The Influence of Immersion in Virtual Environments on Performance of a Modelling Task: An Experiment[1]

David Mizell
The Boeing Company

Bernhard Spanlang and Mel Slater
Department of Computer Science
University College London

1. Introduction

People sometimes have to construct complex objects by copying examples presented as pictures. The most common experience of this is the trend today of buying furniture that needs to be 'self-assembled' - the consumer must reconstruct the furniture usually based on one 3D drawing and 2D plan views. This circumstance is likely to become more and more common as people purchase items over the internet, perhaps without ever initially seeing a physical example of the product. How should the information about the object be presented to people (leaving aside the actual construction issues)? This paper describes an experiment the compares three different modes of object presentation: the real object which could be viewed as in everyday life; an object presented on a 2D monitor, which could be oriented with a joystick device, and the object as presented through a stereo head-tracked head-mounted display.

Intuition suggests that for modelling of complex geometry the stereo head-tracked head-mounted display presentation would result in superior performance to the desktop view, and inferior performance to observation from the real world. However, there have been few studies that have examined this. The goal of this research is to experimentally quantify what differences may exist between visualising geometry on a workstation screen and doing so in a more immersive manner.

A description of an early attempt to experimentally examine this issue was published in (Mizell, et. al., 1995) and followed up by Slater and colleagues (Slater, et. al., 1996). Pausch and his colleagues at the University of Virginia have also done significant work at comparing VR to other visualisation modalities (Pausch, et. al., 1997). Their work compared subjects' ability to locate objects in three-space, however, and did not address the issue of visualising complex 3D geometry.

The earlier work of Mizell et. al., aimed to "experimentally assess and quantify, if possible, a difference in a user's being able to comprehend a complex three-dimensional scene between viewing it in 2-D on a workstation screen and viewing the scene via an immersive VR system...". The Mizell work showed subjects complex 3D shapes in reality, on a conventional workstation display, and through a stereo BOOM head-coupled device. The shapes were of three levels of difficulty. Each subject had to reproduce the shape in reality (using provided basic shapes) while being able to observe the baseline shapes from one of the three sources. Each subject successively used each method (reality, workstation, BOOM) in a pre-assigned randomly determined order. The response variable was time to completion of the shape. The results showed that observing the shape in reality was always the fastest method, and except for the simplest shape, looking at the workstation display was more effective than looking through the BOOM. As the authors noted, however, there were several problems with this experiment, including the fact that looking at the workstation display was a faster operation than looking through the BOOM (which could often get into awkward positions). Moreover, this experiment involved no interactive manipulation of objects in the VE.

The Tri-dimensional chess study carried out by Slater and colleagues, mentioned above, was directly influenced by the earlier Mizell results. In the Tri-dimensional chess study, subjects were asked to reproduce a sequence of moves observed on the virtual chess board. The two independent factors were level of immersion (workstation display or immersive head-tracked) and degree of realism. It was found that the subjects in the VR performed

[1] http://www.cs.ucl.ac.uk/research/vr/Immersion

better than those who learned on the workstation display, and also that the greater degree of realism was positively associated with successful task performance.

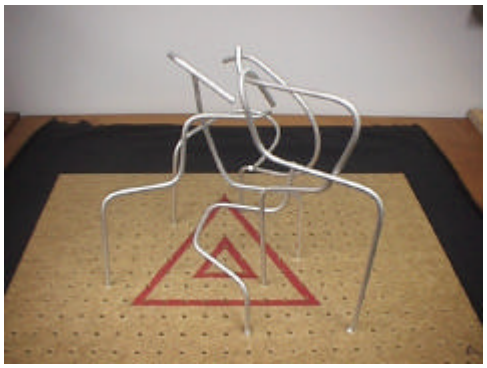
The study reported in this paper is on the largest scale, and most complex to date, but involves a more limited situation. It strictly compares real life, the traditional desktop display ('desktop'), and head-tracked stereo head-mounted displays ('HMD'). Also the scope in the case of this experiment is that of objects that are less than human-size in scale - for example, objects that it is possible to look around by head movements alone while standing in one fixed place.

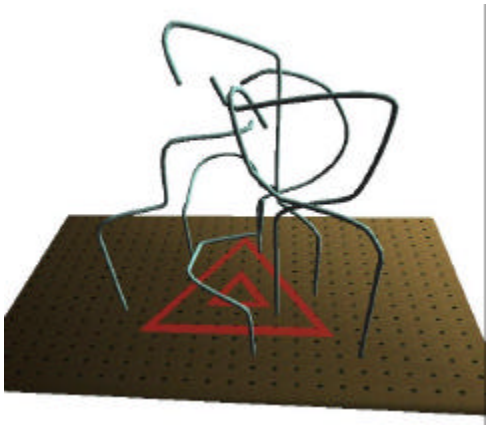
In the next section we discuss the experimental design in detail, followed by the results in Section 3, and conclusions in Section 4.

2. Experiment

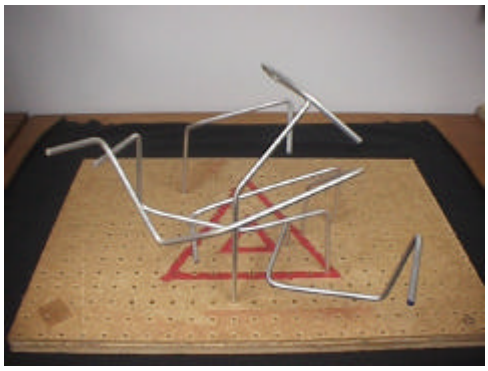
2.1 Procedures

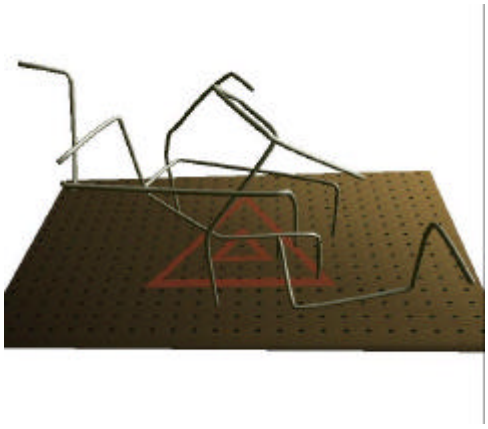
The experimental task was for subjects to reconstruct a complex 3D object from a presentation of a reference object. The objects were constructed as wires inserted into a plane base with a grid of holes to accommodate the wires. The wires each had different shapes and could be placed in different orientations into the holes. The puzzles used are shown in Figure 1.



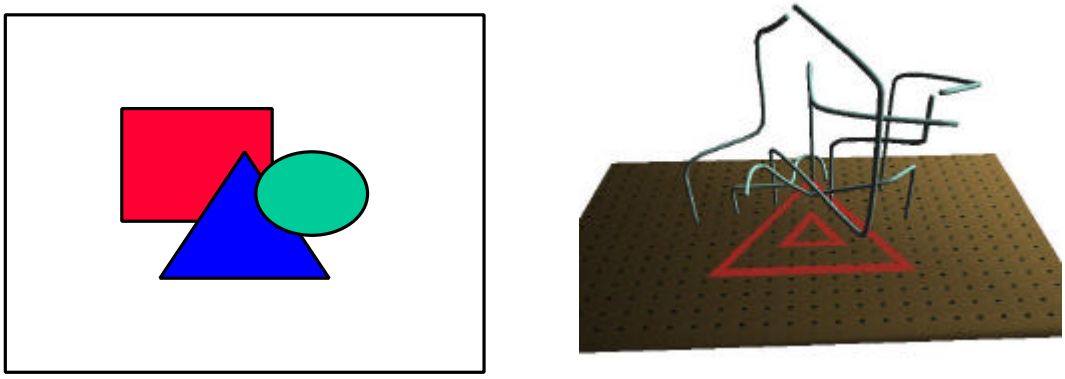


Puzzle 1: Real and Virtual





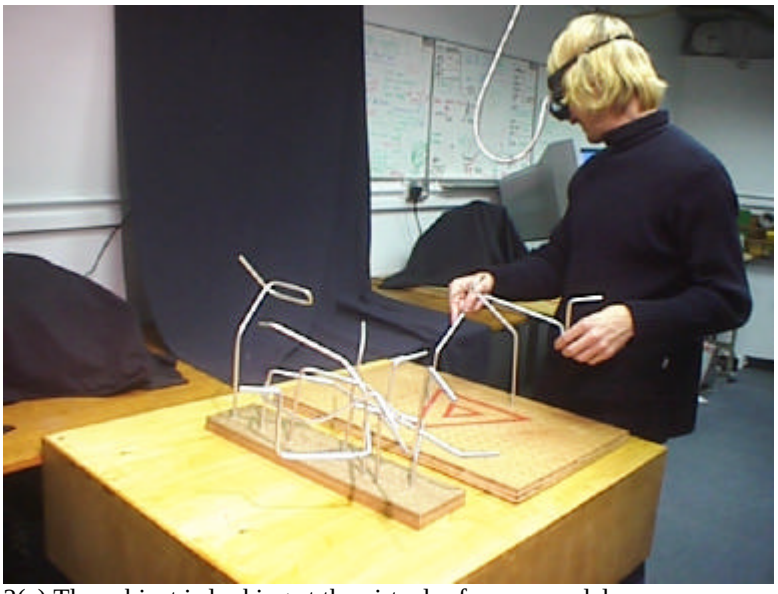
Puzzle 2: Real and Virtual



Puzzle 3: Real and Virtual

Each subject was given another board with holes, and a set of wires, from which it was possible to make all of the presented shapes. Their task was to reconstruct the shape shown to them, the 'reference shape'.

A reference shape was presented to the subjects, in random sequence, under three different conditions: in reality, on a screen, or through a HMD. Hence subjects would observe a reference shape, and copy it. The HMD was partially see-through so that they would not have to remove it when looking back and forth at the reference shape and the shape they were constructing. Figure 2 shows the situation for subjects while wearing the HMD.



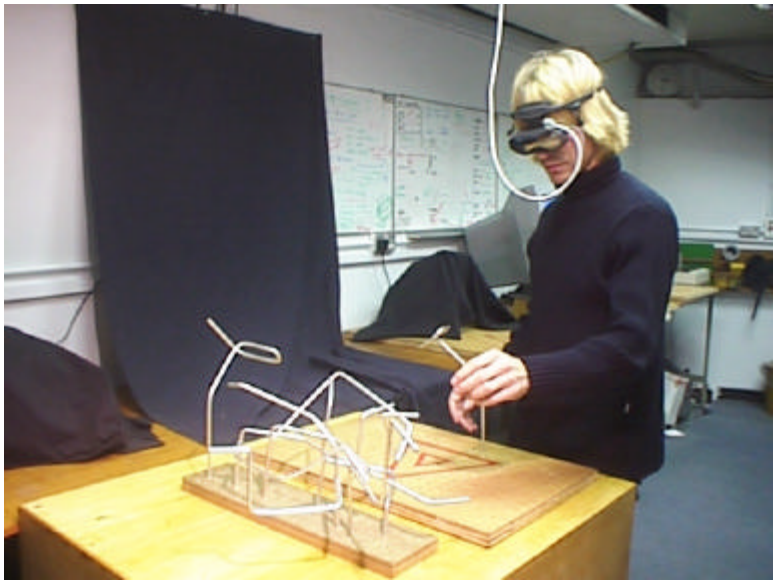
2(a) The subject is looking at the virtual reference model.

In Figure 2(a) the subject is looking at the virtual reference model. In Figure 2(b) the subject is looking back at his copy that is being constructed from the wires at the near side of the picture. The workspace where the subject made the model copy was illuminated by a bright lamp so that it could be clearly seen through the HMD. Models in the real world were placed on the black sheet at which the subject is looking in Figure 2(a). Figure 3 shows the setup for the desktop display situation.

Three different reference puzzles were presented. These had been previously calibrated to be of equal difficulty when presented in reality (see below). They were allocated randomly to the three different display types.

To summarise - each individual subject was assigned a randomised sequence of display types, and a different puzzle was allocated randomly to each display type. The subjects carried out 4 trials. The first 3 trials were with the three display types and puzzles in sequence. The last trial was a repetition of the same display type and puzzle as the first trial - this in order to measure any learning effect.

Since there are 6 different sequences of display type (real, desktop, HMD) and 6 sequences of presentation of the puzzle, there are 36 combinations covering all arrangements of display type and puzzle presentation. 36 subjects were recruited by advertisement in the University, and allocated to one of the 36 combinations at random. Subjects were paid 10 pounds to take part in the experiment. Half of the subjects were women, and allocated randomly over the design.



2(b) The subject is looking back at his constructed model.

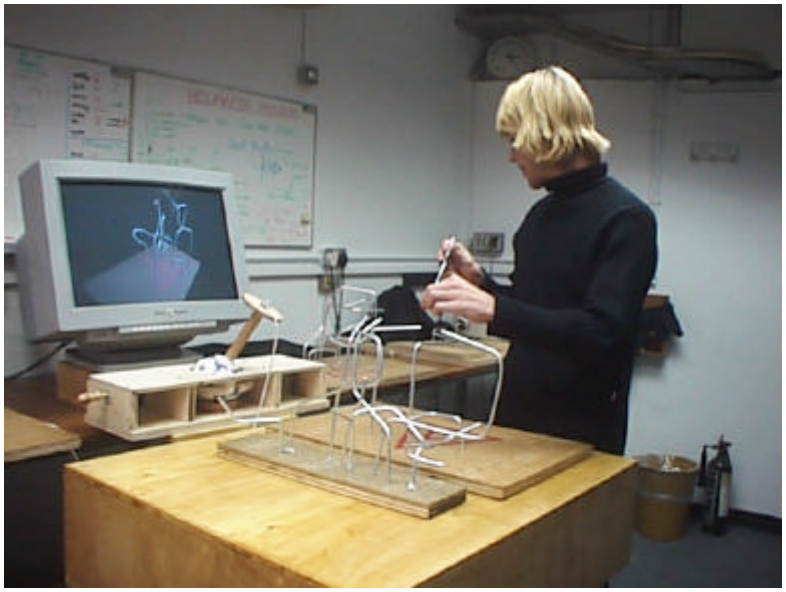


Figure 3 The subject looks at the reference model on the workstation display.

On arrival in the laboratory where the experiment was carried out each was given a standard eye test for acuity[2], and also completed a standard spatial ability test (SAT) (Smith and Whetton, 1988). If they failed the eye examination (with corrected vision) they were excluded from the experiment. They were asked to sign a disclaimer form, which gave information on possible side-effects induced by the wearing of a HMD. Refusal to sign the form would result in exclusion from the experiment, but in fact no one refused.

[2] http://www.web-xpress.com/athens/vatestde.html

After completion of these preliminaries, they were given the a sheet of instructions to read (see web pages) and it was read to them again by the experimenter. The subject then carried out a 3 minute practice session with a puzzle that had been calibrated to be easier than those used in the actual experiment. After checking for further questions they then carried out the experiment, doing the puzzle 4 times. Finally they answered a web-based questionnaire.

2.2 Materials

The scenario was implemented on an SGI Onyx with twin 196 MHz R10000, Infinite Reality Graphics and 64M main memory. The scene was shown on a 21 inch monitor covering the full screen resolution (1920×1080). The HMD used was a Sony Glasstron (LDI-D100B Stereo 800×600) and head tracking with a Polhemus tracker at 60Hz.

For the desktop situation a special controller was built that allowed the virtual puzzle to be rotated and swiveled, as shown in Figure 4. The wooden rod acted as a joystick controller and the disc on top allowed the puzzle to be rotated about the axis given by the wooden rod. However, a software constraint did not allow the puzzle to be completely swiveled around and seen from behind - in order to maintain consistency with what was possible under the other conditions. The tracking for this was with a Polhemus sensor.

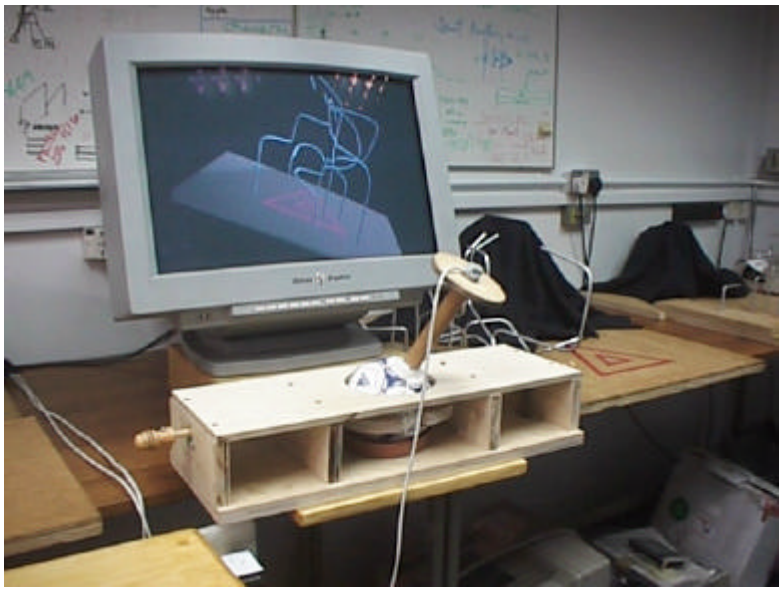


Figure 4 The controller for the workstation display

The virtual puzzles contained 13000 polygons on the average. The frame-rate throughout the main experiment was 40Hz.

2.3 Response Variables

Error

There had to be a way of judging the similarity of the constructed model to the reference one. On reference puzzles 1 and 3 there were 7 rods and on puzzle 2 there were 6 rods. Each rod placed by a subject was evaluated as correct or incorrect. It could be incorrect for the following reasons (which could occur multiply):

(a) the rod was put into the wrong hole.
(b) the rod was oriented incorrectly.
(c) the rod selected did not belong to the puzzle at all.
(d) in some cases a rod could be put below another, when it was supposed to be above it, or vice versa.

The occurrence of each one of these was counted as one error. So for puzzles 1 and 3 there were a possible 28 errors that could be made, and 24 for puzzle 2.

The 3 different puzzles had been tested before the experiment in pilot studies with subjects not involved in the main experiment. This was to choose 3 puzzles that gave on the average the same overall error rates. Earlier, puzzle 2 had resulted in a greater number of errors than 1 or 3, but when one rod was removed from puzzle 2, it resulted in the same average number of errors as 1 and 2. However, all such calibration was carried out by subjects examining the real-world reference models only.

Time

Each puzzle reconstruction was allowed a maximum time of 10 minutes. The actual time taken was recorded.

2.4 Subjective Responses

A number of self-assessed subjective responses were gathered from the questionnaire.

Difficulty

Subjects were asked to 'rate the degree of difficulty in carrying out the task, for each of the three types of viewing' on a 1 to 7 scale, where 1 was 'not at all difficult' and 7 was 'very difficult indeed'.

Presence

This was the sense of being in the same space as the object: 'When you looked at the real model, you were in the same (physical) space as that model. To what extent was your experience similar to this real physical experience for the other two types of display?' The questionnaire asked subjects to respond on a 1 to 7 scale for 'Looking at the screen' and 'Looking through the head-mounted display'.

Performance

Subjects were asked to 'Rate your performance in the tasks (how well do you think you reproduced the models) for each of the three conditions' from 1 = 'not very well at all' to 7 = 'very well'.

2.5 Independent Variables

Display Type

The major independent variable was the display type: real world, screen, or HMD.

Puzzle

The puzzles were labelled as 1, 2 or 3, calibrated to be of equal difficulty.

Session

There were 4 sessions - each of the first three with a different puzzle, and the last session with the same puzzle as the first one.

2.6 Explanatory Variables

The following additional information was recorded - gender, age, status, the extent of prior experience with virtual reality, the extent of prior computer use. The SAT score is a standard measure of spatial ability; higher scores are meant to indicate a greater facility for mentally manipulating complex 3D shapes.

The degree of 'learning' across the 4 trials could be assessed by the repetition of the first puzzle using the same display type in session 4 as had been experienced in session 1. A significant decrease in time or error rate would indicated a learning effect.

The degree of nausea was informally elicited by the following question:

'Did you experience any sensations or feelings of nausea or sickness, or other adverse symptoms during any of the three experiences?'

Each display type was then rated on a 1 to 7 scale.

3. Results

3.1 Summary

The results were analysed separately and independently by two different statisticians, who came to substantially the same conclusions. With respect to both time and number of errors, there is overwhelming evidence of a difference between display types. In particular, the lowest error rates and times occur for the real display, and the highest error rates and times occur for the HMD display. There appears to be a gender bias where women do not do as well as men on this task, whereas spatial ability is positively associated with performance - in other words the higher SAT score results in better performance, independently of all other factors. In spite of the initial calibration, there does seem to be a difference in performance for the 3 puzzles. There is some evidence to suggest that overall performance, taking into account time and number of errors together, significantly improves with the SAT score for the HMD results, but does not do so for the other two display methods. In other words to be successful with the HMD it is necessary to have good spatial ability.

There was not a significant impact of learning, as judged by the different scores for sessions 1 and 4. For the number of errors, the mean and standard deviation for sessions 1 and 4 were 1.3±1.7 and 1.1±1.4 respectively (n=36). For the time the means (in minutes) and standard deviations were 5.2±2.7 and 4.2±2.2. The differences are not significant at the 5% level on a two-tailed test, and on a one-tailed test the time difference is just on the significant side of the boundary. Session number was used as an explanatory factor in the regression analyses described below, but it was never significant.

3.2 Measurement of Task Performance

The measurement of task performance is not straightforward. For example, imagine two results: one which takes 2 minutes and there are 2 errors, compared with another which takes 10 minutes and there are 2 errors. Clearly, we would say that the latter is worse than the former, even though the number of errors is the same. On the other hand, a straightforward ratio of number of errors divided by the time is not appropriate, since this would give a puzzle completion in 1 minute with 1 error the same score as the completion in 10 minutes with 10 errors. Hence three different indicators of performance are used:

- The number of errors, but where time is statistically eliminated from the data.
- The time itself.
- A composite measurement constructed by standardising each of time and error to have 0 mean and standard deviation 1 (called stime and serror) and then taking the length of the vector (stime,serror) as an overall measurement of performance. Clearly good performance is associated with a small distance of (stime,serror) from the origin.

3.3 The Number of Errors (eliminating time)

The raw result is shown in Table 1. It is clear that the mean number of errors is about the same for desktop and HMD which are significantly greater than for the real case.

Table 1
Mean Error by Display (n=48 each column)

	Real	**Desktop**	**HMD**
Mean	0.5	1.4	1.6
Standard deviation	0.95	1.58	1.63

However, the raw data fails to take into account other variables that may be influencing the results (for example, it might just be that those allocated to the ‘real’ group had higher SAT scores than for the other two groups). Hence a regression analysis was carried out using a generalised linear model known as the Poisson log-linear

model (McCullagh and Nelder, 1983). Here the response variable, number of errors, is modelled as a Poisson random variable, since this would be appropriate for the null hypothesis of errors randomly occurring over the time period. Moreover, time was statistically eliminated from the regression model (although it actually is not statistically significant, and the same result is obtained without this elimination). Here we report only the conclusions.

The log-linear regression analysis leads to the following results:

- Independently of anything else the number of errors is higher for females than for males, and is negatively correlated with the SAT score.
- The number of errors is significantly lower for the 'real' display than for the HMD and desktop.
- There is an interaction effect between puzzle and display, so that the impact of display type on number of errors is different for each of the three puzzles. Either the error rate is worse for the HMD (puzzles 1 and 3) or slightly better (for puzzle 2). The difference for puzzle 3 is the significant one.
- There is no significant effect of session number - in other words although the mean error rate is slightly lower for session 4 than for session 1, it is not significantly different.

3.4 Time

A normal regression model was used in this case. The significant variables were display type, puzzle, the SAT score and the self-rating of nausea. There was no interaction effect between puzzle and display type. The overall squared multiple correlation coefficient was 0.56 (meaning that 56% of the variation in time was accounted for by this model). The results were:

- Display type was significant, with the real display resulting in significantly less time than the other two.
- There were differences between the three puzzles, with puzzle 2 accounting for the longest time.
- The SAT score was negatively correlated with time.
- The higher the degree of self-assessed sickness the greater the time.

3.5 Combined Performance

Here we examine the composite score as explained in Section 3.2. A normal regression was used, and it is worth showing the fitted model in full (the squared multiple correlation coefficient was 0.22).

Table 2
Regression Model for Composite Error
(Standard Errors of Estimates in Brackets)

Display	Fitted Model
Real	0.6 + 0.1*nausea + 0.2 (if female) (0.53) (0.04) (0.1)
Desktop	1.5 + 0.1*nausea + 0.2 (if female) (0.48) (0.04) (0.1)
HMD	3.0 + 0.1*nausea + 0.2 (if female) - 0.03*SAT (0.55) (0.04) (0.1) (0.008)

From Table 2 we can see that, other things being equal:

- The composite error increases by about 1.5 for each change in display from real, to desktop to HMD.
- That nausea is positively associated with the composite error.
- That females tend to have a higher composite error than males (this was on the margins of 5% significance).

- That SAT is negatively associated with composite error. However, although statistically significant, the result implies that other things being equal a person using the HMD would have to have a SAT score of about 50 higher than someone using the desktop in order to achieve the same composite error. However, the mean SAT for the whole group was 65±11.

4. Conclusions

This paper is an initial description of an experiment and the results to assess a certain type of task performance comparing desktop and joystick with head-tracked stereo head-mounted displays for a modelling task. The results are not what were predicted, and actually support the earlier result of Mizell. This experiment was repeated because of logistic problems in the use of the BOOM display in the earlier experiment. However, even where a see-through HMD was used, where the conditions were as similar as we could make them to showing the puzzles in the real world, the use of the traditional desktop approach was certainly no worse, and may have been better most of the time.

There is evidence suggest that the actual puzzles used, the SAT score, gender, and the level of sickness influenced the results, and that in fact higher spatial ability was required in the case of the use of the HMD. From the results, and from our anecdotal observations, simulator sickness played a role, and the rate of reported sickness for the HMD experiences was almost twice as high than for the desktop experiences. Our own experiences, and the evidence, suggested that in spite of the high frame rate of 40Hz and the same overall latency as the joystick tracker simulator sickness was still a problem with the HMD. Of course the desktop display was of an order of magnitude more than 4 times greater than the HMD display, which may have been another contributing factor.

The evidence from this experiment suggests that the case for stereo head-tracked HMD producing a better task performance, for this experiment, and with this particular HMD, is not supported by the evidence. It should be recalled that this was for a complex model that was less than human size - this result may not hold for very large structures in comparison to the human scale. The subjective data gathered from the questionnaire remains to be analysed at the time of writing.

Acknowledgements

This project is supported by EPSRC research grant GR/M85692, and EPSRC Visiting Fellowship for Dr David Mizell (GR/M86200). We would like to thank the Boeing Company for their support in this project. We would like to thank the company Virtual Presence, who provided significant support in providing a driver for the HMD used in this experiment.

References

McCullagh, P. and Nelder, J.A. (1983) Generalised Linear Models, Chapman and Hall.

Mizell, D., Jones, S., Jackson, P. and Picket, D. (1995) Is VR Better than a Workstation: A Report on Human Performance Experiments in Progress, Eurographics Workshop on Virtual Environments, Proceedings ed. M. Goebel, Monte Carlo, Jan 31-Feb 1st, 1995.

M. Slater, V. Linakis, M. Usoh, R. Kooper (1996) Immersion, Presence and Performance in Virtual Environments: An Experiment with Tri-Dimensional Chesss, ACM Virtual Reality Software and Technology (VRST), Mark Green (ed.), ISBN: 0-89791-825-8, p 163-172.

Pausch, R., Proffit, D., Williams, G. (1997) Quantifying Immersion in Virtual Reality, Computer Graphics (SIGGRAPH) Proceedings, 13-18.

Smith, P. and Whetton, C. (1988) General Ability Tests (User's Guide) The National Foundation for Educational Research, ASE.