Contents

Chapter 1

Design Evaluation Experiment

1.1 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 **??**). The following experiment (Chapter **??**) 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.

1.2 Methods

1.2.1 Simulator Setup

The simulator workstation as configured for each group is shown and annotated in Figure 1.1. 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

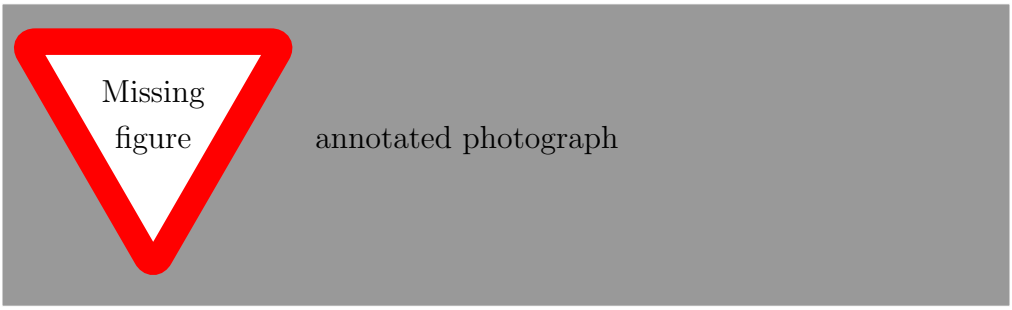

Figure 1.1: Simulator Workstation

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 **??**.

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.

1.2.2 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 1.2. 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 1.1.

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 **??**. 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-2044**NASACR** . This dynamics reference

Figure 1.2: Attitude Indicator Display

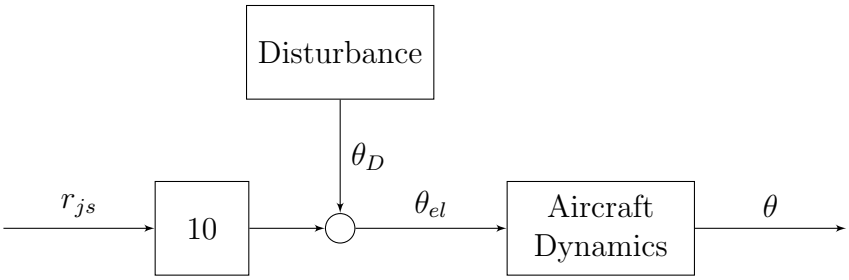

Figure 1.3: Tracking Task Dynamics Block Diagram

model was 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:

$$\frac{\theta}{\theta_{el}} = \frac{-0.572(s+0.553)(s+0.0396)}{(s^2+2\zeta_1\omega_1+{\omega_1}^2)(s^2+2\zeta_2\omega_2+{\omega_2}^2)} \tag{1.1}$$

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

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

The disturbance model is based off the model developed in SweetRef**sweet**. 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:

reference

$$\theta_D = K\sum_{i=1}^{12}\left[a_i\left(\frac{2\pi k_i}{240}\right)\sin\left(\frac{2\pi k_i}{240}t+\phi_i\right)\right] \tag{1.2}$$

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.

1.2.3 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 1.4. 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 1.5. 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

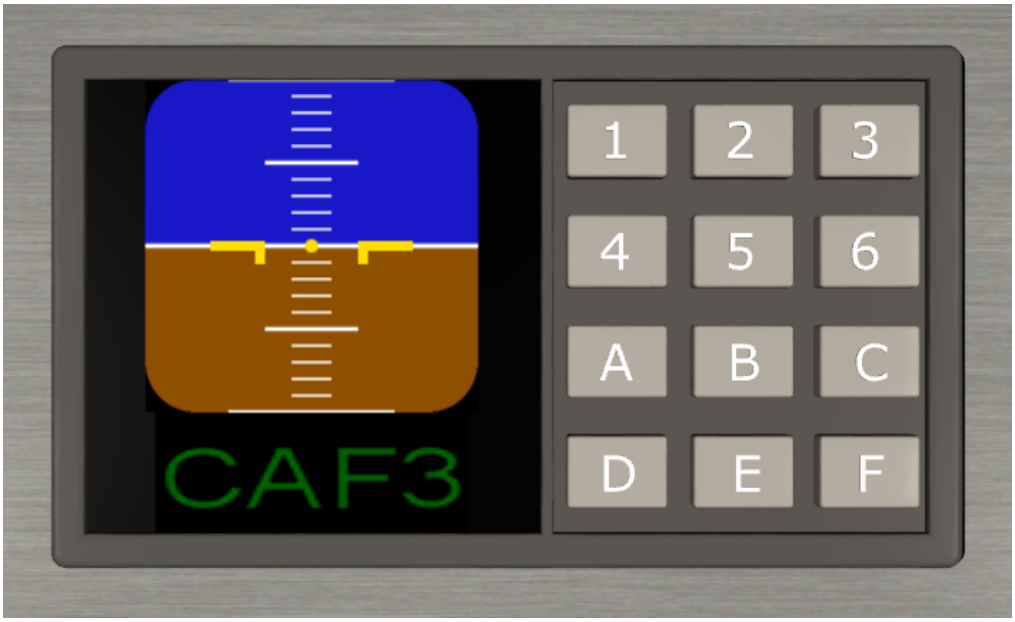

Figure 1.4: Keypad Design

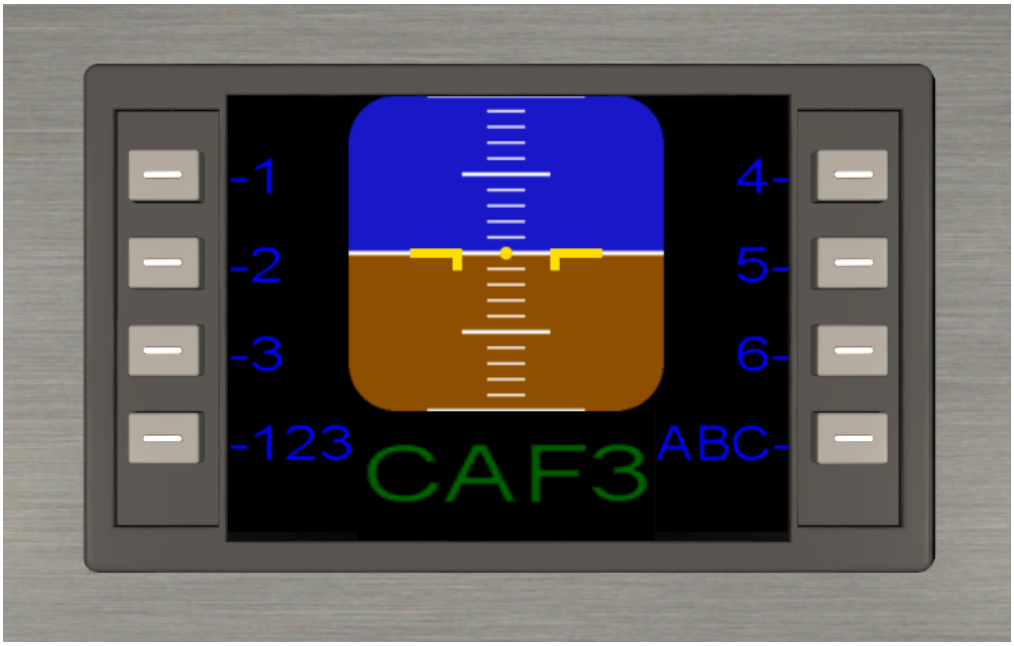

Figure 1.5: Edgekey Design

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 1.6. 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.

1.2.4 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

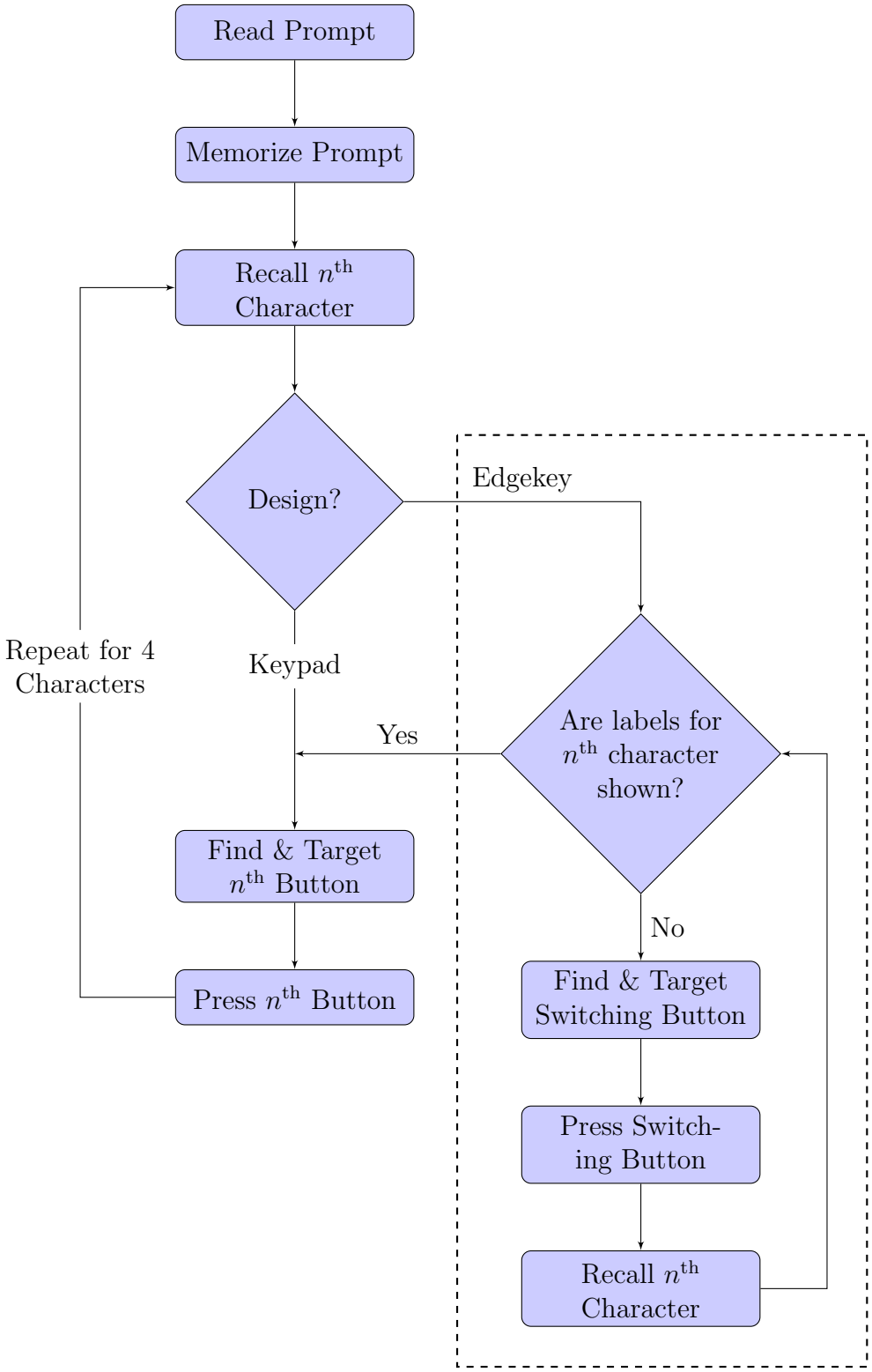

Figure 1.6: Prompting Task Flowchart of Cognitive Work for Each Design. Extra work of Edgekey design encolsed in dashed line box.

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.

1.2.5 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 1.2.2.

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.

1.2.6 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

1.2.7 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.

1.3 Results

1.3.1 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.

1.3.2 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

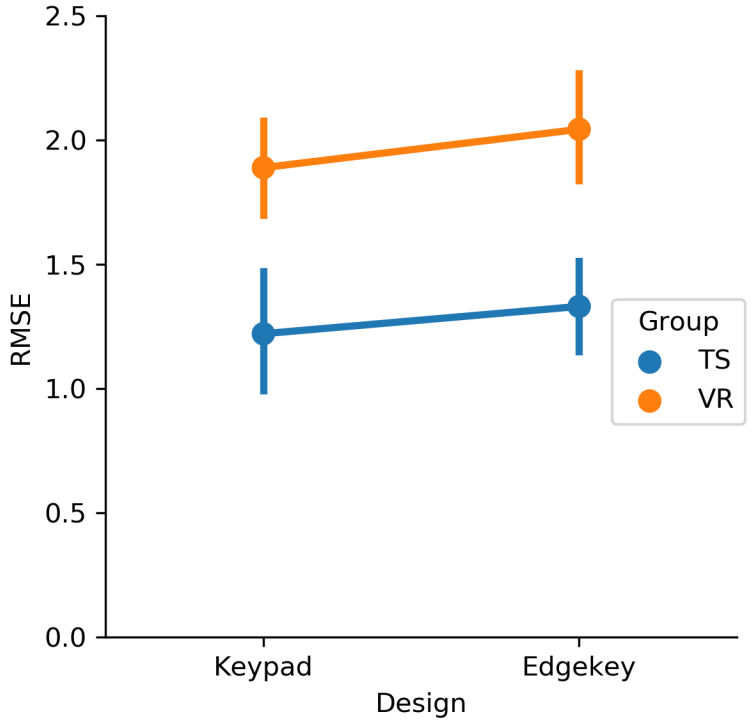

Figure 1.7: Factor Plot of RMSE

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

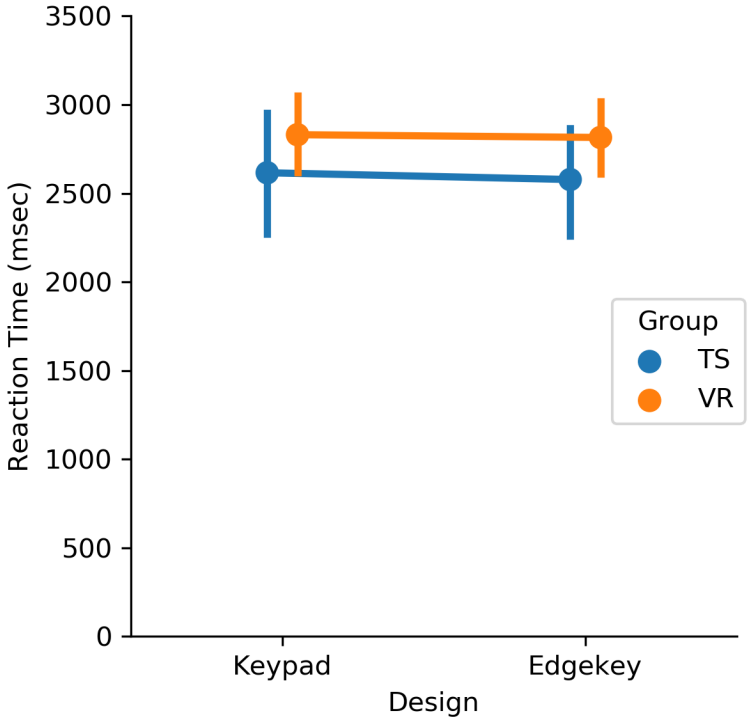

Figure 1.8: Factor Plot of Response Time

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 1.6). 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.

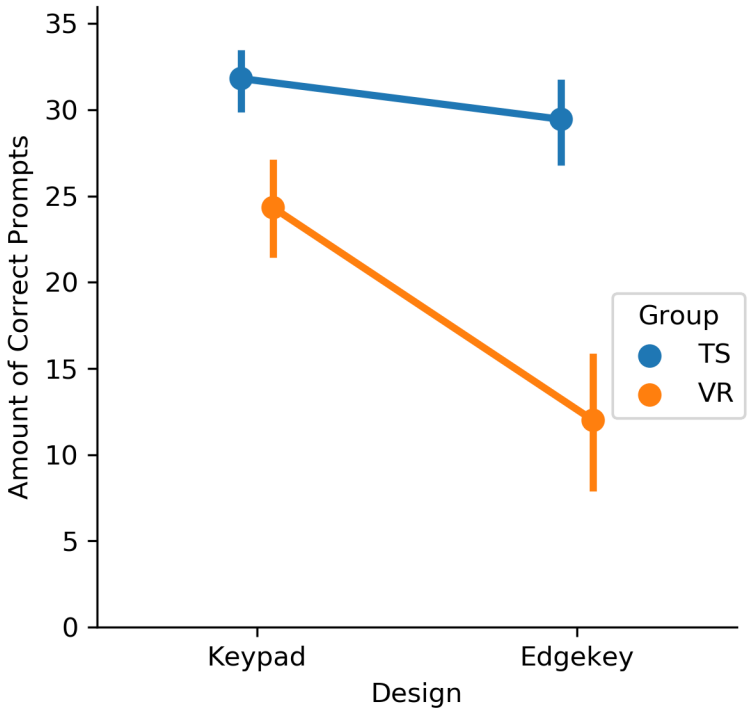

Figure 1.9: Factor Plot of Correct Prompts

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 ($t(10) = 2.28, p =$

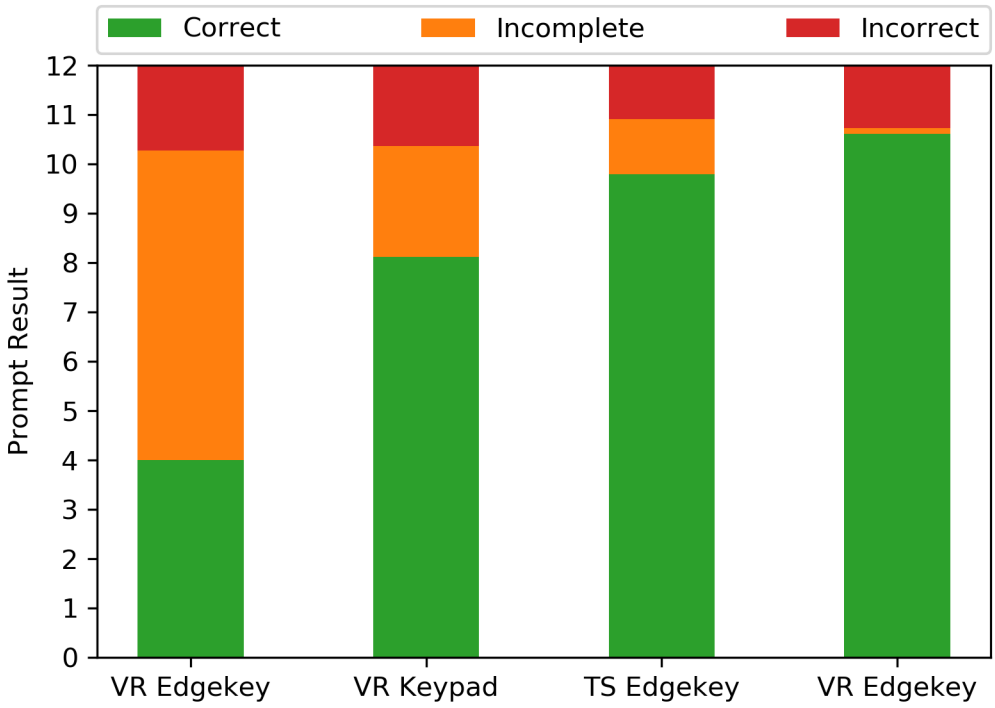

Figure 1.10: Result of prompts

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 1.10 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.

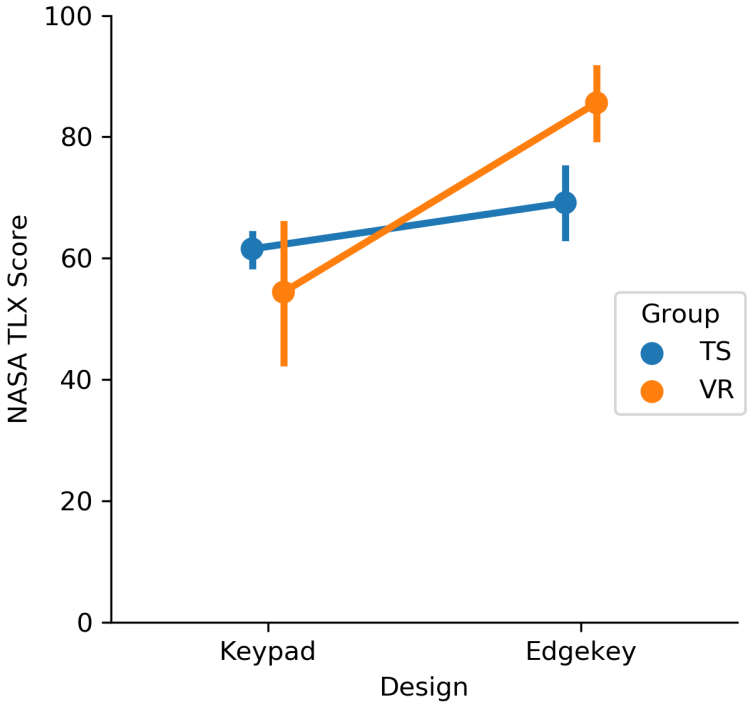

Figure 1.11: Factor Plot of NASA TLX

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 1.10). 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-

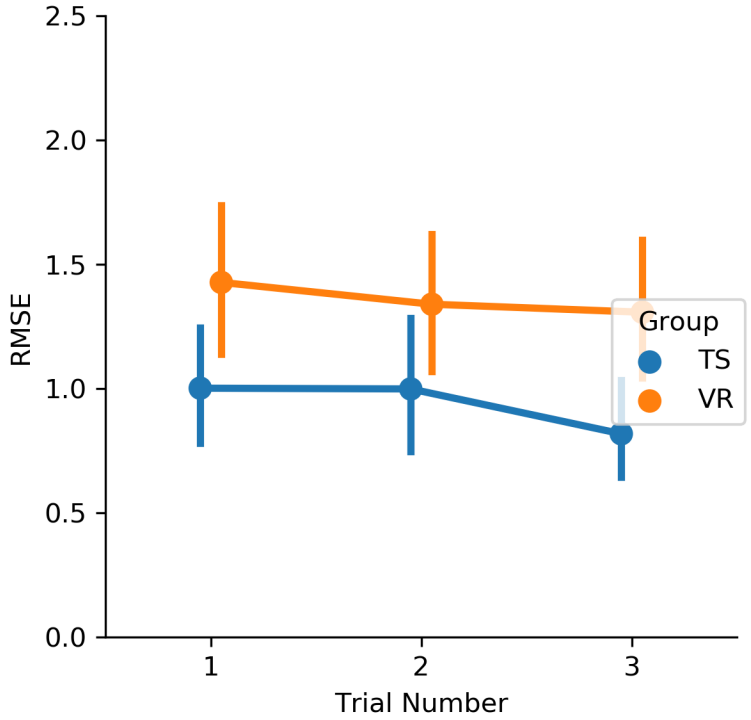

Figure 1.12: Factor Plot of RMSE for Tracking Only Trials

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 1.1. 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 1.1: Statistical Significance Test Results. '∗' indicates significance at the $p < 0.01$ level, '+' indicates marginally significant ($0.01 < p < 0.05$), and '-' indicates no significance.

1.3.3 Design Feedback

As discussed in Section subsection 1.2.5, 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 1.2. Categories which only received one comment are not included in this table in interest of brevity, the full table is shown in Appendix **??**.

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 1.2: 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.

1.4 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 1.9). 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.

Appendices

Appendix A

Result Tables

A.1 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.1: 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.2: 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.3: 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.4: 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.5: 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.6: 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.7: 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.8: 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.9: 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.10: 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.11: 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.12: 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.13: Full Feedback Comments by Category.