Predicting the efficacy of simulator-based training using a perceptual judgment task versus questionnaire-based measures of presence

Guy Wallis*
Centre for Sensorimotor Neuroscience
School of Human Movement Studies
University of Queensland
QLD 4072, Australia
and
Queensland Brain Institute
University of Queensland
QLD 4072, Australia
and
Max Planck Institute for Biological Cybernetics
Spemannstraße 38
72076 Tübingen
Germany

Jennifer Tichon
Centre for Sensorimotor Neuroscience
School of Human Movement Studies
University of Queensland
QLD 4072, Australia

Predicting the Efficacy of Simulator-based Training Using a Perceptual Judgment Task Versus Questionnaire-based Measures of Presence

Abstract

The quality of a virtual environment, as characterized by factors such as presence and fidelity, is of interest to developers and users of simulators for many reasons, not least because both factors have been linked to improved outcomes in training as well as a reduced incidence of simulator sickness. Until recently, most approaches to measuring these factors have been based on subjective, postexposure questioning. This approach has, however, been criticized because of the shortcomings of self-report and the need to delay feedback or interrupt activity. To combat these problems, recent papers on the topic have proposed the use of behavioral measures to assess simulators and predict training outcomes. Following their lead, this paper makes use of a simple perceptual task in which users are asked to estimate their simulated speed within the environment. A longitudinal study of training outcomes using two of the simulators revealed systematic differences in task performance that matched differences measured using the perceptual task in a separate group of control subjects. A separate analysis of two standard presence questionnaires revealed that they were able to predict learning outcomes on a per individual basis, but that they were insensitive to the differences between the two simulators. The paper concludes by explaining how behavioral measures of the type proposed here can complement questionnaire-based studies, helping to motivate design aspects of new simulators, prompting changes to existing systems, and constraining training scenarios to maximize their efficacy.

1 Introduction

Improvements in the fidelity and affordability of virtual reality (VR) technology, combined with a growing awareness of its potential applications, have seen a huge increase in its uptake across a broad range of industries, most notably in the field of staff training. This increase has brought with it a demand for a better understanding of how to make use of the technology in novel applications and new industrial settings. In choosing a suitable platform for conducting simulator-based training, a company must consider numerous design issues including the range of sensory feedback required, levels of interaction supported, and simulator mobility. Equally, once the simulator is ready for use,

Presence, *Vol. 22, No. 1, Winter 2013,* 67–85
doi:10.1162/PRES_a_00135

*Correspondence to gwallis@uq.edu.au.

careful thought must be given to the design of training scenarios that make best use of the simulator's strengths and avoid its weaknesses.

Previous research indicates that one way of maximizing learning outcomes using VR is to increase the trainees' sense of presence (Witmer & Singer, 1998). Witmer and Singer and others have offered a variety of definitions for the term presence, but one can think of it as the extent to which "one feels present in the mediated environment, rather than in the immediate physical environment" (Steuer, 1992). Currently, measurement of the overall simulated experience is primarily based on subjective self-reports or physiological measures which have been shown to correlate with presence under some circumstances (Witmer & Singer; Witmer, Jerome, & Singer, 2005; Schubert, Friedmann, & Regenbrecht, 2001; Freeman, Avons, Meddis, Pearson, & IJsselsteijn, 2000; Dillon, Keogh, Freeman, & Davidoff, 2001; Meehan, Insko, Whitton, & Brooks, 2002).

Recent work in a number of different labs has started to question the effectiveness of self-report and physiological measures, and has instead adopted the use of targeted behavioral tasks to yield more sensitive measures of presence and/or better predictors of training efficacy. This paper describes work undertaken to obtain a sensitive, direct, and more objective predictor of training efficacy based on just such a behavioral measure.

In the study reported here, work was carried out in collaboration with a rail company which uses three types of simulator to train their drivers: a wide-screen interactive system; a smaller, interactive simulator with a restricted field of view and lower image resolution; and a video-based presentation. The purpose of this study was to directly assess the ability of presence questionnaires and the novel behavioral test to: (1) predict the short- and long-term training efficacy of a specific simulator; and (2) identify the weaknesses and strengths of training simulators.

1.1 Simulator Quality and Training Outcomes

When designing a simulator and later designing training scenarios for that simulator, what one would ideally like to know is the extent to which experience in that simulator will evoke the desired outcomes and responses in the real world. Knowing whether to include simulator elements such as a motion platform or sound cues directly impacts on construction costs but also has implications for user presence and simulator fidelity. Understanding the true cost of excluding certain cues, or limits to pixel resolution or scene complexity, can help maximize the quality of the virtual experience, and help maximize training outcomes. This knowledge can potentially not only help produce the best simulator for a given budget, but can also help fashion an appropriate training regime by highlighting the circumstances under which the simulator works best. For example, it might help set limits on the simulated speed of a trainee within the environment (both high and low), or set a limit to rates of heading change that are linked to the refresh rate of the system. Current attempts to quantify simulator quality, in terms of presence and fidelity, are reviewed below.

1.2 Presence

As Witmer and Singer (1998) describe, there appears to be a link between an increased sense of presence and the quality of training outcomes. Specifically, the authors suggest that increased presence increases the similarity of the behavior elicited in the virtual environment to that produced in the real environment. We know that various factors contribute to a sense of presence (Welch, Blackmon, Liu, Mellers, & Stark, 1996), and there have been studies of their relative importance (e.g., Lessiter, Freeman, Keogh, & Davidoff, 2001). Despite this, presence remains an imprecisely defined concept. As mentioned above, in rough terms, one might think of it as the extent to which a user is able to suspend disbelief in the simulated environment, or as the ability of a subject to commit attention to the environment (Steuer, 1992; Loomis, 1992; Biocca, 1997; Coelho, Tichon, Hine, Wallis, & Riva, 2006).

Such descriptions are a starting point, but leave presence extremely hard to quantify. Partly as a reaction to its broad and nebulous definition, authors have made use of the related term, immersion, to highlight what

presence is not. Jennett et al. (2008) point out that immersion refers to a tendency for people to become absorbed or engrossed in their work to the exclusion of the outside world. Although commonly associated with a feeling of presence when in a virtual environment, it can also apply to everyday situations such as reading a book, watching a film, or playing a game—especially games requiring rapid reactions (i.e., computer games such as Tetris) or extended periods of concentration (i.e., chess, cards). Activities of this type are characterized by a distorted experience of time relative to the outside world; that is, losing track of time. The experience relates to expressions such as being lost in one's work. Presence, by contrast, is all about a sense of being transported to another place, having been likened to the extent to which "one feels present in the mediated environment, rather than in the immediate physical environment" (Steuer, 1992). For reviews of the concept of presence, including its history and controversy over its definition, see Biocca (1997) and Lee (2004).

For the most part, presence is assessed using postexposure subjective reports, often based on subjective evaluation scales (e.g., Witmer, Jerome, & Singer, 2005; Nowak & Biocca, 2003). Based on the theoretical work of Sheridan (1992), and Held and Durlach (1992), Witmer and colleagues (1998, 2005) developed the Presence Questionnaire (PQ). This questionnaire acts as a tool for assessing how compelling an environment appears to a specific individual, focusing on his or her opinion about a specific simulator. The questions are quite broad-ranging and cover factors that relate to presence and immersion. The authors also suggested a second questionnaire, called the Immersive Tendency Questionnaire (ITQ), which focuses on a participant's general willingness or capacity to engage in an imagined or projected reality of any type (film, story, play, etc.). The ITQ provides a baseline tendency for each person to become immersed in what they are watching or doing and helps scale responses to the PQ for each individual. Schubert et al. (2001) offer an alternative to the two earlier questionnaires called the IGroup Presence Questionnaire (IPQ). The IPQ focuses directly on the degree to which participants feel present in the environment. It also reposes the core questions in several ways to provide a means for checking internal consistency of the user responses. In practice, a combination of the ITQ and PQ questionnaires has been shown to predict training outcomes (Witmer & Singer), and they have gained a significant level of acceptance. Both they and other similar scales have been tested across a number of studies (Schubert et al.; Lessiter et al., 2001).

Ultimately, however, questionnaires are only as reliable as the subjective reports upon which they are based. They also say nothing about cues that supplement learning but which are beyond superficial, personal reflection. Slater and colleagues (Slater & Steed, 2000; Slater, 2004; Slater & Garau, 2007) have pointed out numerous reasons why questionnaires cannot hope to tell the full story of a subject's experience in a virtual environment. In their work, they offered a means of generating a real-time measure of presence without needing to halt the simulation. They proposed asking a subject to verbally report moments at which he or she disconnects from the environment and becomes aware of his or her real surroundings (Slater & Steed). This approach offers a more objective measure that is also arguably much easier for the subject to judge with confidence. The measure is also not clouded by the vagaries of a subject's memory.

The drive to find alternatives to questionnaires that offer both real-time measurement and objectivity has grown. Several labs have become interested in the use of physiological measures, such as cardiac frequency, skin conductance (GSR, galvanic skin response), reflex motor behavior, and event-evoked cortical responses. Authors generally propose that a sign of high presence would be that physiological reactions to the simulated environment are similar to those observed in a real environment. Meehan et al. (2002) reported reliable changes in a number of physiological measures when participants were confronted with the edge of a simulated pit. As these measures were shown to correlate with reported levels of presence, the authors argued that the physiological measures could serve as an objective indicator of presence (at least in threatening situations). Freeman et al. (2000) assessed presence by measuring postural responses, reasoning that compensatory postural changes (e.g., leaning into a corner, bracing during

acceleration) are an indication of natural, immersive behavior, especially in the absence of physical motion stimuli.

While such measures have the potential to circumvent the problems of self-assessment and subjective report, their use remains patchy, and detailed research of their suitability remains scarce (IJsselsteijn, Riddler, Freeman, & Avons, 2000). In the future it might even be possible to monitor presence using modern brain scanning technology, as trialed by researchers using fMRI (Hoffman, Richards, Coda, Richards, & Sharar, 2003), although such an approach is probably only practicable as an aside during specialized, lab-based research, at least for the foreseeable future. More work is required in the area of physiological measures. Ultimately, however, it seems likely that these measures will, at best, only serve to validate large-scale emotional responses. Such responses may well be crucial in desensitization work or stress inoculation (Meehan, Insko, Whitton, & Brooks, 2002; Coelho, Waters, Hine, & Wallis, 2009), but may prove too coarse to measure all aspects of simulator quality.

A final and promising alternative lies in the use of behavioral measures. Such measures have the immediacy of Slater's approach but can also be tuned or targeted to specific issues of direct interest to the trainer. Work in this area suggests that such measures can act as highly sensitive predictors of the level of presence within an environment. Bailenson et al. (2004), for example, found that a subject's behavior (measured via the proximity of hir or her approach to avatars within the environment) could be manipulated via alterations to the status of an avatar as a tutor or a stranger, producing actions comparable to that seen in real-world interactions. However, despite the measurable changes in behavior, direct questioning about numerous aspects of the avatars proved insensitive to their prescribed status.

1.3 Fidelity

A clearly related and yet distinct concept in assessing the quality of a simulator is its fidelity. Fidelity refers to the extent to which a simulator behaves like its real counterpart. The conclusion of the Fidelity Implementation Study Group, as part of the Simulation Interoperability Standards Organization, was that fidelity can be characterized by:

> The degree to which a model or simulation reproduces the state and behavior of a real world object or the perception of a real world object, feature, condition, or chosen standard in a measurable or perceivable manner; a measure of the realism of a model or simulation; faithfulness. (Gross, 1999, p. 3)

This definition is useful in that it highlights how fidelity captures two distinct aspects of simulator quality: The physical characteristics of the simulator (relating to accuracy, sensitivity, precision, resolution, repeatability, etc.), and its perceptual (user-oriented) impact. In practice, the term is often used to refer to image quality which is affected by a number of factors such as the refresh rate of the simulation, resolution (pixel count), render quality (illumination model, texture resolution), and field of view, among other factors. Image quality is important because, as Kemeny and Paneri (2003) point out, visual cues derived from the visual scene are many and varied. They can impart speed, distance, and size information through a number of perceptual mechanisms such as motion parallax, disparity, and eye vergence. The quality and format (monocular/binocular) of the images presented will affect whether veridical information is available to the user through these various cues. The precise range and veridicality of the cues available will not only have an impact on feelings of presence, but may also serve to enhance training in a subconscious/covert manner by introducing sources of information that are integrated into a trainee's representation of the environment without his or her explicit awareness.

Kemeny and Panerai (2003) also highlight the importance of nonvisual cues which drivers can and do use to estimate the state of the vehicle: cues such as speed related rumble, or the sound of the vehicle traversing a textured driving surface, or vestibular information. The absence of such cues may be most noticeable at high simulated speeds and therefore affect presence at these speeds, but may also play an important role in establishing a suitably information-rich environment across an entire range of speeds.

1.4 The Behavioral Task

The central aim of this study is to go beyond the realm of questionnaires and physiological responses in search of a more objective and more broadly applicable predictor of training outcomes. This section discusses a specific approach and the rest of the paper is dedicated to testing its ability to predict the efficacy of simulators currently in use as part of an established rail operator training program.

Given the previous discussion of presence and fidelity, one encouraging avenue to explore is in the use of a behavioral task, one offering intuitive and easy implementation along with demonstrable predictive power. Although not directed explicitly at the issue of training, several tasks have already been successfully developed for testing the fidelity of a simulator. Waller, Beall, and Loomis (2004), for example, demonstrated how pointing within a virtual environment can give much more realistic estimates of spatial orientation ability than abstracted paper-and-pen tests, suggesting it could be used in assessing the accuracy of acquired spatial knowledge of an environment. Likewise, Knapp and Loomis (2004) studied distance perception within a virtual environment using a range of perceptual and behavioral tasks including verbal report, locomotion, and judgment of perceived size. As distance perception is a function of many visual cues (eye vergence, visual disparity, motion parallax, aerial perspective, perceived size, occlusion, etc.) as well as nonvisual cues (e.g., 3D localization and Doppler effect in sound, proprioception in touch), it offers a broad-ranging insight into the fidelity of a simulator.

The work of Bailenson et al. (2004), mentioned in Section 1.2, demonstrates that behavioral approaches are not limited to physical metrics such as distance or speed, but also extend to tracking social behavior. In social scenarios, one might also consider eye-movement characteristics such as gaze time or pupilometry.

While these are all valid approaches, in this study we chose to focus on speed perception. Humans are quite adept as estimating their rate of forward motion, even at unecological speeds; that is, well beyond those for which evolution has equipped us (possibly through the training gained by observation of a speedometer in fast moving, land-based vehicles). One of the many attractions of this measure is that in order to make speed estimates, the brain relies on integrating a range of sensory cues including proprioception (e.g., vibrations), vision (e.g., optic flow, distance perception), audition (e.g., wind, engine sound), and indeed, any cues that correlate reliably with speed (Lappe, Bremmer, & van den Berg, 1999; Blakemore & Snowden, 1999; Kemeny & Panerai, 2003). A number of studies in motor vehicles suggest that in the presence of rich, natural cues, participants perceive speed reasonably accurately, both when estimating current speed as a passenger, and in obtaining a prescribed speed as a driver (Recarte & Nunes, 1996). Much of this work has been concerned with the role of sound cues in estimating speed. Ironically, with the advent of ever quieter engines and insulated cabins, some researchers are concerned that this useful cue is being lost, leading to a potentially dangerous underestimation of speed (Horswill & Plooy, 2008). The amount of variation in the visual cues obtained in a real vehicle can also vary with factors such as terrain (e.g., sharp bends, steep hills) or climactic conditions (e.g., sun, rain, fog). This is of particular relevance to simulator design, in that the best speed estimation is obtained in the presence of high-contrast, high-spatial-frequency images (Distler & Bülthoff, 1996; Snowden, Stimpson, & Ruddle, 1998).

Because of the polysensory nature of speed perception, a task built around the perception of speed may well offer a convenient metric for the assessment of simulator fidelity. The measure should help ensure that the simulator not only offers a range of cues known to affect behavior in real-world scenarios, but also that these cues are of sufficient fidelity to provide accurate information for the perception and estimation of speed. Such a measure should also help guide training programs by identifying those conditions under which the simulator works most effectively; that is, speed ranges over which artifacts introduced by the simulator are minimal or of an acceptable level.

Several groups have already studied various aspects of simulated egomotion using behavioral measures. Siegle, Campos, Mohler, Loomis, and Bülthoff (2009), for example, used continuous pointing to track perceived motion in a simulated environment, describing how it

Table 1. *Specifications of the Three Simulators Used in the Two Studies Reported in This Paper*

	Simulator		
	WS	Cab	Video
User interaction	Yes	Yes	No
Feedback	Visual, haptic	Visual, haptic	Visual
Visual perspective correct	Yes	Yes	No
Visual angle H × V (approx.)	160 × 40°	50 × 40°	30 × 24°
Image resolution	3 @ 1,280 × 1,024	1,024 × 768	1,024 × 768
Natural visual reference frame	No	Yes	Yes
Refresh rate	60 Hz	60 Hz	25 Hz
Scene update rate	30 Hz	30 Hz	25 Hz

offers a more sensitive and continuous measure of both real and virtual simulator-based motion than traditional measures. This is particularly important if one wishes to distinguish between the fidelity of acceleration versus velocity, for example. In a different, but related, vein, Palmisano and colleagues (e.g., Palmisano & Chan, 2004) have used a perceived egomotion-rating task to help identify how low-level scene characteristics contribute to the sensation of egomotion (or, as they term it, vection).

In the study described here, we adopt a similar approach, but based on actual speed estimates rather than a rating scale. The study takes three types of simulators and tests whether speed perception is a sufficiently sensitive measure to determine differences between them. It then looks at the ability of the measure to predict the effectiveness of training using the simulators as compared with traditional questionnaire approaches. The basic approach will be to compare the ability of the speed test and questionnaire outcomes to predict training outcomes for a separate group of individuals taking part in a real-world appraisal and training program.

2 Environment

2.1 Industry Partner

The studies described here were conducted in collaboration with an Australian rail company which makes extensive use of simulators to train its drivers and guards. The company provides passenger rail transport and is responsible for the safe operation, staffing, and maintenance of passenger trains and stations. It also owns and maintains a metropolitan rail network and provides access to freight operators in the metropolitan area. To provide safety training across its 15,000 widely distributed personnel, the company utilizes a high-fidelity, large-screen simulator and two in-cab simulators. These simulators are centrally located at their training college. None of these simulators is in any way portable, restricting the use of such machines to those staff who can be brought in from around the network. Bringing in staff can be both logistically and financially restrictive, and so the company also makes use of a portable, video-based system which can be taken on the road to staff based across the state. Although the use of video-based training is convenient, it restricts training to passive observation rather than interactive simulation.

2.2 Simulators

The three types of training devices based at the training site (the wide-screen simulator, the cab-based simulator, and the video) are described below and a summary of relevant specifications appears in Table 1.

2.2.1 Wide-Screen Simulator (WS). The largest and most sophisticated simulator offers a large, curved viewing surface, viewed from a control desk which is configured to reproduce the controls found in

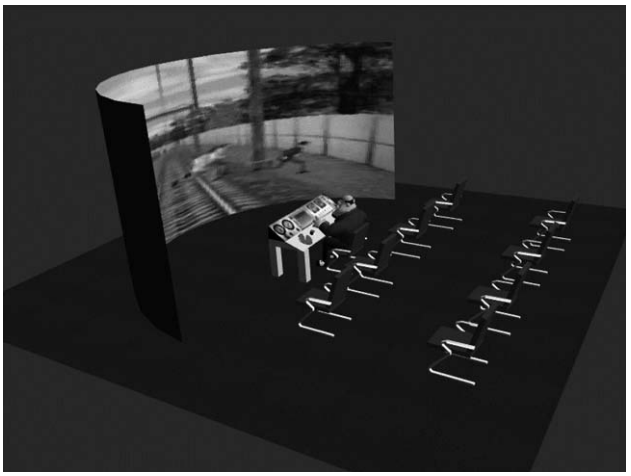

Figure 1. *General layout of the wide-screen simulator. The curved screen affords a large lateral field of view. The driver's console was open and took the form of a control desk.*

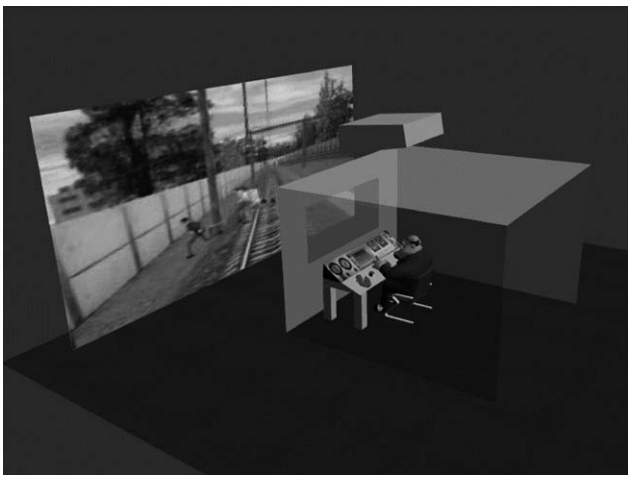

Figure 2. *General layout of the cab-based simulator. The field of view was smaller, but framed by a realistic engine cab window. Controls were modeled on a real vehicle, as were all feedback gauges and dials.*

one of a series of engines run on the company's network, as shown in Figure 1. The simulator includes a force-feedback control lever for regulating speed (both acceleration and braking). Force feedback is modeled on real vehicle behavior, namely, a push and stay control with resistance, requiring the driver to actively push or pull the lever into position, but not related to the actual acceleration or braking effort of the vehicle itself. The display had three SXGA BARCO projectors.

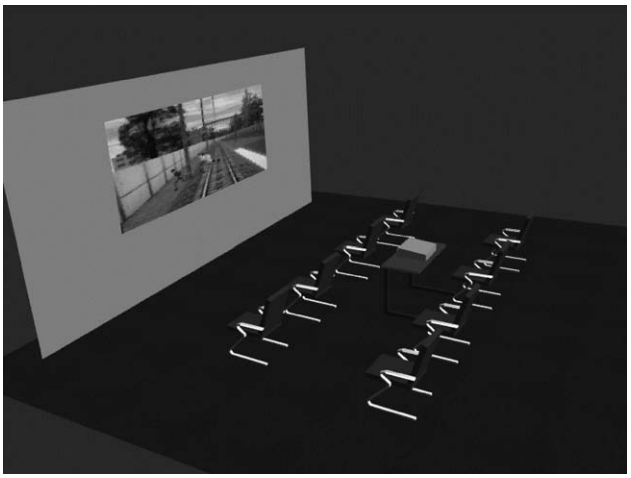

Figure 3. *General format of video presentations. These were conducted in teaching labs with a desk-mounted projector. No effort was made to create correct perspective for the viewers and there was no interaction, only passive viewing of prerecorded runs along a section of track.*

The large field of view provides a strong stimulus to motion-detection systems in the human eye that are highly sensitive to peripheral stimulation (Warren & Kurtz, 1992). Such motion cues have been widely implicated in heading and speed estimation (Lappe et al., 1999).

2.2.2 Cab-based Simulator (Cab). The other two simulators are smaller, full-cab devices with viewing restricted to a flat, frontal portion of the simulated environment, as shown in Figure 2. Although smaller and offering a relatively narrow field of view, the cabs' screen layout provides a naturalistic reference frame, something lacking from the large screen system. The driving interface is very similar to that of the WS simulator, incorporating the same style of realistic force feedback in the speed regulation handle. The display is shown using an XGA BARCO projector.

2.2.3 Video. As described in Section 2.1, the rail operator also makes use of video presentations, as shown in Figure 3. These consist of a wall-projected recording of the simulated environments utilized in the interactive simulators. The video footage was displayed using a standard XGA video projector to project an image with the same aspect ratio as the cab-based simulator, but

with a smaller image, and viewers sat farther from the screen.

3 Study 1: Perceived Speed Test

3.1 Introduction

The main premise of this paper is that speed perception is intrinsically linked to the overall quality of a simulated reality, both in terms of presence and fidelity. The more completely the simulation reproduces a range of sensory input, the better the impression of forward motion and the more precise a user's estimates should be. This seems intuitively appealing, but if the measure is to be of practical use, it should possess several properties. First, it should be consistent across participants (all participants should perform poorly in a low-quality simulator relative to their performance in a high-quality simulator). Second, the measure should be sensitive enough to discriminate variation in the quality of the simulator. This first study includes an experiment designed to test both of these requirements by taking drivers and testing their speed estimates across a range of speeds using all three training devices (WS, Cab, Video).

3.2 Methods

3.2.1 Participants. Twelve expert train drivers, aged from 27 to 52 years (mean 37.8, *SD* 7.6), took part in the experiment. Visual acuity was not explicitly tested, but all drivers were licensed to drive. In order to be licensed, drivers are required to pass an exam for visual acuity every five years, achieving 6/9 vision in at least one eye and no worse than 6/18 in the other eye. Over the age of 50, the regularity of testing increases to every two years. All of the drivers tested had at least 10 years of driving experience with an average of 17.2 years.

3.2.2 Task. Drivers were asked to passively view a section of track being negotiated at a fixed speed. Their task was to estimate their current speed in the absence of any instrument readouts. Four simulated speeds (20, 40, 60, and 80 km/h) were used and they were presented in pseudorandom order, 12 times for each speed, making a total of 48 trials. The experiment was run a total of four times with the order of test speeds varied under each repetition, producing a total of 192 trials. All 12 drivers were exposed to all three training devices, but the order in which they were tested was counterbalanced, in order to counteract any learning effects.

In each trial the drivers were permitted to view the environment for a 5-s period, during which they were required to write down their estimated speed in a response table. The delay between trials was variable and random but was at least 5 s. This random delay helped reduce biases that might have arisen if a correlation existed between a specific intertrial delay and the difference in speed between specific trials.

After completing the speed-perception task in either of the two simulators (but not after the video presentation), drivers were requested to fill out two questionnaires relating to perceived presence: The Presence Questionnaire (version 3; Witmer et al., 2005), and the IGroup Presence Questionnaire (Schubert et al., 2001). The exact questions posed appear in the Appendix. They differed slightly from those outlined in the two papers cited on the basis of relevance and recent findings. For example, questions 26, 27, and 28 of the PQ were not used in the analysis as these items have previously been found to reduce reliability (Witmer et al.). Both the IPQ and PQ were subjected to extensive statistical study, yielding evidence for a consistent grouping of responses into underlying contributory factors (Witmer et al.; Schubert et al.). Drivers exposed to the video presentation were not asked to fill out questionnaires as the majority of questions are not relevant.

3.2.3 Behavioral Results. The entire experiment was analyzed with ANOVA using a 3×4 within-subjects design. Simulator (wide-screen, cab, video) and speed (20, 40, 60, 80) were independent variables, and speed estimation error (in km/h) was the dependent variable. All post hoc analyses were performed using Tukey's HSD test (see Howell, 1997). The analysis revealed significant main effects for both independent variables: Simulator type $F(2, 22) = 6.244$, MSe $= 127.97$, $\eta_p^2 = 0.36$, $p < .01$, and Speed $F(3, 33) = 16.30$, MSe $= 43.24$, $\eta_p^2 = 0.60$, $p < .001$, as well as a significant interaction $F(6, 66) = 3.846$, MSe $= 13.9$, $\eta_p^2 = 0.26$, $p <$

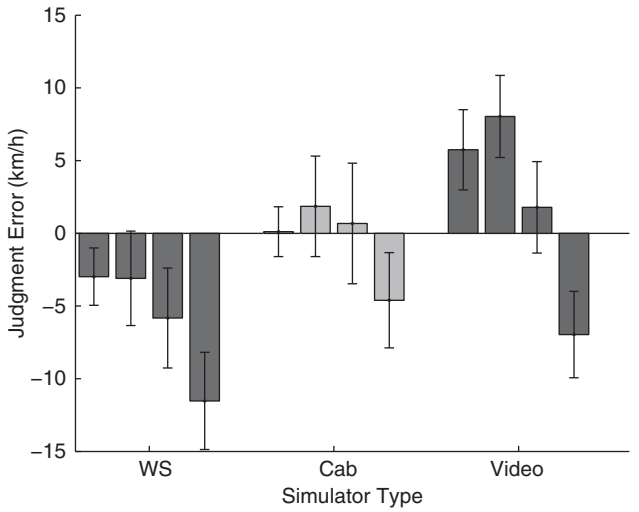

Figure 4. *Speed estimate error expressed in km/h separated into performance at each test speed (20, 40, 60, and 80 km/h). The three bars represent data for the three simulator types (WS: Wide-screen, Cab: Enclosed cab, and Video). Error bars represent the standard error of the mean.*

.01. The results of the experiment appear in Figure 4. It is clear that speed estimation was particularly poor at 80 km/h and post hoc analysis revealed that it was indeed significantly different from all other speeds at the $p < .05$ level. No significant differences in performance emerged between any of the other speeds.

Post hoc analysis also confirmed what is apparent from the figure, namely, that average performance in the wide-screen simulator was significantly worse overall than in the other two simulators ($p < .05$). However, this needs to be interpreted in the context of the significant interaction between speed and simulator type. The errors in the wide-screen simulator are consistently underestimates, leading to a large average error. Errors in the video presentations are large but their sign varies with speed, starting with overestimates and ending with underestimates, yielding a misleadingly small average error. In contrast, errors in the cab simulator remain consistently small across a range of velocities. In fact, averaging the magnitudes of the errors across speeds reveals that the wide-screen and video presentations produced a similar error of just under 6 km/h, which was nearly double that recorded in the cab simulator.

Table 2. *Questionnaire Rating Results for the Two Simulators for the Drivers Also Involved in the Speed Test**

	Simulator			
	Wide-screen		Cab	
Questionnaire	Mean (*SD*)	%	Mean (*SD*)	%
PQ (Overall)	120.8 (7.34)	69.5	112.6 (17.5)	64.3
IPQ	57.5 (2.6)	63.37	50.5 (12.2)	55.5

*Despite a few small discrepancies, the overall scores for the two simulators were both moderate. Although not statistically significant, there was a tendency for the Cab sim to be rated lower on the IPQ than the WS simulator.

3.2.4 Questionnaire Results. Driver responses to the two questionnaires are presented in Table 2. It appears from the results that the drivers regarded both simulators as somewhat compelling, but not fully immersive. The overall scores for the PQ on both the Cab and Wide-screen simulators were very similar and a paired *t*-test confirmed that they were not statistically distinguishable, $t(11) = -1.26$, n.s. The differences on the IPQ were also not statistically significant, although there was a marginal trend toward the Cab simulator receiving lower scores, $t(11) = -1.99$, $p = .073$.

3.3 Conclusions

The results indicate that the speed perception measure is sensitive to differences in a user's simulated experience. The measure was able to discriminate performance in the three training devices and revealed an unexpected effect in the large-screen system, namely, that observers tended to underestimate their speed despite remarkably good performance in the same task when conducted in the cab simulators. In contrast, if they revealed any difference at all, the presence questionnaires favored the WS simulator over the Cab simulators.

One plausible explanation for the apparently counterintuitive result is the lack of a reference frame in the large-screen simulator. The presence of two fixed-reference

edges near the center of vision may well offer a cue for experienced drivers that is lacking in the large simulator. Another possible source of problems is the relatively low simulator refresh rate of 30 Hz. At lower refresh rates, high rotational speeds (negotiating a tight bend at slow speeds or traveling quickly around shallow bends) can result in jerky image motion. A wide-screen display exaggerates these effects in the peripheral visual field, even at relatively low forward speeds. Since the drivers did not drive the train in a jerky manner and did not experience jerky body motion (the simulators did not use a motion platform), this may have enhanced a sense of disconnect between the drivers' actions and their motion through the environment. Indeed, as Lessiter et al. (2001) describe, with respect to feeling present in an environment, the quality and size of a graphic display can be less important than the level of interaction and control that a user has.

4 Study 2: Driving Task

4.1 Introduction

The speed test uncovered differences in the ability of experienced drivers to estimate their speed across the three training environments. Although this satisfies the requirement that the test be both consistent and sensitive, it would be instructive to discover how this translates to training outcomes, not least because one of the results appears to be counterintuitive: the smaller, cab-based simulator appears to produce better estimates than the wide-screen simulator. This section summarizes the findings of a longitudinal field study carried out using the two interactive simulators, conducted as part of a standard annual driver training program. The main aim is to see whether the speed test results or the presence questionnaire results can predict training outcomes in a separate group of drivers both immediately after training and also one year later. Note that video-based training did not form part of the program and hence results are only reported for the two simulators.

4.2 Methods

4.2.1 Participants. The studies made use of currently active drivers undertaking their annual safe-working simulator training for that calendar year. In the first year, 12 participants were trained using the wide-screen simulator, while 51 were trained using the cab simulator. This proportion was determined simply by the practicalities of simulator availability. Since there are two cab-based simulators and only one wide-screen simulator, it was possible to run trainees more efficiently through the cab-based simulators. The choice of which trainee was assigned to which simulator was random. Trainees were rostered beforehand and simply selected when a simulator became free. In order to balance group size, data from 12 of the 51 drivers were selected at random for later analysis. A little over 12 months later, 42 of the original drivers were brought back to the simulation training center for retesting as part of their annual appraisal. Analysis focused on the same group of drivers selected for testing in the previous year. Note that none of the drivers involved in Study 1 participated in Study 2.

4.2.2 Task. The overall purpose of the field study was to assess the effectiveness of training in terms of enhanced decision-making under stress. To achieve this, comparisons among performance outcomes were made over two sessions over the period of one year. All drivers drove in one of the two interactive simulators for 40 min, while trainers completed a checklist of correct actions and errors. The assessors themselves sat outside the simulators, observing behavior via a video monitor. The session involved a range of complex, taxing events that occurred during an everyday run to collect passengers from a station. A simulated worksite necessitated observance of pedestrian signalers and temporary speed restrictions. A later incident involved a failed level crossing, which again required appropriate observance of speed restrictions. A third event involved taking appropriate action when encountering school children trespassing on the track. Each of these events was further complicated through increasing the driver's workload pressure by rapidly changing operational conditions. During retest (a year after initial training), the identical testing scenario was used to gauge retention of the original learning.

After completion of training and testing in the first year, drivers were also asked to complete the PQ and IPQ questionnaires used in Study 1.

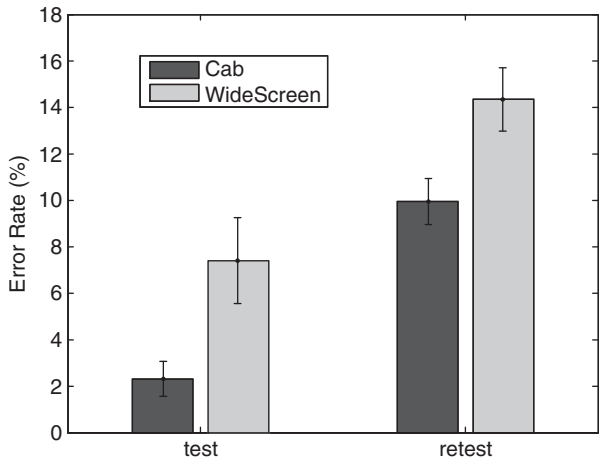

Figure 5. *Summary of driver performance in the two simulators. Data in the test condition were obtained immediately after initial training, and the retest data were obtained one year later. Error bars represent the standard error of the mean.*

4.3 Results

4.3.1 Task. The results from the experiment appear in Figure 5. A two-way, mixed-design ANOVA was performed with Simulator Type and Test Number (i.e., either test, immediately after initial training; or retest, on the same scenario a year later) as independent variables, and the error rate as the dependent measure. Error rates were based on the assessor's score for each driver over the 40-min test session. The analysis revealed a significant main effect of Simulator Type, $F(1, 22) = 13.45$, MSe $= 20.09$, $\eta^2 = 0.38$, $p < .05$. Consistent with the speed perception findings, this difference was due to greater error rates for drivers using the wide-screen simulator (10.9%) than the cab-based simulator (6.1%). There was also a main effect of year in which the test was conducted (test vs. retest), $F(1, 22) = 30.51$, MSe $= 20.91$, $\eta^2 = 0.58$, $p < .05$, due to a drop in performance (increase in error rate) of around 5% over the intervening year across both simulators. There were no other significant effects.

4.3.2. Presence Questionnaires: Group Effects. Questionnaire data from the two groups of drivers are presented in Table 3. The overall scores on both simulators and across both questionnaires are indistinguishable and they are broadly similar to the results obtained in Study 1. That said, the ratings are more similar across simulators, removing evidence for the minor trends reported earlier. For the PQ, $t(22) = 0.61$, n.s., and IPQ, $t(22) = -0.23$, n.s. Hence, it appears that even when asking the actual trainees themselves to rate presence, the questionnaires are unable to predict the difference in training outcomes seen across the two simulators.

The reliability of the two types of questionnaire has been tested before. The IGroup questionnaire, in particular, has been thoroughly tested for reliability and validated through comprehensive factor analysis, with a Cronbach's alpha (α) of .85. We took the opportunity to conduct our own analysis to verify whether the responses of our drivers were also largely consistent. As these measures are sensitive to small sample sizes, we included data from all of our participants in the analysis (Cab: 51, WS: 12). Table 4 presents the measures of reliability calculated using Cronbach's alpha. For both simulators, the value of α for the Presence Questionnaire was high (over 0.7), indicating a highly consistent set of responses. The IPQ data was more mixed, with high reliability for the Cab simulator but more variability in the WS simulator. This variability was mainly due to major discrepancies across subjects driven by inconsistent responses to questions 6, 7, and 13. Of course reliability measures are ordinarily calculated on much larger sample sizes. The relatively small number of drivers who used the wide-screen simulator makes estimating reliability for that simulator difficult, as the measure is highly sensitive to even one or two discrepancies in a single subject's rating behavior. This is particularly true for the IPQ, which contains roughly half the number of questions contained in the PQ.

Nonetheless, knowing that reliability is an issue in the IPQ with the three questions listed, it is possible to go back to the original analysis and see whether excluding these questions alters the conclusions one would draw. The only tangible difference was that the nonsignificant effect for the drivers' responses to the IPQ in Study 1 moved from being marginally significant to being significant: IPQ, $t(22) = -2.57$, $p < .05$, Cohen's $d = -.74$, underlining the fact that if the IPQ detects any difference

Table 3. *Questionnaire Rating Results for the Two Simulators for the Trainees (Study 2)**

	Simulator			
	Wide-screen		Cab	
Questionnaire	Mean (*SD*)	%	Mean (*SD*)	%
PQ (Overall)	121.0 (7.43)	69.1	124.6 (19.07)	71.2
IPQ	59.0 (1.71)	64.8	58.2 (10.97)	64.0

*Despite a few small discrepancies, the overall scores on each simulator were indistinguishable.

Table 4. *Reliability Measures for the Two Questionnaires and Two Simulator Types*

	Questionnaire	Cronbach's alpha α	Number of items
Cab simulator	PQ	0.893	25
	IPQ	0.78	13
Wide-screen simulator	PQ	0.705	25
	IPQ	−0.16	13

between the simulators at all, it is in the opposite direction from that predicted by the speed test; that is, the WS simulator is seen as better that the Cab simulator. This difference was in the same direction for the drivers used in the training program of Study 2, but even after exclusion of the problematic questions, the difference remained far from achieving significance in their data, $t(22) = -0.41$, n.s.

4.3.3 Presence Questionnaires: Individual Effects.

As mentioned in the Introduction, one of the reasons that users of VR technology have been interested in presence is the suggestion that it is linked to learning outcomes. In fact, the evidence for this remains mainly limited to small-scale laboratory experiments, and the results, although promising, have not always been consistent (Witmer & Singer, 1998; Mania & Chalmers, 2001). The study conducted here offers an opportunity to assess the relationship in a practical, applied setting. To that end, a correlation analysis was conducted between driver performance outcomes (measured as error rate) for the 30 drivers who used the Cab simulator, and their reported level of presence (measured using the PQ and IPQ questionnaires) a year after initial training. The results of the analysis are shown in Figure 6. The analysis detected a significant negative correlation, $r^2 = 0.136$ (slope of −4.9, $p < .05$), for the PQ results. For the IPQ, a similar negative correlation emerged, although this did not achieve statistical significance, $r^2 = 0.094$ (slope −5.18, n.s.). A negative correlation is precisely what such authors as Witmer and colleagues would have predicted, since it suggests that performance improves (lower error rate) as presence increases. Although the correlations are modest, it is worth point-

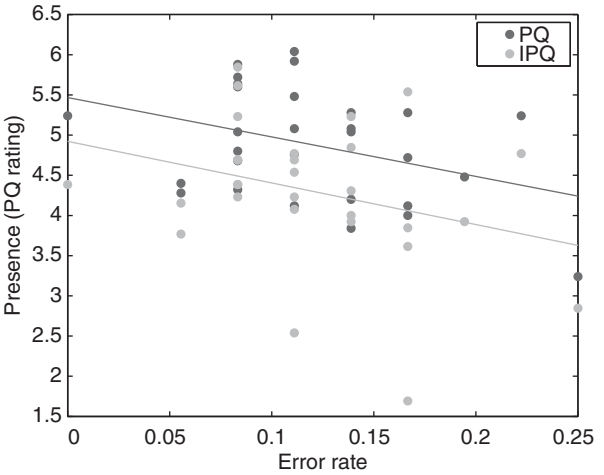

Figure 6. *Correlation analysis of driver performance a year after initial training (error rate), versus their subjective impression of presence, as measured using the IPQ and PQ questionnaires.*

ing out that no attempt was made to scale the questionnaire results from each individual in accordance with general trends or tendencies they might have (q.v. earlier discussion of the ITQ). The fact that correlations have emerged for just 30 individuals answering 25 (PQ) or 13 (IPQ) questions is certainly suggestive of a link between presence, performance, and retention.

One possible problem with interpreting the correlation results is that they may be due to one of two effects, namely, differential levels of recall of earlier training, or differential levels of skill in operating the simulator. Fortunately, we were provided with an opportunity to partially tackle this issue. After the drivers had completed their test a year after original training, they were then retrained (using the scenario they had originally experienced a year earlier) and retested. We took the opportunity to record their new performance levels in order to tease out the source of the original correlations between questionnaire and performance data. In this case, no evidence for correlation between presence ratings and error rate appeared, with a correlation $r^2 = 0.00$ (slope = 0.25, n.s.) for the PQ results, and $r^2 = 0.036$ (slope = -5.1, $p = .032$, n.s.) for the IPQ. It appears, therefore, that drivers who reported low levels of presence in the simulators were able to perform well in the simulator when that training was fresh in their minds. What appeared to be compromised was their ability to retain details of that training a year later, after months of real-vehicle experience.

4.4 Conclusions

The training results summarized above reveal clear discrepancies between the two simulators in terms of performance. These discrepancies mirror the results from the speed test. In contrast, if the questionnaires revealed any evidence for any differences between the two simulators at all (Study 1), it was in the opposite direction from that of the performance outcomes.

What the questionnaires did seem able to capture was long-term retention of learning in individuals using a specific simulator. Hence, the questionnaires were able to predict the efficacy of training within a specific simulator. As mentioned earlier, other authors have reported a link between presence and performance (Witmer & Singer, 1998), but there are also studies that question this link (e.g., Slater, Usoh, & Kooper, 1996). Both Slater, Usoh, and Kooper (1996) and Bowman and McMahan (2007) highlight the usefulness of presence in instantiating natural responses in trainees and of the increased likelihood of transfer of learning to real-world experiences. But in terms of task performance, both sets of authors place more emphasis on the concrete role of simulator fidelity, describing how it is often (though not always) a significant factor in user performance. In many ways, one could argue that both the PQ and IPQ are not restricted to capturing presence, but that they also capture elements of simulator fidelity and immersion, albeit in a subjective manner (see Section 1.2). Hence it may be premature to conclude that presence per se is a predictor of outcomes, but it does appear that scores relating to subjective feelings of immersion, presence, and simulator fidelity do predict task performance and long-term learning in this case.

5 Discussion

The speed test has been trialed as a predictor of simulator performance. It was seen to be a sensitive measure (it detected differences between simulators) and also a consistent one (differences were sufficiently similar across participants to produce a statistically reliable effect). Of equal importance, results were consistent with training outcomes obtained in a separate set of participants, both in terms of the participants' immediate and longer-term performance.

The test relies on a simple behavioral measure that is intuitive for participants to perform and appears to be something that the drivers of trains can perform well if given appropriate cues. Slater (2004) and Slater and Garau (2007), among others, have argued that presence questionnaires are too abstracted from the task to provide meaningful data and at best should be supplemented with alternative, more objective/concrete measures. We would tend to agree. Along with authors such as Bailenson and colleagues (2004), we would argue that behavioral tasks offer a promising alternative means of tracking important aspects of the simulated experience

(both explicit and implicit) with concrete implications for simulator-based training.

As it stands, the current study cannot say precisely why the wide-screen or video-based simulators produced worse training outcomes than the cab-based simulator. We have speculated as to the importance of having a reference frame and to shortcomings of the simulation refresh rate, which may have disproportionately affected the wide-screen display, but without a more systematic study of performance with and without a frame, or at different refresh rates, the exact cause of the discrepancies in performance remain unknown. The relationship between the speed perception test and training outcomes has also only been measured indirectly, relying on scores from one group to predict the performance of a separate set of trainees. One advantage of this approach is that it speaks to the fact that design decisions made during commissioning of the simulator are relevant to the outcomes of later users of the equipment. The downside is that the link between the speed perception test and training outcomes is only suggestive at this stage. Future work should attempt to test the link more directly using a within-subjects design.

The ability of the speed test to predict training outcomes suggests it may be useful in a number of situations. Where learning is poor, for example, it may well help motivate changes to the setup used. In the case of the video presentation studied in the first experiment, altering the viewing distance to suit the recorded viewing perspective, or improvements to video image quality, may well help. For the cab simulator, training at speeds at and above 100 km/h may well benefit from increased sensory input such as cab motion. These suggestions are simply speculative but the speed perception test can provide a sensitive and more objective guide to a simulator's limitations, and help both motivate and test future modifications.

The discussion of speed perception in Section 1.4 offers numerous reasons why it provides a convenient, broadly relevant test due to its sensitivity to a wide range of visual and nonvisual information. It is also relatively easy to administer without the need for further specialist equipment. There are, however, many simulation and training scenarios in which there is no motion or motion over a limited range of speeds. In practice, speed perception is just one of a myriad of possible measures that could be used. As described earlier, distance perception relies on the integration of a large range of cues and lends itself to a range of behavioral measures such as pointing, navigation, and verbal report. For systems that involve arm/hand tracking, such as data gloves, it would be possible to integrate nonvisual cues such as proprioception (i.e., one's sense of the position of one's body parts) into distance estimation for objects within reaching/pointing range. Basic sound localization tasks (including pointing or placing a virtual pointer in space) could likewise be used for assessing the quality of 3D sound generated through auditory equipment. The speed test itself should be further investigated by measuring which improvements to a simulator affect performance in the test, and whether improvements in test performance continue to be reflected in training outcomes. It might also be advantageous to move beyond speed perception to velocity perception (which, strictly speaking, includes a directional component), as this may prove to be a more sensitive measure, with implications for issues such as simulator sickness.

Like distance perception, speed perception can be measured using an array of different tasks other than the verbal report method used here. One might consider an active speed "matching task" in which drivers are asked to attain a specific speed rather than guess their current speed. This would have the advantage that it is closer to the real-world task they have to perform and could help avoid any response biases that might occur when using a small set of discrete test speeds. Three-interval forced-choice paradigms also offer a sensitive measure for the detection of changes in speed, where noticing a change in speed or direction is at least as important as being able to judge speed per se, such as in many sports (Müller et al., 2010) or in piloting an aircraft (Previc & Ercoline, 2004).

What neither the speed test, nor indeed any other basic psychophysical measure, can tell an investigator is whether training errors are due to a lack of presence, or rather to misperceptions caused by failings in simulator fidelity. Measures that speak more directly to the issue of immersion or presence (questionnaires, physiological

measures, Slater et al.'s BIPs) can continue to play an important role in this regard. Hence, it is not envisaged that the speed test or other behavioral measures will supplant all existing measures, but rather supplement them. Presence remains an important separate quantity because it is an indicator of how immersed the trainee feels, which can be important for evoking genuine emotional and other higher-level cognitive responses. This sense of involvement is likely to increase the extent to which trainees become prepared for a broad range of events beyond the scope of any particular training scenario. It is also important to reiterate that although the PQ and IPQ questionnaires were unable to detect shortcomings of the wide-screen simulator, they both successfully predicted retention of learning a year later and hence the long-term efficacy of training. It is also possible, of course, that a larger sample size might have allowed the IPQ and PQ to detect the differences between the two simulators, although given the size and direction of the minor effects reported here, current evidence suggests not.

The main strength of the perceived speed test is that it can help monitor discrepancies between the virtual and real-world experience, and warn trainers that despite reports of user presence, the learning may transfer poorly. Ideally, a simulator should aim to produce high ratings on both introspective (e.g., presence questionnaire) and perceptual (e.g., speed test) measures.

References

Bailenson, J. N., Aharoni, E., Beall, A. C., Guadagno, R. E., Dimov, A., & Blascovich, J. (2004). Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. *Proceedings of the 7th Annual International Workshop on Presence*.

Biocca, F. (1997). The Cyborg's Dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication, 3*(2). Retrieved from http://jcmc.indiana.edu/vol3/issue2/biocca2.html

Blakemore, M. R., & Snowden, R. J. (1999). The effect of contrast upon perceived speed: A general phenomenon? *Perception, 28*, 33–48.

Bowman, D. A., & McMahan, R. P. (2007). Virtual reality: How much immersion is enough? *Computer, 40*, 36–43.

Coelho, C. M., Tichon, J. G., Hine, T. J., Wallis, G., & Riva, G. (2006). Media presence and inner presence: The sense of presence in virtual reality technologies. In G. Riva, M. T. Anguera, B. K. Wiederhold, and F. Mantovani (Eds.), *From communication to presence (*pp. 25–45). Amsterdam: IOS Press.

Coelho, C. M., Waters, A. M., Hine, T. J., & Wallis, G. (2009). The use of virtual reality in acrophobia research and treatment. *Journal of Anxiety Disorders, 23*(5), 563–574.

Dillon, C., Keogh, E., Freeman, J., & Davidoff, J. B. (2001). Presence: Is your heart in it? *Proceedings of the 4th Annual International Workshop on Presence,* 21–23.

Distler, H., & Bülthoff, H. H. (1996). Velocity perception in 3-D environments. *Perception,* 25 ECVP Abstract Supplement.

Freeman, J., Avons, S., Meddis, R., Pearson, D., & IJsselsteijn, W. (2000). Using behavioral realism to estimate presence: A study of the utility of postural responses to motion stimuli. *Presence: Teleoperators and Virtual Environments, 9*(2), 149–164.

Gross, D. C. (1999). *Report from the Fidelity Implementation Study Group:* Simulation Interoperability Standards Organization. Retrieved from http://www.sisostds.org/

Held, R. M., & Durlach, N. I. (1992). Telepresence. *Presence: Teleoperators and Virtual Environments, 1*(1), 109–112.

Hoffman, H. G., Richards, T., Coda, B., Richards, A., & Sharar, S. R. (2003). The illusion of presence in immersive virtual reality during an fMRI brain scan. *CyberPsychology & Behavior, 6*, 127–131.

Horswill, M. S., & Plooy, A. M. (2008). Auditory feedback influences perceived driving speeds. *Perception, 37*, 1037–1043.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Wadsworth.

IJsselsteijn, H., Riddler, J., Freeman, J., & Avons, S. E. (2000). Presence: Concept, determinants and measurement. *Proceedings of SPIE, Human Vision and Electronic Imaging.*

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human–Computer Studies, 66*, 641–661.

Jung, D., Jo, S., & Myung, R. (2008). A study of relationships between situation awareness and presence that affect performance on a handheld game console. *Advances in Computer Entertainment Technology,* 240–243.

Kemeney, A., & Panerai, F. (2003). Evaluating perception in driving simulation experiments. *Trends in Cognitive Sciences, 7*, 31–37.

Knapp, J. M., & Loomis, J. M. (2004). Limited field of view of head-mounted displays is not the cause of distance underestimation in virtual environments. *Presence: Teleoperators and Virtual Environments, 13*(5), 572–577.

Lappe, M., Bremmer, F., & van den Berg, A. V. (1999). Perception of self-motion from visual flow. *Trends in Cognitive Sciences, 3*(9), 329–336.

Lee, K. M. (2004). Presence, explicated. *Communication Theory, 14*, 27–50.

Lessiter, J., Freeman, J., Keogh, E., & Davidoff, J. (2001). A cross-media presence questionnaire: The ITC Sense of Presence Inventory. *Presence: Teleoperators and Virtual Environments, 10*(3), 282–297.

Loomis, J. M. (1992). Distal attribution and presence. *Presence: Teleoperators and Virtual Environments, 1*(1), 113–118.

Mania, M., & Chalmers, A. (2001). The effects of levels of immersion on memory and presence in virtual environments: A reality centered approach. *CyberPsychology & Behavior, 4*(2), 247–264.

Meehan, M., Insko, B., Whitton, M., & Brooks, F. P. (2002). Physiological measures of presence in stressful virtual environments. *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '02,* 645–652.

Müller, S., Abernethy, B., Anderson, T., Eid, M., McBean, R., & Rose, M. (2010). Expertise and the spatio-temporal characteristics of anticipatory visual information pick-up from complex movement patterns. *Perception, 39*, 745–760.

Nowak, K. L., & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 12*(5), 481–494.

Palmisano, S., & Chan, A. Y. C. (2004). Jitter and size effects on vection are robust to experimental instructions and demands. *Perception, 33*(8), 987–1000.

Previc, F. H., & Ercoline, W. R. (2004). Spatial disorientation in aviation. *Progress in Astronautics and Aeronautics.*

Recarte, M. A., & Nunes, L. M. (1996). Perception of speed in an automobile: Estimation and production. *Journal of Experimental Psychology: Applied, 2*, 291–304.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments, 10*(3), 266–281.

Sheridan, T. B. (1992). Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments, 1*(1), 120–126.

Siegle, J. H., Campos, J. L., Mohler, B. J., Loomis, J. M., & Bülthoff, H. H. (2009). Measurement of instantaneous perceived self-motion using continuous pointing. *Experimental Brain Research, 195*, 429–444.

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 13*(4), 484–493.

Slater, M., & Garau, M. (2007). The use of questionnaire data in presence studies: Do not seriously Likert. *Presence: Teleoperators and Virtual Environments, 16*(4), 447–456.

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

Slater, M., Usoh, M., & Kooper, R. (1996). Immersion, presence and performance in virtual environments: An experiment with tri-dimensional chess. *VRST: ACM Symposium on Virtual Reality Software and Technology,* 163.

Snowden, R. J., Stimpson, N., & Ruddle, R. A. (1998). Speed perception fogs up as visibility drops. *Nature, 392*, 450.

Steuer, J. S. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication, 4*, 73–93.

Waller, D., Beal, A. C., & Loomis, J. (2004). Using virtual environments to assess directional knowledge. *Journal of Environmental Psychology, 24*, 105–116.

Warren, W. H., & Kurtz, K. J. (1992). The role of central and peripheral vision in perceiving the direction of self-motion. *Perception & Psychophysics, 51*, 443–454.

Welch, R. B., Blackmon, T. T., Liu, A., Mellers, B. A., & Stark, L. W. (1996). The effects of pictorial realism, delay of visual feedback, and observer interactivity on the subjective sense of presence. *Presence: Teleoperators and Virtual Environments, 5*(3), 263–273.

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

Witmer, B. G., Jerome, C. J., & Singer, M. J. (2005). The factor structure of the presence questionnaire. *Presence: Teleoperators and Virtual Environments, 14*(3), 298–312.

Appendix A

The questionnaires used in our study were based on the Presence Questionnaire (version 2; Witmer et al., 2005) and the IGroup Presence Questionnaire (Schubert et al., 2001). Following the example of Jung, Jo, and Myung

(2008), we chose to exclude certain questions relating to haptic interfaces; we also reworded some questions slightly so as to make them directly relevant to the simulators being used. One or two questions were omitted. Below is a summary of questions that we selected from the original Presence Questionnaire with amendments and exclusions explained in parentheses below those questions that were altered.

A.1 Presence Questionnaire (PQ)

1. How much were you able to control events?
2. How responsive was the environment to actions that you initiated/performed?
   (How responsive was the simulator to actions that you initiated/performed?)
3. How natural did your interactions with the environment seem?
4. How much did the visual aspects of the environment involve you?
5. How much did the auditory aspects of the environment engage you?
   (How much did what you could hear in the environment engage you?)
6. How natural was the mechanism which controlled movement through the environment?
   (How natural did it feel to use the simulator to move through the environment?)
7. How compelling was your sense of objects moving through space?
   (How compelling was your sense of objects moving through the scene?)
8. How much did your experiences in the virtual environment seem consistent with your real-world experiences?
9. Were you able to anticipate what would happen next in response to the actions you performed?
10. How completely were you able to actively survey or search the environment using vision?
    (How completely were you able to actively look around or search the environment visually?)
11. How well could you identify sounds?
12. How well could you localize sounds?
13. How well could you actively survey or search the virtual environment using touch?
    (Not relevant to train simulator.)
14. How compelling was your sense of moving around inside the virtual environment?
    (How compelling was your sense of moving through the virtual environment?)
15. How closely were you able to examine objects?
16. How well could you examine objects from multiple viewpoints?
    (Not relevant to train drivers.)
17. How well could you manipulate objects in the virtual environment?
    (Not relevant to train simulator.)
18. How involved were you in the virtual environment experience?
19. How much delay did you experience between your actions and expected outcomes?
20. How quickly did you adjust to the virtual environment experience?
21. How proficient in interacting with the virtual environment did you feel at the end of the experience?
22. How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?
23. How much did the control devices interfere with the performance of assigned tasks or with other activities?
24. How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those activities?
25. How completely were your senses engaged in this experience?
    (How completely were your senses—sight, hearing, touch—engaged in this experience?)
26. To what extent did events occurring outside the virtual environment distract from your experience in the virtual environment?
27. Overall, how much did you focus on using the display and control devices instead of the virtual experience and driving tasks?
28. Were you involved in the experimental task to the extent that you lost track of time?

29. How easy was it to identify objects through physical interaction, like touching an object, walking over a surface, or bumping into a wall or objects? (Not relevant.)
30. Were there moments during the virtual environment experience when you felt completely focused on the task or environment?
31. How easily did you adjust to the control devices used to interact with the virtual environment? (How easily did you adjust to the simulator in order to interact with the virtual environment?)
32. Was the information provided through hearing and vision in the virtual environment consistent?

The questions are grouped into themes as shown in Table A1.

Table A1. *Themed Categories of Questions from the Presence Questionnaire*

Category	PQ question number
Involvement	1–8, 13, 15
Sensory/fidelity	10, 11, 12, 14
Adaptation/immersion	9, 17, 18, 21–25
Interface	16, 19, 20

A.2 IGroup Presence Questionnaire

The questions used in the IPQ followed those proposed by Schubert et al. (2001). The only difference was that we chose to drop question 11, as drivers complained that it overlapped too closely with question 13. See Table A2.

Table A2. *I Group Presence Questionnaire Used in this Work*

Number	Subscale	English question	English anchors	Copyright (item source)
1	PRES[a]	In the computer-generated world, I had a sense of "being there."	Not at all—very much	Slater and Usoh (1994)
2	SP	Somehow I felt that the virtual world surrounded me.	Fully disagree—fully agree	IPQ
3	SP	I felt like I was just perceiving pictures.	Fully disagree—fully agree	IPQ
4	SP	I did not feel present in the virtual space.	Did not feel—felt present	
5	SP	I had a sense of acting in the virtual space, rather than operating something from outside.	Fully disagree—fully agree	IPQ
6	SP	I felt present in the virtual space.	Fully disagree—fully agree	IPQ
7	INV	How aware were you of the real-world surrounding while navigating in the virtual world? (i.e., sounds, room temperature, other people, etc.)?	Extremely aware—moderately aware—not aware at all	Witmer and Singer (1994)
8	INV	I was not aware of my real environment.	Fully disagree—fully agree	IPQ
9	INV	I still paid attention to the real environment.	Fully disagree—fully agree	IPQ
10	INV	I was completely captivated by the virtual world.	Fully disagree—fully agree	IPQ

Table A2. *(Continued)*

Number	Subscale	English question	English anchors	Copyright (item source)
11	REAL	How real did the virtual world seem to you?	Completely real—not real at all	Hendrix (1994)
12	REAL	How much did your experience in the virtual environment seem consistent with your real-world experience?	Not consistent—moderately consistent—very consistent	Witmer and Singer (1994)
13	REAL	How real did the virtual world seem to you?	About as real as an imagined world—indistinguishable from the real world	Carlin, Hoffman, and Weghorst (1997)
14	REAL	The virtual world seemed more realistic than the real world.	Fully disagree—fully agree	IPQ

[a]PRES = General Presence, SP = Spatial Presence, INV = Involvement, REAL = Experienced Realism. All table elements are from Schubert et al. (2001).