Contents

Chapter 1

Introduction

Chapter 2

Design Evaluation Experiment

2.1 Introduction

After investigating the technical approach and the benefit to including the passive haptics layer, we seek to investigate the use of the Rapidly Reconfigurable Research Cockpit in a more realistic design evaluation study. The advantages of using the R3C system would not be useful if it masked defects in a design study.

2.2 Experimental Design

2.2.1 Task Design

The task the subjects were to perform had a number of requirements.

- Ability to simulate designs for completing task on touchscreen and R3C setup
- Tracking task using a standard attitude indicator display controlled with joystick
- Second task that requires use of multiple button to button movements on the instrument
- Sufficient workload such that subjects have a high workload but are not high but not full workload
- Simple design yet complex enough task to have sufficient workload
- Operationally relevant, or analogous to tasks required in a cockpit

Figure 2.1: Attitude Indicator Display

2.2.2 Instrument Designs

2.2.3 Tracking Task

The tracking task display was a standard attitude indicator display, shown in Figure 2.1.

Subjects controlled the one-dimensional (pitch only) task using a joystick with their left hand. The joystick is pictured in Figure **??**.

The flight dynamics model of the simulator was a stability derivative based mode for a Boeing 747 in a low altitude landing configuration.

The disturbance model is based off the model developed in SweetRef.

Include all the info on the FDM

Explain the disturbance

2.2.4 Prompting Task

The prompting task was designed to be a distraction from the tracking task.

The prompts were seperated into 10 second "windows". The prompt would appear between 2 and 3 seconds of the start of the window. From the time of appearance, a seven (7) second timer will start until timeout. When the subject presses the first button of the prompt, the prompt itself was cleared and asterisk symbols (*) were shown after each button entry by the subject. After the subject has entered 4 buttons or the timeout occurs, whichever comes first, the prompt or entry so far would clear. This process is then repeated every 10 seconds.

2.2.5 VR Group

2.2.6 TS Group

The Touchscreen (TS) group used a 10.1 inch capacitive touch screen with resolution of 1024x768. The two instruments were drawn in a web browser, using standard HTML elements for the buttons. The visuals of the tracker were rendered on top with the same OpenGL rendering code from the VR group.

2.3 Methods

Subjects were divided into the two groups, TS and VR. The overall sequence of the experiment started with a training session on the simulator and the task, then an evaluation session for each of the two designs, finally finishing with questionnaires asking about the 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, and the functionality two designs they were to evaluate. Next, they performed practice trials with just the tracking task and then just the

prompting task.

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. The following three were two minutes each, and were the trials used for the results. Each evaluation session concluded with a two minute trial of just the tracking task. This was included to investigate if the subject had improved or fatigued at the tracking task.

2.3.1 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 above.

The prompting task has two dependent measures, for speed and accuracy. For speed we consider the *response time*, defined as the time between 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, and these measures are meaned per trial and then per design for each subject.

For workload, a NASA Task Load Index (TLX) survey was administered after they completed each design. 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 (50) 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 combinations of pairs of subscales, which of the two 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 goal of 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", it is not a useful exercise, as the only result is to document what potential differences could arise.

To analyze these results, the sentences from the open form responses were first separated into single feedback comments, and reworded to use common language. If a 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.

2.3.2 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 group and the TS group. When the ANOVA showed signif-

icance in the interaction test, post-hoc repeated measured t-tests were undertaken to determine the significance of Design within each Group. All effects were considered statistically significant at the 0.0125 level. Statistical significance level was corrected using the Bonferroni correction considering the 4 different dependent measures being tested ($\alpha = 0.05/4 = 0.0125$).

2.4 Results

2.4.1 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 female subjects 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.

2.4.2 Performance Measures

The performance of the tracking task was measured using the root-mean square average (RMSE) of the error. 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 ($M = 1.97\text{deg}, \sigma = 0.38\text{deg}$). The effect of design indicated no significant difference ($F(1, 21) = 5.94, p = 0.024$) 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 interaction effect was not significant ($F(1, 21) = 0.17, p = 0.69$).

Response time. The effect of group yielded an F ratio of $F(1, 21) = 1.61, p = 0.22$ indicating no significant difference between VR ($M = 2983\text{msec}, \sigma =$ 439msec) and TS ($M = 2737\text{msec}, \sigma = 566\text{msec}$). The effect of design indicated a significant difference ($F(1, 21) = 13.9, p = 0.001$) between Keypad ($M = 2728\text{msec}, \sigma = 512\text{msec}$) and Edgekey ($M = 3002, \sigma = 488\text{msec}$). The interaction effect was not significant ($F(1, 21) = 0.17, p = 0.69$).

Number of prompts correct. The effect of group yielded an F ratio of $F(1, 21) = 43.9, p < 0.001$ indicating a significant difference between VR ($M = 6.06, \sigma = 2.90$) and TS ($M = 10.2, \sigma = 1.23$). The effect of design indicated a significant difference ($F(1, 21) = 64.1, p < 0.001$)

between Keypad ($M = 9.30, \sigma = 1.83$) and Edgekey ($M = 6.78, \sigma = 3.54$). The interaction effect was significant as well ($F(1, 21) = 27.8, p < 0.001$). The post-hoc tests indicated significance between designs for the VR group ($t(11) = 8.0, p < 0.001$) between the Keypad design ($M = 8.11, \sigma = 1.62$) and the Edgekey ($M = 4.00, \sigma = 2.37$) The post-hoc tests indicated no significant difference between designs for the TS group ($t(10) = 2.3, p = 0.05$) between the Keypad design ($M = 9.82, \sigma = 1.38$) and the Edgekey ($M = 10.6, \sigma = 0.96$)

NASA TLX scores. The effect of group yielded an F ratio of $F(1, 21) = 1.69, p = 0.21$ indicating a significant difference between VR ($M = 70.0, \sigma = 22.6$) and TS ($M = 65.3, \sigma = 8.53$). The effect of design indicated a significant difference ($F(1, 21) = 23.6, p < 0.001$) between Keypad ($M = 57.8, \sigma = 15.2$) and Edgekey ($M = 77.7, \sigma = 13.4$). The interaction effect was significant as well ($F(1, 21) = 8.25, p < 0.001$). The post-hoc 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 ($M = 85.6, \sigma = 11.2$) The post-hoc tests indicated no 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)

2.4.3 Design Feedback

2.5 Discussion

2.6 Conclusion

Appendices