Contents

Chapter 1

Prototype Design

Chapter 2

Pointing Experiment

This chapter was originally published in the conference proceedings for AIAA Modeling and Simulation Technologies 2015.

2.1 Introduction

The underlying motivation for this project is to bring high-fidelity simulator response to an earlier stage in the cockpit design cycle, where traditionally, the focus is on layout of displays and controls in a minimally-functional mockup. With the R3C concept, system functionality can be experienced from the start, in the virtual environment via a head-mounted display (HMD), while the user interacts with the cockpit systems by touching the surfaces of a low-fidelity (3-D printed) environment. The user receives the tactile feedback of touching a button or edge-key, and via hand-tracking technology, sees the response in the virtual displays. Re-

quired changes to the user-interface environment or even "what-if" ideas can be made quickly and at low cost by modifying the 3-D printed surfaces, and then reflecting these physical changes in the virtual visual scene. By creating a high-fidelity virtual simulator on top of a low-fidelity geometric mockup, this approach may allow more design/layout iterations at lower cost, which could lead to a final design that is closer to optimum than would be feasibly attained with traditional approaches.

We have developed a proof of concept of the R3C system, which was used to perform an initial, simple, targeting study to validate our approach, and which will be the basis for future work. Figure 2.1 shows an over-the-shoulder view of a user of the R3C prototype with annotations to call out the major components of our current technical approach. The user is wearing an immersive virtual reality head mounted display (Oculus Rift Development Kit 2), which presents them a virtual scene (shown in Figure 2.2) that is spatially stabilized with respect to the physical instrument panel. As the user reaches out toward the inert 3D printed instruments, a LeapMotion hand tracker reads the position of each finger and the pose (attitude and configuration) of the hand. A simple collision detection algorithm, described below, uses this information to determine when a user has touched a button. Since it is important for the user to visually track their own hand during pointing and actuation tasks, we also use the hand tracker to render the dynamic position of the hand in the virtual scene (Figure 2.2). On certain buttons we have added copper pads that are connected

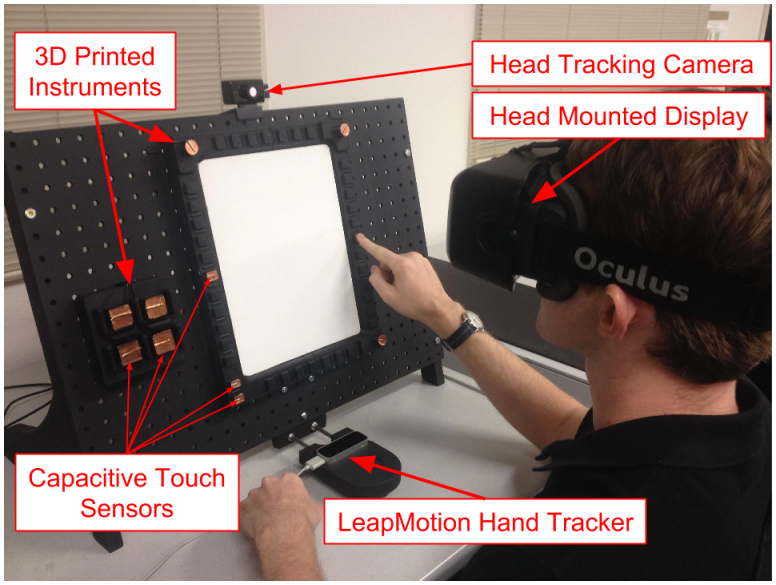


Figure 2.1: Prototype R3C System

to a capacitive touch sensor as a secondary sensor for button actuation. The camera mounted above the panel is an infrared camera that senses the position of the HMD. The display on the screen can be configured as desired; it is currently showing an external camera view. The behaviors of each button are controlled in software, and can be easily reconfigured.

The use of virtual or augmented reality in aerospace has been extensive, starting with the first flight simulators, and even in the cockpit in research studies[10, 1]. Significant recent work investigates virtual reality tools in a simulator[12, 15, 27]. However, our concept of merging an accurate but low-cost tactile environment with the high-fidelity virtual view is so far untested. Accurate haptic feedback has been a goal of virtual reality (VR) researchers since the emergence of VR. Providing dynamic haptic feedback, however, still proves challenging to date[26, 19]. Our approach

Figure 2.2: Virtual view of user from Figure 2.1

allows accurate haptic feedback for the case of a static workstation, such as found in a real cockpit. Combined with 3D printing and our virtual simulator overlay, this provides an inexpensive platform to create a functioning simulator, able to adapt quickly to large-scale design changes.

2.2 Technical Approach

This section details the components of our prototype and how they are integrated into the R3C system, as well as rational for selection, improvements and lessons learned.

2.2.1 3D Printed Instruments

A central tenant of our technical approach is to use physical, geometrically accurate instrument shapes that provide no functionality (i.e. no screens, no working buttons, etc.). This is intended to imitate the fidelity

of a typical early design stage mockup. In order to achieve this we have produced 3D-printed "instruments" that can be easily rearranged on a panel mount (pegboard). Since the devices are rapidly prototyped, they can be redesigned in a much smaller time frame than typical simulator instruments. By using the geometrically accurate instruments at the correct cockpit locations, the user is provided with accurate tactile and proprioceptive feedback without the need for entering the challenging field of virtual haptic feedback.

2.2.2 EDGE Rendering Engine

We are using a NASA developed rendering engine named EDGE8 to provide the visuals for the virtual scene rendered in the head-mounted display (HMD). EDGE is highly customizable and extendable through C/C++ plugins, Tcl scripts, or networking functions. Existing dynamic simulations (e.g. Matlab) can also be integrated.

2.2.3 Oculus Rift

The Oculus Rift virtual reality head mount display is currently only available as a developer kit. The lightweight headset provides an immersive virtual reality experience by combining a wide field-of-view scene with accurate head tracking, giving a stable virtual world. The small display (cell phone sized) is viewed through a single set of lenses, and the barrel distortion of the lens is corrected for in the rendering engine by a pincush-

ion distortion of the rendered scene. The orientation of the head is tracked using internal sensors as well as an infrared camera to give an accurate pose for rendering the scene. The software developers kit9 (SDK) exposes the head position and orientation and provides the appropriate distortion for a rendered scene (which is done for each eye), and this has been developed into an EDGE plugin. The technology of these head mounted displays is improving rapidly, driven by technology developed for the cell phone and demand from the gaming industry.

The head tracking provided in the newest development kit version gives a significant improvement over the original Oculus Rift in the registration between the real world and the virtual world. The head-tracking camera is mounted on a known location on the panel (see Figure 2.1), and the tracking software provides a measurement of the relative location between the head and the camera. The virtual world can then be rendered relative to the location of the camera.

2.2.4 LeapMotion Hand Tracker

The LeapMotion hand tracker is a small, consumer orientated optical hand tracker. It uses two infrared cameras as the source of its proprietary tracking algorithm. As of version 2.0 of their SDK10, they expose to the application developer the positions of all the joints and bone positions of the hand. An EDGE plugin was developed in-house at UC Davis to position the nodes in the scene from this information, which is the source

for rendering the virtual hand as well as the button detection algorithm.

We decided to use the LeapMotion as it provides logistical simplicity compared to professional motion capture devices that require multi-camera setups with lengthy calibrations as well as high cost. A main goal of our system is that it can be used without much setup in the physical world, leading us to the new small optical trackers. There are other promising devices (i.e. Kinect2 and other 3D camera systems) that could be used in the future.

LeapMotion Calibration

Early in the integration process, we discovered that the LeapMotion provided precise and repeatable measurement of the hand positions throughout its tracking volume. The accuracy as the hand got further away from the sensor, however, was insufficient. Put another way, the position of the hand in the virtual world was offset from the true button position in the physical world, yet the offset was consistent between movements. This led us to develop a calibration to provide a more accurate registration between the virtual and physical hand positions. Since we accurately know the position of the panel and instruments, it is only required to solve the following linear algebra equation for the transformation matrix, $\mathbf{T}$:

$$\vec{x}_{known} = \mathbf{T}\vec{x_{measured}} \tag{2.1}$$

Using a simple least squares approach to find the coefficients of the

matrix, the registration between the virtual and physical worlds is vastly improved. The transformation matrix is not constrained to a simple rotation (i.e. not assumed orthogonality or other special properties) so the solution is found by expanding and solving the general homogenous coordinates transformation matrix.

$$\begin{bmatrix} x_{known} \\ y_{known} \\ z_{known} \\ 1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{31} & T_{32} & T_{33} & T_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{measured} \\ y_{measured} \\ z_{measured} \\ 1 \end{bmatrix} \tag{2.2}$$

Typical least squares approaches would attempt to find $\mathbf{x}_{measured}$ in Eqn. 2.1, however we desire to find the matrix T itself. It can be shown that expanding the matrix equation (Eqn. 2.2) for multiple points (i.e. $\vec{x}_{known,1}, \ldots, \vec{x}_{known,n}$ and $\vec{x}_{measured,1}, \ldots, \vec{x}_{measured,n}$) and then collecting like terms will convert the problem into three different least squares problems. They are shown here, dropping the subscripts to k and m for known and measured.

$$\begin{bmatrix} x_{k1} \\ x_{k2} \\ \cdots \\ x_{kn} \end{bmatrix} = \mathbf{X}_M \begin{bmatrix} T_{11} \\ T_{12} \\ T_{13} \\ T_{14} \end{bmatrix}, \quad \begin{bmatrix} y_{k1} \\ y_{k2} \\ \cdots \\ y_{kn} \end{bmatrix} = \mathbf{X}_M \begin{bmatrix} T_{21} \\ T_{22} \\ T_{23} \\ T_{24} \end{bmatrix}, \quad \begin{bmatrix} y_{k1} \\ y_{k2} \\ \cdots \\ y_{kn} \end{bmatrix} = \mathbf{X}_M \begin{bmatrix} T_{31} \\ T_{32} \\ T_{33} \\ T_{34} \end{bmatrix},$$

$$\text{where } \mathbf{X}_M = \begin{bmatrix} x_{m1} & y_{m1} & z_{m1} & 1 \\ x_{m2} & y_{m2} & z_{m2} & 1 \\ & \cdots & & \\ x_{mn} & y_{mn} & z_{mn} & 1 \end{bmatrix}$$

At least 4 points are needed to solve this system, and it has been found that a calibration with small least squares residuals can be achieved with 10-20 well chosen points. The panel and instrument edges can provide well-known points, as well as button locations. The calibration matrix is then used to find the offset of the tip of the index finger for each hand, and this offset is used to move the entire hand.

LeapMotion Mounting Position

The LeapMotion software recently (as of version 2.2) began officially supporting a "VR Mode" which allowed users to mount the device on the front of a VR HMD and track the hands relative to the head. This was

tested in our system and although we experienced better tracking, the hand would not stay stable relative to the panel during head movement. This led us to develop a mount that would hold the LeapMotion above the panel looking down, and putting the software in "VR Mode" we were able to achieve better tracking but with a fixed known position, the hand remained stable relative to the panel.

2.2.5 Button Detection Algorithm

The button detection algorithm is a simple collision detection model. A rectangular box is defined that extends outside the button, including a tolerance zone to account for misalignment and poor tracking. When a fingertip enters and stays in the box for 50ms then a button event is triggered. An event is also triggered when a finger enters the box, which we use to change the color of the button to indicate proper alignment to the user. The advantage to using the optical tracking to determine when a user has selected a button is that it has the potential to significantly reduce the complexity of the system. If the pilot interactions with the panel can be determined solely by tracking his/her hands from the external sensor, then the cockpit panel needs only to provide physical feedback, and does not require any wiring. Of course, the primary drawback of this approach is that the efficacy is limited to the accuracy and reliability of the hand tracker, and the algorithms tracking this movement for button inputs. Nonetheless, as we show in our experiment, the LeapMotion coupled with

this algorithm can be used effectively.

For the experiment described below, the collision detection box was set to extend 0.5 inches inward and outward from the button, including a 0.1-inch tolerance on all dimensions. Aslandere [27] found no significant effect with button selection ability based on changing this volume in a purely virtual environment, so we did not change the collision detection box size for this first test of the R3C (although it is easily modified in software).

2.2.6 Capacitive Touch Sensors

For this early research phase, we needed a baseline method to determine if a button was touched, for comparison with the LeapMotion IR camera hand tracking method. A simple method was first developed using copper tape electrodes on top of certain buttons (as can be seen in Figure 2.1). These electrodes are connected to a Freescale MPR121 capacitive touch controller that registers when a finger touches them. The capacitive touch controller is connected to an Arduino, acting as a serial communicator between the microchip and the computer. This setup provides a binary state of whether a finger is touching the button. This is used to help determine whether the button detection algorithm using the hand tracker is working correctly.

Since touch accuracy can be important in safety-critical applications, we also desired the ability to record where on the button the finger press was located. This would help to determine if we had any registration

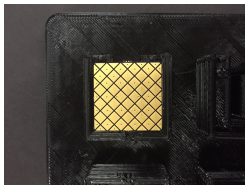


Figure 2.3: Capacitive touch array button. Each row and column of diamond pads are connected as one electrode each and together can provide location information of where the user presses.

biases in our system. To accomplish this, a custom printed circuit board was developed that provides an electrode array of 5 rows and 5 columns over a 1-inch by 1-inch square. This board is show in Figure 2.3, where it is mounted in a 3D printed instrument. The capacitive state of each row and column can provide a measurement of the center of the finger press on the grid created by the rows and columns. With this configuration finger-location accuracy of under 0.1-inch can be achieved. The location of the finger press can help provide a measure of the accuracy of the registration between the optical sensors and the real world location.

2.2.7 Summary and Lessons Learned

We've discovered a few qualitative effects of various technologies and methods throughout the time weve spent building this prototype, resulting in the following "lessons learned".

The registration between the virtual world and real world was a concern in the first prototype that did not include the head tracking setup of the Oculus Rift DK2, and it required precise measurement between the location of the user (including seat height) and the desk and panel. This would cause

problems when the head orientation yaw sensors drifted as well, where the user would not be lined up with the panel after a few minutes of use. The upgrade to the head-tracking camera to anchor the virtual world to the real world greatly improved the both stability and accuracy, and also significantly improved the feeling of presence in the virtual environment.

The registration between the real world and the hand tracker position in the virtual world is even more important, however. Since the user can feel the proper place for their finger when the finger arrives at the target, if the hand tracker is not lined up they may not be able to activate the button detection algorithm easily, or at all if there is physical interference. This problem created the need for the calibration scheme, and having the calibration is essential for using the LeapMotion in a task where the real and virtual world needs to be aligned. This problem also highlights the importance of testing people who are new to the system. New users were fixated on placing their finger where the physical feedback indicated the button was and did not adjust for misalignments with the hand tracker, while expert users learned to ignore the physical feedback and activated the hand tracker by finding the misaligned location. Additionally, we discovered that hand pose and speed can influence the performance of the hand tracker, such that expert users can perform in the system quicker and with greater accuracy. This is discussed again later in the context of the experiment performed.

The purely optical button detection algorithm is prone to failures, espe-

cially without a complex algorithm performing heuristics on the movement to determine the users desired selection. The simple collision detection model works well, but is difficult for the user without the loop-closing visual indication of when the detection box is activated. In our case we added the button changing color. It was also important to have a slight delay in activation, as this not only provides a failsafe for false positives, but it also adds a way to explore the tactile surface without consequences.

The latency of the entire R3C system has not yet been quantified, and there are two significant aspects where this might be interesting: from hand movement to hand tracker measurements through the rendering pipeline to the image shown to the user; as well as from head movement through to the image shown to the user. However, latency is clearly low enough to make measuring it low priority, as the hand movement delay in the virtual world is unnoticeable to the average user. Additionally, the latency of the Oculus Rift head mounted display is drastically improved from earlier technologies such that it does not cause a concern. Finally, the manufacturers of both LeapMotion and Oculus are actively improving on the latency on both fronts.

We also initially found tracking performance degradation when the finger approaches the instrument panel that was due to optical interference with the panel objects (i.e. it was happening only when the panel was present). This was improved when we moved the LeapMotion to the top looking down, but the larger effect was found when looking at the image

captured by the LeapMotion infrared cameras, discovering our black 3D printed instruments were highly reflective in infrared and showing up the same color as the hand. Applying a matte finish greatly improved this, but it is still a problem we are dealing with to date. We have also found better results when controlling entire backdrop of the LeapMotion field of view with a dark matte material, as this helps provide a greater contrast between the hand and the background. There still remains work to improve this, as a few subjects in the experiment indicated they found degraded tracking with the panel present.

2.3 Experimental Methods

2.3.1 Experiment Goals and Motivation

We performed a brief pilot study to validate our technical approach and take some measurements of the effects of using the R3C system. Specifically, we were interested whether a new user could accurately target the correct button in an instrument panel, and how the different technologies used affected the targeting task. There were three main effects of the targeting task under study in this experiment:

- The effect of the use of the VR HMD versus no VR HMD (i.e. virtual vs. real world).
- The effect of the physical panel versus no physical panel present (i.e.

having the tactile feedback vs. no tactile feedback)

- The effect of the optical tracking button detection versus the capacitive touch button detection

The use of the virtual reality headset also implies that subjects have to rely on the visual feedback from the hand tracker virtual hand to target the button.

We were concerned about the effect of both touch-selection accuracy and movement time, but the context for our design is an aerospace cockpit, where accuracy in selecting the intended button is typically paramount over movement time. Previous research in targeting tasks in virtual environments has not always reported on success rates, indicating incorrect trials were re-performed. We did not do this, as we wanted to ensure we recorded the success rate.

Previous work has investigated various 3D pointing tasks in the real and virtual world without tactile feedback[18], or with tactile feedback but no immersive virtual reality[29], or with virtual haptic feedback[8], or pure virtual worlds[4, 11]. Our work differs in the use of the tactile feedback provided by the panel combined with an immersive virtual world.

2.3.2 Experiment Design

Figure 2.4 shows a diagram of the experimental setup. The panel was configured with a single four-button keypad, arranged in a 2x2 grid. Three

different starting pads were placed on the desk near the edge. The buttons on the keypad were $1'' \times 1''$ and were equipped with the capacitive touch sensor arrays as described above, allowing the measurement of position of finger press. The start pads were capacitive touch copper tape electrodes, also $1'' \times 1''$, but had no position detection capability. The participants were asked to start with their index finger on the proper start pad and then target the correct button on the panel. Audio prompts were given through headphones, which indicated the starting pad and panel button goal before each task. Subjects were instructed to target the center of the correct button, but were given no instructions on speed. Subjects targeted the button using their index finger of their dominant hand, which was the right hand for all subjects. Each participant performed 48 targeting tasks under each of the four different conditions:

1. Wearing the HMD, button selection registered by capacitive sensors, panel physically present

2. Wearing the HMD, button selection registered by hand tracking sensors, panel physically present

3. Wearing the HMD, button selection registered by hand tracking sensors, panel not physically present

4. Not wearing the HMD, button selection registered by capacitive sensors, panel in position

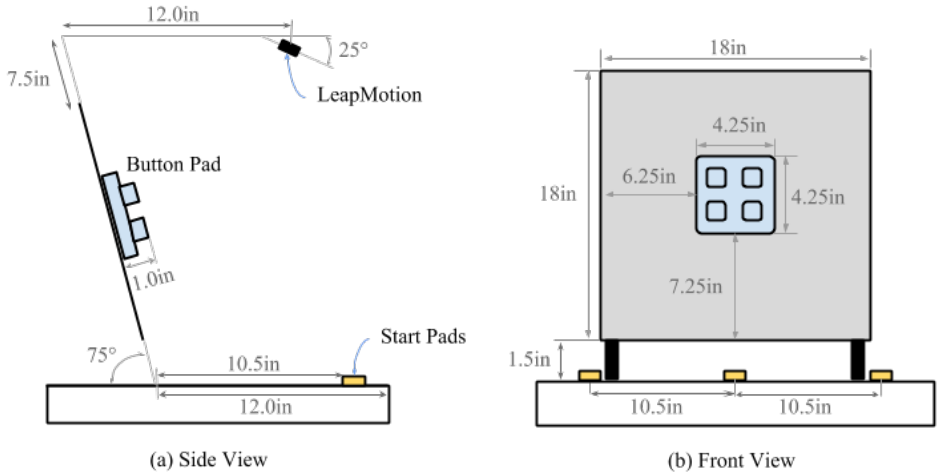


Figure 2.4: Diagram of Experiment Setup

Before each condition the subjects were given one set of 12 trials for familiarization of that particular condition. During the "no panel" condition, the hand tracker remained mounted at the same point, but the panel was moved out of the way, such that the subjects were targeting the buttons in the virtual world only. The button selection in every condition was indicated by a sound played in the headphones when the appropriate sensor measured a button press. In addition, during the optical sensor trials, the button color changed color to indicate they had entered the selection zone (as described in Section **??**). The tasks would time out after 10 seconds. The task was not repeated for an incorrect button selection or timing out, instead the subject moved on to the next task. The condition order was randomized for each subject, and the tasks were performed in a pseudo-random order.

The subjects were all briefed on the technology from a prewritten speech explaining how to activate each of the sensors. No information was given on how to best perform the task or advice on working with the various sensors.

The dependent variables measured were time from starting pad to trial goal button, accuracy measured on the capacitive touch sensors (except for condition 3 where this was not possible), as well as success rate. The raw data from the LeapMotion hand tracker was also recorded for each trial, allowing for post analysis of hand trajectory.

2.3.3 Subject Pool

The experiment was performed with 8 subjects, recruited from the university student population, all with no or limited previous exposure to the R3C setup. All subjects indicated they had either never used a virtual reality headset or briefly tried them once or twice. Ages ranged from 18-31 with 6 male, 2 female. They all had correctable vision to 20/20 and none indicated difficulty seeing the image in the virtual reality HMD. After the experiment, every subject indicated no motion sickness, and only mild eyestrain was reported.

2.4 Results

The time was recorded for each trial from release of starting pad to the successful button press using the correct button registration. Trials which timed out or where the incorrect button was targeted were discarded for the time analysis. The time for each trial was corrected for the varying distances of the movements by multiplying by the average distance (18.4 in) over the straight-line distance between the start and goal button for that trial (which varied from 15.5in to 20.5in). A boxplot of the distribution of the corrected time grouped by condition across subjects is shown in Figure 2.5. The effects of each were determined using a two sample T-test on the corrected time measure between the appropriate conditions. The effect of the use of the HMD between condition 1 and 4 (both using capacitive and panel) was significant (p ¡ 0.01), and on average caused the corrected time to drop 1.31 seconds. The optical detection algorithm also has a negative effect on corrected time (p ¡ 0.01), comparing conditions 1 and 2, but the drop in performance was only 0.55 seconds. However, it was found that the panel had no significant effect on performance across all subjects (conditions 2 and 3).

When comparing the results of the corrected time within subjects it was found that 3 subjects had a significant positive effect with the panel while only 1 subject had a significant negative effect, the rest showing no effect. Similarly, the optical tracking versus capacitive tracking had no effect on 3 subjects, while the rest display the same as the between subjects finding

Condition	Mean Corrected Time (s)	Average Speed (in/s)	Success Rate
1. HMD, Capacitive, Panel	1.88 ($\sigma = 1.47$)	14.51 ($\sigma = 7.32$)	98.6%
2. HMD, Optical, Panel	2.43 ($\sigma = 1.42$)	9.80 ($\sigma = 4.61$)	97.4%
3. HMD, Optical, No Panel	2.49 ($\sigma = 1.68$)	10.21 ($\sigma = 5.17$)	96.8%
4. No HMD, Capacitive, Panel	0.59 ($\sigma = 0.18$)	34.04 ($\sigma = 9.16$)	99.4%

Table 2.1: Mean results across subjects. Standard deviations are reported as σ.

of negative effect.

The selection of the correct button was performed almost without error, which is a promising result for new users of the system. There was no significant effect between the different conditions on success rate. Over 8 subjects and 1517 trials, only 12 trials selected the wrong button and only 12 trials timed out, giving an overall 98.4Observations made during the experiment indicate that at least some of the wrong button selections were due to misheard prompts or loss of attention.

The accuracy of the button press itself was also measured, and Figure 2.6 shows the distribution of the locations registered by the capacitive touch sensors on the first press of the trial (over the first 100ms). In the plots, (X,Y) = (0,0) corresponds to the bottom left corner of the button. The two conditions shown (1 and 4) were the two trials where capacitive touch registered the button selection. There is a smaller distribution for the no HMD condition (Figure 2.6(b)), but with the HMD (Figure 2.6(a)) subjects were still within error to the center of the button. It should be noted again here that in the HMD condition, subjects had to rely on the

virtual hand from the hand tracker for any visual feedback of hand position.

A Fitts law analysis of the data was not the goal with the experiment design, and since we did not vary the button width a Fitts analysis could only be performed with varying distance. The index of difficulty would only range from 4.0 to 4.5, providing an insufficient range for a proper Fitts analysis. Including extra parameters for the 3D nature of the task [22, 7] did not improve the fit, especially considering the symmetry of the setup. Furthermore the time recorded of the tasks were typically more than in a Fitts' task, since many subjects spent time honing based on the accuracy instructions.

2.5 Discussion

The start pads and hand tracker were located relative to each other in such a way that the hand would not be in view of the tracker at the beginning of the task, thus the user would have to wait for the hand tracker to acquire the hand when in view. This meant for the trials using the HMD the subject would typically have an open loop (blind) ballistic phase followed by the closed loop homing phase once the virtual hand appeared. Across subjects for all HMD conditions this hand acquisition by the tracker took on average 660ms ($\sigma = 649$ms) from leaving the start pad. An expert user who performed the experiment was able to get the hand to appear in the tracker in 330ms ($\sigma = 133$ms). This could indicate that training on how to get the best hand tracker performance could improve results. The

same expert also only experienced a 416ms slowdown between using the HMD and not using the HMD (conditions 1 and 4), compared to the 1.31s slowdown measured across the 8 new subjects.

We had originally hypothesized that having the panel would improve the selection performance of the button, but for this task it had no effect. This may turn out to be due to the instructions given, which called attention to accuracy. It is also possible that while the panel did not improve targeting performance in this task, it may provide a benefit to the subjective feeling of presence in the virtual world, which may be beneficial in a more demanding situation with multiple tasks.

Another variable that was not measured was fatigue. Having the panel as a backstop likely decreases the amount of fatigue for a repeated targeting task, as floating the arm in free space to target a button in a purely virtual environment with no panel can cause a buildup of arm fatigue. If the virtual world and physical world are correctly aligned, then the finger can rest on the button while the hand tracker works to detect a button press.

Overall, the lack of an effect of having the panel there could also be interpreted as a positive result. As discussed before, a significant problem weve been working on is the improvement of tracking performance when the hand nears the panel. The results indicate that this problem has been mitigated at least as much to provide performance on par with the purely virtual world.

2.6 Future Work

The promising results from the targeting pilot study indicate that users had no trouble accurately selecting the 1 inch square buttons. This would be a fairly large button area for an aerospace cockpit. A future study would investigate the limits of this system in terms of button size, and if smaller buttons pose any significant effect in accuracy. By varying the button size, this would provide a better experimental design to perform a Fitts Law analysis of the R3C system.

As previously discussed, an expert user can handle the inaccuracies of the system better than a novice, thus implying that the pilot study results did not reach an asymptote in training effects, or that there are long term training effects. The pilot study did not provide any training to subjects so as to not bias a certain condition, but future experiments may include a longer training session before subjects perform tasks in the R3C setup.

2.7 Conclusion

Although this initial application of the R3C system is to aviation/space vehicle cockpits, any sufficiently complex human-system interface could be designed with this system, such as telerobotics, air traffic control, robotic surgery, etc. The technology that supports our proof of concept system is rapidly improving, and further iterations of our technology integration will provide higher fidelity and an easier user experience.

We have showed that new users can accurately select buttons in a simple targeting task in our system. All of our measured effects showed no significance on correct button selection. The use of the head mounted display to provide visuals of hand position provides a small time penalty in button selection. The use of the physical panel provided no significant effect in time compared to having the subjects target purely virtual buttons. Our optical tracking algorithm had a slight negative effect in time compared to using the capacitive touch sensors. We are hopeful that training and improvements in the technology will reduce the performance gap.

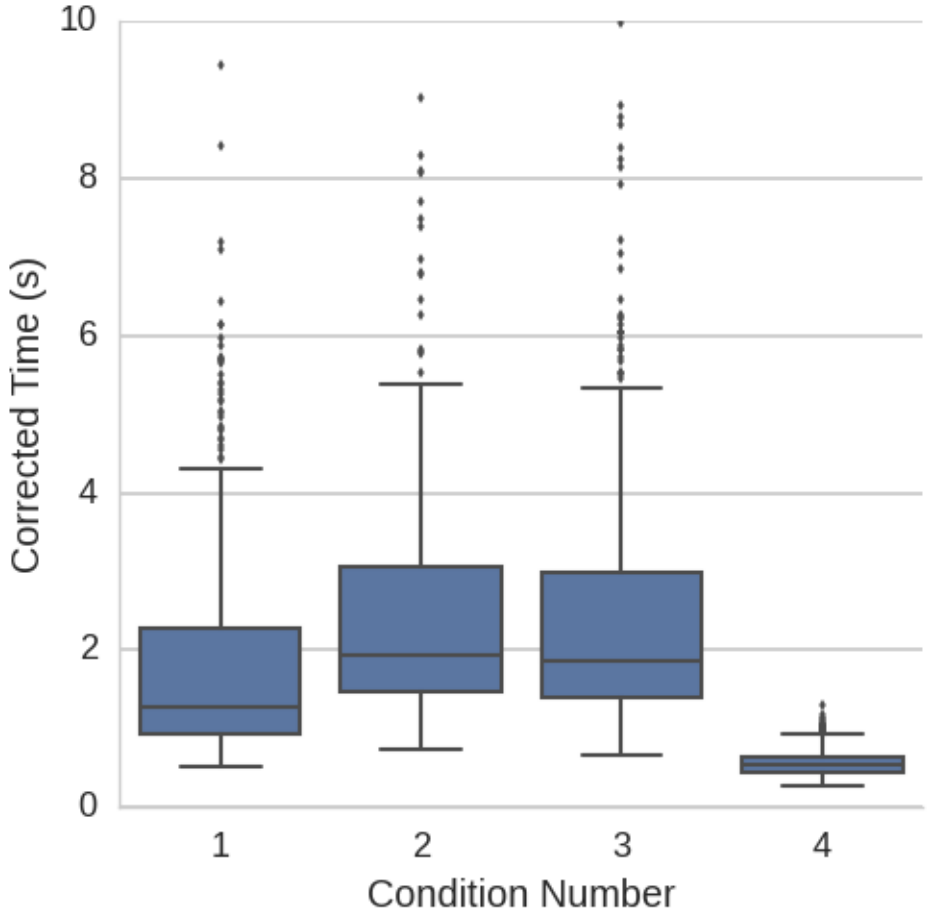


Figure 2.5: Boxplot of normalized time distribution by condition. Condition numbers are: 1. HMD, Capacitive, Panel; 2. HMD, Optical, Panel; 3. HMD, Optical, No Panel; 4. No HMD, Capacitive, Panel. The median is indicated by the line in the box, the boxes contain the inter-quartile range (IQR), and the whiskers extend to 1.5*IQR with outliers plotted.

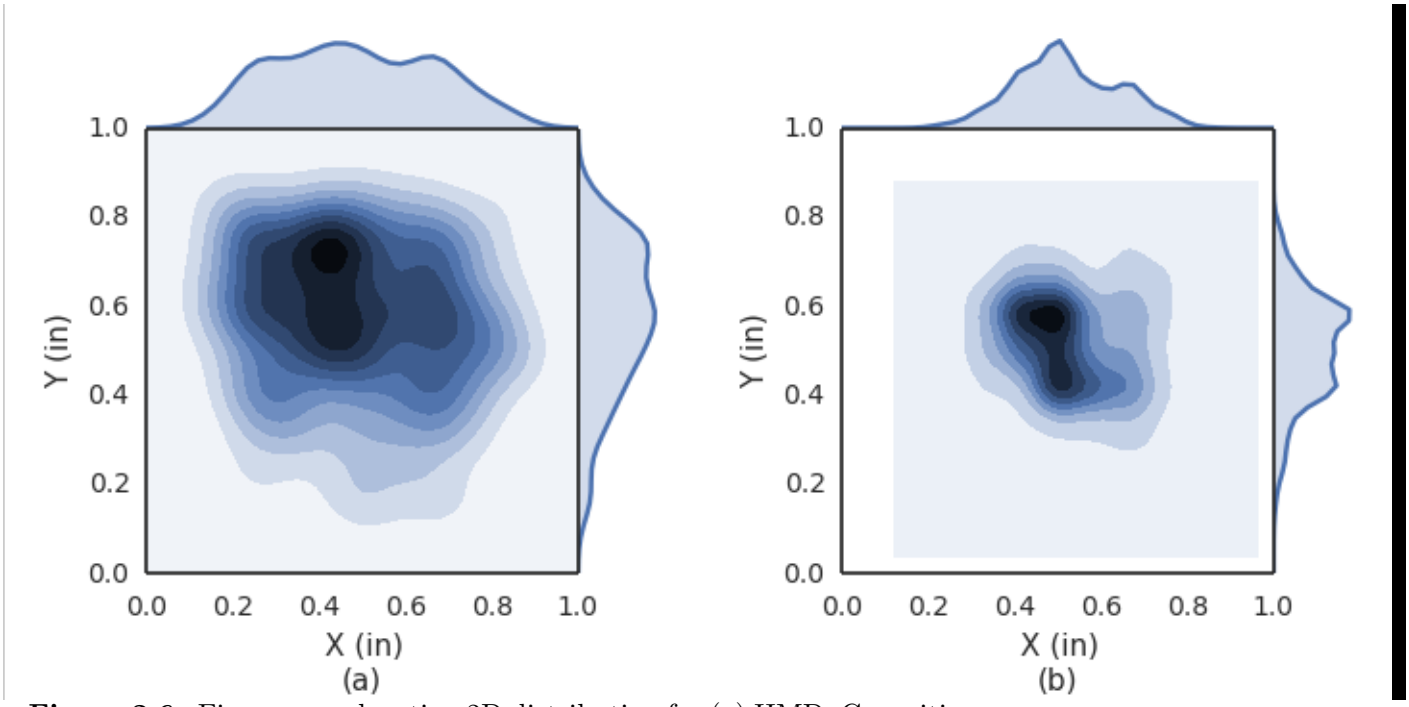


Figure 2.6: Finger press location 2D distribution for (a) HMD, Capacitive, Panel and (b) No HMD, Capacitive, Panel. The distribution is calculated as a Gaussian kernel density estimate. The mean (x, y) location and standard deviation for each is (a) $(0.48 \pm 0.18, 0.56 \pm 0.17)$ and (b) $(0.52 \pm 0.12, 0.51 \pm 0.12)$.

Chapter 3

Passive Haptics Experiment

3.1 Introduction

Passive haptics is a term that has been used to describe a variety of technologies or techniques to provide the sense of touch to a user of a virtual environment. It is often defined by its distinction from active haptics, which simulate the sense of touch with energy exchange, typically electromechanical. A common active haptic technology used in immersive virtual environments is a haptic glove, often utilizing small motors at the fingertips. In contrast, passive haptics often utilize proxy objects placed in the physical world to co-incide with the virtual environment experience. The proxy objects can be simple or complex. They can be colocated and accurate with the virtual world or purposefully designed to trick the user. In our paper, we utilize a simple colocated passive haptic device and measure its effect on the presence and performance of subjects using a 2D panel

in a 3D immersive virtual environment.

The advantages to using a simple passive haptic can be easily understood: less cost and complexity compared to most active haptic solutions. However, the disadvantage comes with its inflexibility. Due to their nature, passive haptics often have to be purpose built for a single or limited experience. While past research has aimed to address this, by either actively positioning a proxy object or simplifying the proxy object to fool the user, our application does not suffer from this limitation. The motivation for our research comes from the application of designing aerospace cockpits, complex human-machine interfaces where the user is stationed at their workspace. For the purpose of evaluating a cockpit design, the user does not need a dynamic tactile environment. Furthermore, many cockpit design processes already create a physical mockup which can provide the passive haptics for this evaluation.

We present our findings in testing passive haptics versus no haptics in an immersive virtual reality environment. Using a head-mounted display and a hand tracker, the subjects performed the same Fitts' Law style task under these two haptic conditions. The passive haptics was a flat surface placed at an angle on a desk in front of their seating area. Their performance on the Fitts' task was recorded as well as their responses to a presence survey, a self reported arm fatigue score and a general questionnaire.

3.2 Background

3.2.1 Haptics

Passive haptics has been a topic of research since the early immersive virtual environments. Robotic passive haptics were used to ameloriate the inflexibility of a proxy object by utilizing a robotic arm to position the proxy object in the virtual environment where the user was reaching[28, 21]. Insko[16] found increased presence using passive haptics for a maze, and also found that subjects trained with the passive haptics performed better after they were removed than the group that never used them. Another track of work combined active haptics with passive haptics, using a haptic glove with a physical panel to create mixed haptics [3]. Again, performance was increased with the haptics, but minimal differences were found between using the mixed haptics and the passive haptics alone. Similar to our motivation and work, Schiefele et al.[24] replaced a cockpit panel with a flat panel in an immersive head-mounted virtual environment, and found that users could activate buttons and switches in less time with the panel present than without. While much of the research involving passive haptics indicates an increase in the presence of the user, some have questioned whether active haptics provides benefits. Pontonnier et al.[23] discovered that subjects had decreased presence ratings in a virtual assembly task when using a haptic glove, versus both a real environment and a virtual environment without haptics. We build on this previous work by inves-

tigating the effects of passive haptics with the lastest virtual enviroment technology, as well as performing a complete Fitts' Law characterization between no haptics and passive haptics.

3.2.2 Fitts' Law

Fitts' originally devised a relationship between movement time and the distance and size of targets for a human performing rapid aimed movements[9]. This has since become known as Fitts' Law, and later work has refined the index of difficulty (ID) as:

$$ID = \log_2 \left(\frac{D}{W} + 1 \right) \tag{3.1}$$

where D is the distance to the target from the starting location and W is the width of the target. This formula for index of difficulty is known as the Shannons' formulation[20].

Commonly, the index of difficulty is related to movement time (MT) through a linear regression. However, in this work we are concerned with the measurement known as throughput (TP). Throughput has been recommended as the dependent measures for comparisons between experimental conditions[25]. As the name suggests, it can be thought of as the rate of information the human can input with the particular experimental setup or input device. It is defined as the index of difficulty over the movement

time, and has the units of "bits per second."

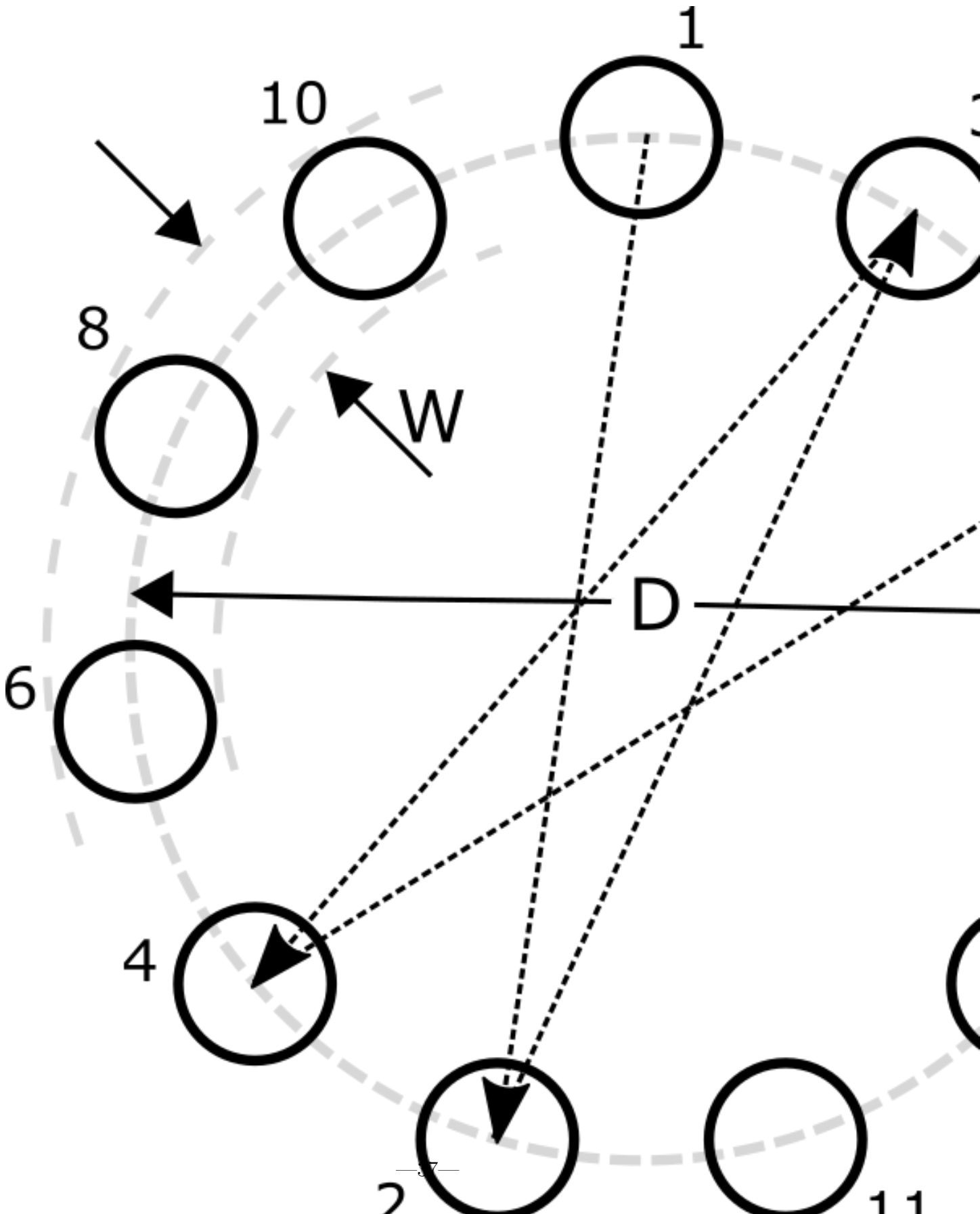
$$TP = \frac{ID}{MT} \tag{3.2}$$

The use of Fitts' Law as a tool for human-computer interface research began with the research of Card et al.[6] for the evaluation of different input devices for text entry. In 2000, The ISO 9241-9 standard was published with guidance on using Fitts' Law as an evaluation of pointing devices[17]. Along with the ISO standard, there have been calls to standardize the use of Fitts' Law so that results can be compared across literature[25]. For a 2-dimensional task (where all the buttons exist on a single plane), it is reccomended to use the circle layout as shown in Figure 3.1. This layout is referred to as the "Fitts' circle" within this article.

In this work, we use the effective width for the targets, which is defined as:

$$W_e = 4.133\sigma \tag{3.3}$$

where σ is the standard deviation of the end point positions. This is known as the adjustment for accuracy[30]. This correction accounts for the performance of the subject, especially on lower index of difficulty conditions where they may aim for the inside edge of a target. Hence, the use of the effective width provides the index of difficulty for the task that the subject performed, not the task presented to them. The effective width is calculated per subject per distance and width configuration, and subsequently


1
10
8
W
D
6
4
2
11

used in the index of difficulty equation.

$$ID_e = \log_2\left(\frac{D}{W_e} + 1\right) \tag{3.4}$$

Fitts' Law has been used in evaluating virtual environments and their input devices. Most of the work has been focused on 3D stereo displays[18] or . Chun et al. (2004) evaluated a set of 3D stereo displays with a single haptic-enabled stylus using a Fitts tapping task. They did not have well-fit regressions, but this could have been due to a small range of ID values used (2-3), targets being placed in 3D space (a 3D task with 3D movement) with random order, and averaging across subjects before completing the regression. One condition performed a mid-air 2D planar task with 3D movement, and no significant difference in throughput was found compared to the same task with constrained 2D motion. The trajectories were analyzed to determine where the extra movement time was spent, which is discussed in a later section (Trajectories in Virtual Environments).

The use of Fitts' Law with haptics has mostly focused on active haptics[8].

- fitts law with haptics
- collect references

3.2.3 Presence

The feeling of presence is often specified as a goal of a virtual environment. A definition from Witmer and Singer[31] reads:

> "*Presence* is defined as the subjective experience of being in one place or environment, even when one is physically situated in another."

Increased presence can lead to increased performance in a virtual environment task[33].

3.2.4 Arm Fatigue

Despite the concern of arm fatigue in virtual environments[5], it was surprising that most results in literature were anecdotal or for mitigations without quantification of the fatigue. Since fatigue is a subjective quantity, it can be hard to measure it between subjects, and sometimes even within. The negative impact of arm fatigue on using virtual environments makes it worth investigating. The arm fatigue scale used within this experiment is a Borg Rating of Perceived Exertion (RPE) scale that ranges from 6-20[2]. Hincapie-Ramos et al.[14] proposed a model for quantifying and predicting the amount of arm fatigue that correlated well with a Borg scale.

3.3 Methods

The purpose of the experiment described within this paper is to answer the following research questions:

1. Will the throughput be higher with passive haptics?
2. Do subjects learn the task quicker with passive haptics?
3. What are the differences between the formation of reaching motion trajectories with passive haptics?
4. Does the use of passive haptics lower arm fatigue?
5. Does the use of passive haptics cause greater presence?

3.3.1 Experimental Setup

For the experimental setup, subjects were seated at a desk with a blank panel mounted on an angle in front of them. The plywood panel (45cm x 45cm) was used to provide only the backstop of the virtual buttons for the "Passive Haptics" condition. The button selection is registered by the subject moving their index finger into a hover zone (cylinder for the circle buttons) in front of the button that extends outward 0.5in. Their entrance into the hover zone is indicated to them by the button changing color. A successful button press is registered after 150ms, and is indicated by the color turning off and a button click noise being played over speakers.

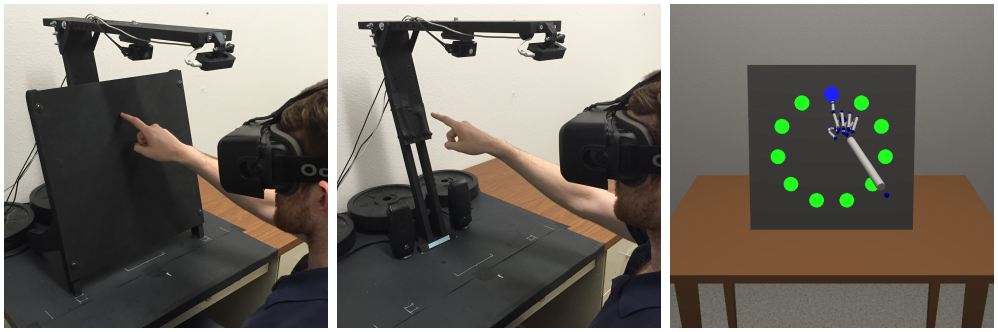
The equipment used consists of an Oculus Rift DK2 (Development Kit 2) head-mounted display (HMD) and a LeapMotion hand tracker. The low-persistence OLED display has a resolution of 1920x1080, with a referesh rate of 75Hz. The field of view is approximately 100°. It utilizes internal trackers and an external infrared camera for head tracking.

The LeapMotion is a markerless hand tracker which utilizes dual infrared cameras to provide a skelatal level position of hands and fingers in view. This position is used to provide an image of the hand position in the virtual environment, as well as for determining when a button is pressed. Instead of using the LeapMotion in its original face-up configuration, it was mounted above the working area and pointed down. Our pilot studies indicated that hand tracking from the LeapMotion was improved utilizing this face-down setup with the software using the head mounted configuration. However, the hand tracker could not be mounted on the head mounted display as it required a fixed position relative to the passive haptics to maintain appropriate registration between the virtual world and the passive haptics.

A custom calibration scheme was developed for the hand tracker as the initial registration between physical and virtual worlds was not very accurate. Despite the inaccuracy, the LeapMotion software was very precise, so after performing the calibration the registration was kept stable. The calibration performed a least squares claculation to solve for a transformation matrix between known real world locations and the reported location from the LeapMotion.

3.3.2 Experimental Task

The experimental task was a Fitts' circle in the virtual environment, performed by subjects in two haptic conditions. The subjects were seated



(a) Passive Haptics condition **(b)** No Haptics condition **(c)** View of virtual world

Figure 3.2: Experimental conditions and view of virtual environment.

at a desk for the experimental task, and the circle was located on a panel mounted on the desk. The two haptic conditions were "No Haptics (NH)" and "Passive Haptics (PH)." These conditions are pictured in Figure 3.2. For the "Passive Haptics" condition a physical panel was co-located with the panel in the virtual world, which was removed for the "No Haptics" condition.

There were no differences to the task itself or the method of button activation. The only difference between the conditions was the removal of the physical panel. The hand tracker remained in the same location, preserving the location of the buttons in the virtual environment. The dimensions of the virtual world were no different for either condition. In fact, there was no change to the software between the conditions.

Subjects performed the Fitts' circle for three different distances (20cm, 30cm, 40cm) and five different button widths (5mm, 10mm, 15mm, 20mm, and 25mm). These configurations were chosen to span a wide range of in-

dices of difficulty (3.2—6.4). For each configuration of distance and width, subjects had to complete the full pattern of 11 buttons three times consecutively. This set of 33 movements for a single configuration is referred to as a single trial for a subject. The distance was kept constant for the consecutive trial until all five button widths were complete. The distances were presented in either smallest to largest or vice versa, which was counterbalanced among subjects.

This set of 15 trials was repeated for each haptics condition and the order was kept the same within subjects. The sequence that the two conditions were presented to each subject was also counterbalanced.

3.3.3 Experimental Design

As described in the previous section, the experiment was performed with a within-subjects design. Subjects were asked to complete the same experimental task for both conditions of haptics: Passive Haptics (PH) and No Haptics (NH). The two different haptic conditions are the main independent variables.

It was expected we would find a large amount of skill transfer between the two conditions, so the order in which subjects performed the two conditions was counterbalanced. This created a second independent variable that is between subjects. The subjects who performed the conditions with the PH being their first condition were one group, and the subjects who performed NH as their first condition were a second group. We call this

grouping "sequence" and refer to the two groups as "PH First" and "NH First".

For a Fitts' Law evaluation, it is often recommended that the data collection only begins when the subject is fully trained on the task. However, one of the goals of the experiment is to investigate the learning rate of the subjects. For that reason, the subjects were given no seperate training time for the task or virtual environment.

In lieu of collecting the data from fully trained subjects, the throughput analysis will be carried out on movements that are determined to be composed of mostly ballistic motion. The filtering parameters were determined post-hoc from the trajectory recordings. Their development and parameter selection are discussed in the Results section.

3.3.4 Dependent Measures

The main dependent measure is the Fitts' throughput, measured through the movement time between button presses. Additionally, the trajectory of each movement is recorded from the hand tracker for analysis. To determine the arm fatigue, subjects were asked to rate their arm fatigue on the Borg scale from 6 to 20. The scale was presented with anchors as shown in Table A.10. The arm fatigue rating was collected at the beginning of each condition, and then after every other configuration of distance and width combination (and after the final trial, due to an odd number of trials). At the completion of each condition the subject was given a presence question-

naire. At the end of the experiment, an additional condition comparison survey was given to ask for opinions on the two haptic conditions.

3.3.5 Trajectory Phases

Human reaching movements have long been known to consist of two distinct phases[32]. To seperate the trajectories into the various phases, a simple algorithm was developed. First, the local minima are found throughout the velocity profile of the movement to seperate the movement into various submovements. The "ballistic phase" is then classified as the submovement which contains the peak velocity of the entire movement. The various submovements after the ballistic phase are classified as the "corrective phase". Any movement before the ballistic phase is classified as a "reaction time". An example of the results of this classification is shown in Figure 3.3. This mathematical definition does break down for certain cases where a subject might have two submovements in the ballistic phase due to a mid-course correction or similar, however one of the main purposes of this classification was to find movements which are appropriate to use for the Fitts' Law calculations.

3.3.6 Trajectory Filtering

We used a low-pass filter on the trajectory recordings to reduce the amount of noise. The LeapMotion processes data at a variable frequency, thus creating a variable rate for recording. The frequency typically varies

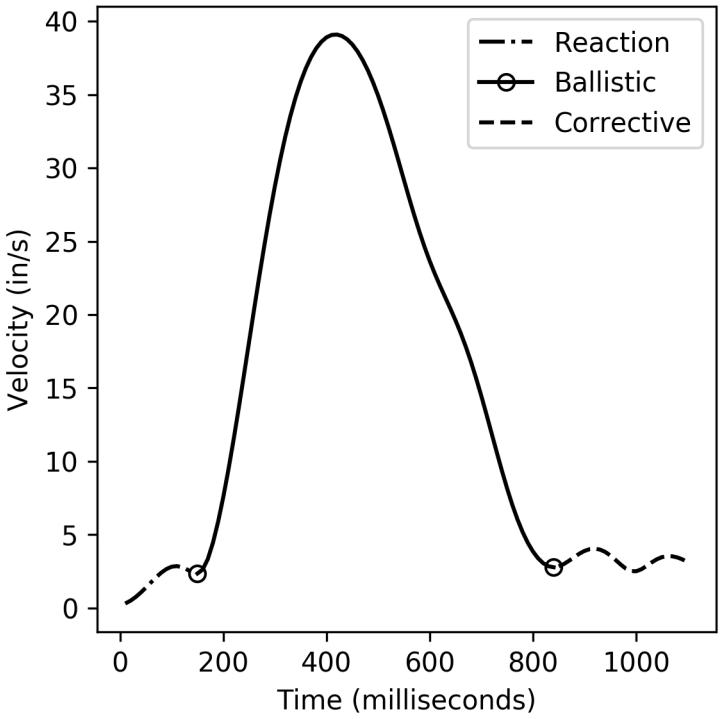


Figure 3.3: Example trajectory with three phases indicated.

from about 100Hz to 120Hz. To perform the filtering, the data was first resampled to a fixed rate of 100Hz. The filter used is a fourth-order Butterworth filter with a cut-off frequency of 5Hz. The cut-off frequency was chosen as voluntary hand movements have been shown to be below such a rate.

citation needed

3.3.7 Statistical Methods

The throughput, arm fatigue rating and presence score were all statistically tested using a two-way ANOVA with one within-subjects factor (Haptics) and one between subjects factor (Sequence). When the ANOVA

showed an interaction effect, the two Sequence groups were seperated and a repeated measures t-test was performed for the Haptics factor with each group. The statisitical significance level was corrected using the Bonferroni correction given the three dependent measures being tested. This leads to effects being considered statistically significant at the 0.0167 level ($\alpha = 0.05/3 = 0.167$). Effects between $0.05 < p < 0.0167$ are noted as marginally significant. Additionally, the presence questionnaire was tested using Cronbach's alpha for internal consistency.

3.4 Results

3.4.1 Participants

Twenty (20) subjects were recruited from the UC Davis engineering student population, both undergraduate and graduate students. The age range was 19—29 ($M = 22.95, \sigma = 3.0$) with 16 males and 4 females. The genders were balanced amongst the counterbalanced groups. All subjects indicated either less than one hour or no prior experience with virtual reality.

3.4.2 Throughput

The throughput is calculated per movement using Equation 3.2 and meaned per trial before being meaned per subject and condition. Throughput is used to investigate the first two research questions: do the subjects

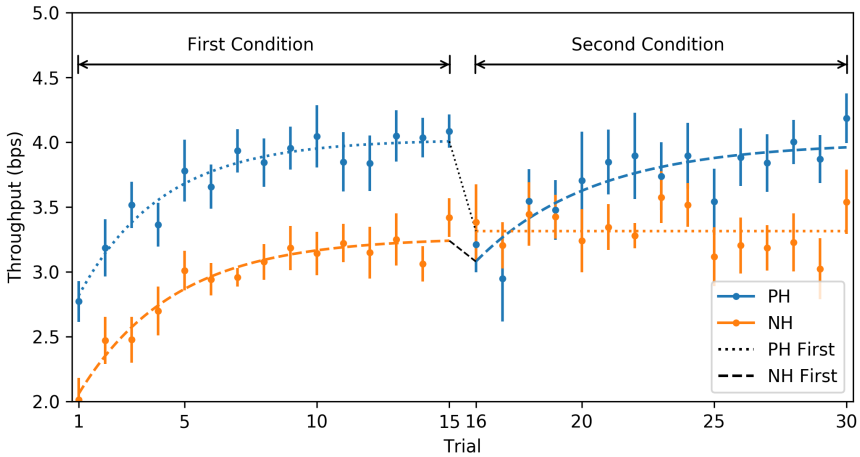


Figure 3.4: Throughput per trial. The learning curve exponential fit is given by Eqn. 3.5 with parameters from Table 3.1.

learn quicker and does their throughput performance improve with passive haptics.

Rate of Learning

The average throughput for each trial, seperated by haptics condition, is shown in Figure 3.4. An exponential rise to a learned state can be fit to each haptics condition to model the learning curve of the subjects. The equation used is given as:

$$\mathrm{TP}(T) = \mathrm{TP}_\infty - (\mathrm{TP}_\infty - \mathrm{TP}_0)e^{(-T/\tau)} \tag{3.5}$$

where T is the trial number, TP_∞ is the asymptotic learned value of throughput, TP_0 is the initial value at $T = 0$ and τ is the time constant. The time constant is the amount of trials for the throughput to rise 36% of

the difference between the fully learned throughput (TP_∞) and the initial throughput (TP_0). The parameters of the fit for each condition is shown in Table 3.1, as well as the standard error of the estimate (SEE). The value of throughput from the regression fit as a percentage of TP_∞ for certain trials are listed in Table 3.2.

The rate of learning is very similar amongst both groups in the first condition. For the first condition it can appear that the subjects performing NH learned quicker than their counterparts performing PH by looking at the time constants. However, as the values in Table 3.2 show, the NH First group started their first trial at a lower percentage of their learned state. The shape of the learning curves are very similar and both groups reached approximately 90% of their fully learned state by the 5^{th} trial.

The learning curves are quite different for the second condition. The NH condition does not have a learning curve, with a straight line being a better fit than the exponential function. This indicates the transfer of training from the PH condition allowed the group who did PH first to immediately perform in NH at the same level as the fully learned state of the subjects who learned NH in their first condition. This transfer of training to the second condition did not occur as strongly for the group who did NH first. Their initial performance of PH did start out at a slightly higher level than the subjects who did PH first (2.8 bps vs 2.4 bps), but after 5 trials this difference had converged (3.6 bps vs 3.7 bps).

These results indicate that subjects did not learn faster with Passive

Sequence	Condition	Haptics	TP_∞	TP_0	τ	SEE
PH First	First	PH	4.0	2.4	3.2	0.58
PH First	Second	NH	3.3	3.3	0.0	0.65
NH First	First	NH	3.3	1.7	3.6	0.49
NH First	Second	PH	4.0	2.8	4.5	0.78

Table 3.1: Exponential fit parameters of Eqn. 3.5. Curves are shown in Figure 3.4.

Sequence	Condition	Haptics	TP_1	TP_5	TP_{10}	TP_{15}
PH First	First	PH	70.0%	91.5%	98.3%	99.6%
PH First	Second	NH	100.0%	100.0%	100.0%	100.0%
NH First	First	NH	63.0%	87.8%	97.0%	99.2%
NH First	Second	PH	77.0%	90.6%	97.0%	99.0%

Table 3.2: Percentage of fully learned state for various trials for each group and condition. TP_i is $TP(i)/TP_\infty$ using Eqn. 3.5.

Haptics. The only differences between learning rates is the positive transfer of training from performing Passive Haptics first and No Haptics second. The answer to the research question *Do subjects learn the task quicker with passive haptics?* appears to be that the passive haptics does not make subjects learn faster, but they are able to learn the task quicker without passive haptics afterward. It does appear that the fully learned state is different between the haptic conditions, which is investigated further in the next sections.

Ballistic Movement Filtering

Before they could be used for the Fitts' Law analysis the trajectories were filtered so that only movements which were direct to target were included. A well-learned movement appropriate for Fitts' Law is one which moves directly towards the target and does not have much extranous movement or idle time beyond what is required to complete the task. A number of deviations from these well learned movements were observed in the dataset which we aimed to filter out for the final throughput calculations. We describe in this section the three metrics developed that were used to determine whether a movement was direct to target.

The first metric was the ratio of path distance travelled in the ballistic phase to the distance between the targets for that movement (i.e. 20cm, 30cm or 40cm). Ideally, the ballistic portion would cover the majority of the distance between the targets. A movement which covered too little of the target distance could mean the subject slowed or stopped in the middle of movement, and one that covered more could mean the subject overshot or had an indirect trajectory. A sample movement that gets flagged by this filter is shown in Figure 3.5, which has a ratio of 1.49. This is an example of a movement that does seemingly have a direct movement toward the target, but includes a large deviation perpindicular to the movement axis. This deviation could mean the subject initially aimed their ballistic portion in the wrong direction but performed a correction during the movement. The limits for this filter were chosen as having a ratio between 0.90 and

1.10, i.e. within 10% of the target distance. This filtered out 4577 of the 17970 movements.

The second filtering metric was also based on the ballistic phase path distance. For this metric it was compared to the total path distance the subject travelled for their entire movement. This filter targetted movements where after the ballistic phase the subject moved away from the target, or had a smaller but significant movement before the main ballistic movement. An example is shown in Figure 3.6 which shows a movement which passed the first metric (the ballistic phase distance ratio to the target distance was 1.03), but the ballistic phase was only 56% of the total distance travelled during that movement. This was a common problem where a subject would have a false start and move towards the next target before their button press was activated. The threshold was set at 0.80 which has 4264 movements lower than the threshold, though only 1675 were unique from the target distance filter.

The last filter took a time based approach, and looked a the ratio of the time spent in the ballistic phase over the total movement time. This filter removed movements where the subject spent an inordinate amount of time either before or after the ballistic phase. If they were not moving during the non-ballistic phases, it would not have been caught by the distance-based filters either. The threshold of 0.40 meant that 4496 movements were filtered, however only 873 of those are unique of the other two filters. Figure 3.7 illustrates a movement where the subject waited before initiating the

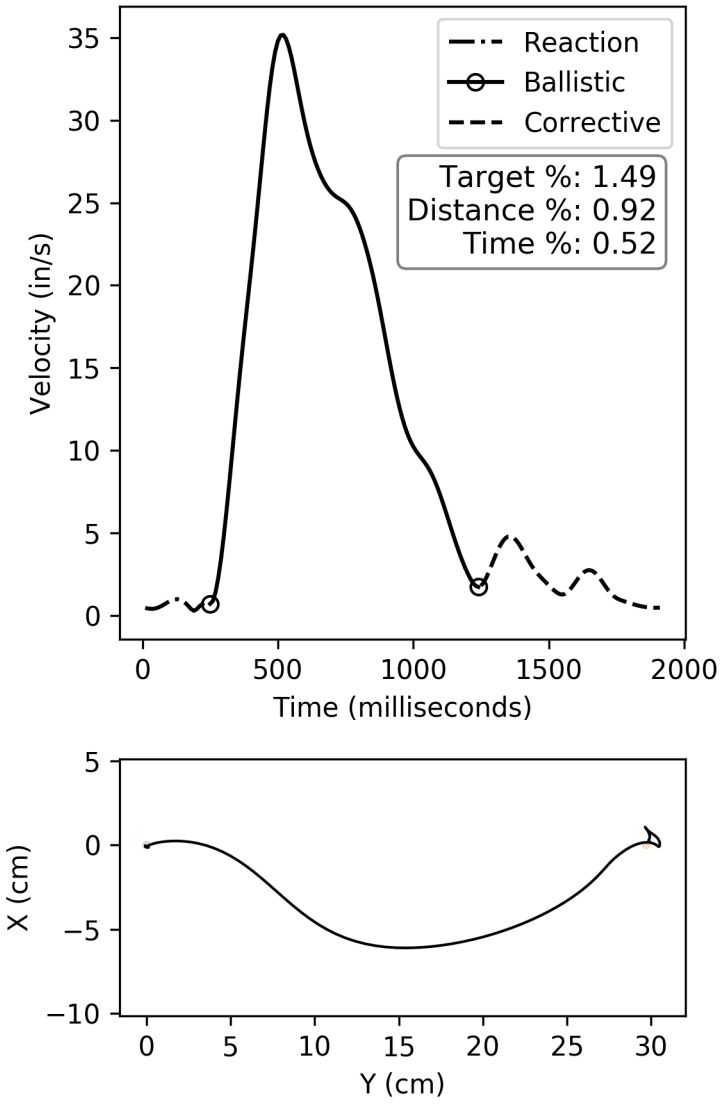


Figure 3.5: An example of a movement with a large ballistic distance to target distance ratio. The bottom plot shows the projection of the trajectory on the plane of the Fitts' circle.

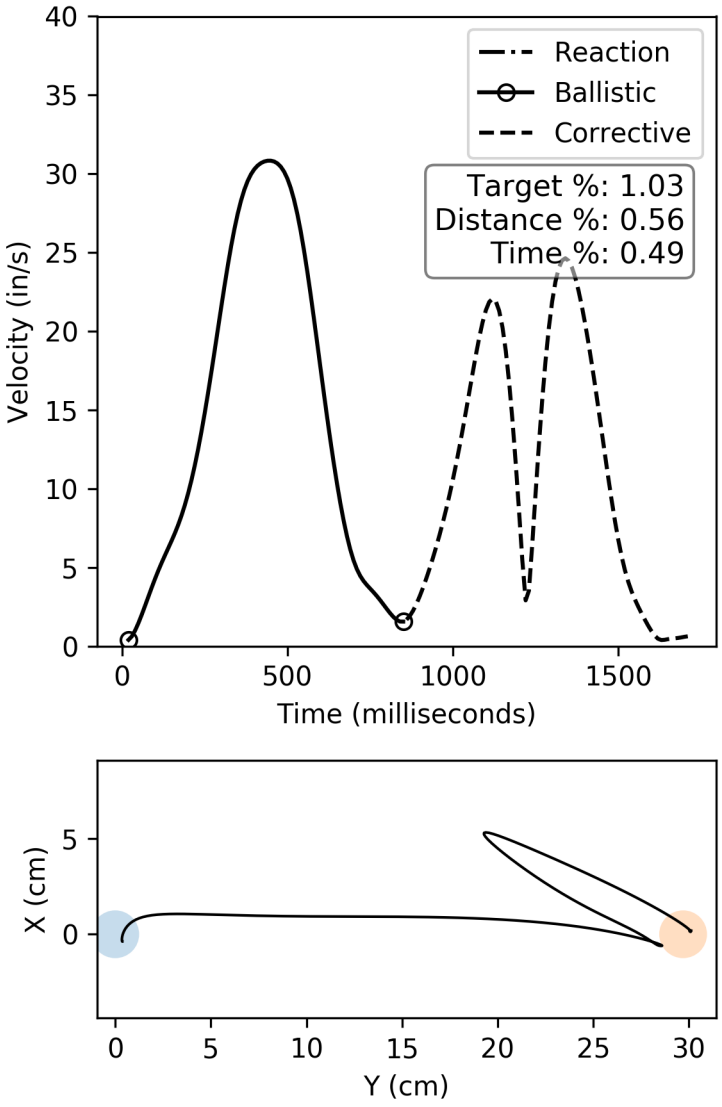


Figure 3.6: An example of a movement with a small ballistic distance to total path distance ratio. The bottom plot shows the projection of the trajectory on the plane of the Fitts' circle.

ballistic movement. Since the subject did not move during this idle time, the other two filters did not flag this movement.

The combination of these three filters led to 7125 of 18000 movements being filtered out, leaving 60% of the movements. For each of the metrics, the threshold was determined by investigating the distribution and looking at sample movements on either end of the threshold to determine if it was an approriate value.

The final check before performing the Fitts' calculation was to ensure that each trial had enough data for the adjustment for accuracy calculation. The adjustment for accuracy is based on the distribution of endpoint data from a single trial (which is one distance and width configuration). One trial consists of 30 consectutive movements, but the ballistic filtering could diminish the amount remaining in each, so a trial was only included if at least half (15 of 30) movements were considered to be good movements by the ballistic filters. This means that movements from a trial that did not have enough good movements were also filtered out from the Fitts' calculation. Not only is this important to make sure the adjusted width is valid, it also removes trials where the subject likely did not reach a fully learned state, as most of their movements were not primarily ballistic movements direct to target. On average, 11 of 15 trials per condition from each subject ($M = 10.98, \sigma = 3.25$) had enough good movements to be included. This left 9218 movements for the Fitts' calculation, just over half of the total movements (51.3%). Slightly more movements were

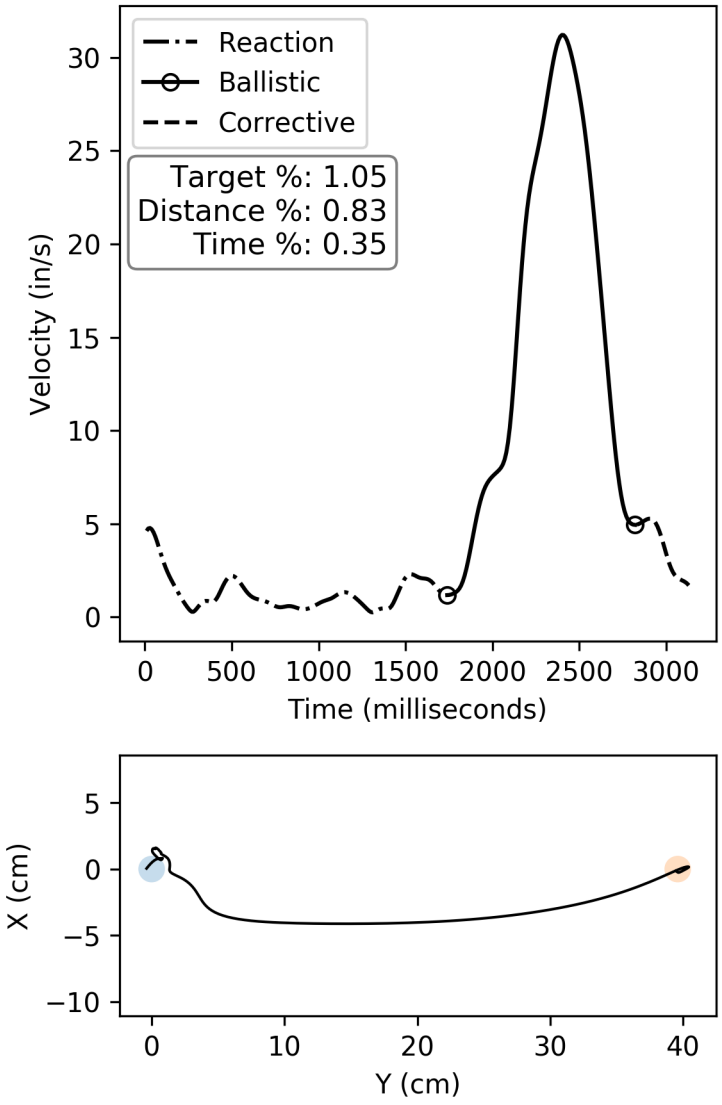


Figure 3.7: An example of a movement with a small ballistic time to total time ratio. The bottom plot shows the projection of the trajectory on the plane of the Fitts' circle.

filtered from the No Haptics condition, with 47.1% of NH movements left after the filtering, compared to 55.5% of the PH movements.

Throughput

The throughput was found to be higher in the PH condition, at 4.25 bps compared to the 3.76 bps of the NH condition. A two-way mixed ANOVA was performed to determine the effect of haptics condition. Since we already expected order effects and have seen them with the transfer of training seen in the Rate of Learning section, the sequence the subjects performed the haptic conditions was a between subjects factor. The effect of haptics was found to have a significant effect on the throughput ($F(1, 18) = 35.59, p < 0.001$) between the PH condition ($M = 4.25, \sigma = 0.44$) and the NH condition ($M = 3.76, \sigma = 0.38$). There was no effect on throughput based solely on sequence group ($F(1, 18) = 0.53, p = 0.47$), but there was a marginally significant interaction effect between the sequence and haptics ($F = (1, 18) = 4.48, p = 0.048$).

As can be seen in Figure 3.8, this marginal interaction effect appears to indicate that both groups had improved performance but the PH First group performed better at the NH condition. The post-hoc repeated measures ttest between haptic conditions for the subjects who performed PH First was significant ($t(9) = 4.62, p < 0.001$), with the PH condition ($M = 4.23, \sigma = 0.34$) outperforming the NH condition ($M = 3.91, \sigma = 0.40$). The mean of the differences between subjects was 0.32 bps. The

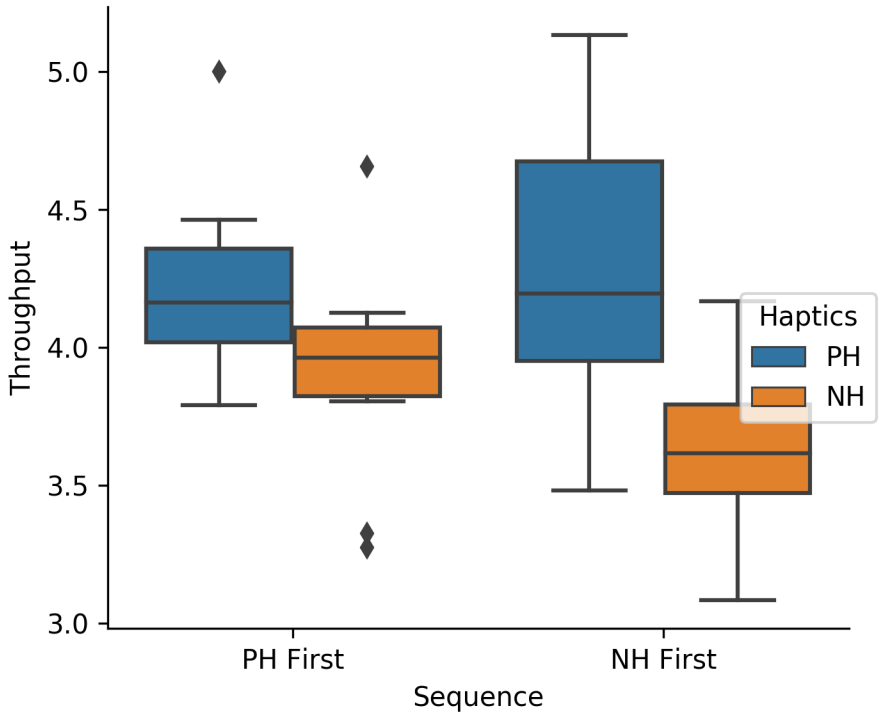


Figure 3.8: Throughput boxplot by haptics and sequence.

group of subjects who performed NH first also had a significant effect ($t(9) = 3.96, p < 0.001$) the PH condition ($M = 4.28, \sigma = 0.54$) and the NH condition ($M = 3.62, \sigma = 0.32$), with a higher mean of differences of 0.66 bps. These post-hoc tests confirm that both groups had a significant effect of haptics, though the NH First group had a larger difference between the conditions.

It is worth noting that without using the ballistic filter, the major conclusions found do not change. The only major difference is the magnitude of the throughput and size of the differences. The statistical tests have the same results as well. The results by haptics for both filtered and unfiltered

	Filtered		Unfiltered	
Haptics	Mean	SD	Mean	SD
PH	4.25	0.44	3.86	0.50
NH	3.76	0.38	3.25	0.37

Table 3.3: Throughput scores by haptics condition.

are shown in Table 3.3. These results indicate that subjects do have higher throughput with passive haptics, answering our first research question.

3.4.3 Trajectory Phases

As described in Section 3.3.5, each movement of the subjects can be dissected into three distinct phases: reaction time, ballistic phase, and corrective phase. We have already seen that, overall, the subjects took more time to complete a movement without the passive haptics in place with the results of throughput. In this section we investigate the differences in time spent in the three phases. We report here the means of time spent in each of the three phases. Times reported are all milliseconds. The results also include the filtered and unfiltered results, where the filtered results only include movements that were deemed purely ballistic by the filtering methods in Section 3.4.2. The unfiltered results include all movements. The phases were meaned per subject first, and then by condition. The time spent in each phase is listed in Table 3.4. Each phase was tested for the effect of haptics and sequence through a mixed two-way ANOVA.

The reaction time had no effect by haptics ($F(1, 18) = 0.32, p = 0.58$)

or sequence ($F(1, 18) = 0.001, p = 0.98$). However, for the interaction between the two a significant effect was found ($F(1, 18) = 18.56, p < 0.001$). The interpretation of this interaction effect without main effect significance is that the first and second condition had a different reaction time, without dependence on the haptics condition or group. The mean reaction time in the first condition was 130.9 milliseconds ($\sigma = 41.0$), but in the second condition it was just over 20 milliseconds faster, with an average of 109.1 milliseconds ($\sigma = 36.6$). The reaction time for both conditions was lower than generally accepted values for reaction time to a visual or aural stimulus. This is not surprising, as the task was a serial task which the subjects would likely learn the pacing of throughout the experiment. They would learn to anticipate the activation of a button (which was also the start of the next movement) as it would activate 160 msec after the subject entered the zone of the previous button. In fact, this interaction effect tells us that subjects did learn how to anticipate the activation event independent of the haptics or the order they performed the seqeunces.

citation needed

The ballistic phase time had a significant effect of haptics ($F(1, 18) = 24.14, p < 0.001$) between PH ($M = 772.6, \sigma = 67.7$) and NH ($M = 719.5, \sigma = 60.0$). There was no effect of sequence or the interaction effect between haptics and sequence. Since the ballistic phase should be mostly independent of the use of passive haptics, it was not expected to see the ballistic phase have an effect of haptics. It is unclear the exact mechanism that led to this, but it could likely be an artifact of the passive haptics

		Filtered			Unfiltered		
		Mean	SD	p	Mean	SD	p
Reaction Time	PH	118.58	45.32	0.57	241.57	205.39	0.33
	NH	121.44	34.80		291.97	169.19	
Ballistic Phase	PH	719.54	59.99	< 0.001	679.48	44.79	< 0.001
	NH	772.57	67.62		727.53	60.01	
Corrective Phase	PH	199.13	48.10	< 0.001	564.73	253.56	< 0.001
	NH	256.58	42.08		837.22	293.15	

Table 3.4: Time in each movement phase by haptics conditions.

causing the subjects to learn the movement, and thus allowing them to move quicker. There is little difference between the filtered movements and unfiltered movements results, the difference of the means were within a few milliseconds.

The corrective phase time also had a significant effect of haptics ($F(1, 18) = 22.46, p < 0.001$) between NH ($M = 256.6, \sigma = 42.1$) and PH ($M = 199.1, \sigma = 48.1$). There was no effect of sequence or the interaction effect between haptics and sequence. This result was expected as one of the main benefits of the passive haptics is that the subject does not have to 'find' the target along one dimension. There was a more noticeable difference between the results of the unfiltered and filtered movements for the corrective phase. The corrective phase time was much higher in both conditions, with PH having a mean of 564.7 ($\sigma = 253.6$) and NH having a mean of 837.2 ($\sigma = 293.2$).

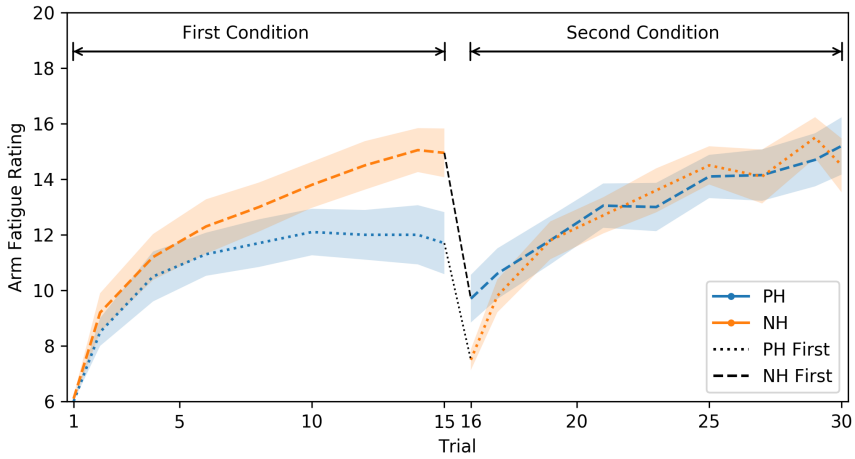


Figure 3.9: Arm Fatigue by Trial.

3.4.4 Arm Fatigue

The subjects were asked for a rating of their arm fatigue every other trial, as well as before the first trial and after the last trial of each condition. One trial lasted for 30 movements and consisted of a single distance and width configuration. The scale ranged from 6 to 20, and subjects were allowed to record decimal ratings. The full scale with anchors are shown in Appendix **??**. The average rating at each trial, seperated by haptics condition, is shown in Figure 3.9.

where is this

There is an evident difference between the two haptic conditions for the first condition performed, with the NH condition subjects accumulating more fatigue throughout the trials. At the end of the first condition, the subjects who performed the PH condition rated their arm fatigue 3.25 points lower on average than the NH condition subjects. Both groups have a similar rate of recovery, but the second condition quickly converges and

shows no apparent difference between the two haptics conditions.

A within-subjects repeated measure (haptics) with two between-subjects measures (sequence and trial) ANOVA was performed to test the significance of haptics and the interaction effect of haptics and sequence. The interaction effect of haptics of sequence was found to be significant ($F(1, 18) = 22.6, p < 0.001$) as well as the main effect of haptics ($F(1, 18) = 5.47, p = 0.03$). As a result of the significant interaction effect, a post-hoc ANOVA with Haptics and trial as the two within subjects repeated measures was run on both sequence groups. The NH First group had no effect due to haptics ($F(1, 9) = 2.08, p = 0.18$), consistent with the observations from Figure 3.9. The subjects rated the same trial between conditions an average of only 0.69 ($\sigma = 2.1$) points higher for the PH condition, which was their second condition. The PH First group did have a significant effect due to haptics ($F(1, 9) = 42.37, p < 0.001$). This group rated the PH condition an average of 2.0 ($\sigma = 2.0$) points lower within trials. As expected, the effect of trial number was significant for all of these tests ($p < 0.0001$).

These results show that the subjects only had reduced arm fatigue using the passive haptics for the first condition. For all other conditions the arm fatigue ratings reached the same level by the end of the condition.

Haptics	Score	Std. Dev	Cronbach's alpha
PH	77.7	9.56	0.718
NH	71.0	9.70	0.711

Table 3.5: Presence Score Summary

3.4.5 Presence

The presence survey was adminsistered after each haptics condition. The questions had a 7-point Likert scale response with anchors at either end and the middle. The score given in this section is a sum of the responses, on a scale of 1 to 7, where a higher score indicates higher presence. A few questions were asked with an inverted scale (i.e. a score of 1 indicated higher presence) and were reversed before the score was calculated. The internal consistency of the presence questionnaire was tested per condition using Cronbachs' alpha, and was found to be consistent in both conditions ($\alpha = 0.72$ and $\alpha = 0.71$ for PH and NH, respectively). The full survey questions and average responses per condition are listed in Table 3.6. The item total correlation (the correlation between the questions' score and the total score) is also listed for each question.

The average scores per condition are given in Table 3.5. The presence scores were tested with a mixed within subjects repeated measures (haptics) and between subjects measures (sequence) ANOVA. The score had a marginally significant effect between Haptics conditions ($F(1, 18) = 6.08, p = 0.024$), with Passive Haptics having a slightly higher mean ($M = 77.7, \sigma = 9.56$) than No Haptics ($M = 71.0, \sigma = 9.70$). There was no signif-

icant effect of Sequence ($F(1, 18) = 4.01, p = 0.58$) nor for the interaction effect between Sequence and Haptics ($F(1, 18) = 0.71, p = 0.41$).

Discuss results of items.

3.4.6 Condition Comparison

The condition comparison survey asked the subjects five questions directly comparing the two conditions. A summary of their answers are shown in Table 3.7. The subjects overwhelmingly responded that they preferred the Passive Haptics (PH) condition (Q5), with all but two subjects choosing it. In fact, no subject preferred the No Haptics (NH) condition, the two subjects who did not choose PH responded that neither was more preferred.

The other questions had responses similar to the results from the other sections. The majority of subjects responded that they were more accurate and faster in the PH condition (Q1 and Q2), which agrees with the throughput results. The subjects who chose Neither or NH were usually not actually faster or more accurate in the NH condition. In fact, only two subjects had a throughput that was higher in the NH condition, and neither subject chose NH as the condition they performed faster in, though one did say they performed more accurately in the NH condition.

Question 4 asked subjects directly about their feeling of presence, and 13 subjects chose PH, with the remaining split between 4 saying neither and 3 saying NH. The results of the presence questionnaire suggested that

Question	NH Score	NH ITCorr	PH Score	PH ITCorr
1. How much were you able to control events?	4.8	0.59	5.3	0.74*
2. How natural did your interactions with the environment seem?	4.3	0.45	5.0	0.81*
3. How much did the visual aspects of the environment engage you?	5.4	0.36	5.3	0.65
4. How much did the auditory aspects of the environment engage you?	6.2	0.32	6.0	0.19
5. How much did the tactile (sense of touch) aspects of the environment engage you?	2.6	0.31	5.6	0.53
6. To what extent did you associate the computer generated arm and hand with being "your body" while in the virtual environment?	4.5	0.35	5.0	0.65
7. How natural was the mechanism which controlled movement through the environment?	4.0	0.48	4.5	0.72*
8. How much did your experiences in the virtual environment seem consistent with your real-world experiences?	4.0	0.49	4.5	0.66
9. How involved were you in the virtual environment experience and the task you were performing?	5.8	0.50	5.8	0.22
10. How distracting was the control mechanism?†	4.0	0.52	4.7	0.29
11. How much delay did you experience between your actions and expected outcomes?†	4.7	0.38	4.9	0.04
12. How quickly did you adjust to the virtual environment experience?†	2.9	-0.08	2.1	-0.31
13. How proficient in moving and interacting with the virtual environment did you feel at the end of the experience?	4.9	0.65	5.2	0.25
14. How much did the control devices interfere with the performance of assigned tasks?	4.0	-0.40	4.0	-0.45
15. How well could you concentrate on the assigned tasks rather than on the mechanisms used to perform those tasks?	4.7	0.42	5.5	0.17
16. Were you involved in the experimental task to the extent that you lost track of time?	4.3	0.22	4.2	0.33

Table 3.6: Presence questions.

Question	NH	Neither	PH
Q1. In which condition did you feel you performed faster?	1	1	18
Q2. In which condition did you feel you performed more accurately?	3	2	15
Q3. Did you feel that your arm fatigued more or quicker in one condition over the other?	13	3	4
Q4. Did you feel that you were actually in the virtual room more in one condition over the other?	3	4	13
Q5. Which condition did you prefer?	0	2	18

Table 3.7: Condition comparison survey summary of results.

subjects felt more present with the passive haptics, which this agrees with. The arm fatigue question (Q3) was worded to ask which condition they felt provided more or quicker arm fatigue, and 13 subjects felt this was the case with the NH condition. The remaining were split between neither (3) and PH (4). The results of the arm fatigue questionnaire during the experiment found a similar result as the condition comparison survey.

3.5 Discussion

- time of each button press
- trajectory information for filtering

- arm fatigue rating every other circle
- presence questionnaire after each condition
- condition comparison questionnaire at end of experiment
- naive button registration
- arm fatigue not validated

3.6 Conclusion

Chapter 4

Design Evaluation Experiment

4.1 Introduction

The final experiment combines the lessons of the previous experiment to investigate the use of the Rapidly Reconfigurable Research Cockpit (R3C) in a design evaluation study. The goal of this experiment is to determine if the R3C system can be used in the place of a more traditional evaluation tool. As previous chapters have discussed, there are a number of self-evident advantages to using the R3C system. However, there remain some technical limitations to the technology that could hinder adoption. We found that a button targetting task took more time in our virtual environment than in the real world (Chapter 2). The following experiment (Chapter 3) found that a Fitts' Law task produced a higher throughput using a passive haptics layer, mitigating some of the time increase of targeting buttons in a cirtual environment. In the experiment described in

this chapter, we used the R3C system as the simulation tool for a design evaluation study of a cockpit instrument. The purpose in undergoing this evaluation study is to understand if these limitations would interfere with the metrics that might be used in evaluating a new cockpit design.

We designed an experiment which asks for feedback from subjects who take the role of design evaluators for a cockpit instrument. The subjects were divided into two groups: one group used an R3C setup to operate the instrument, while the other used a more traditional setup: a touch-screen simulator of the instrument. This seperation of groups will allow a comparison of the feedback from subjects between groups. Both groups evaluated the same two instrument designs, and subjects were asked to provide feedback using the same questionnaires. We hypothesize that the R3C system could be used in place of a traditional simulator if the two groups provide similar responses to the designs. Additionally, we utilized common quantitative metrics to evaluate performance to determine if the conclusions that would be drawn from these change between groups.

4.2 Methods

4.2.1 Simulator Setup

The simulator workstation as configured for each group is shown and annotated in Figure 4.1. It was designed to have as much as possible to be the same between the two configurations. The joystick and instrument were

this is not done yet

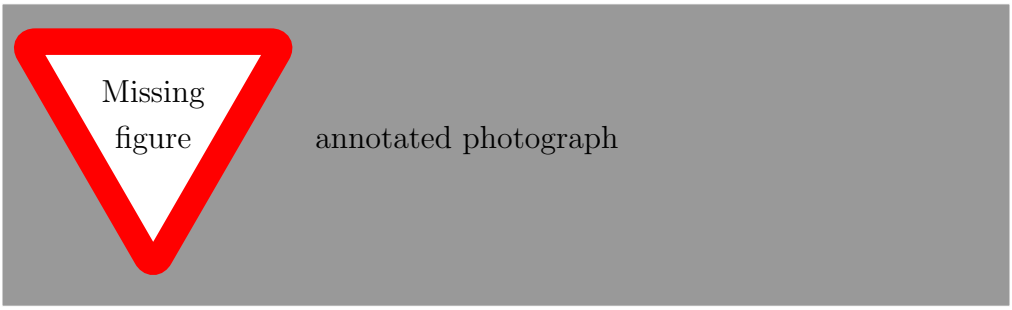


Figure 4.1: Simulator Workstation

positioned in the same location for each group. Neither group had out the window visuals, relying only on the attitude indicator on the instrument. For the Virtual Reality (VR) group, the visuals showed a plain interior of a cockpit, but the out-the-window view was black. Both groups had an aural indication (a click noise of a button being pressed) when a button was activated on the instrument, using the speakers mounted behind the instrument panel.

Beyond the VR group using a virtual reality headset for the visuals, the main difference between the two groups was the method for pressing the buttons on the instruments. The VR group used the hand tracker activated system previously described in Chapter chapter 1.

Include a short overview of the hand tracker system

For this experiment, the buttons were configured to highlight a blue color when the hand tracker registered a finger within the zone. The zones were extended 0.1in around the border of the button, raised a height of 0.5in above the surface of the button. When the button was activated after the 150 millisecond delay, the highlight would disappear and the button

in the virtual world would move inwards as if it were being pushed in[1], the press sound would play, and the behavior on the instrument associated with pressing that button would occur. A separate release sound would play when the finger left the zone after a successful press, and for the VR group the button would move back to its starting position.

The Touchscreen (TS) group used a 10.1 inch capacitive touch screen with resolution of 1024x600. The active area of the screen was 8.8in by 5.1in, with outside dimensions of 10.4in by 6.7in. The two instruments were rendered in a web browser using standard HTML elements. Javascript press and release events were used to simulate the same behavior as described for the VR group, except for the highlighting before a button press. The visuals of the tracker were rendered on top of the browser window with the same OpenGL rendering code used for the VR group.

4.2.2 Task Design

Based on the technology available for the simulator base, a number of requirements were laid out that would guide the design of an appropriate task and instrument designs. The instrument and task required:

- Flight task using a standard joystick
- Additional task that requires use of multiple buttons on the instrument

[1] Of course, the physical button could not and did not move.

- Able to develop simulator for both touchscreen and R3C setup
- Able to design two different layouts with one design having distinct flaws
- Simple design, yet complex enough task to have sufficient workload
- Operationally relevant tasks analogous to those required in a cockpit

Ultimately, we designed a task that required number and letter inputs using the buttons, while simultaneously flying a pitch disturbance profile.

Tracking Task

The tracking task display was a standard attitude indicator display, shown in Figure 4.2. Each tick corresponds to 1 degree in the dynamics simulation, with major ticks at intervals of 5 degrees. The attitude indicator was rendered to the size of 3.4 inches square on the instrument. Subjects controlled the one-dimensional (pitch only) task using a joystick with their left hand. The joystick is pictured in Figure 4.1.

The flight dynamics model of the simulator was a stability derivative based model for a Boeing 747 in low altitude flight. The block diagram of the dynamics is shown in Figure 4.3. The dynamics model was updated and recorded at a rate of 125Hz. The output of the joystick, r_{js}, varies from -1.0 to 1.0, and the gain of 10° was chosen to ensure the pilot had enough control authority to complete the task. The flight condition is listed as "Flight Condition 2" in NASA CR-2144[13]. This dynamics model was

Figure 4.2: Attitude Indicator Display

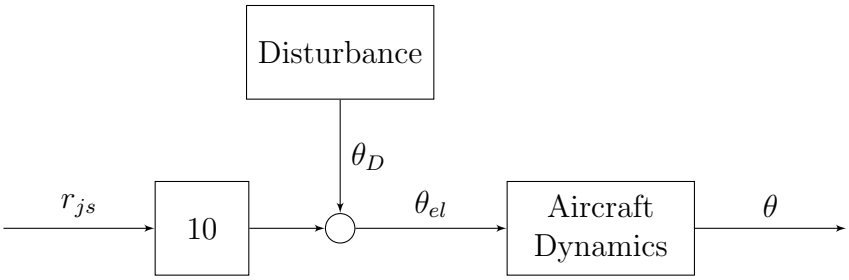


Figure 4.3: Tracking Task Dynamics Block Diagram

chosen due to availability. The specifics of the dynamical model were not important other than providing a response similar to an aircraft in flight. The model was linearized from sea-level flight at an airspeed of 335 ft/s. The configuration of the airplane had the gear up, no flaps, total weight of 654,000 lbs, and an angle of attack of 7.3°. The transfer function of the aircraft dynamics is given as:

$$\frac{\theta}{\theta_{el}} = \frac{-0.572(s+0.553)(s+0.0396)}{(s^2+2\zeta_1\omega_1+{\omega_1}^2)(s^2+2\zeta_2\omega_2+{\omega_2}^2)} \tag{4.1}$$

$$\omega_1 = 0.0578 \qquad \zeta_1 = 0.0160$$

$$\omega_2 = 1.12 \qquad \zeta_2 = 0.798$$

The disturbance model is based off the model developed in SweetRef[**?**]. reference It is designed to provide a broad spectrum of frequencies that the human controller needs to respond to. The disturbance is a sum of sines described by:

$$\theta_D = K\sum_{i=1}^{12}\left[a_i\left(\frac{2\pi k_i}{240}\right)\sin\left(\frac{2\pi k_i}{240}t+\phi_i\right)\right] \tag{4.2}$$

The k_i terms are given as,

$$k_1 = 7, \qquad k_2 = 11, \qquad k_3 = 16$$

$$k_4 = 25, \qquad k_5 = 38, \qquad k_6 = 61$$

$$k_7 = 103, \qquad k_8 = 131, \qquad k_9 = 151$$

$$k_{10} = 181, \qquad k_{11} = 313, \qquad k_{12} = 523$$

The amplitude terms are $a_i = 0.5$ for $i <= 6$ and $a_i = 0.05$ otherwise. The phase terms, ϕ_i, were randomly selected on the $(-\pi, \pi)$ interval en-

suring a uniform distribution. This random selection was pre-calculated for each trial, however the order was repeated for each subject so there was no between subject variance in the disturbance signal. Furthermore, each subject received the same sequence of disturbance signals for each instrument design. The disturbance amplitude, K, was chosen such that the root-mean square (RMS) of the signal was 3.5 degrees. The value of the was chosen through pilot studies to ensure the task was challenging but not overwhelming.

Prompting Task

The prompting task was designed to be both a realistic task for a cockpit as well as a demanding task when done in addition with the tracking task. The task developed required the subjects to read and memorize a short string of characters and enter it back using the buttons on the instrument. To limit the task physically (by number of buttons) and mentally, the characters used were the number 1 through 6 and the letters A through F. The prompts were 4 characters long and once the subject started entry the prompt would disappear, forcing them to hold it in short term memory.

The sequence of the prompts was separated into 10 second "windows". The prompt would appear randomly between 2 and 3 seconds of the start of the window. From the time of appearance, subjects were given seven seconds until timeout. When the subject pressed the first button of the prompt, the prompt itself was cleared and asterisk symbols ($*$) were shown

in place of the prompt for each button entry by the subject. If the subject ran out of time, the text in entry area would return to black. Although subjects were briefed on the timeout and given practice to learn the pace, no warning or indication of time left was shown during the trials. Whether they completed the prompt within the time limit, or they timed-out, this process was repeated every 10 seconds. This meant that subjects had at least 3 seconds of time with no prompt.

The prompts themselves were always composed of three numbers followed by a letter or three letters followed by a number. This structure was decided upon to provide a consistent pattern, yet still utilize both letters and numbers in every prompt. The prompts were randomly chosen but were not allowed to have repeat numbers or letters. The selection of letters or numbers as the first three characters was randomly chosen as well, with an equal weight to each.

4.2.3 Instrument Designs

The two different designs used were developed to be both realistic as a cockpit instrument design that would be under consideration, yet still have one design with flaws that would be found in a design evaluation. We developed a 'Keypad' design with the prompting task button keys on the right side and the tracking task on the left, and an 'Edgekey' design with the prompt buttons split on either side of the tracking task display. The tracking task display was the same size on the display for both designs.

The prompting task text was placed below the tracking task display, and the same font, size and color was used for both designs. The prompting task text font was approximately 0.62in tall. These were kept consistent to limit the number of possible variables between the two designs. The prominent difference is the placement and behavior of the buttons which is described in this section.

The Keypad design is pictured in Figure 4.4. The buttons are 1in by 0.75in, with about 0.26in between buttons horizontally and 0.38in vertically. Each button has the label directly on the top of the button. The 3D-printed instrument used for the VR group had the buttons raised a height of 0.31in from the surface of the instrument. The button labels were also raised to provide a tactile feedback. The font was approximately 0.36in tall, and the labels were embossed above the button surface 0.05in.

The Edgekey design is pictured in Figure 4.5. In this design, there is not a single button for every number and letter. Instead, the bottom button on either side would switch the behavior (and labels) of the remaining six buttons from being 1 through 6 to A through F. In other words, the bottom "switching" buttons would change the rest of the buttons from the numbers to the letters, and vice-versa. The labels were placed offset from the button on the "screen" portion of the instrument, allowing them to change dynamically. The fonts were approximately 0.32in tall, and were blue on the screen. The buttons are slightly smaller in this design, at 0.76in by 0.55in. A smaller button size was needed to fit the labels and the

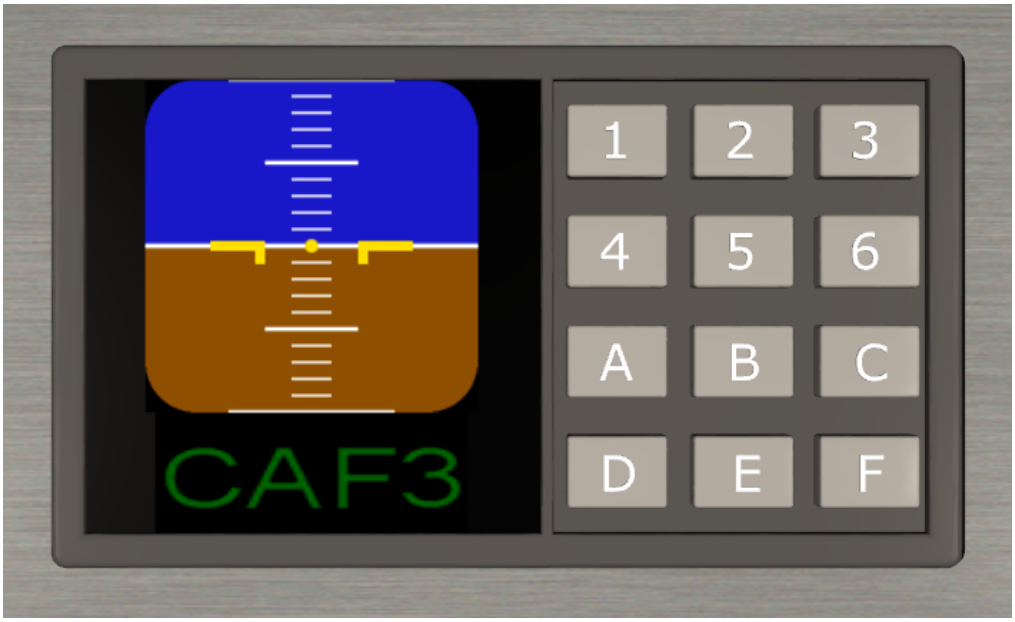


Figure 4.4: Keypad Design

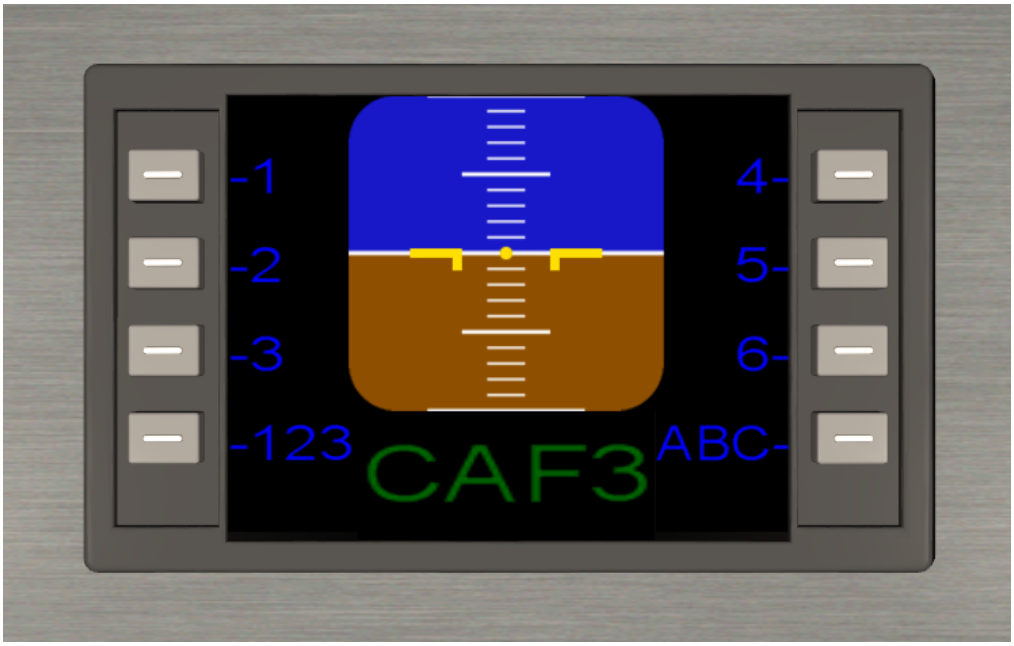


Figure 4.5: Edgekey Design

buttons side by side. The spacing between buttons vertically is the same as the Keypad design at 0.38in. The center to center distance between the two sides of the button rows is 7.3in. The 3D-printed instrument version had raised nubs on each button covering half the width, 0.08in tall and raised 0.05in. As with the Keypad design, the buttons had the same height of 0.31in from the surface.

While some of the more subtle differences were expected to be noted by the evaluation study (e.g. having smaller buttons, different position of the flight task), the major flaw designed into the Edgekey design was the switching key to change from letters to numbers and back. This additional action fundamentally changed the demands of the task, as the subjects now had to press this additional button to change labels at least once per prompt. The cognitive workflow of the subject is diagrammed in Figure 4.6. The additional mental effort of the Edgekey design is shown in the dashed box, where the subject has to verify the state of the instrument and possibly press the switch button before they press the buttons of the prompt. Since the prompts were kept as a consistent format of three of the same type and fourth of the other, this extra work was easily skipped for most buttons, and anticipated between the third and fourth button. There was no guarantee that the next prompt would start with the instrument in the correct state for the new prompt, so there was always an additional cognitive load in determining whether a switch was necessary at the beginning of the prompting window, which would be accompanied with the

physical effort if the switch was needed.

4.2.4 Experiment Design

Subjects were divided into the two groups, Touchscreen (TS) and Virtual Reality (VR). The overall sequence of the experiment started with a training session on the simulator and the task, followed by an evaluation session for each of the two designs, finishing with questionnaires asking subjects to evaluate the two designs. The timeline of the experiment was the same for each subject, except for counterbalancing the order that the designs were evaluated. The training portion started with a slide deck explaining the tasks, the simulator that the subject was using (depending on which group they were in), and the functionality of the two designs they were to evaluate. Next, they performed practice trials with just the tracking task and then just the prompting task. The practice trials of the tracking task were 60 seconds long and repeated until the subjects' performance had flatlined. This took between three to six trials for each subject.

For the evaluation sessions with each design, they performed six trials with both tasks. The first three were a minute long, and were considered practice trials, and not included in the data analysis, though this was not communicated to the subjects. The following three trials were two minutes each, and were used for the analysis. Each evaluation session concluded with a two minute trial of just the tracking task without the prompting

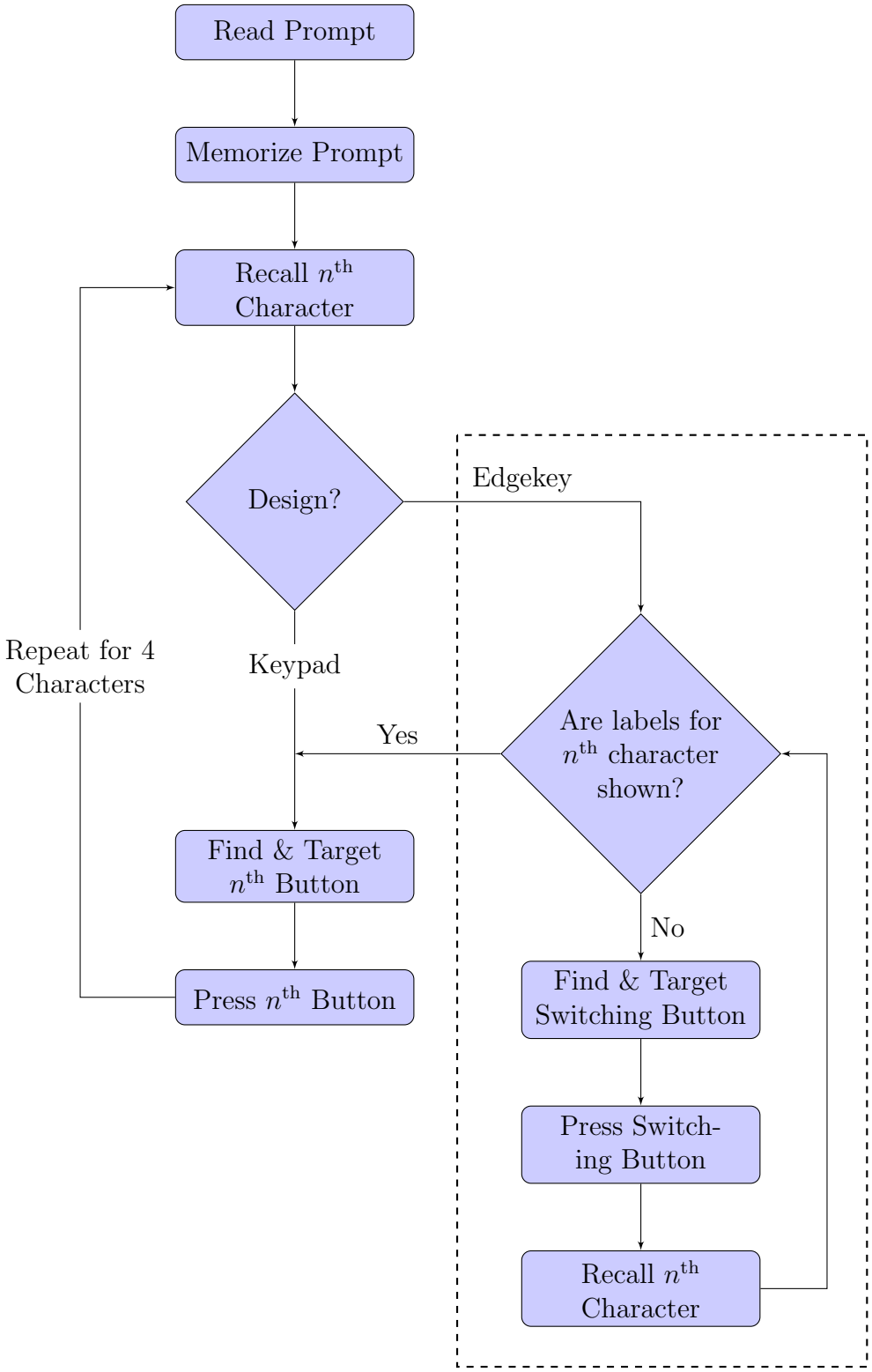


Figure 4.6: Prompting Task Flowchart of Cognitive Work for Each Design. Extra work of Edgekey design encolsed in dashed line box.

task. This was included to investigate if the subject had improved or fatigued at the tracking task throughout the experiment.

The independent variables of the experiment are Group and Design. The Group is the simulator the subject used, a between subjects factor, and either TS or VR. The Design is a within subjects factor, the two instrument designs that every subject evaluated — Edgekey and Keypad.

4.2.5 Dependent Measures

The dependent measures were chosen to evaluate the performance of each task individually as well as the workload of the subject. For the tracking task, the root-mean square error (RMSE) was calculated for each trial. The error in this case is simply the pitch shown to the subject, the output of the flight model described in Section 4.2.2.

The prompting task has two dependent measures, for speed and accuracy. For speed we consider the *response time*, defined as the time between when the prompt is first shown to the subject and when they press the first button of their response entry. The accuracy is measured by how many prompts they complete correctly. Twelve prompts are shown to the subject within each trial. The response time was meaned per trial first and then per design for each subject, and the number of correct prompts is meaned per design for each subject.

A NASA Task Load Index (TLX) survey was administered after they completed each design to measure the workload of the subject. The TLX

survey asks for a rating of their workload between 0-100 for the following subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Our implementation allowed selection of the ratings within increments of 5, and included anchors of "Low" and "High" at the extrema of 0 and 100, respectively (except for Performance, which uses "Good" and "Bad"). The midpoint was also visually indicated with a larger tick. The ranked pairs modification was used and completed for both times the subject took the survey. This modification asks the subject, for each of the pairwise combinations of subscales, which they felt contributed more to their workload. The number of times they select each subscale is used a weight to calculate a weighted mean for the total TLX score.

Finally, the subjects were given a questionnaire asking for their feedback on each instrument design. For each design, the subjects were asked the following questions:

- Please comment on any difficulties you had performing the prompting task with this design especially in contrast to the other design.
- Please comment on anything you liked in this design.
- Please comment on anything you did not like in this design.
- Any other comments?

Additionally, the following questions were asked:

- Which instrument design did you prefer? Why?
- Did you experience any physical fatigue during the experiment? Where?
- Any other comments?

An open form text box was used for the response field for each of these questions.

In a standard design evaluation study, the feedback received from the users in this questionnaire (and other debriefing interviews) would often be the main source for carrying out re-design. The purpose of this feedback in this experiment is to determine and document in which ways does this feedback differ. For example, if most subjects in one group noted issues with the size of a button, while no one in the other group found an issue with that button, this would indicate that using this VR system may not highlight the same issues regarding button sizes. The groups were purposely left ambiguous in the example, as it does not matter which group found the flaw and which group did not comment on it. Although we could postulate as to which group are "correct" in their evaluation of the instrument, it is not a useful exercise, as the only result is to document what potential differences could arise so that users of this system can be aware.

With that goal in mind, the analysis of the feedback questions seeks to find differences between the groups. The sentences from the open form responses were first separated into single feedback comments, and summarized using common language. If a single subject repeated the same

comment in the answers to multiple questions, they were only counted once. Each of these simplified feedback comments were assigned to a category or overall summary of their feedback. This process was completed separately for each group. We aim to look for feedback that is unique to a certain group or feedback that receives a higher frequency of comments in one group. This will provide a summary of where the groups provide the same feedback and where they provide differing feedback.

4.2.6 Hypotheses

The main hypothesis of this experiment is that the use of a VR/R3C simulator will not affect the conclusions of a design evaluation study, compared to a traditional touchscreen simulator. We do expect that some of the dependent measures may have a significant difference in Group or a significant difference in Design. The more important measure for us, however, is the interaction effect. This will test if the change between Designs is similar for the two Groups. If this is the case, then it may indicate that an evaluation study using one of these simulators could draw differing conclusions of an evaluation study using the other. Statistically, we will test the hyptothesis that there exists no interaction effect between Group and Design for any of our dependent measures.

Additionally, the two tracking only trials performed at the end of each evaluation session, as well as the final tracking only training trial, will be used to investigate if the subjects were still learning the tracking task. The

concern if subjects became more trained in the tracking task is that it could lower their attentual needs to that portion of the task, causing a change in performance on the prompting task that was not due to the design change. These hypotheses are enumerated here:

H1. The tracking task RMSE will have no interaction effect between Group and Design

H2. The prompt response time will have no interaction effect between Group and Design

H3. The number of correct prompts will have no interaction effect between Group and Design

H4. The NASA TLX scores will have no interaction effect between Group and Design

H5. The tracking task RMSE for the last training trial and the tracking only trials will not change throughout the experiment

4.2.7 Statistical Tests

The quantitative dependent measures are tested with a two-way ANOVA, with one within subjects factor (Design) and one between subjects factor (Group). The Design factor contains two levels, the two designs each subject tested, Edgekey and Keypad. The Group factor also contains two levels, the VR (Virtual Reality) group and the TS (Touchscreen) group.

When the ANOVA showed significance in the interaction test, post-hoc repeated measured t-tests were undertaken to determine the significance of Design within each Group. Independent samples t-tests were used to test the significance of Group within each Design. The last hypothesis testing the effects of learning on the trials with only the tracking task will be tested with a two-way ANOVA, with the Group as a between subjects factor, and the trial number as a within subjects factor. The trial number is chronological in the order the subjects performed them. The first trial was the last tracking only training trial, and the next two were tracking only trials at the end of each design evaluation.

Statistical significance level was corrected using the Bonferroni correction considering the 5 hypotheses being tested. All effects were considered statistically significant at the 0.01 level ($\alpha = 0.05/5 = 0.01$). Effects which have a significance level between $0.05 < p < 0.01$ are considered to be marginally significant.

4.3 Results

4.3.1 Demographics

Twenty-three subjects were recruited from the UC Davis engineering undergraduate and graduate student population. Twelve subjects were placed in the VR group, and the remaining eleven in the TS group. The mean age was 21.0 ($\sigma = 3.14$), with 19 male and 4 female subjects. The

genders were balanced between the two groups. Most subjects had no flight experience (two were student pilots), and all of the VR group subjects indicated that they had less than one hour of experience using virtual reality headsets. It should be noted that the subjects are not the beneficial population of the research. The task and experiment was designed with this in mind and mitigated through training and the simplicity of the task design.

4.3.2 Performance Measures

Tracking Task RMSE

The performance of the tracking task was measured using the root-mean square error (RMSE) of the pitch. The effect of Group yielded an F ratio of $F(1, 21) = 21.4, p < 0.001$ indicating a significant difference between VR ($M = 1.28\text{deg}, \sigma = 0.38\text{deg}$) and TS ($M = 1.97\text{deg}, \sigma = 0.38\text{deg}$). In both groups, subjects were performing the tracking task using the same joystick. The most direct factor that could contribute to the decreased performance in the tracking task for the VR group is the loss of visual acuity in the tracking task display due to the technical limitations of the VR head-mounted display. Indirectly, the additional workload of the prompting task could be taking attention away from the tracking task. The effect of Design indicated a marginally significant difference ($F(1, 21) = 5.94, p = 0.024$) for the tracking task RMSE between Keypad ($M = 1.57\text{deg}, \sigma = 0.51\text{deg}$) and Edgekey ($M = 1.70\text{deg}, \sigma = 0.52\text{deg}$). The only change in the tracking task

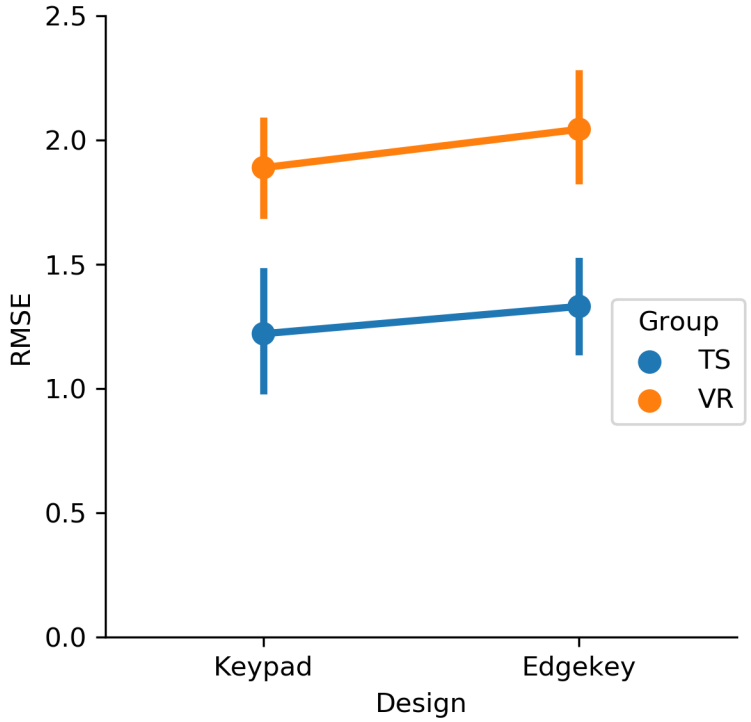


Figure 4.7: Factor Plot of RMSE

display between the two instrument designs is a small change in position. It moves from being on the left side for the Keypad to the middle for the Edgekey. Since there was no change otherwise, this suggests that any difference on the tracking task performance between the designs would be related to additional workload from the prompting task. The interaction effect was not significant ($F(1, 21) = 0.17, p = 0.69$).

We can investigate the trials where the subjects were only doing the tracking task to further investigate the change in performance between the two groups. The subjects ran a single trial that was just the tracking task without the prompting task at the end of each evaluation session. These

trials were included to be used as a test of the assumption that the subjects were no longer learning, but can also be used as a test of the Group factor on the tracking task performance. The effect of group on RMSE for the tracking-only trials yielded a marginally significant difference ($F(1, 21) = 4.81, p = 0.039$) between the VR Group ($M = 1.32, \sigma = 0.50$) and the TS Group ($M = 0.91, \sigma = 0.43$). There was no significant difference for the effect of design ($F(1, 21) = 0.068, p = 0.80$). The interaction effect between group and design was also not significant ($F(1, 21) = 3.21, p = 0.087$).

Although the tracking only trials found a marginally significant difference for the group, the difference was much more distinct for the trials with both tasks. This indicates that when the subjects were focused on the single task, they were able to mitigate most of the visual resolution differences between using a touchscreen and the virtual reality screen. Additionally, the marginally significant difference between the designs for the trials with both tasks was reduced to no significance when the additional prompting task was removed. This also points to the additional workload of the prompting task causing a performance drop on the tracking task. The factors leading to the added workload of the prompting task are investigated in the next performance measures discussed.

Prompt Response Time

The first measure of the prompting task is the response time of the subject. The response time is defined as the time from the prompt is

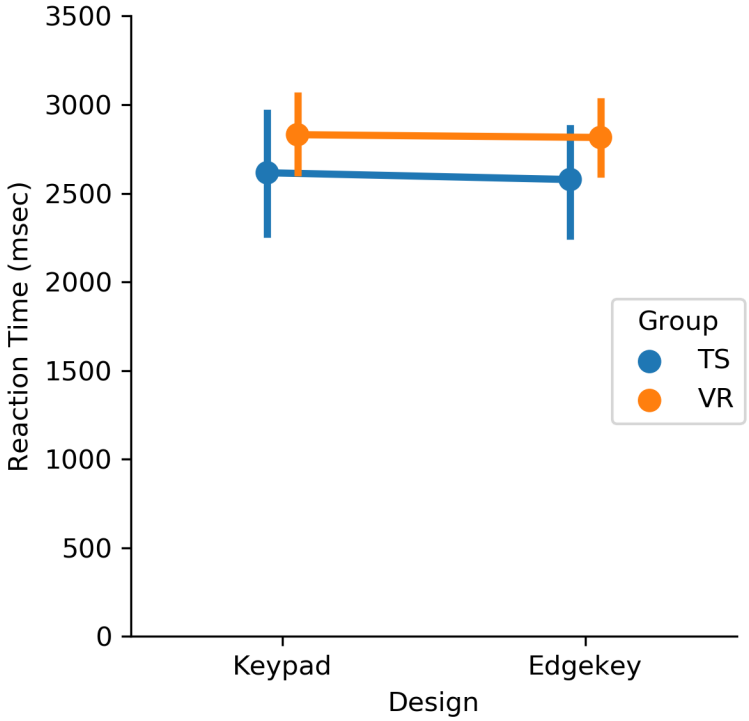


Figure 4.8: Factor Plot of Response Time

shown to each subject until they press the first button of the prompt. For the Edgekey design, it would be possible that the subject had to start with the switching button if the new prompt did not start with the same mode (letters or numbers) as the previous prompt (see Figure 4.6). Since this button would not clear the prompt when it was pressed, it is not considered the first button of their entry. However, this would still require an additional movement of the subject, adding additional time. For this reason, the prompts which required the subject to start with the switch key are filtered out of this analysis. After filtering, 885 of the total 1700 prompts recorded for the Edgekey design were kept.

The response time was unique among the dependent measures, as all tests were insignificant. The effect of group yielded an F ratio of $F(1, 21) = 1.19, p = 0.29$ indicating no significant difference between VR ($M = 2812$msec$, \sigma =$ 383msec) and TS ($M = 2594$msec$, \sigma = 567$msec). One factor that could influence the response time between groups is the additional time to activate a button in the VR environment versus the touchscreen. The touchscreen subjects were using a familiar interface for activating the buttons, while the VR subjects needed to activate the button with the virtual hand. However, a large portion of the response time for the subject is their cognitive processing of the prompt – recognizing the new prompt has appeared, reading it, then memorizing it. Beyond potential differences in the visual environment, the cognitive portion should not take more time for one group or the other. A potential reason that there could be a lower than expected difference between the group means is that some VR subjects learned to keep their hand closer to the instrument so that the hand tracker could keep it in view. When the hand tracker lost view of the hand, the re-acquisition time could be significant, so holding it close to the instrument would prevent this from happening. This issue comes up again when looking at the subjects' response to questions about fatigue.

The effect of design was also insignificant ($F(1, 21) = 0.68, p = 0.42$) between Keypad ($M = 2728$msec$, \sigma = 512$msec) and Edgekey ($M = 2687, \sigma = 471$msec). The biggest difference between the two designs is the switching key on the Edgekey design. As described above, the need

for an additional switch press before the first prompt button was filtered out, so we are only comparing prompts where the first button was available right away to the subject. Even though the physical requirements were filtered out, subjects still need to verify that the labels are in the correct state for starting entry. Since the Edgekey design had more time pressure due to the need for the switch key, subjects could have learned to respond quicker to adapt for this. However, these differences in the design did not appear to have a significant effect on the response time. Finally, the interaction effect was not significant ($F(1, 21) = 0.001, p = 0.96$).

Prompts Correct

The second measure of the prompting task is the accuracy of the subjects in correctly completing the prompt. To get the prompt correct includes two important components for the subject. First, they must remember the prompt as they enter it, and second, they must be able to physically press the buttons within the seven second response window. For the statistical test, we are using the count of how many prompts each subject completed successfully per trial. Among the incorrect prompts, we can differentiate between whether the subject entered the prompt incorrectly (failure to remember the prompt) or whether the subject ran out of time (failure to physically press the buttons). These counts are reported to help analyze the results, but are not used in the statistical tests. There were 12 prompts per trial, and every subject completed three trials for each design.

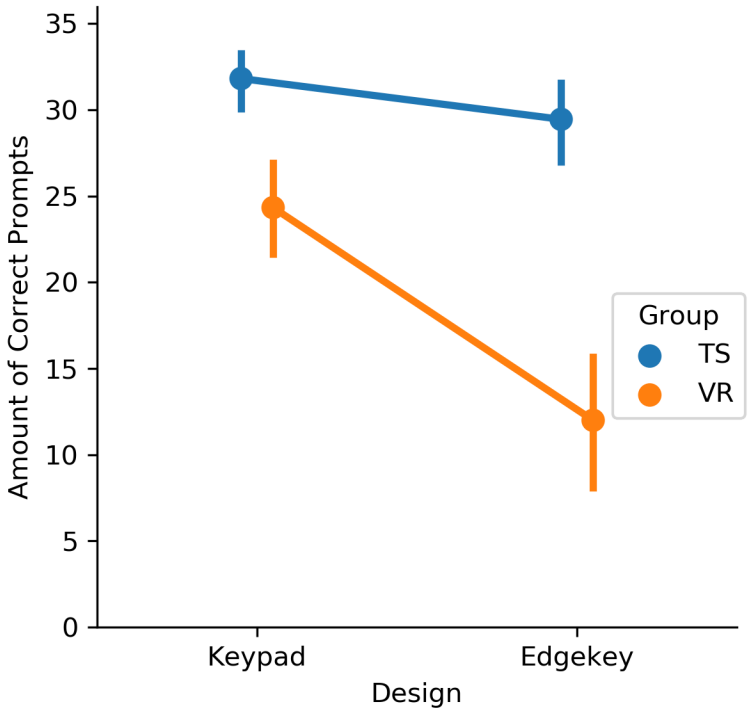


Figure 4.9: Factor Plot of Correct Prompts

The number of correct prompts had a significant interaction effect between group and design ($F(1, 21) = 27.8, p < 0.001$), meaning the main effects must be interpreted with the post-hoc tests as well. Both main effects were significant, the effect of group yielded an F ratio of $F(1, 21) = 43.9, p < 0.001$ while the effect of design yielded an F ratio of $F(1, 21) = 64.1, p < 0.001$.

For the effect of design on the VR group, the repeated measured t-test indicated a significant difference ($t(11) = 8.0, p < 0.001$) between the Keypad ($M = 8.11, \sigma = 1.62$) and the Edgekey ($M = 4.00, \sigma = 2.37$) The TS group had a marginally significant difference ($t(10) = 2.28, p =$

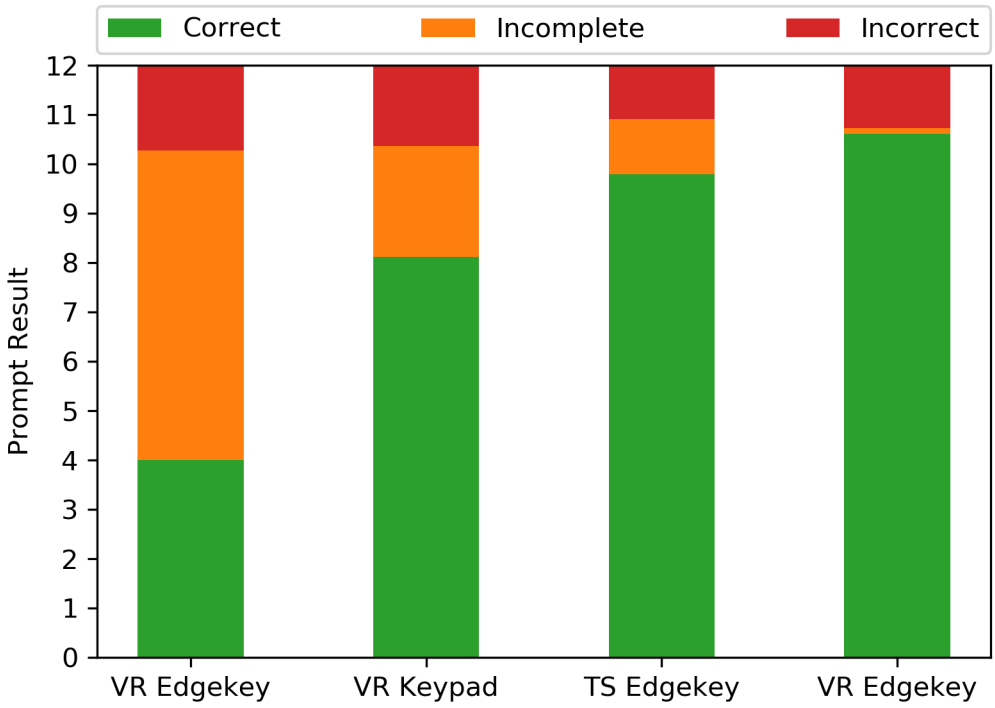


Figure 4.10: Result of prompts

0.045) between Keypad ($M = 10.6, \sigma = 0.96$) and the Edgekey ($M = 9.82, \sigma = 1.38$). These results indicate that both groups had trouble with the additional time pressure caused by the Edgekey design requiring the use of the switch key. The TS group performed a lot closer to their performance in the Keypad design, however, only getting approximately 1 fewer prompt correct. The VR group had much more difficulty in the Edgekey design, correctly completing about half as many as they completed in the Keypad design. However, they had more difficulty in both designs compared to the TS group.

This agrees with the post-hoc tests for differences between groups within

each design. These tests had significant effects for both the Keypad design ($t(21) = 4.44, p < 0.001$) between the VR group and the TS group, and the Edgekey design ($t(21) = 7.05, p < 0.001$) between the VR group and the TS group. The main effect of group clearly has a meaningful effect, which found the VR group ($M = 6.05, \sigma = 2.88$) had significantly fewer correct prompts than the TS group ($M = 10.2, \sigma = 1.2$). This difference is largely due to subjects not being able to complete the prompt. Figure 4.10 shows the breakdown of the mean result of each trial for each group and design.

Across all groups and designs, very few prompts were completed that were incorrect, and most of the difference in number completed correctly is due to the incomplete prompts. A contributing factor for this would be the method of button activation used for the VR group combined with the time pressure. Another contribution would be the limitations of the hand tracker. When the hand tracker lost tracking or gave bad information, it became hard or impossible for the subject to activate a button until the hand tracker returned to normal. When this happened in the middle of a prompt, the amount of time it took to recover from the bad tracking would lead to a timeout on the prompt entry, causing an incomplete prompt. The variance of number correct was also much larger in in the VR group, which could be caused by some subjects adapting to the unfamiliar VR environment more rapidly.

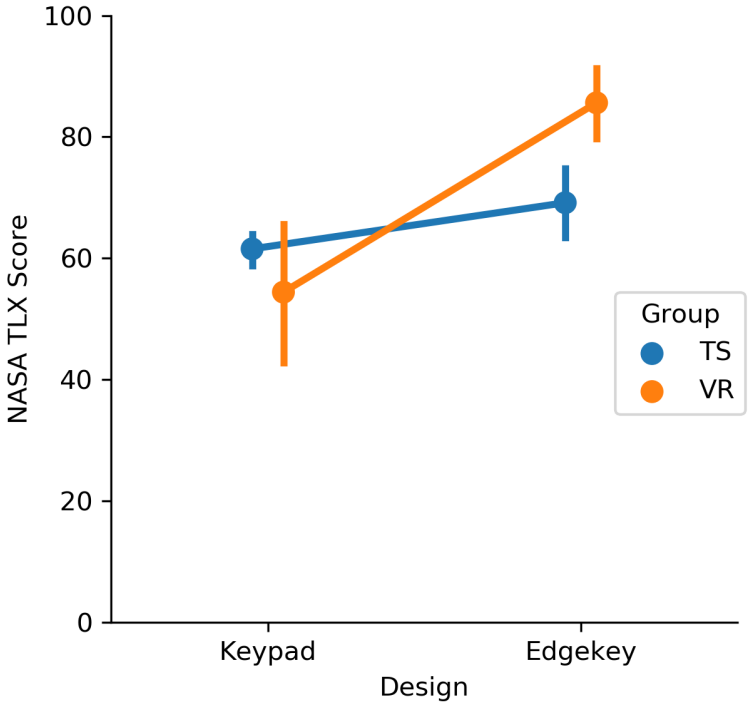


Figure 4.11: Factor Plot of NASA TLX

NASA TLX

After the subject completed their trials for each design, they filled out a NASA TLX workload survey. Their scores, weighted means by the pairwise comparisons, are used here as a measure of their self-reported workload. The interaction effect between group and design was found to be significant ($F(1, 21) = 8.25, p < 0.001$). The main effects showed a significant difference in design ($F(1, 21) = 23.6, p < 0.001$), but not in group ($F(1, 21) = 1.69, p = 0.21$). This could mean that the group did not affect the TLX score, but in the presence of an interaction effect, the post-hoc tests guide the interpretation.

The repeated measures t-tests indicated significance between designs for the VR group ($t(11) = -4.20, p = 0.001$) between the Keypad design ($M = 54.4, \sigma = 20.4$) and the Edgekey ($M = 85.6, \sigma = 11.2$) There was a marginally significant difference between designs for the TS group ($t(10) = -2.72, p = 0.02$) between the Keypad design ($M = 61.5, \sigma = 4.46$) and the Edgekey ($M = 69.2, \sigma = 10.1$). The effect of design was much stronger in the VR group, but both groups indicated respectively higher workload on the TLX scores for the Edgekey design. This follows from the experimental design which predicted that the Edgekey design would be more difficult. One factor that could have contributed to a larger difference in scores for the VR group could be the increased difficulty subjects had in completing the prompt, as seen in the results of the number of incorrect and incomplete prompts for the VR group using the Edgekey design (Figure 4.10). The effect of group was not shown to be significant in the ANOVA analysis, but the independent samples t-test showed a significance for the Edgekey design ($t(21) = 3.69, p < 0.01$) between the VR Group ($M = 85.6, \sigma = 11.2$) and the TS Group ($M = 69.2, \sigma = 10.1$). With the Keypad design, The effect of group was not significant ($t(21) = -1.13, p = 0.27$) between VR ($M = 54.4, \sigma = 20.4$) and TS ($M = 61.5, \sigma = 4.46$). These tests further illustrate that the VR group found a higher workload for the Edgekey design specifically, as both groups rated the workload in the Keypad design similarly.

Tracking Task Learning

Throughout the experiment the subjects did trials with only the tracking task, instead of both the tracking task and the prompting task. Initially, they performed a number of training trials at the beginning with only the tracking task, and then after each evaluation session there was a single trial of just the tracking task. In this section we will test the RMSE of their final training trial and the two after-evaluation trials for any significant learning effects. The trial number is chronological throughout the timeline of the experiment for each subject. This means that due to the counterbalancing, the second and third trial are done with different designs based on the subject. Since the visual environment of the tracking task was quite different for each group, the Group factor is included as a between subjects factor.

The two-way ANOVA found Group to be a marginally significant factor ($F(1, 21) = 4.94, p = 0.37$) between the VR Group ($M = 1.36, \sigma = 0.51$) and the TS Group ($M = 0.94, \sigma = 0.43$). The effect of group on the tracking task was already established, so the marginal significance found here is not unexpected. Trial number was found to have no significant effect ($F(1, 21) = 3.65, p = 0.069$) between the three trials. The means of the three trials, in order, are $1.23^{circ}(\sigma = 0.54^{circ})$, $1.18^{circ}(\sigma = 0.51^{circ})$, and $1.07^{circ}(\sigma = 0.51^{circ})$. Even though the statistical test indicates no significance, the means do decrease as trial number increases. This combined with the large variance suggest that some subjects were experienc-

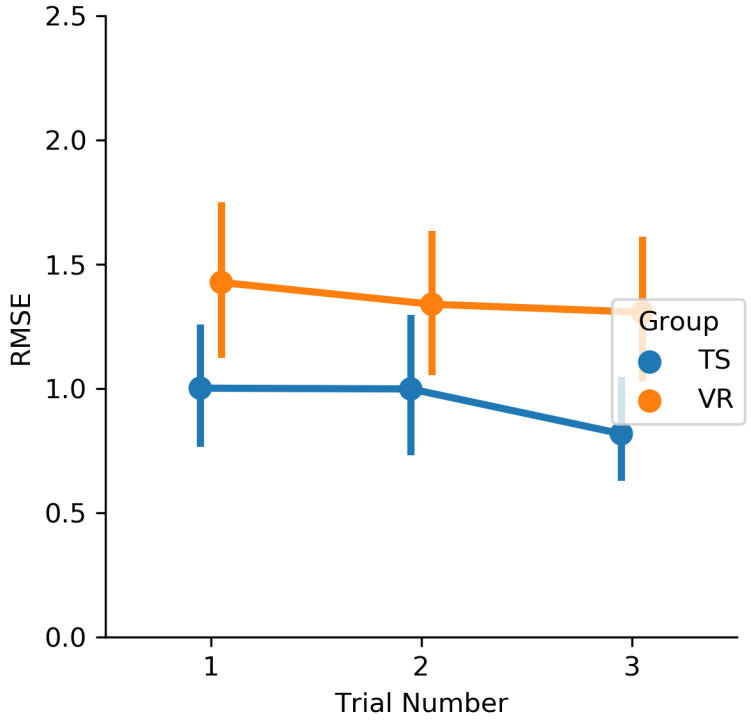


Figure 4.12: Factor Plot of RMSE for Tracking Only Trials

ing some training effects, but overall the effect of training is not significant. The interaction effect of Group and trial number had no significance ($F(1, 21) = 0.16, p = 0.69$).

Summary

A summary of the significance results from the ANOVA and post-hoc t-tests for all the performance measures are shown in Table 4.1. The significance is indicated by '$*$' for $p < 0.01$, '+' for $0.01 < p < 0.05$, and '-' for no significance. For the measures with significant interaction effect, the post-hoc t-tests are shown per group and per design.

	ANOVA		
	Group	Design	Group:Design (Interaction)
Tracking RMSE	∗	+	-
Response Time	-	-	-
Prompts Correct	∗	∗	∗
NASA TLX	-	∗	∗

	t-tests			
	Design		Group	
	VR Group	TS Group	Keypad Design	Edgekey Design
Prompts Correct	∗	+	∗	∗
NASA TLX	∗	+	-	∗

Table 4.1: Statistical Significance Test Results. '∗' indicates significance at the $p < 0.01$ level, '+' indicates marginally significant ($0.01 < p < 0.05$), and '-' indicates no significance.

4.3.3 Design Feedback

As discussed in Section subsection 4.2.5, the long-form feedback questions were synthesized and summarized into categories. The categories and the counts of comment occurrence for each group is summarized in Table 4.2. Categories which only received one comment are not included in this table in interest of brevity, the full table is shown in Appendix A.23.

By far the issue that received the most feedback was the difficulty of using the switch key (Edgekey, Switch Difficult). Most of the complaints stated the extra difficultly of having to press another button. Some of the other complaints from this category were: it took extra time (with no extra time given), it added to the mental demands of the task, and it was

Topic	Feedback Summary Category	VR Group	TS Group
Edgekey	Switch Difficult	14	12
Keypad	Familiar	6	11
Edgekey	Centered Flight Task Better	3	13
Keypad	Buttons Proximal	6	7
Keypad	Buttons Always Visible	5	5
Other	Hand Tracking Issues	9	0
Edgekey	Hand Blocks View	3	4
Fatigue	Prompting Arm	4	1
Edgekey	Clean Design	3	2
Fatigue	Fatigue from Joystick	0	4
Edgekey	Easier	0	4
Keypad	Buttons Confusable	0	4
Other	Colors Disliked	2	2
Fatigue	Eye Fatigue	3	0
Keypad	Easy Focus Switch	2	1
Keypad	More Mistakes	1	2
Edgekey	Accuracy Worse	1	2
Keypad	Buttons Bad Layout	2	0

Table 4.2: Counts of Design Feedback Comments per Group. Sorted by sum of comments.

difficult to see which mode the instrument was in. Both groups disliked the switch key, and mentioned it just as frequently.

> "Switching from numbers to letters was hard, especially if I was trying to compensate for turbulence and was struggling at the time." (TS Subject)

> "I did not like how much extra work it was. It took so much extra focus that I forgot I was flying with the joystick" (VR Subject)

Many subjects noted the familiarity of the Keypad design (Keypad,

Familiar) and that having the buttons close together (Keypad, Buttons Proximal) as things they like about that design. The familiarity was noted more often for the TS Group, but both were some of the more frequent comments within each group.

One comment about the Edgekey design that got more frequent mentions from the TS Group was that they found having the flight task in the middle of the display, centered between the buttons, was preferred (Edgekey, Centered Flight Task Better). The subjects who chose the Edgekey as their preferred design nearly unanimously cited this as their reason for their preference[2]. The comments that fed into this category also included subjects who noted the difficulty of splitting their focus back and forth with the Keypad design. Interestingly, two of the TS Group subjects noted that they would have found the Keypad easier if they had tactile feedback to guide their input. This could suggest that the reason the VR Group subjects did not find the centered flight task advantageous is because with the tactile feedback of the 3D-printed instruments they were able to keep visual focus on the left half of the screen in the Keypad design, thus not seeing benefit from the centering of the flight task display.

> "[The Edgekey design] forced me to pay more attention to what I was typing, this wouldn't have been a problem if the keypad was a physical device that allowed me to locate the numbers and letters without looking, much like the dots on a computer keyboard." (TS Subject)

> "I like that the flight control was cent[e]red, so you could see it even when you were looking at the buttons." (VR Subject)

[2]The one holdout did not explain why they preferred the Edgekey design.

The most notable exceptions to providing similar feedback between groups are the categories that relate to fatigue issues. Many subjects in the TS group noted fatigue caused from using the joystick, yet none in the VR group did, despite using the same joystick setup, and seated in the same location. The VR group did note more fatigue in their other arm that was used for the prompting task. This fatigue seemed to be caused by the additional effort needed to have the hand tracker recognize the hand. For example, one subject wrote:

> "My right wrist was somewhat fatigued. Though I think this is mostly from positioning my hand for the simulator to recognize my input." (VR Subject)

Some of this additional effort was due to subjects learning to hold their prompting task hand "hovering" while waiting for the next prompt. This was done to keep the hand in view of the hand tracker as when the hand leaves the field of view, the re-acquisition will slow down the entry of first button. Many subjects organically learned this, and kept their arm in front of the instrument between prompts.

Similar to the fatigue issues being different, there were some comments that were due to the technology being used more-so than the designs themselves. Obviously, the subjects who noted difficulty using the hand tracker, or the one subject who mentioned touchscreen issues, are specific to the simulator technology the used. However, some of the other categories had comments that may have been indirectly caused by the different technologies and their limitations. For example, some subjects noted the keypad

design caused them to make more mistakes. For the TS Group, this was due to the touchscreen being too responsive to the button presses:

> "[S]ince I was able to go more quickly with this layout, I had more mistakes in the entry." (TS Subject)

One subject in VR who complained of more mistakes in the Keypad design, identified a common problem caused by the hand tracker. When the hand tracker was having registration issues it would sometimes mistakenly place the other fingers in the activation zone of the buttons underneath the one being targeted, causing multiple buttons to be pressed in a short period of time.

> "There's more unintended register since other fingers might trigger the buttons." (VR Subject)

Although only one subject noted this, it was observed happening to many subjects. In fact, for the VR group, eight of the twelve subjects had the wrong button register within 200 milliseconds of the last button in the Keypad design. In the other designs and groups this happened to only one or two subjects.

4.4 Discussion

The motivation of this experiment was to determine the differences between using an R3C simulator system and a traditional simulator system to perform a design evaluation experiment. We had two groups of subjects

perform the same evaluation task on two different designs of a cockpit instrument, one group using the R3C system and the other a touchscreen system. The evaluation task included a pitch disturbance tracking task and a call and response prompting task. In addition to the quantative performance measures of the task, subjects were asked for their feedback on the two designs at the conclusion of the experiment.

The results are summarized using their two independent variables: Group and Design. Group, a between subjects factor, refers to the technology the subject used: either Virtual Reality/R3C (VR) or Touchscreen (TS). Design is a within subjects factor, and is the instrument design the subject was evaluation: Edgekey or Keypad.

The VR Group had worse performance than the TS Group with the RMSE of the tracking task. Subjects from both groups had a marginally significant difference in tracking task performance due to Design, with subjects performing better with the Keypad design. It was also shown that, on control trials that had only the tracking task (no prompting task), the effect of Group was reduced to marginally significant. The response time of the prompting task had no significant effect based on Group nor Design. Neither of these two previous measures had interaction effects between Group and Design. The number of correct prompts had a significant interaction effect. While the TS Group was able to complete significantly more prompts correctly overall than the VR group (averages of 10.2 vs. 6.1, respectively) the VR group had a significant effect with the Design and the

TS group only had marginal significance. This interaction can be clearly seen in the factor plot of correct prompts (Figure 4.9). The NASA TLX workload scores also had an interaction effect between Group and Design. The TLX scores for the VR group had a significant effect in Design, with subjects rating the Edgekey design over 30 points higher than the Keypad design (averages of 54.4 to 85.6, respectively). However, like the number of prompts correct measure, the TLX score was found to be only marginally significant for the TS group, rating the Keypad at 61.5 to the Edgekey's 69.2.

Our results suggest that tasks or performance measures which are dominated by a cognitive portion, such as the prompt response time, provide similar results. Tasks which rely on visual resolution or time pressured responses may not produce the same results between designs using the R3C system. None of the effects reversed slope between designs, however, and the only change is in magnitude of the effect. In fact, for both the number of prompts correct and the workload ratings, which had significant interaction effects, the use of the VR system amplified the effect of design within the groups from a marginally significant effect to a significant effect.

The results of the subjective feedback analysis found that there was no omission of major feedback items on the design of the two instruments from either group. The only feedback comments that did not transfer were the fatigue issues, and of course technology-specific issues. We did discover that some issues were mentioned at differing frequencies, which is to say,

one group would have more subjects mention it than the other. These results suggest that the use of the R3C system for receiving feedback from a design would be appropriate.

Many design evaluation studies would be concluded with both paper questionnaires as well as open interviews to receive the feedback from the subject. Our experimental design avoided the use of the interview for two reasons. First, since our subjects were not subject domain experts or experienced evaluators, we wanted to ensure that the prompting of the questions were consistent. Second, the primary goal of the design feedback for this experiment was not to evaluate the designs, but rather to compare evaluations. The use of a proctor interviewing the subjects could introduce accidental bias into the responses of the subjects. This can often be useful when evaluating a new interface, for example, an interviewer could ask subjects about a flaw they had not mentioned yet to determine if they did not notice it or did not care about it. However, in our case, we forgoed this additional information to ensure no bias was introduced in the collection of their opinions.

Compare to other literature

This was a limited study of the utility of VR/R3C for design evaluation purposes. The task and instrument design was kept simple in nature for this study in order to limit the amount of confounding variables as well as keep it easy to learn for the subject population. Future studies could investigate this system in a more involved design study, with multiple in-

struments or designs, or more complex behavior in the cockpit. At this point, it would become more essential to use subject domain experts (i.e. experienced pilots) in order to validate these results.

References

[1] E. Bachelder, D. H. Klyde, N. Brickman, S. Apreleva, and B. Cogan. Fused Reality for Enhanced Flight Test Capabilities. In *AIAA Atmospheric Flight Mechanics (AFM) Conference*, Guidance, Navigation, and Control and Co-located Conferences. American Institute of Aeronautics and Astronautics, Aug. 2013. DOI: 10.2514/6.2013-5162.

[2] G. Borg. *Borg's perceived exertion and pain scales.* Human Kinetics, Champaign, IL, US, 1998.

[3] C. W. Borst and R. A. Volz. Evaluation of a haptic mixed reality system for interactions with a virtual control panel. *Presence: Teleoperators and Virtual Environments*, 14(6):677–696, 2005.

[4] G. Bruder, F. Steinicke, and W. Sturzlinger. To touch or not to touch?: comparing 2d touch and 3d mid-air interaction on stereoscopic tabletop surfaces. In *Proceedings of the 1st symposium on Spatial user interaction*, pages 9–16. ACM, 2013.

[5] G. C. Burdea and P. Coiffet. *Virtual reality technology*, volume 1. John Wiley & Sons, 2003.

[6] S. Card, W. K. English, and B. J. Burr. Evaluation of Mouse, Rate-Controlled Isometric Joystick, Step Keys, and Text Keys for Text Selection on a CRT. *Ergonomics*, 21(8):601–613, Aug. 1978.

[7] Y. Cha and R. Myung. Extended Fitts' law for 3d pointing tasks using 3d target arrangements. *International Journal of Industrial Ergonomics*, 43(4):350–355, July 2013.

[8] K. Chun, B. Verplank, F. Barbagli, and K. Salisbury. Evaluating haptics and 3d stereo displays using Fitts' law. In *Haptic, Audio*

and Visual Environments and Their Applications, 2004. HAVE 2004. Proceedings. The 3rd IEEE International Workshop on, pages 53–58. IEEE, 2004.

[9] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology: General*, 47(3):262, 1954.

[10] D. C. Foyle, A. D. Andre, R. S. McCann, E. M. Wenzel, D. R. Begault, and V. Battiste. Taxiway Navigation and Situation Awareness (T-NASA) System: Problem, Design Philosophy, and Description of an Integrated Display Suite for Low-Visibility Airport Surface Operations. SAE International, 1996.

[11] T. Grossman and R. Balakrishnan. Pointing at trivariate targets in 3d environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 447–454. ACM, 2004.

[12] H. Wan, S. Zou, Z. Dong, H. Lin, and H. Bao. MRStudio: A mixed reality display system for aircraft cockpit. In *2011 IEEE International Symposium on VR Innovation*, pages 129–135, Mar. 2011.

[13] R. K. Heffley and W. F. Jewell. Aircraft handling qualities data. Technical Report 2144, NASA, 1972.

[14] J. D. Hincapi-Ramos, X. Guo, P. Moghadasian, and P. Irani. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. pages 1063–1072. ACM Press, 2014.

[15] I. Yavrucuk, E. Kubali, and O. Tarimci. A low cost flight simulator using virtual reality tools. *IEEE Aerospace and Electronic Systems Magazine*, 26(4):10–14, Apr. 2011.

[16] B. E. Insko. *Passive haptics significantly enhances virtual environments*. PhD thesis, University of North Carolina at Chapel Hill, 2001.

[17] International Organization for Standardization. ISO 9241-9:2000, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 9: Requirements for non-keyboard input devices. Technical report, 2000.

[18] L. Liu, R. Van Liere, C. Nieuwenhuizen, and J.-B. Martens. Comparing aimed movements in the real world and in virtual reality. In *Virtual Reality Conference, 2009. VR 2009. IEEE*, pages 219–222. IEEE, 2009.

[19] A. Lcuyer. Simulating haptic feedback using vision: A survey of research and applications of pseudo-haptic feedback. *Presence: Teleoperators and Virtual Environments*, 18(1):39–53, 2009.

[20] I. S. MacKenzie. A note on the information-theoretic basis of Fitts' law. *Journal of Motor Behavior*, 21(3):323–330, 1989.

[21] W. McNeely and others. Robotic graphics: a new approach to force feedback for virtual reality. In *Virtual Reality Annual International Symposium, 1993., 1993 IEEE*, pages 336–341. IEEE, 1993.

[22] A. Murata and H. Iwase. Extending Fitts' law to a three-dimensional pointing task. *Human movement science*, 20(6):791–805, Dec. 2001.

[23] C. Pontonnier, G. Dumont, A. Samani, P. Madeleine, and M. Badawi. Designing and evaluating a workstation in real and virtual environment: toward virtual reality based ergonomic design sessions. *Journal on Multimodal User Interfaces*, 8(2):199–208, June 2014.

[24] J. Schiefele, O. Albert, V. van Lier, and C. Huschka. Simple force feedback for small virtual environments. In *Aerospace/Defense Sensing and Controls*, pages 100–110. International Society for Optics and Photonics, 1998.

[25] R. W. Soukoreff and I. S. MacKenzie. Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts law research in HCI. *International Journal of Human-Computer Studies*, 61(6):751–789, Dec. 2004.

[26] R. Stone. Haptic feedback: A brief history from telepresence to virtual reality. *Haptic Human-Computer Interaction*, pages 1–16, 2001.

[27] T. Aslandere, D. Dreyer, and F. Pankratz. Virtual hand-button interaction in a generic virtual reality flight simulator. In *2015 IEEE Aerospace Conference*, pages 1–8, Mar. 2015.

[28] S. Tachi, T. Maeda, R. Hirata, and H. Hoshino. A Construction Method of Virtual Haptic Space. In *Proceedings of the ICAT'94*, Tokyo, Japan, 1994.

[29] R. J. Teather, D. Natapov, and M. Jenkin. Evaluating haptic feedback in virtual environments using ISO 92419. In *Virtual Reality Conference (VR), 2010 IEEE*, pages 307–308. IEEE, 2010.

[30] A. T. Welford. *Fundamentals of skill*. 1968.

[31] B. G. Witmer and M. J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3):225–240, June 1998.

[32] R. S. Woodworth. Accuracy of voluntary movement. *The Psychological Review: Monograph Supplements*, 3(3):i, 1899.

[33] C. Youngblut and O. Huie. The relationship between presence and performance in virtual environments: Results of a VERTS study. In *Virtual Reality, 2003. Proceedings. IEEE*, pages 277–278. IEEE, 2003.

Appendices

Appendix A

Result Tables

A.1 Passive Haptics Experiment

Sequence	Haptics	Mean	Std. Dev.
NH First	—	3.877	0.5265
PH First	—	3.99	0.3684
—	NH	3.71	0.3805
—	PH	4.157	0.412
PH First	NH	3.879	0.3892
PH First	PH	4.1	0.3284
NH First	NH	3.541	0.3007
NH First	PH	4.213	0.4934

Table A.1: Throughput Means

Factor	F ratio	p value
Sequence	0.5297	0.4761
Haptics	35.53	1.221×10^{-5}
Sequence:Haptics	9.062	0.007516

Table A.2: Throughput ANOVA

Sequence	Haptics	df	t	p value
PH First	—	9	2.547	0.03135
NH First	—	9	5.503	0.0003787
—	PH	18	−0.602	0.5547
—	NH	18	2.176	0.04309

Table A.3: Throughput t-tests

Sequence	Haptics	Mean	Std. Dev.
NH First	—	77.5	10.56
PH First	—	71.2	8.764
—	NH	71	9.701
—	PH	77.7	9.565

Table A.4: Presence Score Means

Factor	F ratio	p value
Sequence	4.089	0.05828
Haptics	6.079	0.02396
Sequence:Haptics	0.7164	0.4084

Table A.5: Presence Score ANOVA

Haptics	Cronbach's alpha
No Haptics	0.9139
Passive Haptics	0.9191

Table A.6: Presence Score Cronbachs alpha

Sequence	Haptics	Mean	Std. Dev.
NH First	—	15.07	2.948
PH First	—	13.1	3.523
—	NH	14.72	2.854
—	PH	13.45	3.762
PH First	NH	14.5	3.064
PH First	PH	11.7	3.529
NH First	NH	14.95	2.773
NH First	PH	15.2	3.259

Table A.7: Arm Fatigue Ratings Means

Factor	F ratio	p value
Sequence	2.119	0.1627
Haptics	9.717	0.00595
Sequence:Haptics	13.9	0.001538

Table A.8: Arm Fatigue Ratings ANOVA

Sequence	Haptics	df	t	p value
PH First	—	9	−4.583	0.001323
NH First	—	9	0.4596	0.6567
—	PH	18	−2.304	0.03336
—	NH	18	−0.3443	0.7346

Table A.9: Arm Fatigue Ratings t-tests

Please circle the number that best describes the amount of fatigue you feel in your arm. Try to consider the overall sensation of fatigue, soreness, and exertion level. Consider the arm to include shoulder muscles that you use to move the arm.	
6	No fatigue
7	Very, very lightly fatigued
8	
9	Very lightly fatigued
10	
11	Lightly fatigued
12	
13	Somewhat fatigued
14	
15	Heavily fatigued
16	
17	Very heavily fatigued
18	
19	Very, very heavily fatigued
20	Maximal fatigue

Table A.10: Borg RPE Scale as used

A.2 Design Evaluation Experiment

Group	Design	Mean	Std. Dev.
TS	—	1.277	0.3789
VR	—	1.967	0.378
—	Edgekey	1.704	0.5188
—	Keypad	1.57	0.5068

Table A.11: RMSE Means

Factor	F ratio	p value
Group	21.42	0.000145
Design	5.944	0.02374
Group:Design	0.1669	0.687

Table A.12: RMSE ANOVA

Group	Design	Mean	Std. Dev.
TS	—	2595	567.3
VR	—	2813	383.2
—	Edgekey	2688	471.1
—	Keypad	2729	512.5

Table A.13: Response Time Means

Factor	F ratio	p value
Group	1.199	0.2859
Design	0.6814	0.4184
Group:Design	0.00184	0.9662

Table A.14: Response Time ANOVA

Group	Design	Mean	Std. Dev.
TS	—	10.2	1.242
VR	—	6.056	2.889
—	Edgekey	6.768	3.525
—	Keypad	9.304	1.831
VR	Edgekey	4	2.37
VR	Keypad	8.111	1.616
TS	Edgekey	9.788	1.393
TS	Keypad	10.61	0.964

Table A.15: Correct Prompts Means

Factor	F ratio	p value
Group	43.56	1.552×10^{-6}
Design	63.93	8.309×10^{-8}
Group:Design	26.89	3.872×10^{-5}

Table A.16: Correct Prompts ANOVA

Group	Design	df	t	p value
VR	—	11	8.039	6.234×10^{-6}
TS	—	10	2.287	0.04526
—	Keypad	21	4.441	0.0002262
—	Edgekey	21	7.053	5.839×10^{-7}

Table A.17: Correct Prompts t-tests

Group	Design	Mean	Std. Dev.
TS	—	65.35	8.535
VR	—	70.01	22.65
—	Edgekey	77.74	13.4
—	Keypad	57.83	15.18
VR	Edgekey	85.61	11.21
VR	Keypad	54.42	20.4
TS	Edgekey	69.15	10.06
TS	Keypad	61.55	4.468

Table A.18: NASA TLX Means

Factor	F ratio	p value
Group	1.688	0.208
Design	23.57	8.455×10^{-5}
Group:Design	8.252	0.009113

Table A.19: NASA TLX ANOVA

Group	Design	df	t	p value
VR	—	11	−4.205	0.001474
TS	—	10	−2.718	0.02164
—	Keypad	21	1.132	0.2703
—	Edgekey	21	−3.693	0.001351

Table A.20: NASA TLX t-tests

Group	Trial	Mean	Std. Dev.
TS	—	0.9402	0.425
VR	—	1.359	0.5176
—	1	1.225	0.5421
—	2	1.177	0.5144
—	3	1.074	0.5053
VR	1	1.428	0.5754
VR	2	1.34	0.4877
VR	3	1.308	0.5244
TS	1	1.002	0.4223
TS	2	0.9994	0.5038
TS	3	0.8188	0.3488

Table A.21: Tracking Only Trials RMSE Means

Factor	F ratio	p value
Group	4.937	0.0374
Trial	3.649	0.06987
Group:Trial	0.1621	0.6913

Table A.22: Tracking Only Trials RMSE ANOVA

Topic	Feedback Summary Category	VR Group	TS Group
Edgekey	Accuracy Worse	1	2
Edgekey	Busy Design	0	1
Edgekey	Centered Flight Task Better	3	13
Edgekey	Clean Design	3	2
Edgekey	Easier	0	4
Edgekey	Familiar	1	0
Edgekey	Hand Blocks View	3	4
Edgekey	Labels Easy to Read	1	0
Edgekey	Labels Hard to Read	1	0
Edgekey	Switch Difficult	14	12
Edgekey	Switch Neutral	0	1
Fatigue	Eye Fatigue	3	0
Fatigue	Fatigue from Joystick	0	4
Fatigue	Prompting Arm	4	1
Keypad	Buttons Always Visible	5	5
Keypad	Buttons Bad Layout	2	0
Keypad	Buttons Confusable	0	4
Keypad	Buttons Proximal	6	7
Keypad	Easy Focus Switch	2	1
Keypad	Familiar	6	11
Keypad	Labels Good	1	0
Keypad	Labels Hard to Read	1	0
Keypad	More Mistakes	1	2
Keypad	More Successful	0	1
Keypad	Tracking Easier	1	0
Other	Colors Disliked	2	2
Other	Hand Tracking Issues	9	0
Other	Prompt Location Bad (Both)	1	1
Other	Single Finger Hard	1	1
Other	Touchscreen Issues	0	1

Table A.23: Full Feedback Comments by Category.