3 Results

3.1 Overview of framework used in data analysis

The complexity of the tasks that participants undertook necessitates the establishment of a framework within which results may be described and interpreted.

The experiment was partitioned into two phases, and only data from the latter test phase was considered for in-depth analysis. Although data was recorded from the training phase, during this phase the participant was at various times able to stop and ask questions, or was instructed to try procedures not directly related to the completion of the task, hence the data was not universally suitable for performance comparisons.

The change in performance requirements moving between the three tasks encountered during the test phase was enough to suggest that considering each separately in the analysis would increase the strength of any conclusions drawn from each individual task, at the expense of being able to draw more general conclusions across all three. Hence, the three test tasks are considered separately in most of the analyses below.

Furthermore, each task naturally partitioned into two phases, the first being movement of the arm from its initial resting position and alignment with and latching to the payload, and the second being movement of the arm and connected payload to return the payload to the payload bay and overlay it on top of the payload target. The first phase will be referred to henceforth as the *unladen* phase and the latter the *laden* phase.

3.2 Task completion rate

3.2.1 Motivation for and derivation of measure

The most important raw performance measure was considered to be the time it took participants to complete each trial. In the experimental task, time-to-completion acted as a composite measure of several other performance measures: It incorporated the effects of discrete manipulation errors such as collisions and arm limit conditions, through the 3-s time penalty associated with these errors. It also indirectly incorporated a measure of the control effort required to carry out the manipulation, since time spent erroneously manoeuvring along sub-optimal trajectories, time spent looking about for the correct direction of travel, and time spent backtracking out of areas near the arm limits all consumed control effort, but equally consumed time and hence appeared indirectly in the time-to-completion. Time-to-completion in the experiment thus acted as a composite measure of several competing real-world performance requirements encountered in the work domain of space telerobotic manipulation.

However, raw measures of time-to-completion were of limited use, as many participants found Tasks 2 and 3 difficult and in some cases this difficulty was manifest in the results as trials where the participants failed to complete the manoeuvre within the time limit of 180 s (Figure 7). Thus raw time-to-completion was not a suitable basis for performance comparisons owing to the relatively high number of missing scores. Instead, a measure of the rate at which the participant proceeded through the task was derived and used in analyses. The numerator of this elapsed *Completion Rate* is a standardised measure of the distance the arm had moved from its initial position towards the target. It was calculated from the projection of the position of the point-of-resolution (POR, the end-effector during the unladen phase or the centroid of the payload during the laden phase) along an axis extending between the initial and final positions of the POR in each phase. The denominator of the rate was the time-to-completion or, in cases where the time limit was breached, the time limit.

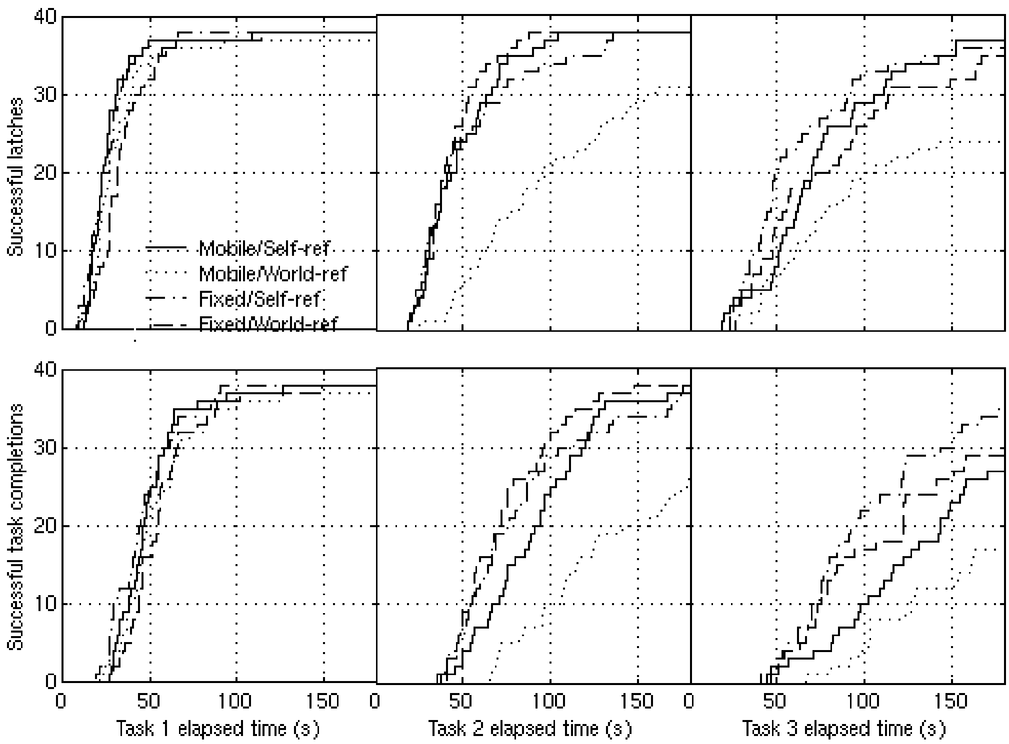

Figure 7. Plot of number of participants (out of 38) who had successfully latched the payload (top row) and completed the task (bottom row) at the elapsed time indicated on the x axis, for Tasks 1, 2 and 3. A greater number of participants breached the 180-s time limit in Tasks 2 and 3, particularly in the mobile/self-referenced condition.

An illustration of the quantities involved in calculation of the progress during a manoeuvre in the unladen phase of a trial in test Task 2 appears in Figure 8. The length of the solid line extending from the initial position of the arm to the latch surface on the payload was standardised at 1.0 for each phase in each task. That is, a complete manoeuvre from the arm's initial position to the payload and back to the target was a progress measure of 2.0, regardless of the task. In Figure 8 the arm progress measure is equal to the length of the arrow labelled "p" divided by the length of the long black axis. The arm progress rate is thus p / $t_{elapsed}$ where $t_{elapsed}$ is the total time taken in the task since the first control action.

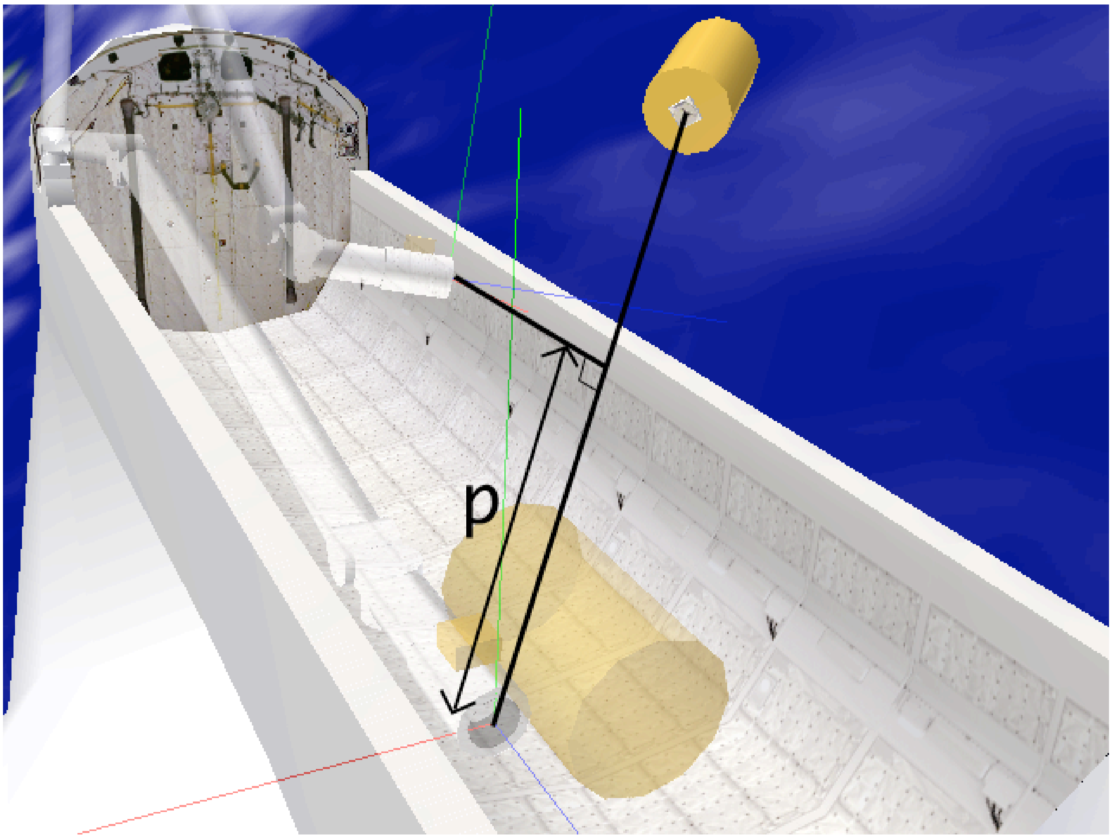

Figure 8. For each task phase, Completion Rate at elapsed time *t* was calculated by determining the projection *p* of the position of the arm at time *t* onto an axis extending between the POR initial and target positions, and dividing by *t*.

The Completion Rate measure as defined above takes into account only the *position* of the POR; progress made by the participant in correctly *orienting* the POR does not contribute to Completion Rate. The reason for this is that an examination of the orientation data suggested that progress made by the participants in orientating the POR varied much less linearly with time elapsed, was far more "bursty" (rapid progress in orientation occurred over much smaller timescales) and was far more prone to indicating reversals away from completion that did not actually affect the time taken to complete the manoeuvre. In short, progress in correctly orienting the payload could be rapidly undone, and reversals in progress could be rapidly corrected. Thus it was felt that the weighting of position versus orientation in a composite completion rate would be somewhat arbitrary and would thus have reduced the meaningfulness of the Completion Rate measure.

3.2.2 Score statistics

The distributions of participants' completion rate scores for the three test tasks appear in Figure 9. Each histogram represents the scores of 38 participants divided into 10 groups. The distributions were approximately normal and of homogenous variance, thus allowing the application of common inferential statistical procedures based on these assumptions.

Three separate 2 × 2 (Viewpoint Frame of Reference × Control Frame of Reference) factorial ANOVAs with repeated measures on both factors were conducted for the three separate test tasks, and are presented in Table 4 in Appendix E.

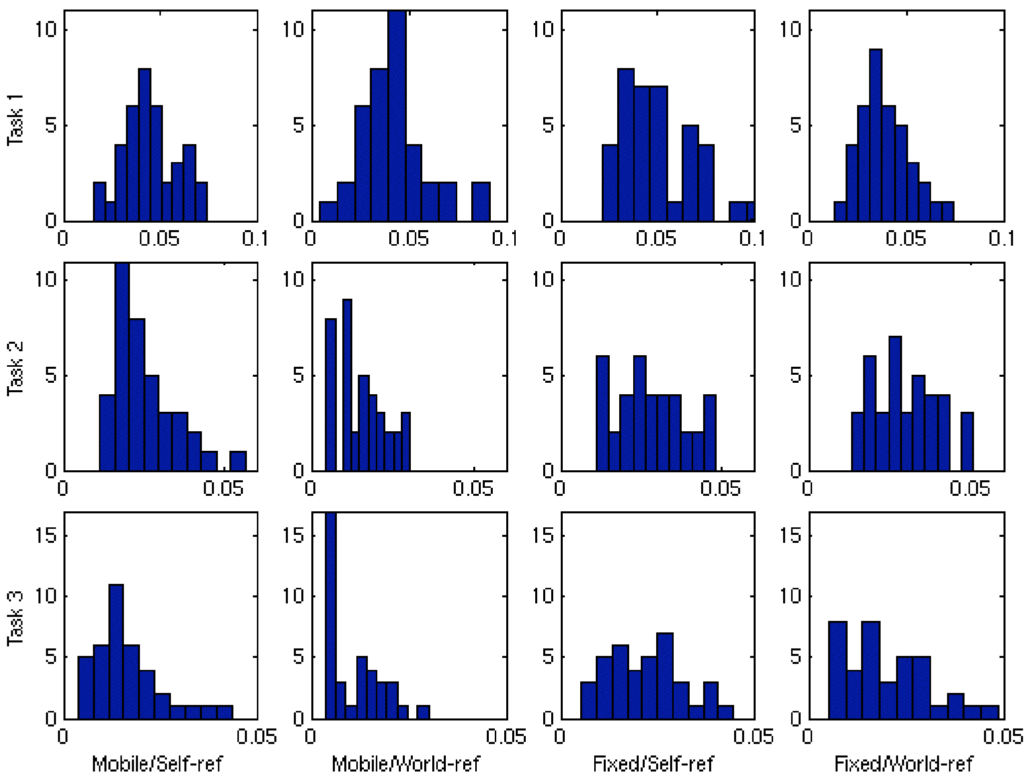

Figure 9. Distributions of participants' task Completion Rate scores. Rows are ordered by test task, columns by experimental condition. (NB: Scale varies between rows.)

3.2.3 Summary of results

The mean Completion Rates for the three tests are displayed in Figure 10 as a function of experimental condition. Within each panel, the four conditions are further grouped by the viewpoint factor and the control frame of reference factor.

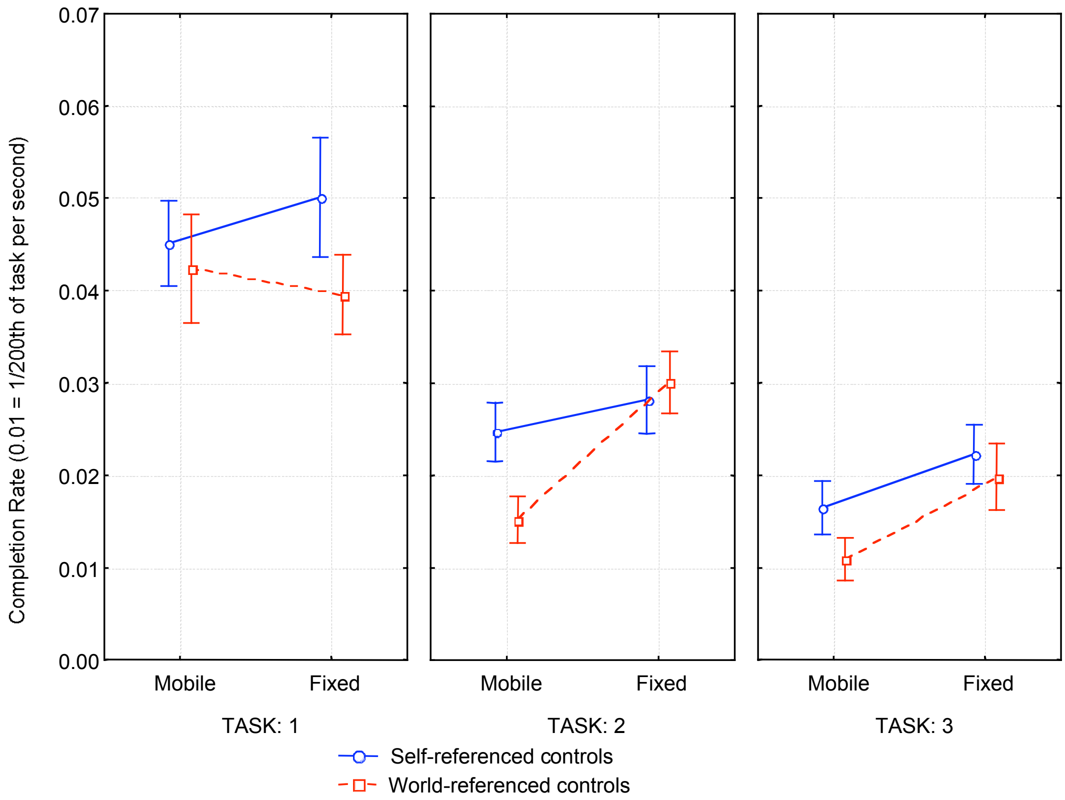

Figure 10. The marginally significant, highly significant and non-significant interaction effects of experimental condition upon mean Completion Rates for test Tasks 1, 2 and 3 respectively. Vertical bars indicate 95% confidence intervals.

In Task 1, participant performance was better when using the self-referenced control frame rather than the world-referenced control frame, a finding supported by the significant main effect for control frame of reference (Figure 11). However, as can be seen from the left-hand panel of Figure 10, there was also a marginally significant interaction between the viewpoint and control frame of reference for Task 1 [F(1,37) = 3.641, p < 0.064]. The interaction indicates that for Task 1, performance in the mobile viewpoint was little affected by the choice of control frame of reference, whereas in the fixed viewpoint, there was a strong performance decrement imposed by the world-referenced control frame. Application of Tukey's Honestly Significant Different (HSD) post-hoc tests provided support, indicating a highly significant difference between the fixed/self-referenced and the fixed/world-referenced conditions and a marginally significant difference between the fixed/self-referenced and mobile/world-referenced conditions. No other differences were significant.

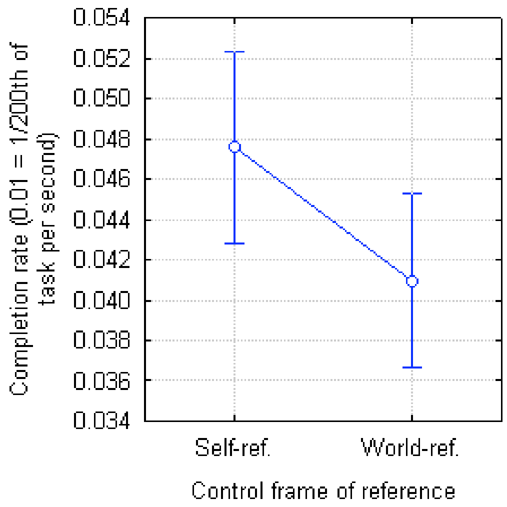

Figure 11. The significant main effect for control frame of reference in Task 1 [F(1,37) = 9.342, p < 0.004]

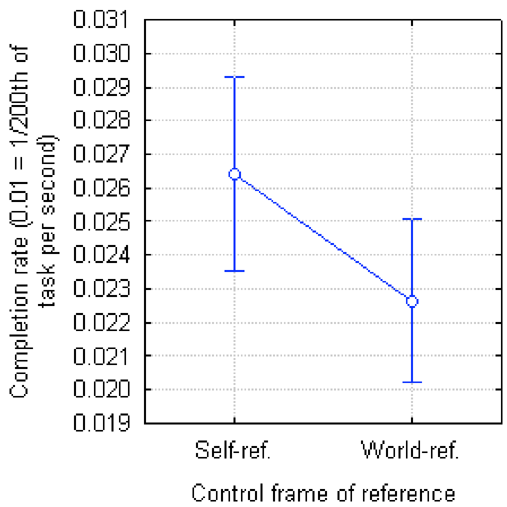

Figure 12. The significant main effect for control frame of reference in Task 2 [F(1,37) = 12.43, p < 0.0011]

In Task 2, participant performance was again better when using the self-referenced control frame, as indicated by the significant main effect for control frame of reference (Figure

12). Performance was also better from the fixed viewpoint, as can be seen from the main effect illustrated in Figure 13. However, the main effect of viewpoint is again complicated by the interaction of viewpoint and control frame of reference. As can be seen in the centre panel of Figure 10, the interaction indicates that in Task 2, performance from the *fixed* viewpoint was unaffected by the control frame of reference used in performing the manipulation, whereas performance from the mobile viewpoint was strongly affected. A post-hoc application of Tukey's HSD test supported this, indicating highly significant differences between the mobile/world-referenced and the other conditions, and a significant difference between the mobile/self-referenced and fixed/world-referenced conditions, but no other significant differences.

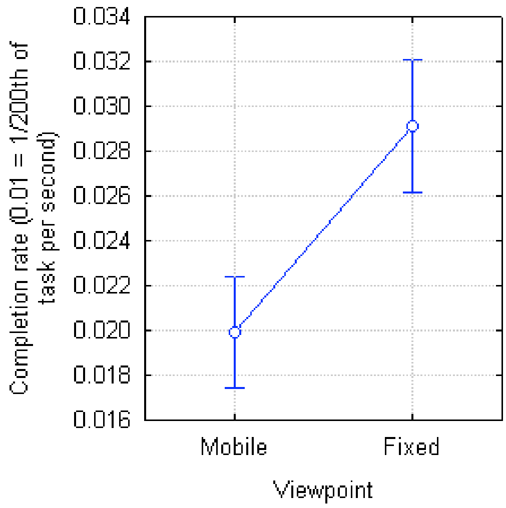

Figure 13. The highly significant main effect for viewpoint in Task 2 [$F(1,37) = 55.71$, $p < 0.000$]

Performance in Task 3 was similar to that in Task 2, proving superior on both the self-referenced control frame and the fixed viewpoint, as indicated in Figure 14 and Figure 15 respectively. The interaction was not significant.

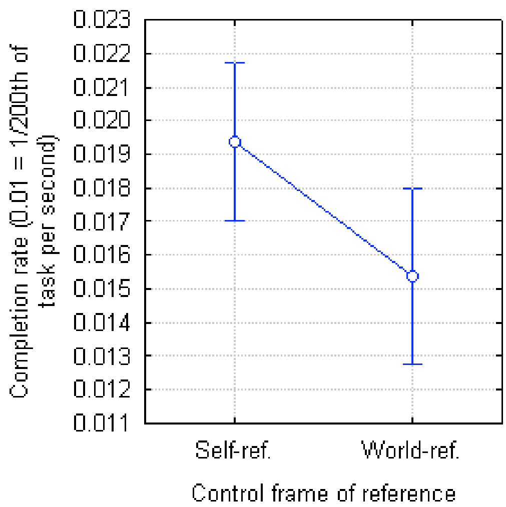

Figure 14. The significant main effect for control frame of reference in Task 3 [$F(1,37) = 10.21$, $p < 0.0029$]

Post-hoc tests indicated caution in making individual comparisons between conditions in Task 3 however. Although the main effect for control frame of reference was significant, Tukey's HSD test for the difference between fixed/self-referenced and fixed/world-referenced was not significant, and neither was the difference between fixed/world-referenced and mobile/self-referenced. In fact, when the Completion-Rate scores were converted into projected time-to-completion scores (by taking the reciprocal and multiplying by 2) it was found that the top three performing conditions were separated by only 32 s (with projected time-to-completion scores of 89 s, 101 s and 121 s) whereas the worst performing condition of mobile/world-referenced was separated from this group by almost twice as much, with a projected time to completion score of 182 s.

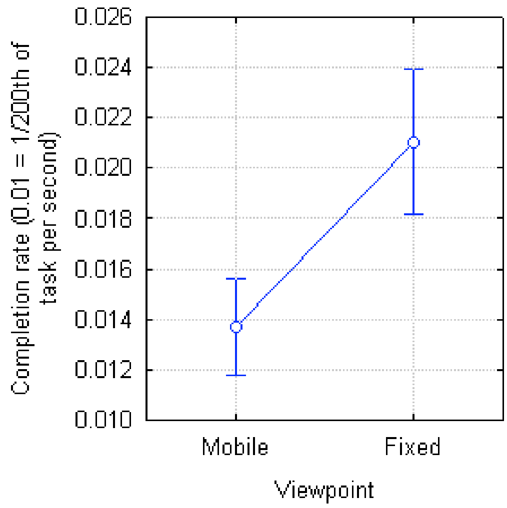

Figure 15. The highly significant main effect for viewpoint in Task 3 [F(1,37) = 36.51, p < 0.0000]

Overall, participant performance in terms of Completion Rate was strongly dependent on the task, which indicates that there is an interaction between the physical layout of the task and the relative contributions of the individual factors to performance for operating the SRMS in that task. Furthermore, a general trend is discernable across all three tasks towards better performance with the fixed viewpoint and the self-referenced controls.

Across all three tasks there was a great deal of variability in Completion Rates between participants. This is evident in the broad spread in the distributions of scores in Figure 9, and is also emphasised by the highly significant F-values for the between-participants intercept terms in the ANOVAs. The standard deviations, expressed as a proportion of the mean, were large, with the average across the four conditions in each of Tasks 1, 2 and 3 equal to 0.37, 0.41 and 0.54 respectively. Further analyses presented below resolve some of this between-participants variability in terms of participants' styles of control.

3.2.4 Task Completion Rate pre-latch versus post-latch

Exploratory analysis of task Completion Rates over the time course of individual trials for a number of participants indicated that the advantage of certain experimental conditions was felt more at some stages of each trial. The length of time it took participants to manoeuvre the arm from its rest position and latch the payload was generally more than the time it took them to return the latched payload and align it with the target, over all the conditions. However it appeared some conditions favoured post-latch manoeuvres to a greater degree than others.

The ratios of Completion Rate pre-latch to post-latch across Tasks 1, 2 and 3 were derived and were approximately normally distributed. The means of these ratios are plotted in Figure 16 as a function of experimental condition. A higher ratio on a given condition indicates that that condition favoured the post-latch manoeuvre or imposed some handicap upon the pre-latch manoeuvre.

Three separate 2×2 (Viewpoint Frame of Reference $\times$ Control Frame of Reference) factorial ANOVAs with repeated measures on both factors were conducted for the three separate test tasks, and are presented in Table 5 in Appendix E. In Task 1 there was a marginally significant main effect of viewpoint [$F(1,37) = 4.37$, $p < 0.04$], however post-hoc tests indicated that only the fixed/world-referenced condition was different from the other conditions by a significant margin (Tukey's HSD, $p < 0.050$). However, the effects of the ratio were far more evident in Task 2. Under the mobile/world-referenced condition, participants generally had considerable difficulty aligning the arm with the latch plate, owing mainly to the fact that during the approach to the payload latch plate the required hand controller actions were in the opposite direction to the onscreen motion they produced. The delays caused by the incorrect control actions made as a result of this display-control incompatibility are evident in the significant main effect for control frame of reference and the highly significant interaction effect. Post-hoc tests indicated that the difference between the fixed viewpoint conditions was not significant, whereas the differences between the pair of fixed viewpoint conditions and each of the individual mobile viewpoint conditions were significant (Tukey's HSD, $p < 0.02$ for both differences).

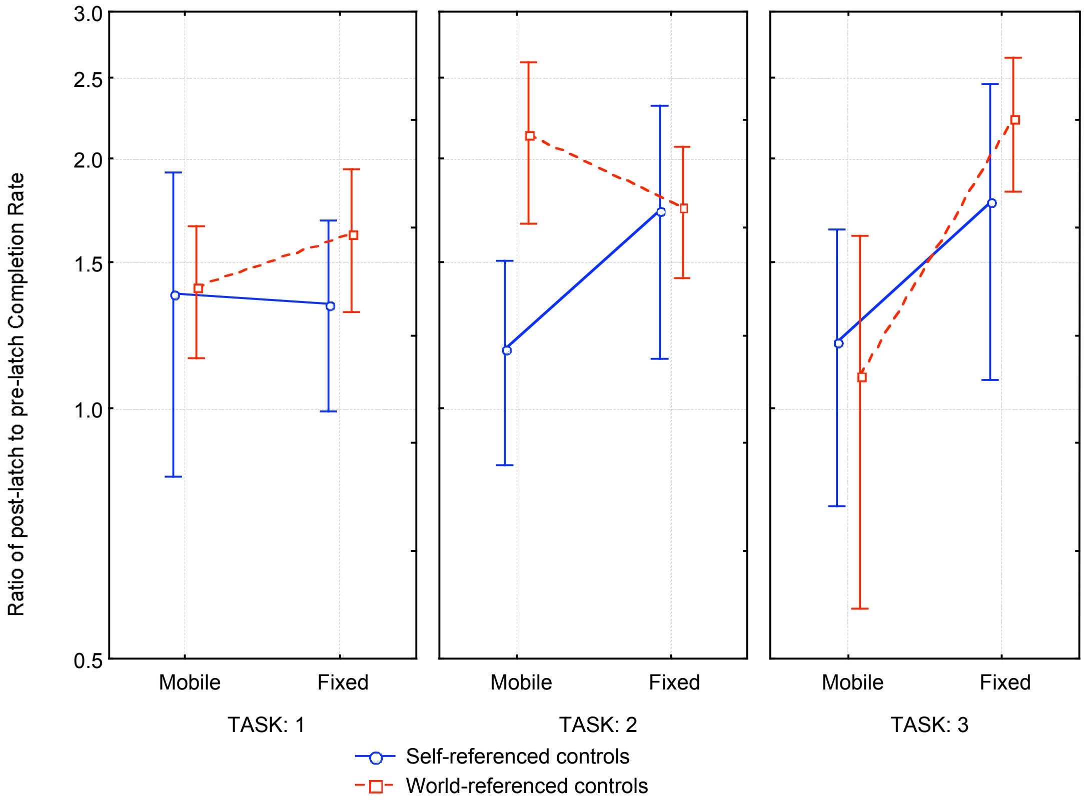

Figure 16. The non-significant, highly significant [$F(1,37) = 19.79$, $p < .000$], and non-significant interaction effects of condition upon the ratio of Completion Rate pre-latch to Completion Rate post-latching for Tasks 1, 2 and 3 respectively. The y axis is plotted on a logarithmic scale in order to represent ratios larger than 1.0 on the same scale as a ratios less than 1.0. Vertical bars represent 0.95 confidence intervals.

The effects were different again in Task 3, with the highly significant effect for viewpoint indicating that the fixed viewpoint offered an advantage for the post-latch manoeuvre relative to the pre-latch manoeuvre.

3.3 Manipulation errors: arm limits

3.3.1 Motivation for and derivation of measure

Although commands made through the hand controllers expressed changes in the position and orientation of the point-of-resolution attached to the end of the arm, the motion of this point was bounded by the possible kinematic configurations of the articulated SRMS arm. The bounded kinematics create a *work envelop*e of points that the SRMS may reach and move through. Like other articulated manipulators such as the human arm, the SRMS work envelope is irregular in shape. The extent of the envelope is further restricted under particular orientations of the end-effector, for example facing back towards the shoulder of the arm[1]. At certain points, the SRMS also reaches singularities, where linear motion of the end-effector in a particular direction cannot be maintained because none of the axes of any of the joints has any component acting in the desired direction of travel. [See Nguyen & Hughes (1994) for a more complete discussion of the SRMS singularities.]

Successful manipulation required traversal around the boundaries of the SRMS work envelope and singularities. It is not easy to represent the SRMS work envelope since it is a rather convoluted region in six-dimensional space. Because of this, in the experiment there was no visual aid provided in the virtual environment to represent the SRMS work envelope, and the arm limits and singularities were not directly perceivable. They were indirectly perceivable though, through observing the amount of travel left in each joint. Thus, in order to avoid the arm limits, participants had to either observe the configuration of the arm joints and note when a particular joint was approaching its limit, or else use a trial-and-error strategy. One such trial and error strategy observed being employed by a few participants

[1] This phenomenon also occurs with the human arm, where reach of the tip of the finger is restricted if the hand is required to be held at a particular angle. By way of illustration, try touching the tip of your nose with your little finger while holding your palm facing upwards.

involved the participant identifying areas of difficulty using small-scale exploratory movements and working around these.

Failure to avoid the arm limits not only hindered progression in the task, it also incurred a time and motion penalty in the form of the 3-s backtrack manoeuvre. This particularly affected a small proportion of participants whose control style was typified by rapid poorly-specified exploratory movements of the arm, because the end-effector could travel a considerable distance along an optimal path, only to diverge from optimality at the last instant and strike an arm limit, causing the arm to backtrack over a substantial portion of the path travelled in the seconds prior.

3.3.2 Score statistics

The distributions of participants' *arm reach limit violation count* scores for the three test tasks appear in Figure 17. Each histogram represents the scores of 38 participants divided into 10 groups. The distributions are highly non-normal, enough to prevent the application of common inferential statistical procedures that are based on assumptions of an underlying normal distribution and homogeneity of variance.

In order to assess quantitatively the effect of experimental task and condition on differences in rates of participant arm movement errors, the Friedman ANOVA, a non-parametric equivalent of a one-way ANOVA for repeated measures, was applied to the data. The null hypothesis under test was that the rankings in the scores when grouped by experimental condition were not significantly different from condition to condition, i.e. that the differences in the ranks between conditions were zero.

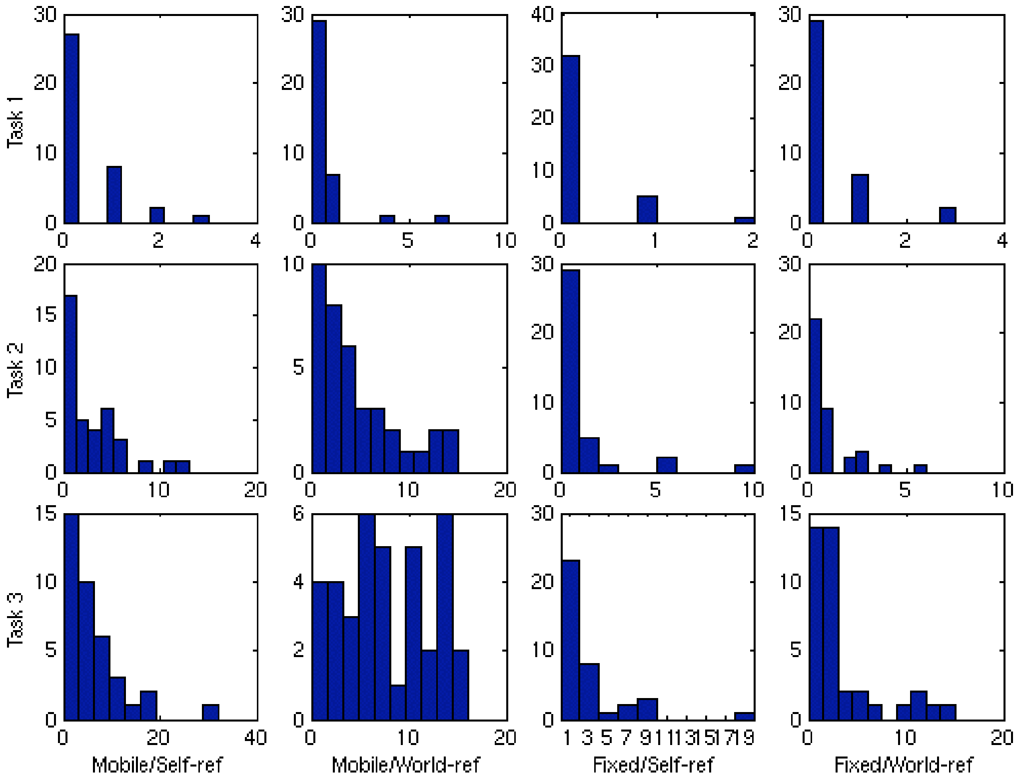

Figure 17. Distributions of participants' *arm reach limit violation count* scores. Rows of graphs are ordered by test task, columns of graphs are ordered by experimental condition. (N.B. Scale varies.)

3.3.3 Summary of results

The distribution of participants' arm limit count scores for the three test Tasks 1 – 3 are displayed in Figure 18, as a function of experimental condition.

The differences in arm limit counts between conditions were not significant for Task 1 [$\chi^2(3) = 2.25$, N = 38, $p < .52$] but were highly significant for both Task 2 and Task 3 [$\chi^2(3) = 48.14$, N = 38, $p < .000$ and $\chi^2(3) = 32.86$, N = 38, $p < .000$ respectively]. Examination of the confidence intervals around the median scores, indicated by the notches in the boxes in Figure 18 indicates that performance in Tasks 2 and 3 was worst under the mobile/world-referenced conditions, followed by the mobile/self-referenced condition, and that the two fixed viewpoint conditions were best, but not significantly different from each other.

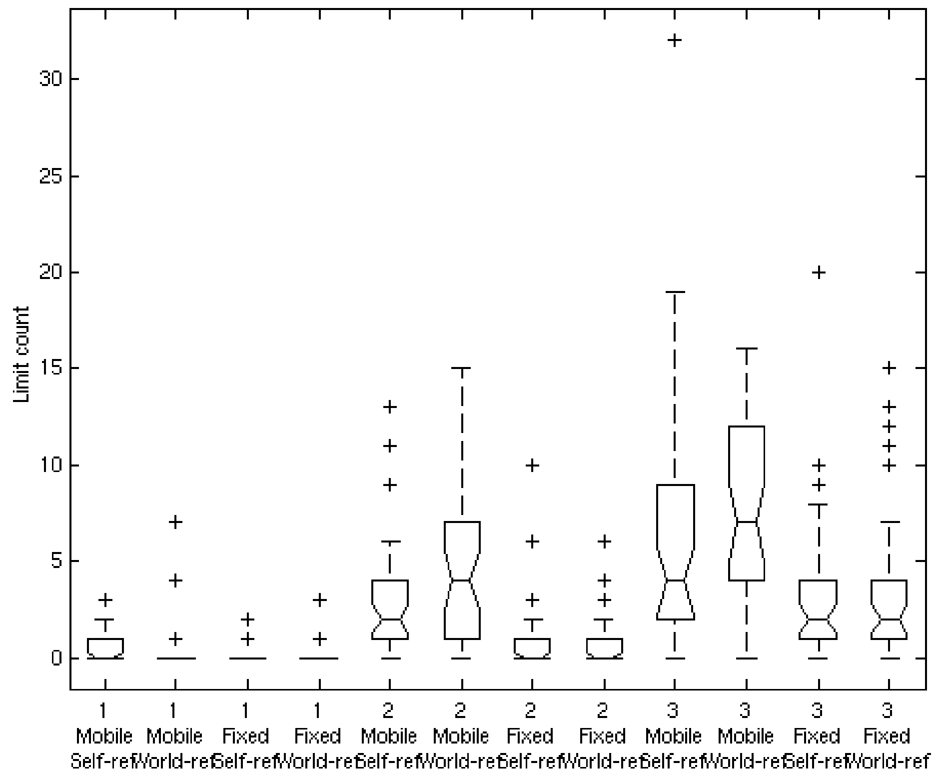

Figure 18. Multiple notched box-plots show distribution of participant *arm reach limit violation count* scores. Each box represents a different combination of task, viewpoint frame of reference and control frame of reference, as indicated by the label beneath. Central line of box is median; upper and lower bounds of box are inter-quartile range (IQR), and whiskers represent spread (excluding outliers). Outliers are defined as values greater than ±1.5 □ IQR above or below the box, and are marked by + symbols. The notch in each box represents a .95 confidence interval for the median.

These comparisons of medians effectively amount to multiple planned comparisons, hence should be viewed cautiously with respect to their robustness to Type I errors.

3.3.4 Arm limit violation correlation with time-to-completion

There was a moderate degree of correlation between the number of arm reach limit violations made by participants in each task and performance as measured by projected time to completion. *Projected time to completion* was obtained by taking the reciprocal of Completion Rate and multiplying by 2. The measures were moderately positively correlated, with the values of Pearson's *r* for the correlations equal to 0.66, 0.65 and 0.65 for Tasks 1, 2 and 3 respectively. Some degree of linkage between these two factors is quite understandable considering that each arm limit violation was penalised by 3 s and thus lengthened the time to completion by 3 s. It is useful to quantify the degree of linkage however, and $r^2 = 0.42$ of the variance in projected time to completion in each trial was accounted for by arm limit violations, uniform across all three test tasks.

3.4 Collisions

3.4.1 Motivation for and derivation of measure

Participant performance, as measured by Completion Rate provides us with a useful index of the compatibility between different display and control conditions and the task at hand. However, it does not completely account for one of the most critical determinants of participant performance, that of errors. In the experiment, although the two different types of discrete manipulation errors that the participant could make were penalised in similar ways, they have radically different consequences in real manipulation tasks. Whereas arm limit violations merely impede progress of a manoeuvre, collisions can have catastrophic consequences.

A secondary motivation for analysing the number of collisions occasioned in each condition is that in many manipulation tasks there is a speed-accuracy trade-off. In order to assess whether a speed-accuracy trade-off also occurred in this experiment, the number of collisions between the manipulator arm, the payload and the body of the space shuttle served as a discrete measurement of manipulation accuracy.

3.4.2 Score statistics

The distributions of participants' *collision count* scores for the three test tasks appear in Figure 19. Each histogram represents the scores of 38 participants divided into 10 groups. Similarly to the scores for the arm reach limit violation measure, the distributions are highly non-normal. The Friedman ANOVA was again utilised to assess quantitatively the effect of experimental task and condition on differences in rates of participant collisions errors. As was the case for the arm-limit measures, the null hypothesis under test was that the rankings in the scores when grouped by experimental condition were not significantly different from condition to condition, i.e. that the differences in the ranks between conditions were zero.

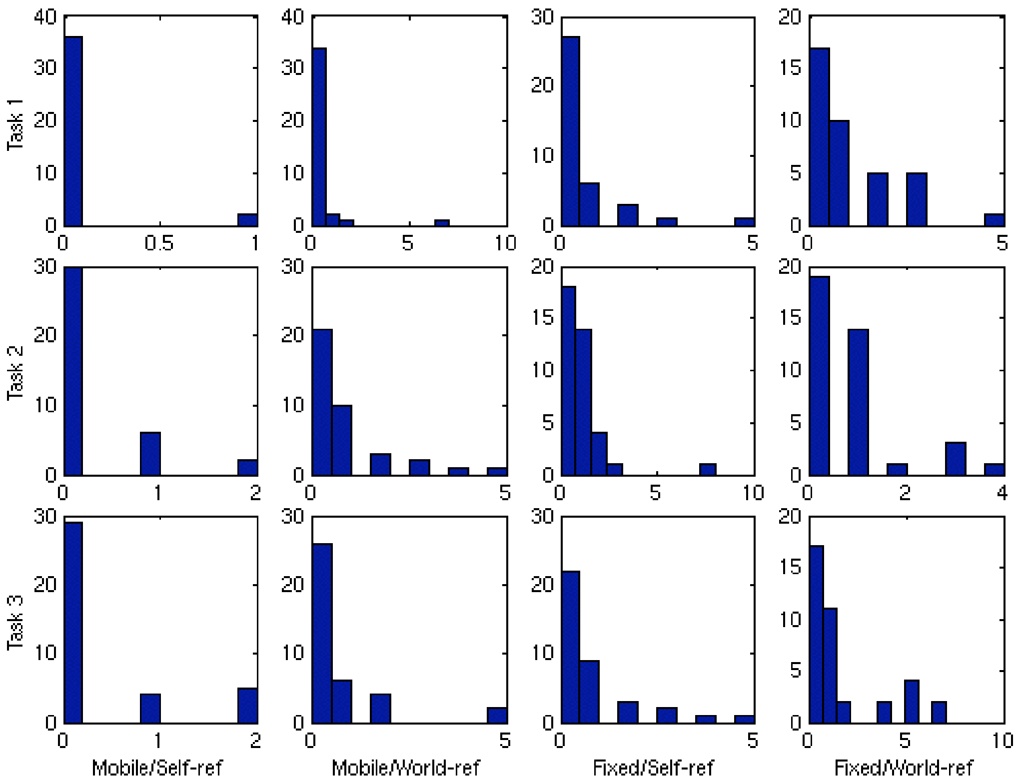

Figure 19. Distributions of participants' *collision count* scores. Rows are ordered by test task, columns are ordered by experimental condition. (N.B. Scale varies.)

3.4.3 Summary of results

The distributions of participants' collision count scores for test Tasks 1 – 3 are displayed in Figure 20, as a function of experimental condition.

The differences in collision counts between conditions were highly significant for Task 1 [$\chi^2(3) = 29.29$, $N = 38$, $p < .000$], marginally significant for Task 2 [$\chi^2(3) = 9.38$, $N = 38$, $p < .024$] and significant for Task 3 [$\chi^2(3) = 12.57$, $N = 38$, $p < .0057$]. Examination of the confidence intervals around the median scores (indicated by the notches in the boxes in Figure 20) revealed a strong first-order effect for viewpoint frame of reference, with the number of collisions under the mobile viewpoint clearly fewer than under the fixed viewpoint in all three tasks.

There was also some degree of second-order interaction between factors in Tasks 1 and 3. In Task 1 the two mobile viewpoint conditions were best for avoiding collisions and not significantly different from each other, whereas the fixed viewpoint condition experienced a higher number of collisions and there was a simple effect of control frame of reference across the two fixed viewpoint conditions. Although there was an increase in collisions in Task 3 under the world-referenced control frame (evident as a rise in the upper quartile and range for both mobile and fixed levels of the viewpoint factor), this increase was potentiated under the fixed viewpoint.

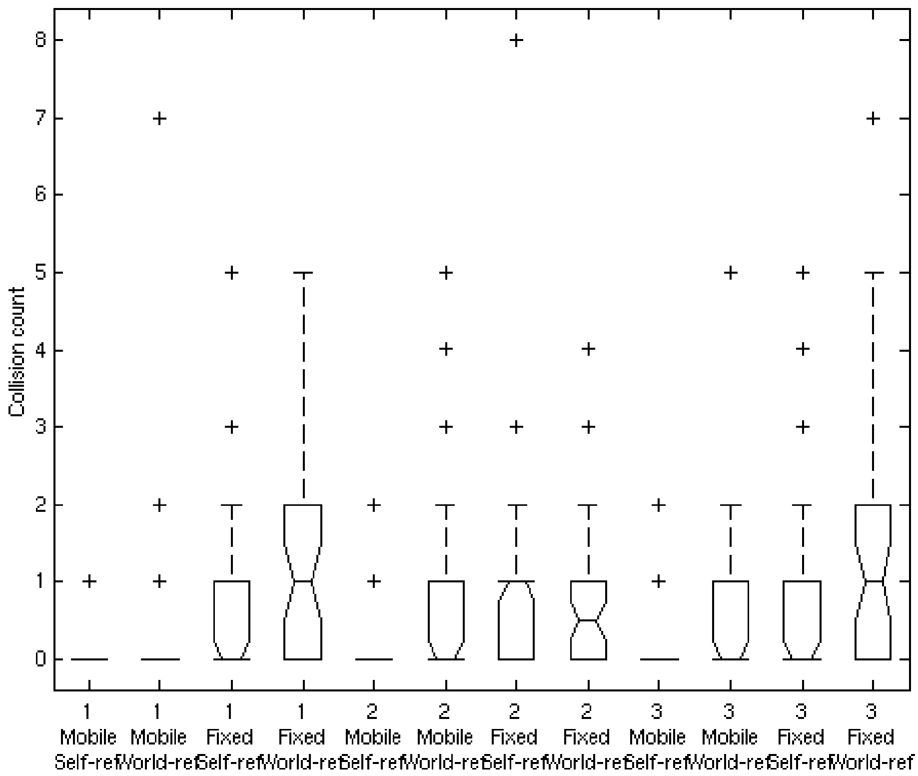

Figure 20. Multiple notched box plots show the distribution of participant *collision count* scores. Each box represents a trial with a different combination of task, viewpoint and control frame of reference, as indicated by the label beneath. Central line of box is median; upper and lower bounds of box are inter-quartile range (IQR), and whiskers represent spread (excluding outliers). Outliers are defined as values greater than ±1.5 IQR above or below the box, and are marked by + symbols. The notch represents a .95 confidence interval for the median.

The prior caution mentioned with respect to visual comparisons of medians effectively amounting to multiple planned comparisons and the resultant increase in familywise Type I error rate also applies equally here.

3.5 Head movements

Although the Completion Rate measures for Tasks 1 – 3 generally indicated a performance advantage for the fixed viewpoint, observations of the participants working in the fixed versus the mobile viewpoint indicated that the performance advantage might have been related to the increased demands upon visual search behaviour necessitated by the mobile viewpoint. An analysis of participant head movement data was conducted to determine the degree to which the amount of head movement made by a participant was differentially influenced by the experimental conditions.

Head movements during each trial were recorded in raw form as a time series of head orientation and position data, and rotational head movement data was derived by calculating the difference in head orientation at each point in the time series. A measure of total rotational work done by the head during each trial was obtained by summing the absolute values of the rotational head movement time series data for that trial. Translational head movement was not considered because translational head movements were primarily a function of the participant's seated posture at the experimental console. Thus they held little relevance to the task requirements and were more likely to be an artefact of natural variability in the head position of a seated participant. Henceforth, references to total head movement will refer to only the rotational component of total head movement work.

The distributions of participant head movement scores for each trial were found to have marked positive skew, and a variance proportional to their mean. A log transform was applied to the data (the natural logarithm, $\log_e$) and this produced distributions with approximately normal shape and uniformity of variance.

A 3 × 2 × 2 (Task × Viewpoint Frame of Reference × Control Frame of Reference) factorial ANOVA with repeated measures on all three factors was conducted, and is presented in Table 6 in Appendix E. Head movements increased with task difficulty (in order of Tasks 1 – 3) and were substantially larger under the mobile viewpoint, as indicated by the highly significant main effects for task and viewpoint frame of reference [$F(2, 74) = 303.1$ $p < .000$, and $F(1,37) = 851.1$, $p < .000$ respectively]. Post-hoc analysis of differences between the means using Tukey’s HSD test indicated that all differences were significant between these main effects.

Additionally the first-order interaction of task and viewpoint frame of reference was significant. The interaction is plotted in Figure 21 and appears to indicate a slightly non-uniform difference in head movements between the two viewpoints from Task 2 to Task 3. This difference, although statistically significant is too small to be of any practical significance.

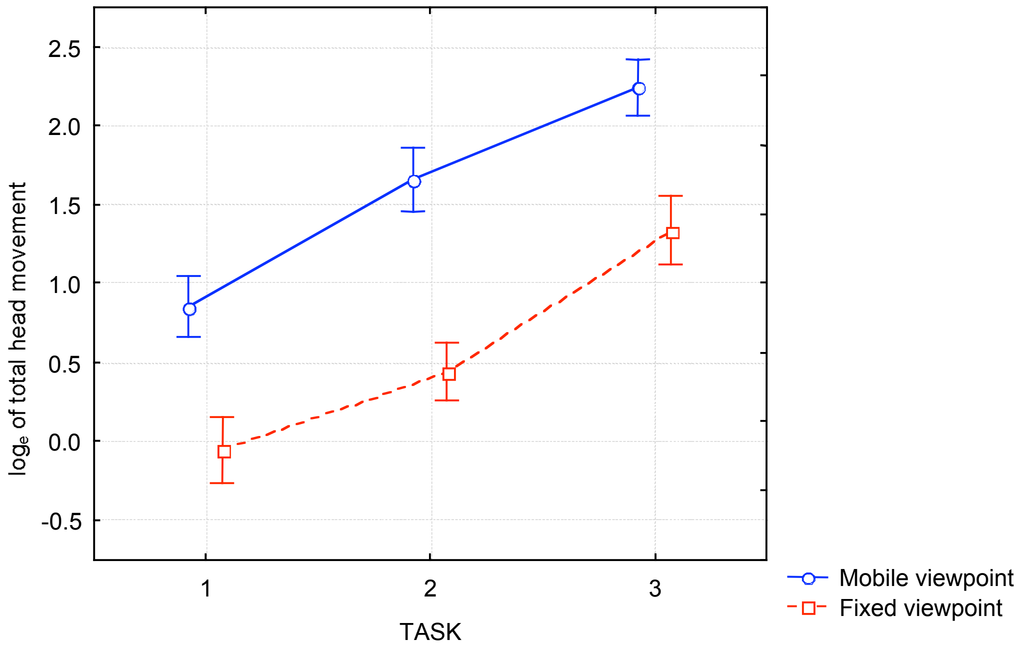

Figure 21. The significant interaction [$F(2,74) = 19.79$, $p < .000$] of task and viewpoint frame of reference on $\log_e$ of total head movements per trial. Vertical bars represent 0.95 confidence intervals.

3.6 Hand controller work

A second direct measure of the physical work done by the participant in each trial was that done by the hand controllers. Analysis of hand controller data was performed in order to test the hypothesis that the total work done on the hand controls was a function of the experimental condition.

Hand controller actions were recorded in raw form as a time series representing the excursion of each hand control from its rest point. The hand controllers were both rate controllers, where a constant excursion from the rest position specified a constant rate of change in position or orientation. A discrete-time integration operation (Equation 2) was performed on the hand controller data over the duration of each trial in order to obtain a measure of the total work done by the hand controller during that trial. Raw readings from both hand controllers were standardised so that maximum excursion in one axis was represented by the value 1.0 for that axis. This allowed both position and orientation measures to be summed to produce a composite total bimanual control work measure for each participant in each trial.

A 3 × 2 × 2 (Task × Viewpoint Frame of Reference × Control Frame of Reference) factorial ANOVA with repeated measures on all three factors was conducted, and is presented in Table 7 in Appendix E. There were significant main effects for task and viewpoint frame of reference, and a marginally significant main effect of control frame of reference [$F(2,74) = 152.3$, $p < .000$; $F(1,37) = 20.76$, $p < .000$; and $F(1,37) = 5.70$, $p < .022$ respectively.] There were also significant first-order interaction effects of task and viewpoint frame of reference and a significant three-way interaction between task, viewpoint frame of reference and control frame of reference [$F(2,74) = 7.34$, $p < .0012$; and $F(2,74) = 5.18$, $p < .0078$ respectively]. A plot of the three-way interaction appears in Figure 22.

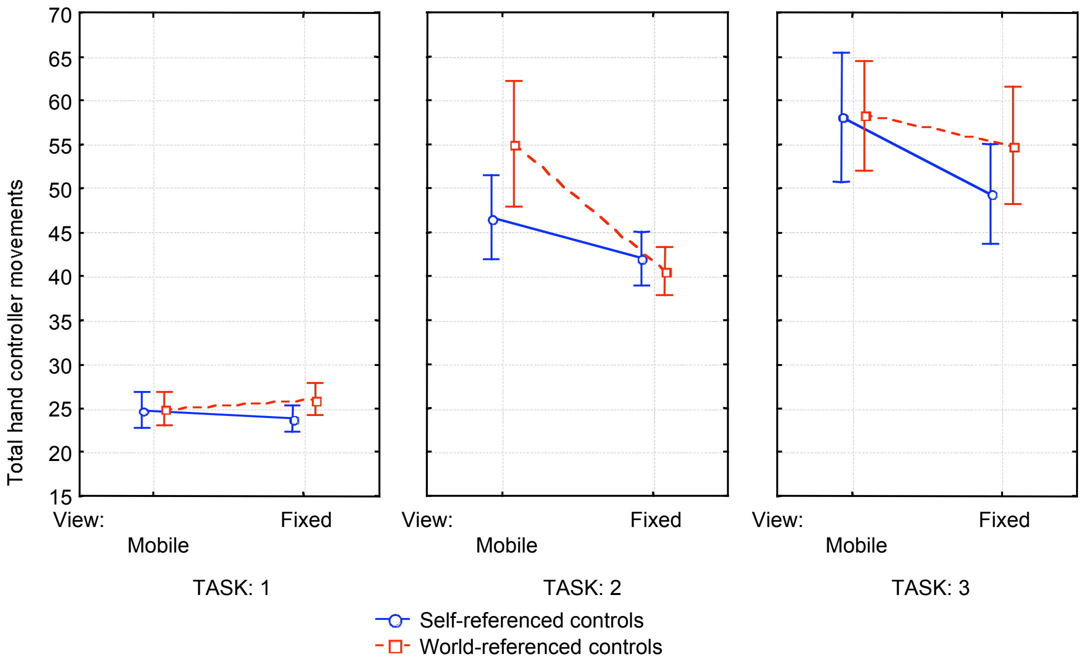

Figure 22. The significant three-way interaction effect of task, viewpoint frame of reference and control frame of reference on total hand control work. [F(2, 74) = 5.18, p < .0078]. Vertical bars represent 0.95 confidence intervals.

Post-hoc individual comparisons were made using Tukey's HSD test in order to identify the significant differences between conditions within each task. None of the comparisons between conditions in Task 1 were significant, indicating that participants did not expend more control effort manoeuvring the payload in any one condition in Task 1.

The results for Tasks 2 and 3 appear to be the mirror image of the results found in Tasks 2 and 3 on the Completion Rate measure (see Figure 10, p36) and this reciprocal relationship is supported by the fact that in Tasks 2 and 3, the set of significant differences between conditions that emerged matched pair-for-pair those observed in post-hoc comparisons of the Completion Rate results. This reciprocal relationship indicates that those conditions that were associated with slower rates of completion were also associated with increased total hand control work.

To further characterise this negative linkage between Completion Rate and total hand control work, Pearson's r measure of correlation was calculated across all participants and all trials for Tasks 1 – 3 and a moderate negative correlation was found, $r = -0.67$, $r^2 = 0.45$.

3.7 Simulator sickness effects

3.7.1 Simulator sickness statistics

Intolerable simulator sickness effects were felt by 6 of the 45 participants in the sample. These participants withdrew from the experiment prior to completing the test phase. No participants withdrew prior to the commencement of the test phase.

Subjective simulator sickness effects were evaluated using a slightly modified version of the Simulator Sickness Questionnaire (Robert S. Kennedy, Lane, Berbaum, & Lilienthal, 1993). The change in SSQ scores from pre-experiment to post-experiment measured over all 45 participants was found to be an increase with a mean of 142.7 points on the SSQ scale and a standard deviation of 242.8. This includes eight participants who experienced a *decline* in SSQ-related symptoms over the duration of the experiment. These participants verbally reported a high degree of engagement with the simulation.

3.7.2 SSQ correlates

Over all 45 participants there were no significant correlations between change in SSQ from pre-experiment to post-experiment and either reported level of prior simulator experience ($r = -0.1$), or gender ($r = 0.24$).

For the 38 participants who completed the experiment, there were low positive correlations between change in SSQ and total rotational head movements made by the participant (r = 0.45, r^2 = 0.20), and total control work made by the participant (r = 0.38, r^2 = 0.14). There was no correlation between change in SSQ and overall participant performance as measured by mean Task Completion Rate over the 12 test trials (r = –0.1) and a low positive correlation with participant performance as measured by total number of arm limit errors and collisions (r = 0.39).

3.7.3 Verification of SSQ efficacy against objective measures

Although the SSQ has proved efficacious in its use elsewhere (Robert S. Kennedy et al., 1993), it remains a self-report measure, and thus potentially open to subjectivity and misreport. It has been noted that it has shortcomings in accounting for some motion sickness effects (Gianaros, Muth, Mordkoff, Levine, & Stern, 2001). In one strong example where the SSQ failed to account for a participant's simulator sickness symptoms in this experiment, a participant who withdrew because of intense discomfort did not rate any of the SSQ symptoms higher than *slight* on the post-experiment SSQ. Yet this participant withdrew earlier than anyone else and also experienced the longest duration after-effects, reporting verbally to the experimenter that he had felt sustained dizziness and disorientation for 14 hours after the experiment.

Although no direct objective measurements of simulator sickness-related factors were made in this experiment, one objective measure was available in the form of a categorical division of participants into those who withdrew part way through and those who completed the experiment. The correlation between changes in SSQ scores from pre-experiment to post-experiment and early withdrawal from the experiment was calculated in terms of Pearson's point-biserial coefficient, and change in SSQ scores was found to account for only 16% of the variance in early withdrawal, $r_{pb} = 0.396$, N = 45. The correlation is indicated graphically in Figure 23.

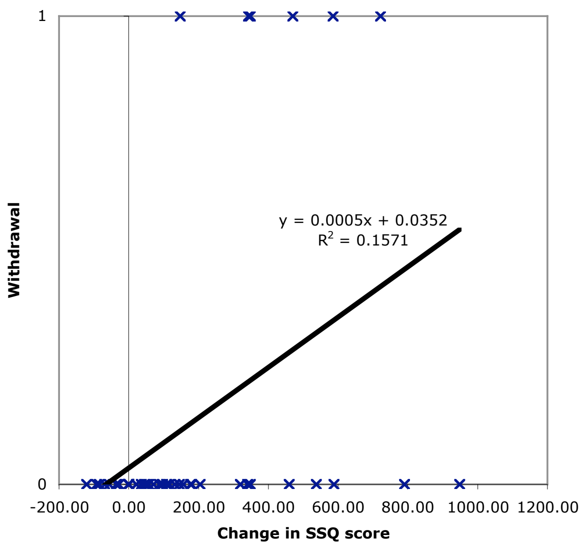

Figure 23. Graphical determination of point-biserial correlation between change in SSQ scores and early withdrawal due to simulator sickness. Early withdrawl is scored as 1.0 on the y axis, completion without withdrawal as 0.0. The slope of the regression line is equal to Pearson's point-biserial measure.

Based on this analysis, the utility of using change in SSQ scores as a measure to identify participants experiencing intolerable simulator sickness effects is questionable.

3.7.4 Gender effects

Gender differences in susceptibility to simulator sickness have been reported in the literature (Robert S Kennedy, Lanham, Massey, & Drexler, 1995). In this experiment, although females accounted for 10 of the 45 participants tested, they accounted for 3 of the 6 participants who *withdrew* from the experiment due to the intensity of simulator sickness effects they were experiencing. However, the one-tailed χ^2-Test of Homogeneity of the hypothesis that females were more likely to *withdraw* due to simulator sickness than males was not significant at the 5% level ($\chi^2(1) = 3.091$, $p < .079$), indicating that random sampling variation cannot be ruled out as the source of the higher proportion.

3.8 Demographics

3.8.1 Prior simulation experience

Distribution of participants prior experience with 3D computer generated environments was assessed via self-reported number of hours, and was categorised into five broad categories each varying by an order of magnitude. The distribution of the 38 scores for those participants who completed the experiment is depicted in Figure 24. As can be seen, there was a broad spread of experience levels, and the experience levels were reasonably well distributed amongst the categories, with the exception of the 50 – 500 hours category. The relatively high number of participants in the upper categories owed perhaps to the relatively high proportion of participants who were engineering students.

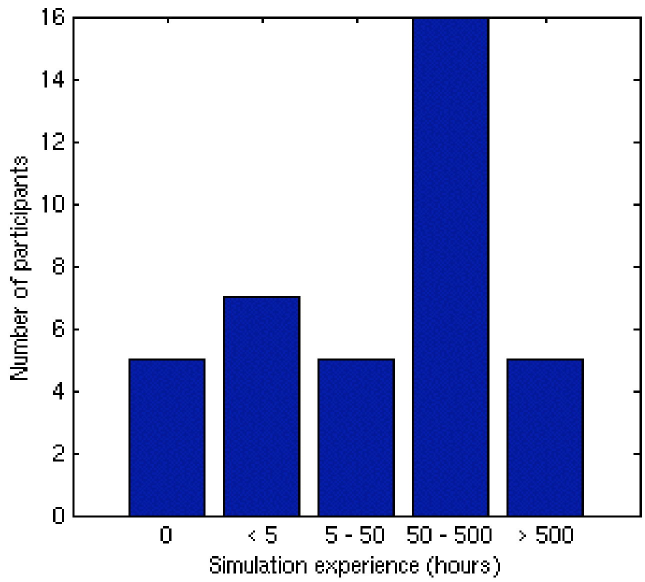

Figure 24. Distribution of participants' reported prior simulation experience.

Participants' reported level of simulation experience was compared against three different measures. Performance on task Completion Rate, averaged for each participant over all trials in Tasks 1, 2, and 3, was not correlated, *r* = 0.23. Performance as measured by total number of arm limit errors and collisions for each participant was moderately negatively correlated, (*r* = –0.51, r^2 = 0.26) indicating that to a small degree participants who were more experienced made fewer manipulation errors. Lastly, in order to determine whether more experienced participants were more efficient in respect of expending less control effort to complete the task, a measure of *control intensity* was formed for each participant by summing the total hand controller movements over all trials in Tasks 1, 2 and 3 and dividing by the reciprocal of the averaged Completion Rate over the same trials. This measure showed no correlation with simulator experience, *r* = –0.20.

3.8.2 Gender effects

Comparisons were made between males and females to test for significant performance differences on the basis of gender. Two measures that appear above were reused: average task Completion Rate for each participant over all trials in Tests 1, 2 and 3, and total number of manipulation errors over the same trials. The distributions by gender of participant scores on these measures are illustrated using notched box plots in Figure 25 and Figure 26.

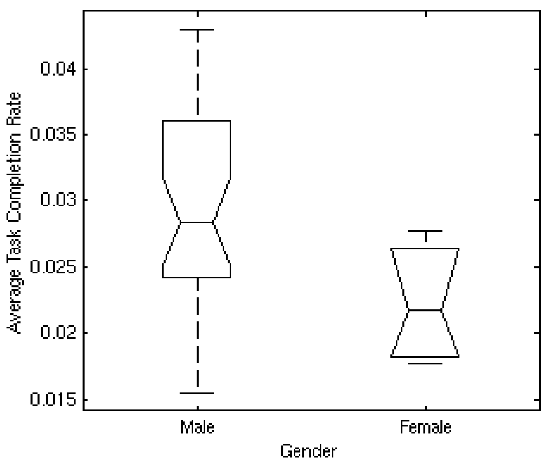

Figure 25. Notched box plot indicating distribution of participants' performance scores on average Completion Rate. Higher y values indicate better performance.

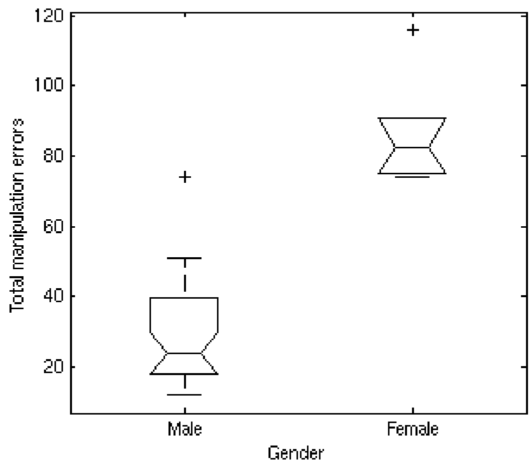

Figure 26. Notched box plot indicating distribution of participants' performance scores on total number of manipulation errors. Lower y values indicate better performance.

Male performance was found to be significantly higher than female performance on both measures. One-tailed t tests for two independent samples were applied to the performance measures to test the directional hypothesis that male performance was better than female performance and were significant on both measures, $t(36) = 2.48$, $p < .009$ for average Completion Rate, and highly significant, $t(36) = -8.57$, $p < .000$, for total number of manipulation errors.