWEEN-GROUPS DESIGN

We conducted the multiple exposures, passive haptics, and frame rate studies as within-participants to avoid the variance due to natural human differences. That is, each participant experienced all of the conditions for the study in which he participated. This allowed us to

look at relative differences in participant reaction among conditions and to overcome the differences among participants in reporting and physiological reaction. The latency study was between-groups—each participant experienced only one condition (either 50 ms or 90 ms of lag).

The UCL questionnaire has been used successfully between-groups (Usoh et al., 1999). However, we expected the variance among participants would mask, at least in part, the differences in physiological reaction evoked by the different conditions. We investigated this hypotheses by analyzing the data using *only the first task* for each participant in the multiple exposures, passive haptics, and frame rate studies—eliminating order effects and treating the reduced data sets effectively as between-groups experiments. That is, we treat each experiment as if only the first task for each participant was run. This means that the analysis uses only 10 data points (10 participants—first exposure only) for the multiple exposures study, 52 data points for the passive haptics study, and 33 data points for the frame rate study. In addition, we looked at the latency study data as recorded since it was performed between groups.

Reliability Between-Groups: Physiological Reaction in the Pit Room

Even between groups, we expected that there would be a consistent physiological reaction to the pit room, because we expected such a reaction for every exposure to the VE. We expected the significance to be lower, however, because of the reduced size of the data set. We found exactly that. The right half of Table II shows the values of the physiological measures averaged across conditions for the between-groups analysis. As compared to the full dataset, the between-groups data have lower significance values, but participants still have strong physiological reactions to the pit room. Table II demonstrates that the physiological orienting effects caused the averages for the first exposures to be *higher* than for subsequent exposures (regardless of experimental condition).

Validity Between-Groups: Correlation with Established Measures

We expected correlations with the reported measures to be lower when taken between groups because there were fewer data points and individual differences in physiological reaction and reporting would confound the correlations. This was the case. The only significant correlation between any self-reported measure and any physiological measure was with ΔSkin conductance and Reported Fear in the latency study ($r = .275$, $p = .024$).

Multilevel Sensitivity Between-Groups: Differentiating Among Presence Conditions

We expected interparticipant variation in physiological reaction to mask the differences in physiological reactions evoked by the presence conditions (e.g., various frame rates). Contrary to this expectation, however, we found significant differences in the physiological measures among conditions in both the passive haptics and frame rate studies. (The condition was not varied in the multiple exposures study.)

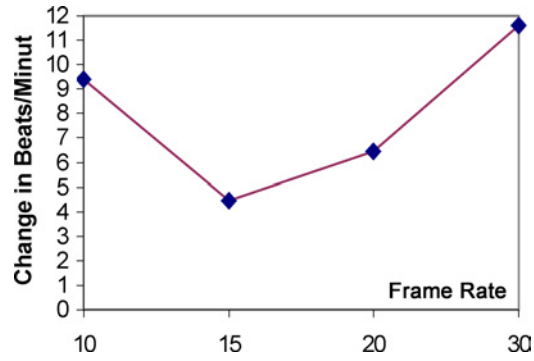

Fig. 8. Between-groups analysis: ΔHeart rate.

In the passive haptics study, both ΔHeart rate and ΔSkin conductance both varied in the expected direction nonsignificantly (3.3 bpm, $p = .097$; 1.0 μS, $p = .137$, respectively).

In the frame rate study, ΔHeart rate followed hypothesized patterns, but ΔSkin conductance did not. After the anomalous reaction at 10 FPS (as in full data set compare Figs. 6 and 8), ΔHeart rate differentiated among presence conditions: at 30 FPS it was higher than at 15 FPS, and this difference was nearly significant (7.2 bpm; $p = .054$).

As noted above, ΔHeart rate differentiated with borderline significance between the two latency conditions (50 and 90 ms) in the latency study (3.1 bpm; $p = .050$).

Overall, ΔHeart rate shows promise as a between-groups measure of presence. Though it did not correlate well with the reported measures, it did differentiate among the conditions with some statistical power in passive haptics and frame rate. ΔSkin conductance did not show as much promise as a between-groups measure. For more discussion of physiological reactions as between-groups measures of presence, see Meehan (Meehan, 2001; Meehan et al., 2003).

VE EFFECTIVENESS RESULTS: FOUR STUDIES

Above we described the experiments as they related to the testing of the physiological presence measures, below we discuss each experiment with respect to the aspect of VEs it investigated.

Effect of Multiple Exposures on Presence

As described in Experimental Procedure section, 10 users go through the same VE 12 times (over 4 days) in order to study whether the presence inducing power of a VE declines, or becomes unusably small, over multiple exposures. We did find significant decreases in each presence measure (reported and physiological) in either this experiment or one of the subsequent two experiments (see Table III). However, none of the measures decreased to zero nor did any become unusably small. The findings support our hypothesis that all presence measures decrease over multiple exposures to the same VE, but not to zero.

Effect of Passive Haptics on Presence

Our hypothesis was that supplementing a visual–aural VE with even rudimentary, low-fidelity passive haptics cues significantly increases presence. We found significant support for the hypothesis in that, with the inclusion of the 1.5-in. ledge, presence as measured by ΔHeart rate, Reported behavioral presence, and ΔSkin conductance was significantly higher at the $p < .05$ level. Reported presence also had a strong trend ($p < .10$) in the same direction.

Effect of Frame Rate on Presence

Our hypothesis was that as frame rate increases from 10, 15, 20, 30 frames per second, presence increases. For frame rates of 15 frames/second and above, the hypothesis was largely confirmed. It was confirmed with statistical significance for 15 to 20 FPS and 15–30 FPS. Twenty to thirty frames per second though not statistically significant was in the same direction. 10 FPS gave anomalous results on all measures except Reported presence, which increased monotonically with frame rate with no statistical significance.

Effect of Latency on Presence

Participants experiencing lower end-to-end latency (50 ms) in our stress-inducing VE had significantly more heart rate reaction to the stressful pit room than did those in the higher latency condition (90 ms). Based on this, we believe that end-to-end latency, even when below 100 ms, is an important parameter in understanding the effectiveness of the VE. Like frame rate, it should be measured, controlled and reported in all VE research.

FUTURE WORK

Given a compelling VE and a sensitive, quantitative presence measure, the obvious strategy is to degrade quantitative VE quality parameters in order to answer the questions: What makes a VE compelling? What are the combinations of minimum system characteristics to achieve this? Our future work will explore these issues and continue to investigate the construct of presence using physiological and self-report measures in order to better understand how it is effected by such system factors as lighting realism and realistic physics in one's interactions with virtual objects. As well, we are interested in how presence may affect user interactions with other "virtual" people or agents cohabitating a VE and on the impact of having yourself exist in the VE as a virtual human representation or avatar.

ACKNOWLEDGMENTS

This work was supported by grants from the Office of Naval Research, the NIH National Center for Research Resources (P41 RR 02170), and the National Institute for Biomedical Imaging and Bioengineering. Other support was supplied by NVIDIA, and the Link Foundation.

REFERENCES

Abelson, J. L., & Curtis, G. C. (1989). Cardiac and neuroendocrine responses to exposure therapy in height phobics. *Behavior Research and Therapy*, *27*(5), 561–567.

Andreassi, J. L. (1995). *Psychophysiology: Human behavior and physiological response*. Hillsdale, NJ: Erlbaum.

Cowings, P., Jensen, S., Bergner, D., & Toscano, W. (2001). *A lightweight ambulatory physiological monitoring system*. California: NASA Ames.

Dillon, C., Keogh, E., Freeman, J., & Davidoff, J. (2001). *Presence: Is your heart in it*?. 4th International Workshop on Presence, Philadelphia.

Ellis, S. R. (1996). Presence of mind: A reaction to Thomas Sheridan's "Further musings on the psychophysics of presence." *Presence: Teleoperators and Virtual Environments*, *5*(2), 247–259.

Emmelkamp, P., & Felten, M. (1985). The process of exposure in vivo: Cognitive and physiological changes during treatment of acrophobia. *Behavior Research and Therapy*, *23*(2), 219.

Guyton, A. C. (1986). *Textbook of medical physiology* (pp. 688–697). Philadelphia, PA: W.B. Saunders.

Heeter, C. (1992). Being there: The subjective experience of presence. *Presence: Teleoperators and Virtual Environments*, *1,* 262–271.

Hodges, L., Rothbaum, B., Kooper, R., Opdyke, D., Willford, J., Meyer, T., et al. (1994). *Presence as the defining factor in a VR application*. Technical Report GVU-94-06. Georgia Tech University, Graphics, Visualization, and Usability Center.

Kleinbaum, D., Kupper, L., Muller, K., & Nizam, A. (1998). *Applied regression analysis and other multivariate methods* (3rd ed.). Pacific Grove, CA: Duxbury Press.

Lipsey, M. W. (1998). Design sensitivity: Statistical power for applied experimental research. In L. Brickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 39–68). Thousand Oaks, CA, Sage.

Lombard, M., & Ditton, T. (1997). At the heart of it all: The concept of presence. *Journal of Computer Mediated Communication*, *3*(2). Available at: http://www.ascusc.org/jcmc/vol3/issue2/lombard.html

Meehan, M. (2001). *Physiological reaction as an objective measure of presence in virtual environments*. Doctoral dissertation, Computer Science, University of North Carolina, Chapel Hill, NC.

Meehan, M., Insko, B., Whitton, M., & Brooks, F. P. (2002). *Physiological measures of presence in stressful virtual environments*. In Proceedings of ACM SIGGRAPH 2002.

Meehan, M., Razzaque, S., Whitton, M., & Brooks, F. (2003). Effects of latency on presence in stressful virtual environments. *Proceedings of IEEE Virtual Reality 2003* (Los Angeles, CA, March 2003) (pp. 141–148). IEEE Computer Society.

Schubert, T. (2003). The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie*, *15,* 69–71.

Sheridan, T. B. (1996). Further musings on the psychophysics of presence. *Presence: Teleoperators and Virtual Environments*, *5*(2), 241–246.

Singleton, R. A., Straits, B. C., & Straits, M. M. (1993). *Approaches to Social Research*. New York, Oxford University Press.

Slater, M. (1999). Measuring Presence: A Response to the Witmer and Singer Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, *8*(5), 560–565.

Slater, M., Usoh, M., & Steed, A. (1994). Depth of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *3*(2), 130–144.

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* (*TOCHI*), *2*(3), 201–219.

Slater, M., & Steed, A. J. (2000) A virtual presence counter, *Presence: Teleoperators and Virtual Environments*, *9*(5), 413–434.

Slater, M. (2003). How colourful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *13*(4), 484–493.

Slonim, N. B. (Ed.). (1974). *Environmental physiology*. Saint Louis: C. V. Mosby.

Sutherland, S. (1996). *The international dictionary of psychology*. New York: The Crossroads.

Usoh, M., Arthur, K., Whitton, M., Bastos, R., Steed, A., Slater, M., et al. (1999). Walking > walking-in-place > flying in virtual environments. Proceedings of ACM SIGGRAPH99. pp. 359–364.

Weiderhold, B. K., Gervirtz, R., & Wiederhold, M. D. (1998). Fear of flying: A case report using virtual reality therapy with physiological monitoring. *CyberPsychology and Behavior*, *1*(2), 97–104.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, *7*(3), 225–240.