Assessing the Impacts of Immersive, Adaptive Virtual Reality Training for Entry, Descent, and Landing Tasks on Long Duration Exploration Missions

by

Sandra Tredinnick

A Culminating Experience Project submitted to the
Faculty of the Graduate School at the University of Colorado Boulder
for fulfillment of the requirement for the degree of
Master of Science
Department of Applied Mathematics
2023

Research Advisors:
Allison P. Anderson
Eric Vance

Abstract

This paper seeks to explore the impacts of immersive, adaptive virtual reality training (VR-A training condition) on entry, descent, and landing (EDL) tasks by investigating the following research questions: 1) how well do subjects learn EDL skills in training as a function of their training condition?, 2) how did each training condition perform in the physical mock-up, what skill level did each training condition attain, and how do they compare across training conditions?, 3) how does the subjective perceptions of each training condition differ by training condition in training and in the physical mock-up?, and 4) for each training condition, how does the subjects' subjective perceptions of their assigned training condition change as training progresses? Relying on System Usability Scale (SUS), Affect Grid, and Flow Short Scale survey responses, as well as the skill and performance dataset, generalized logistic mixed effects modeling and appropriate hypothesis testing were performed. Interesting results found that the VR-A training condition plateaued at the highest integrated skill level in training. Additionally, results showed the usability of the VR-A system increased significantly as training progressed, and the VR-A system was the most usable in training overall. In terms of flow experience, the VR-A condition had an extensive increase in flow experience from the start to end of training, whereas the other training conditions' stayed relatively stagnant. In the matter of perceived task importance, the VR-A training condition considered the EDL tasks at hand the most important in training. Finally, in the physical mock-up, results demonstrated that the VR training conditions had substantially higher perceived task importance then the 2D training condition.

1 Introduction

On long-duration exploration missions (LDEMs), there is a significant gap in time between when astronauts are trained pre-flight to perform mission-critical tasks, and when they are asked to perform them. This duration, coupled with non-immersive and/or non-adaptive training methods, can potentially lead to degraded performance and skill that threatens mission success. Virtual reality (VR) has recently become widely adopted as a tool for learning as well as entertainment. As a low mass, considerably understudied, customizable, and off-the-shelf hardware option, VR may be useful for retaining mission-critical skills during LDEMs more efficiently. To understand the effects of adaptive, immersive virtual reality training on training efficacy for LDEMs, the Trinity team developed 3 different training conditions: 1) adaptive VR (VR-A), 2) non-adaptive VR (VR-NA), and 3) non-adaptive screens (2D-NA); these training conditions are combinations of task environments and training algorithms.

The training algorithms are adaptive (A) and non-adaptive (NA), and refer to the method for adjusting difficulty to facilitate flow, learning development, and learning effectiveness. The training task environments are virtual reality (VR) and screens (2D), and encompass the conditions that an astronaut trains and completes a given set of tasks. For purposes of this paper, we focus specifically on Entry, Descent, and Landing (EDL) relevant tasks such as landing site

selection (ls), manual control (mc), and terminal descent engine (de). The Trinity team also makes use of the ARES cockpit physical mock-up at the University of Colorado Boulder, which closely resembles the true spaceflight environment experienced by astronauts when completing EDL tasks on LDEMs.

Accordingly, this paper seeks to understand the effects of the training conditions (with specific focus on the immersive, adaptive trainer: VR-A) on EDL tasks by investigating the following research questions: 1) how well do subjects learn EDL tasks in training, as a function of their training condition, 2) how did each training condition perform in the physical mock-up, what skill level did each training condition attain, and how do they compare across training conditions?, 3) how does the subjective perceptions of each training condition differ by training condition in training and in the physical mock-up?, and 4) for each training condition, how does the subjects' subjective perceptions of their assigned training condition change as training progresses? For organizational reasons, the research questions can be divided into 3 categories of analyses, namely: 1) learning as assessed in training, 2) skill and performance as assessed in the physical mock-up, and 3) subjective perceptions of a given training condition.

2 Materials and Methods

2.1 The Data and Testing Procedures

Following ethical practices, 24 healthy subjects who attended the University of Colorado Boulder were randomly recruited for participation in this experimental research study. Subjects were excluded from the study if their age was not between 18 to 65 and/or if the subject scored above the 90th percentile on the Motion Sickness Susceptibility Questionnaire. The 24 willing subjects were then randomly assigned a training condition (i.e. VR-A, VR-NA, or 2D-NA) - each training condition comprised 8 subjects.

Spread out between 18 to 48 hours, subjects completed 3 training sessions using their assigned training condition. Then, following training and within the corresponding 48 hours, the subject was asked to complete a final session in the physical mock-up. Note that all 24 subjects (regardless of their assigned training algorithm) were tested using the same algorithm in the physical mock-up. Prior to beginning their first trial, subjects watched a task introduction video that explained the goals of the EDL scenario, each task, the interactions needed to be successful, and how each task would be scored. After watching the training video, the operator verified that subjects could identify essential information on the flight display and knew important button interactions before beginning the first trial. At this point, subjects were treated as "Earth-independent", with no further questions answered about the training.

Each session consisted of 10 trials, and each participant underwent 4 sessions. For every trial, subjects were tasked with first selecting a landing site on Mars, then controlling/guiding their aircraft to the landing site, and finally descending onto the surface of Mars. After completing the trial, subjects were scored on their performance, skill, and number of crashes for each task. This data was collected and stored on MongoDB Compass.

Subjects were also asked for their subjective perceptions of the training conditions by participating in 3 well known surveys, which are the: 1) System Usability Scale (SUS) Survey, 2) Flow Short Scale Survey, and 3) Affect Grid Survey. Participants were requested to complete the SUS survey and the Flow survey once after every session, whereas the participants responded to the Affect Grid survey before and after every session. The SUS and Flow surveys were administered through Google Forms and the Affect Grid survey was administered via pencil-and-paper. For reference, the SUS survey and Flow survey are scored on a likert scale, while the Affect Grid survey uses a 9-point scale. For further reference on the 3 surveys, see Ref[1, 2, 5, 6, 13, 15, and 16].

All data in this study was collected in the Bioastronautics Labs in the Aerospace Engineering Department at the University of Colorado Boulder for purposes of Trinity team research. This experimental design was approved by the University of Colorado at Boulder Institutional Review Board under protocol #21-0349 and all subjects signed a written informed consent form.

2.2 Training Algorithms

By definition, a training algorithm is adaptive (A) if the level of difficulty to successfully complete a task (ls, de, mc) changes depending on the subject's skill and performance scores for a given trial. For instance, each EDL task has a difficulty staircase with 25 different levels of difficulty; a difficulty level of 1 means the subject encounters the easiest presentation of the task, while level 25 is the hardest difficulty level. An example of increasing difficulty level in the manual control task would be increasing wind speed as one pilots to their landing site. Conversely, a training algorithm is non-adaptive (NA) if the level of difficulty is fixed or static for every task and trial regardless of the subject's skill or performance score on a given trial.

2.2.1 Adaptive Training Algorithm

The adaptive training algorithm utilized in this paper takes on the independent 2-up-1-down adaptive staircase approach for each task. In this approach, 2 consecutive "excellent" performance scores on a given task results in a single step up in difficulty for that corresponding task. Additionally, any score of "poor" for a given task induces a step down in difficulty level for that respective task in the next trial. A score of "adequate" on a task kept the difficulty level the same.

A lockstep was also used in this algorithm to prevent unbalanced skill acquisition. In the lockstep, a task could only increase in difficulty if it was no more than 1 level of difficulty higher than the lowest difficulty task. For example, if the landing site selection task was at level 12, but the manual control and terminal descent engine tasks were both at level 10, the landing site selection task would remain at level 12, even if subjects received two consecutive excellent ratings for the task. Only when the manual control and terminal descent engine tasks had made it to level 12 as well could the landing site task increase to level 13.

If a subject received a "poor" rating for a task at level 1 (e.g., if they descended to "Level 0"), they rewatched the introduction video for that task. Level 18 was a skip level in training for the adaptive training algorithm to prevent any subject from having previous experience with the difficulty level as it is the difficulty level used in the physical mock-up (i.e. physical mock-up uses a fixed difficulty level of 18 for all tasks).

The training condition which trained with an adaptive training algorithm is the VR-A group.

2.2.2 Non-Adaptive Training Algorithm

The non-adaptive training conditions utilized in this paper implements a static staircase approach

for each task. Every subject trained with this training algorithm stays at the same difficulty level of 12 for every task, regardless of performance on those tasks. The training conditions which trained with the non-adaptive training algorithm are the VR-NA and 2D-NA groups.

2.3 Hardware Used During Testing

2.3.1 Hardware in Training

Subjects assigned to the VR training task environment completed their training using the HTC VIVE Pro VR headset. Subjects assigned to the 2D training task environment completed their training seated in front of a desktop monitor which displayed the same content (view of the cockpit) as the VR subjects.

For all subjects, the tasks were performed with a Logitech X52 Flight and Space Simulator joystick and hand thruster. Interactions with the joystick and hand thruster were identical for the VR and 2D training task environments with one exception: since VR subjects had the ability to lean forward in the spacecraft cockpit, 2D based subjects were given two unique buttons on the joystick to zoom into the flight display and landing site map; this replicated the behavior seen with the VR training task environment.

2.3.2 Hardware in the Physical Mock-Up

Concluding training, all subjects were evaluated in a physical mock-up of the cockpit in which they trained. This mock-up is called the ARES mock-up and is located at the University of Colorado Boulder. ARES is a composite shell of the HL-20 spacecraft cockpit, with multiple displays and interfaces, as well as out-the-window views as seen in Figure 1. As in training, all subjects performed the tasks in the physical mock-up using a Logitech X52 Flight and Space Simulator joystick and hand thruster.

Figure 1. ARES Physical Cockpit for evaluating skill and performance for each EDL task.

2.4 Scoring and Difficulty Control

As briefly mentioned above, for every trial, the EDL game simulator collects raw skill metrics about a subject's performance and assigns a performance summary value for each task: poor(-1), adequate (0), excellent (1). The raw metrics carry detailed information about proximity to optimal landing sites, alignment between the user's flight path, as well as the optimal one, and gentleness of landing touchdown; they also describe crash conditions such as landing on steep terrain or running out of fuel. However, the raw metrics are neither interpretive nor directly comparable as they have different orders of magnitude and different sensitivities. Additionally, the performance metric ignores the fact that achieving a skill of x at difficulty level y is not the same as achieving a skill of x at level y + 1. The original performance summaries offer interpretation of the raw skill metrics and a method for direct comparison between numerical values, but their coarseness prevents detection of small effects. Thus, to answer the research questions discussed in section 1, we sought a composite metric, denoted skill, that brings together the features of the raw metrics and binned performance scores without the drawbacks.

In summary, our constructed skill metric: 1) takes in a raw metric, 2) performs a linear transformation such that the error function of the ending result maps onto a threshold of poor being between $(0 \ - \ 0.0125)$ and excellent being between $(0.875 \ - \ 1)$, 3) applies an error function to compress extreme values to scale near ±1, and 4) implements difficulty scaling by taking the result from the error function and multiplying it by $(difficulty\ level \ / \ 25)$.

Important to understand is that, unlike the landing site selection and terminal descent engine tasks, the linear transformation performed for the manual control task depends on the difficulty level. For instance, different difficulty levels will yield different poor and excellent thresholds for the manual control task only; otherwise, poor performance are skills between $(0 - 0.0125)$ and excellent are scores between $(0.875 - 1)$. Note that when calculating the skill metric for each trial and task, we assign a failure (i.e. crash) a skill metric of 0.

Ultimately, we were able to specify a continuous skill metric that is comparable across tasks, clearly interpretive, sensitive to small changes in skill, and controls for difficulty. Consequently, the statistical analyses that follow will take advantage of both the performance metric and the newly constructed skill variable.

2.5 Performance and Skill Evaluation in the Physical Mock-up

As stated above, after training is completed, all subjects were evaluated in the physical mock-up of the cockpit in which they trained. Evaluation in the physical mock-up is considered session 4, and consists of 10 trials. All participants, regardless of their assigned training condition, completed the 10 trials at difficulty level 18. Likewise to training, subjects were scored on performance, skill, and the number of crashes for each task. Additionally, subjects were asked to complete the Affect Grid survey before and after the physical mock-up session, as well as, Flow and SUS surveys at the end of session 4.

2.6 Task Selection

The EDL environment was developed in Unity Game Engine, and it allows for the application of the three complex tasks for which subjects were scored: Landing Site Selection (ls), Manual Control (mc), and Terminal Descent engine (de). These tasks were chosen as they are complex, mission-critical tasks that most closely resemble the true tasks that crew are asked to perform on LDEMs. These simulated EDL tasks allow subjects to learn vital information on how to execute manual control inputs and monitor gauges/flight display information. Training of these types of tasks are usually conducted pre-mission in Earth-bound simulators or physical mock-ups, furthering these tasks as a perfect candidate for exploring the impacts of training conditions.

2.7 Survey Selection

Mentioned above, the SUS, Affect Grid, and Flow Short Scale surveys were administered to subjects in this study. The SUS survey was chosen as it is commonly used, easy to score, and allows us to uncover valuable information such as the usability of the different systems, where the systems are defined as the different training conditions (i.e. VR-A, VR-NA, and 2D-NA). The Affect Grid was implemented as it allows us to understand how a subject was feeling in a given moment. The grid measures how a participant is feeling in two ways: arousal and pleasure.

The 2 conflicting axes of arousal are sleepiness vs anxiety, whereas the 2 conflicting axes of pleasure are positive vs negative feelings. Finally, the Flow survey was utilized in this study as the survey measures flow during a particular activity (i.e. the 3 EDL tasks). Measuring flow allows us to investigate whether or not one training condition had significantly higher motivation, commitment, immersion, peak enjoyment, energetic focus, and creative concentration than the other training conditions when participating in the EDL tasks. Flow is measured formally according to seven attributes, which are flow experience, perceived task importance, demands, skills, perceived fit of demands and skills, fluency of performance, and absorption by activity.

2.8 Statistical Methods Used for the Statistical Analyses

As mentioned above, this paper seeks to understand the effects of the training conditions (focusing on the VR-A training condition) on EDL tasks by investigating the following research questions: 1) how well do subjects learn EDL tasks in training, as a function of their training condition?, 2) how did each training condition perform in the physical mock-up, what skill level did each training condition attain, and how does it compare across training conditions?, 3) how does the subjective perceptions of each training condition differ by training condition in training and in the physical mock-up?, and 4) for each training condition, how does the subjects' subjective perceptions of their assigned training condition change as training progresses? Given the complex nature of these research questions, they can be divided into 3 overarching categories of analyses, specifically: 1) learning as assessed in training, 2) skill and performance as assessed in the physical mock-up, and 3) subjective perceptions of a given training condition. The 'learning as assessed in training' category consists of a skill evaluation analysis using all 30 training trials. The 'skill and performance as assessed in the physical mock-up' category includes the: achieved integrated skill in 1st trial (or technically the 31st trial), achieved mean integrated skill in all 10 trials, all excellent trials count, and integrated poor performance count analyses. Finally, the 'subjective perceptions of training conditions' category consists of six analyses; these include: 1) system usability in training, 2) system usability in the physical mock-up, 3) change in affect during training, 4) change in affect in the physical mock-up, 5) flow in training, and 6) flow in the physical mock-up. Recall that 'affect' is measured according to two attributes, which are arousal and pleasure. Additionally, recall that 'flow' is measured according to seven attributes, including flow experience, perceived task importance, demands, skills, perceived fit of demands and skills, fluency of performance, and absorption by activity; however, we limit our analysis to only 2 dimensions of flow which are flow experience and perceived task importance. For clarity, an overview of the analyses and categories discussed can be seen below in Figure 2.

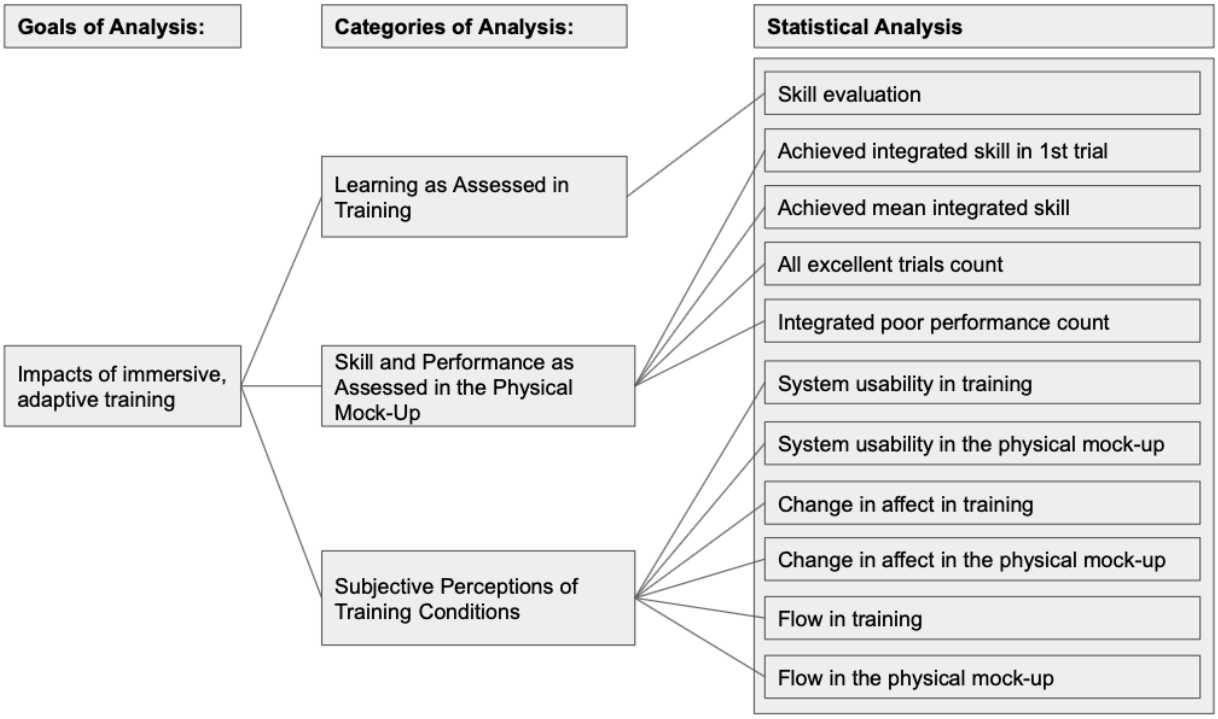

Figure 2. High level overview of the 3 categories of analyses and the various statistical analyses performed under those categories to understand the impacts of the 3 different training conditions, with specific focus on the adaptive, immersive training condition (VR-A).

To perform the statistical analyses listed above, we utilize 1 of the following 3 statistical methods: Nonlinear Mixed Effects modeling, Kruskal-Wallis hypothesis testing, and/or Mixed-Anova. In other words, we utilize one of the methods listed above to perform each analysis. The analysis which applied Nonlinear Mixed Effects modeling is skill evaluation in training, while the analyses that applied Mixed-Anova are: system usability in training, change in affect in training, and flow in training. Consequently, the analyses which utilize Kruskal-Wallis are achieved mean integrated skill, achieved skill in the 1st trial, all excellent trials count, integrated poor performance count, system usability, change in affect, and flow, *all in the physical mock-up*. All hypothesis tests were conducted with a global alpha level set to 0.05 (i.e., $\alpha = 0.05$).

Nonlinear Mixed Effects modeling was used as it allows the explicit modeling of random effects on a hierarchy of groups. We are able to estimate the natural variation between subjects, and then also the variation induced by the subjects' membership to a training condition in training. By analyzing the impact of these random effects on the parameters of our models, we can draw conclusions about the effect of each training condition in training. For this paper, we applied a multilevel generalized logistic growth curve as the nonlinear model; our multilevel model varies its asymptote and scale coefficients by both subject and training condition, while fixing the xmid (inflection point) parameter at 0.112 to help with convergence issues. Additionally, the model

constructed and used in the paper has “progress in training” as the predictor variable. The progress in training variable can be described as the percentage of training completed (which is the given trial number divided by 30). Having data that resembles learning curves (period of rapid learning and then a plateau), we believe that the generalized logistic growth curve mixed effects model, which was originally developed for growth or learning, is a plausible model to use. Additionally, we implied a multilevel model as it is effective in dealing with repeated measures and high variation among subjects. For more information on the generalized logistic growth curve and Nonlinear Mixed Effects modeling see Ref[4, 7, 11, and 14].

In regards to Mixed-Anova, this statistical method was exploited as it allows us to incorporate both within-subject factors (within-group) and between-subject factors (across groups). In other words, Mixed-Anova allows us to understand how continuous or interval variables of interest (for example, skill or usability of the system) change across training sessions for a given training condition, as well as allows us to determine overall differences among training conditions on average. Note that it is common practice to consider a discrete variable with more than 6 levels continuous. Mixed-Anova is beneficial because it acknowledges that different subjects train with different training conditions. Although a parametric test, Mixed-ANOVA is robust to moderate violations of its general statistical assumptions. For more information on Mixed-Anova see Ref[9 and 12].

Finally, Kruskal-Wallis testing was administered as it permits the investigation of differences in training conditions while making no assumptions about the distribution of the data. Unlike Anova testing, Kruskal-Wallis determines whether different training conditions have different median values for a variable of interest (dependent variable). Kruskal-Wallis was an appropriate choice for most analyses involving physical mock-up data as Kruakal-Wallis allows for ordinal dependent variables (i.e. counts with a hierarchy to them) as well. In cases where One-Way Anova was possible, Kruskal-Wallis was applied instead due to harsh violations of One-Way Anova assumptions. For more information on Kruskal-Wallis see Ref[8].

For every analysis run, all assumptions of the corresponding statistical method used were checked. The assumptions of Mixed Effects Anova are normality of random effects, continuous random variable, homoscedasticity of variance, orthogonality, and independent data, while the assumptions of Mixed-Effects Anova are no significant outliers, homogeneity, normality, homogeneity of variance, sphericity of variance, and homogeneity of covariances. The minimal assumptions of the Kruskal Wallis test are independent data and having ordinal/continuous dependent variables. Note that no assumptions were substantially violated. See my code in Ref[17] for further reference to assumption checking and well as Ref[3 and 10].

Following significant results, pairwise t-tests or Dunn’s testing was performed. Dunn’s test was used when the normality assumption did not hold (i.e. analyses using Kruskal Wallis), whereas

pairwise t-testing was used when there was minimal violation of the normality assumption (i.e. analyses using Mixed Anova and Mixed Effects modeling).

Note also, to perform the statistical analyses involving skill above and to best answer our research questions, skill level was integrated across the 3 EDL tasks. For transparency, integrated skill level can be defined as the summed skill level for all 3 tasks for a given trial and subject. Ultimately, we want to employ a training condition that trains subjects or astronauts to have a high skill level on all 3 tasks as opposed to just 1 or 2 of the tasks. Thus, our analyses consider integrated skill level only for analyses involving skill (i.e. our dependent variable). Important to recognize is that integrating skill level is helpful in being concise in the number of analyses we run (i.e. we do not run one test for each task).

One last remark: recall that the Affect Grid survey is administered to subjects before and after every session. To interpret the results of the analyses that use this survey data most easily, delta scores were calculated. Definitively, a ‘change in pleasure/arousal’ score or a ‘delta pleasure/arousal’ score is the change in pleasure or arousal from before the session starts to when it's completed (i.e. post pleasure or arousal score minus the pre pleasure or arousal score).

All analyses were performed using R Studio version 2022.12.0+353 and Python version 2.8.2.

3 Results

3.1 Results for Learning as Assessed in Training

3.1.1 Skill Evaluation: Integrated Skill vs Progress in Training

Modeling integrated skill as a function of progress in training using a varying asymptote and scale generalized logistic growth curve mixed effects model resulted in Figure 3. A summary output of this model is seen in Table 1.

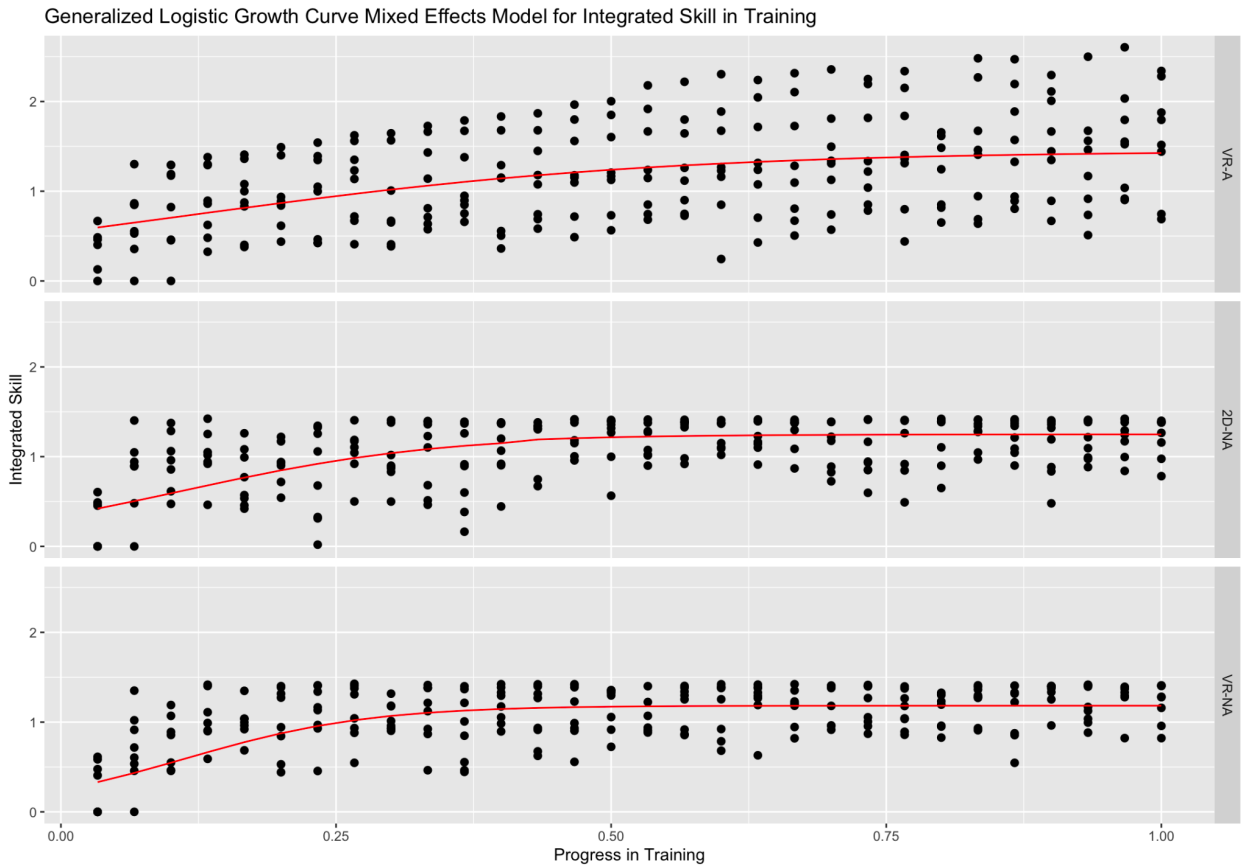

Figure 3. Generalized logistic growth curve mixed effects model (red curve) at the training condition level, predicting integrated skill as a function of progress in training for each of the 3 training conditions. Recall the model varies its asymptote and scale parameters by subject and training condition. One datapoint represents the integrated skill achieved for a given subject during a specific trial in training; each subject is assigned to a training condition. Integrated skill ranges from 0-3.

VR-A	**VR-NA**	**2D-NA**
Coefficient Estimates	*Coefficient Estimates*	*Coefficient Estimates*
Asymptote $\hat{A}_{VR-A} = 1.426$	Asymptote $\hat{A}_{VR-NA} = 1.183$	Asymptote $\hat{A}_{2D-NA} = 1.248$
Scale $\hat{S}_{VR-A} = 0.079$	Scale $\hat{S}_{VR-NA} = 0.023$	Scale $\hat{S}_{2D-NA} = 0.056$

Table 1. Estimated coefficients (asymptote and scale) at the training condition level using the generalized logistic growth curve mixed effects model. The asymptote and scale parameters are denoted as $\hat{A}_J$ and $\hat{S}_J$, respectively, and J are the different training conditions.

Note that at the training condition level, the standard deviation of $\hat{A}_J$, denoted $\hat{\sigma}_{A,J}$, is 0.122, whereas the standard deviation of $\hat{S}_J$, denoted $\hat{\sigma}_{S,J}$, is 0.062.

Figure 3 and Table 1 shows that the VR-A training condition reaches a higher integrated skill then the other two training conditions (higher asymptote estimate) by the end of training. Conversely, we notice that the VR-A training condition has the slowest rate of learning (i.e.

largest scale parameter - learning rate is the inverse of the scale parameter), while the VR-NA training condition has the fastest rate of learning. Moreover, relative to the scaling metrics, we notice a comparatively large standard deviation in the asymptote and scale parameter estimations across training conditions (looking at $\hat{\sigma}_{A,J}$ and $\hat{\sigma}_{S,J}$). In other words, when comparing the 3 asymptote parameter estimations (for example) we get a large standard deviation in estimated values, indicating good separation across training conditions.

3.2 Results for Skill and Performance as Assessed in the Physical Mock-Up

3.2.1 Achieved Integrated Skill in the 1st Trial of the Physical Mock-Up

The distribution of the achieved integrated skill for the 1st trial in the physical mock up for each training condition is depicted in Figure 4. Running Kruskal Wallis on this data, we found no significant difference in median achieved integrated skill for the 1st trial across training conditions ($p = 0.887$).

Figure 4. Distribution of achieved integrated skill for the 1st trial in the physical mock-up by training condition (colors). One datapoint represents the integrated skill achieved for a given subject during the 1st trial of the physical mock-up. The white triangles represent the average achieved skill for the 1st trial among subjects, partitioned by training condition.

Despite being insignificant, the VR-A training condition had the highest mean and median achieved integrated skill for the 1st trial in the mock-up, while the 2D-NA training condition had the lowest.

Notice that there are only 7 data points for the VR-A training condition; subject 27 crashed on the manual control task during the 1st trial of the physical mock-up, resulting in a NaN value for integrated skill. The mc task is a uniquely nested task in which landing site selection and terminal descent engine scores for skill and performance are not recorded when the subject crashes. In spite of this, subject 27 is dropped from this specific analysis.

3.2.2 Achieved Mean Integrated Skill in the Physical Mock-Up

The distribution of achieved mean integrated skill for a given subject across all 10 trials in the physical mock-up is seen in Figure 5. A Kruskal Wallis test was run on this data, and results found no significant differences in the median value of achieved mean integrated skill across training conditions in the physical mock-up ($p = 0.928$).

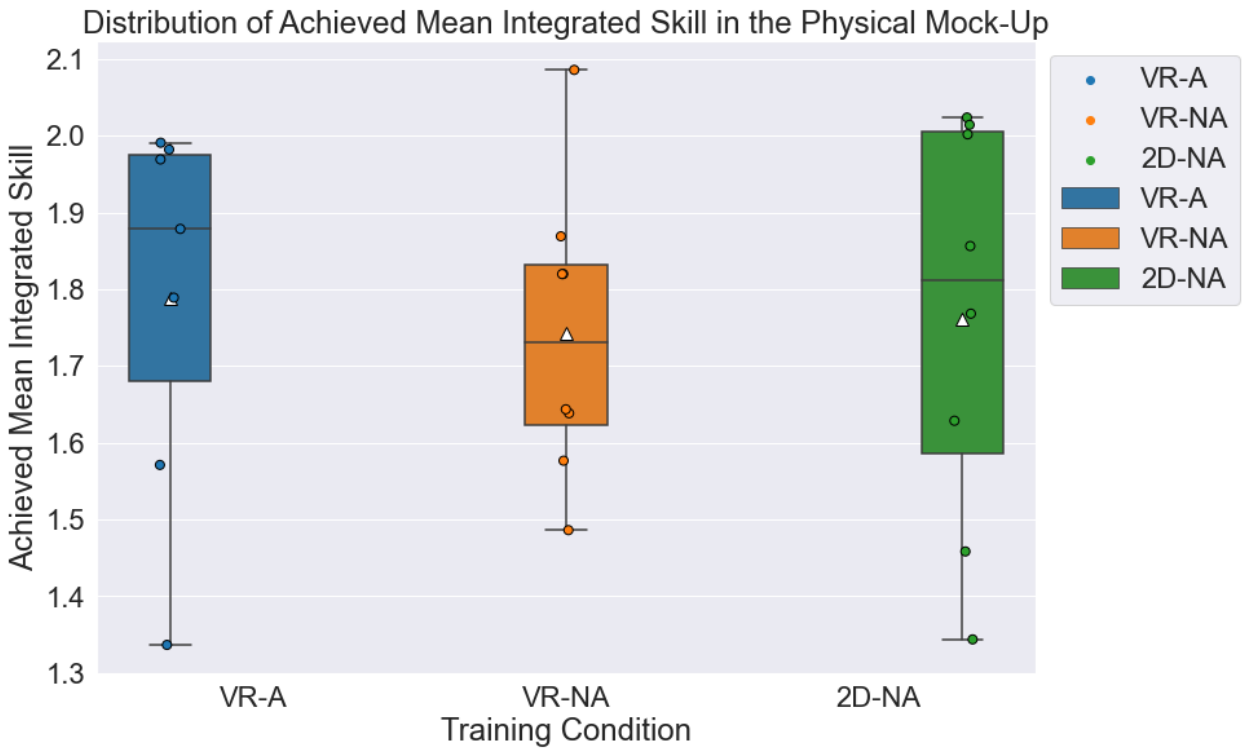

Figure 5. Distribution of achieved mean integrated skill for a given subject across all 10 trials in the physical mock-up by training condition (colors). One datapoint represents the integrated mean skill achieved for a given subject in the physical mock-up (i.e. averaged across all 10 trials). The white triangles represent the average achieved mean integrated skill among subjects, partitioned by training condition.

From Figure 5, we can see that all 3 training conditions have nearly the same mean value of achieved mean integrated skill level in physical mock-up. However, the VR-A training condition has a noticeably higher median value of achieved mean integrated skill across all 10 trials in the physical mock-up. Again, these results are insignificant.

Note that subject 27 is dropped from this analysis due to the fact that subject 27 crashed on the manual control task during the 1st and 2nd trials of the physical mock-up. This crash resulted in a NaN value for integrated skill during trials 1 and 2 for subject 27, and thus an overall achieved mean integrated skill of NaN in the physical mock-up.

Additionally, note that we averaged integrated skill levels for a given subject across all 10 trials to avoid the complexity of Mixed Effects hypothesis testing. Our procedure using Kruskal-Wallis accounts for repeated measures, as well as avoids making unnecessary assumptions about whether trial is continuous or categorical.

3.2.3 All Excellent Trials Count in the Physical Mock-Up

The count of all excellent trials in the physical mock-up partitioned by training condition is illustrated in Figure 6. A Kruskal Wallis test was performed on this data, which resulted in no significant differences in the median number of all excellent trials across training conditions in the physical mock-up ($p = 0.778$).

Figure 6. Distribution of all excellent trials count in the physical mock-up by training condition (colors). One datapoint represents the summed number of trials in which a given subject received an “excellent” rating on all 3 tasks. The white triangles represent the average number of all excellent trials among subjects in a given training condition.

Figure 6 depicts that the 2D-NA training condition and the VR-A training condition have identical distributions of all excellent trials counts. The VR-NA training condition has the highest mean and median values of all excellent trials count in the physical mock-up. Again, these results are not significant.

3.2.4 Integrated Poor Performance Count in the Physical Mock-Up

The distribution of integrated poor performance count across all 10 trials in the physical mock-up and across all 3 EDL tasks is expressed in Figure 7. Implementing a Kruskal Wallis test on this data, we found no significant difference in the median number of integrated poor performances across the 3 training conditions in the physical mock-up ($p = 0.841$).

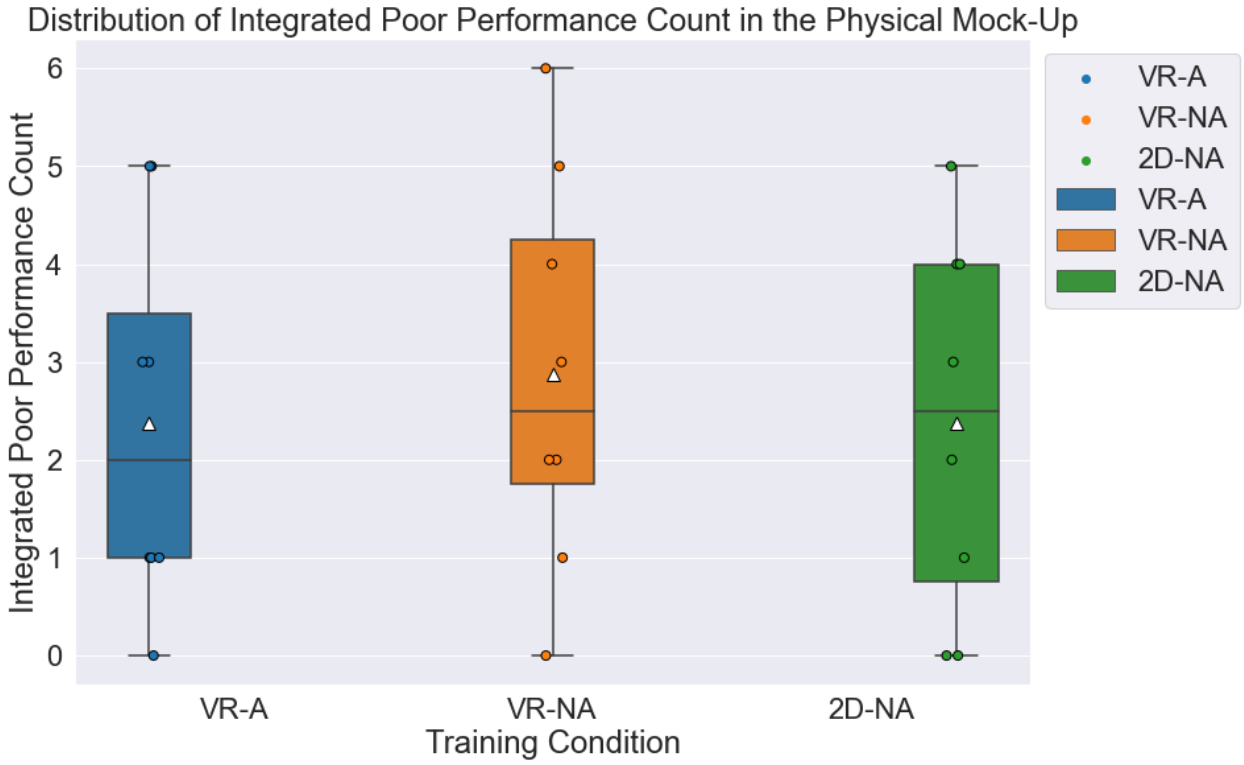

Figure 7. Distribution of integrated poor performance count in the physical mock-up by training condition (colors). One datapoint represents the summed number of poor performances across all 3 tasks and all 10 trials for a given subject in the physical mock-up. The white triangles represent the average number of integrated poor performances among subjects in a given training condition.

Figure 7 depicts that the 2D-NA training condition and the VR-A training condition have nearly the same mean value of integrated poor performance counts in the physical mock-up. Although,

the VR-A training condition noticeably has the lowest median number of integrated poor performance counts in the physical mock-up. Recall we want a lower integrated poor performance count. Again, these median values are not significant.

3.3 Subjective Perceptions of the Training Conditions

3.3.1 System Usability in Training

The spread of SUS scores across the 3 training sessions for each training condition is seen in Figure 8. A Mixed-Anova test was performed with the factors being training condition, session, and the interaction between the training condition and session. The dependent variable is system usability. No significant results were found for the training condition factor ($p = 0.379$); although, there were significant results for the trial factor ($p = 0.000$) and the interaction factor ($p = 0.044$). Note that the session factor is not meaningful on its own in our analyses as it collapses the training conditions into one group (i.e. ignores the training condition factor) to determine differences in SUS scores for each session on average. Given the interaction is significant, a pairwise t-test was performed with a Bonferoni correction (having checked assumptions already). The post hoc test revealed: 1) the SUS score in session 1 is significantly different from the SUS score in session 2 for the VR-A training condition ($p = 0.001$) on average, 2) the SUS score in session 1 is significantly different from the SUS score in session 3 for the VR-A training condition ($p = 0.027$) on average, and 3) the SUS score in session 1 is significantly different from the SUS score in session 3 for the 2D-NA training condition ($p = 0.033$) on average. These significant results are illustrated in figure 9 below.

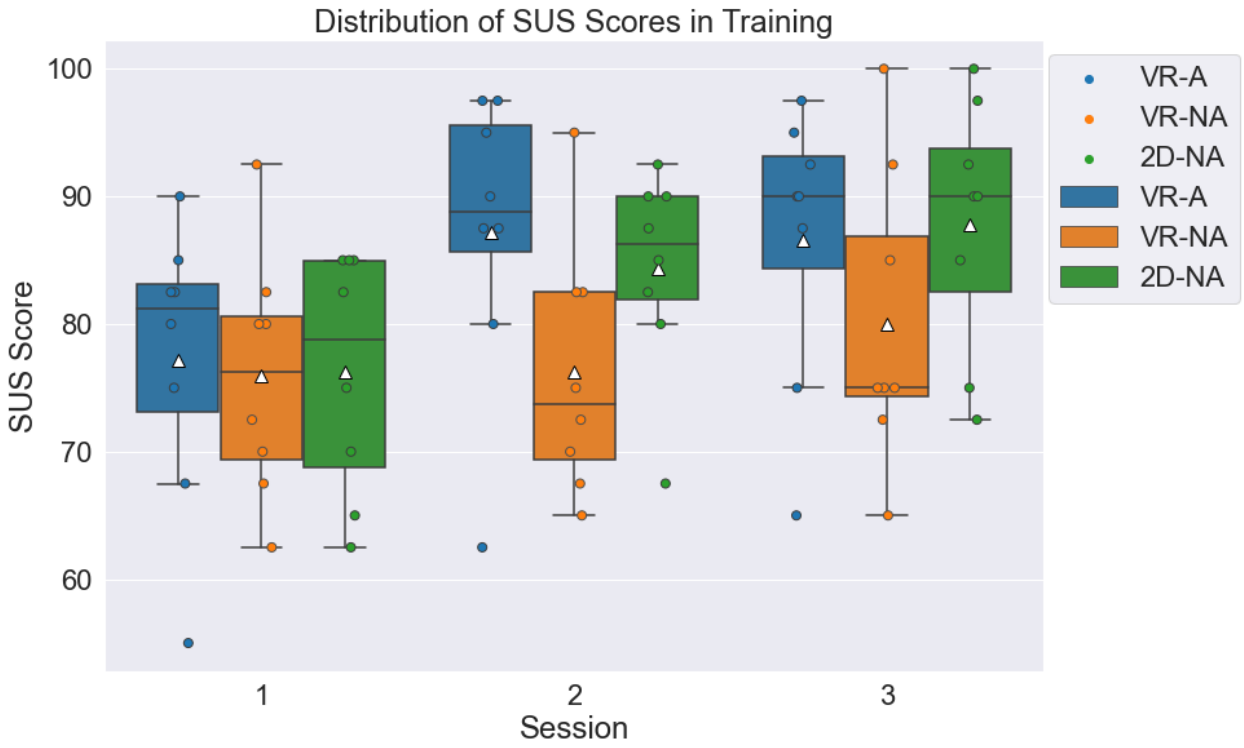

Figure 8. Distribution of SUS scores for each session in training, partitioned by training condition (colors). One datapoint represents the SUS score for a given subject in a given session; each subject is assigned to a specific training condition. The white triangles represent the average SUS score among subjects in a given training condition and session.

From Figure 8 we also observe that the VR-A training condition rates their training condition the most usable in sessions 1 and 2 when compared to the other training conditions (median and mean). The VR-NA training condition consistently rates their training condition the least usable in every training session. Finally, there is a large increase in subjective usability of a training condition from session 1 to session 3 for the VR-A and 2D-NA training conditions.

3.3.2 System Usability in the Physical Mock-Up

The dispersion of SUS scores in the physical mock-up for each training condition is seen in Figure 9. A Kruskal-Wallis test run on this data showed no significant difference in median SUS scores across the training conditions ($p = 0.480$).

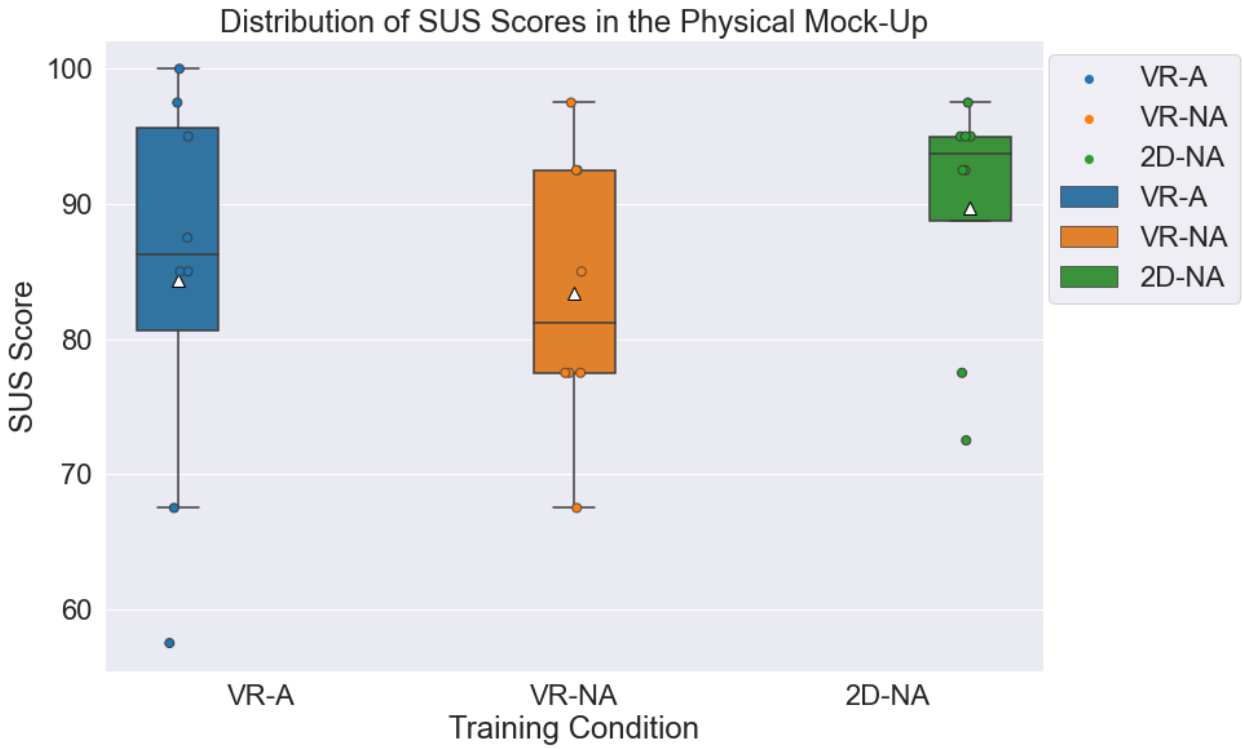

Figure 9. Distribution of SUS scores partitioned by training condition (colors) in the physical mock-up. One datapoint represents the SUS score for a given subject in the physical mock-up; each subject is assigned to a specific training condition. The white triangles represent the average SUS score among subjects in a given training condition.

Figure 9 depicts that the 2D-NA training condition has the highest mean and median SUS score in the physical mock-up, while the VR-NA training condition has the lowest. These differences in medians were not significant.

3.3.3 Change in Affect in Training

Recall that affect is measured according to 2 attributes: pleasure and arousal.

3.3.3.1 Change in Arousal in Training

The distribution of perceived change in arousal scores (or delta arousal scores) across the 3 training sessions for each of the 3 training conditions is seen in Figure 10. A Mixed-Anova test was run on this data, with the factors being training condition, session, and the interaction between training condition and session. The dependent variable is delta arousal scores. No significant results were found for the training condition factor ($p = 0.632$), session factor ($p = 0.983$), or interaction factor ($p = 0.513$).

Figure 10. Distribution of delta arousal scores for each session in training, partitioned by training condition (colors). One datapoint represents the delta arousal score for a given subject in a given session; each subject is assigned to a specific training condition. The white triangles represent the average delta arousal score among subjects in a given training condition and session.

Note that negative delta arousal scores indicate arousal went down during that session of training, whereas positive delta arousal scores indicate arousal went up during that session of training. We can notice that the mean and median delta arousal scores for the VR-A training condition are positive for all 3 sessions. More specifically, out of all 3 sessions, only one individual from the VR-A training condition group had a negative delta arousal score; this isn't true for the other training conditions.

Note that there are only 7 data points for the 2D-NA training condition on session 2. No arousal score was collected post session 2 for subject 3, resulting in a NaN value for delta arousal.

3.3.3.2 Change in Pleasure in Training

The distribution of perceived change in pleasure scores (or delta pleasure scores) for the 3 training sessions, partitioned by each of the 3 training conditions is seen in Figure 11. Mixed-Anova was performed, with the factors being training condition, session, and the interaction between training condition and session. The dependent variable is delta pleasure scores. This Mixed-Anova test revealed no significant results for the training condition factor ($p = 0.941$), session factor ($p = 0.407$), or interaction factor ($p = 0.839$).

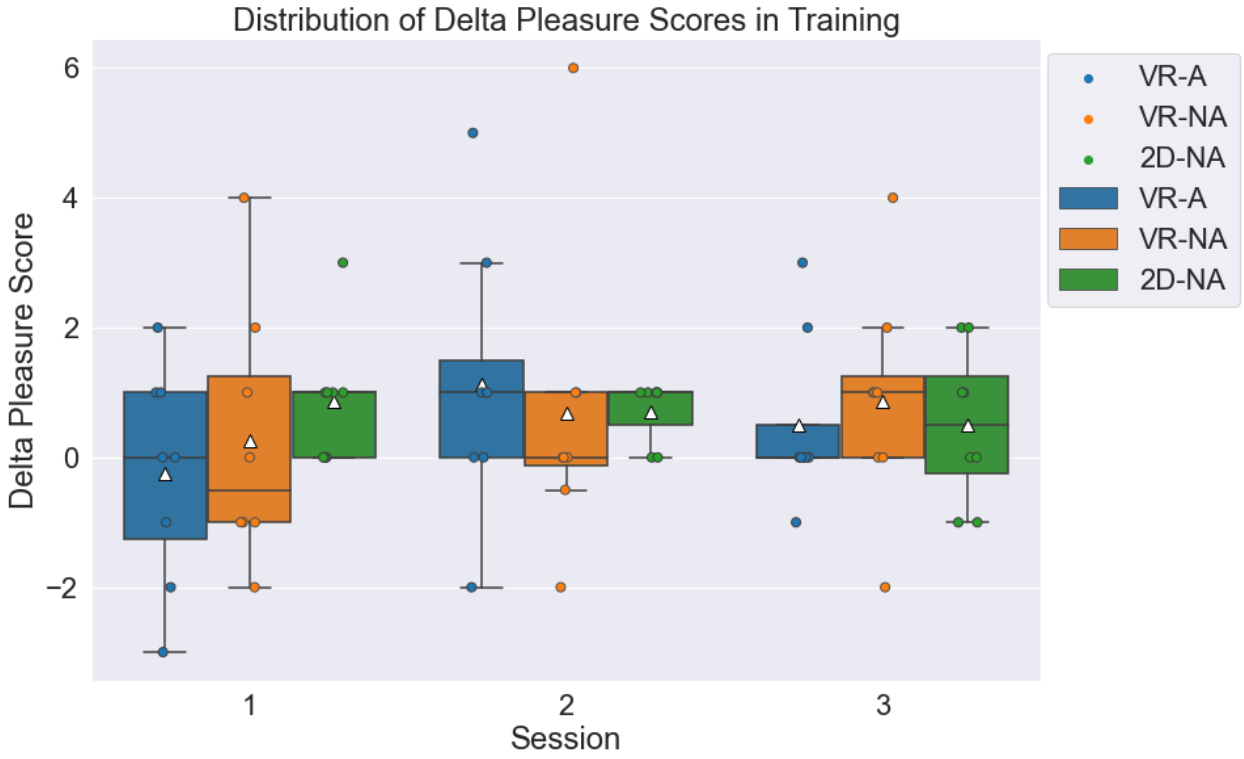

Figure 11. Distribution of delta pleasure scores for each session in training, partitioned by training condition. One datapoint represents the delta pleasure score for a given subject in a given session; each subject is assigned to a specific training condition. The white triangles represent the average delta pleasure score among subjects in a given training condition and session.

Notice in Figure 11 that the standard deviation of delta pleasure responses decrease as training progresses for the VR-A training condition.

Note that there are only 7 data points for the 2D-NA training condition on session 2. No pleasure score was collected post session 2 for subject 3, resulting in a NaN value for delta pleasure.

3.3.4 Change in Affect in the Physical Mock-up

As said above, affect is measured according to 2 attributes: pleasure and arousal.

3.3.4.1 Change in Arousal in the Physical Mock-Up

The dispersion of delta arousal scores in the physical mock-up for each training condition is seen in Figure 12. A Kruskal Wallis test on this data determined no significant difference in median delta arousal scores across the training conditions in the physical mock-up ($p = 0.311$).

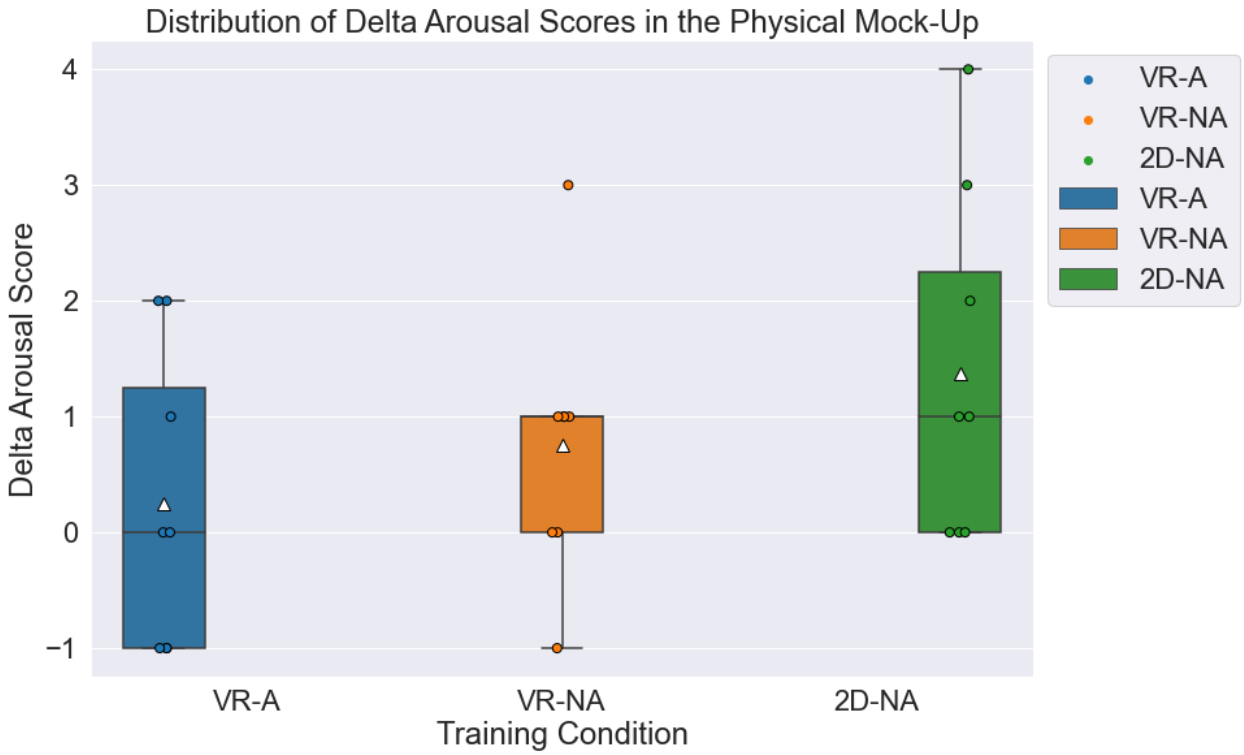

Figure 12. Distribution of delta arousal scores partitioned by training condition (colors) in the physical mock-up. One datapoint represents the delta arousal score for a given subject in the physical mock-up; each subject is assigned to a specific training condition. The white triangles represent the average delta arousal score among subjects in a given training condition.

Looking at Figure 12, we see that the 2D-NA training condition has the highest delta arousal score in the physical mock-up on average. Conversely, the VR-A training condition has the lowest delta arousal score in the physical mock-up on average. Also note that the 2D-NA training condition had the highest standard deviation (skewed upward) in delta pleasure responses in the physical mock-up. A delta arousal score of 0 in the physical mock-up (i.e. “true LDEM scenario”) is ideal, see discussion section below. Again, these results are not significant.

3.3.4.2 Change in Pleasure in the Physical Mock-Up

The dispersion of delta pleasure scores in the physical mock-up for each training condition is seen in Figure 13. A Kruskal Wallis test implemented on this data resulted in no significant difference in median delta pleasure scores across the training conditions in the physical mock-up ($p = 0.932$).

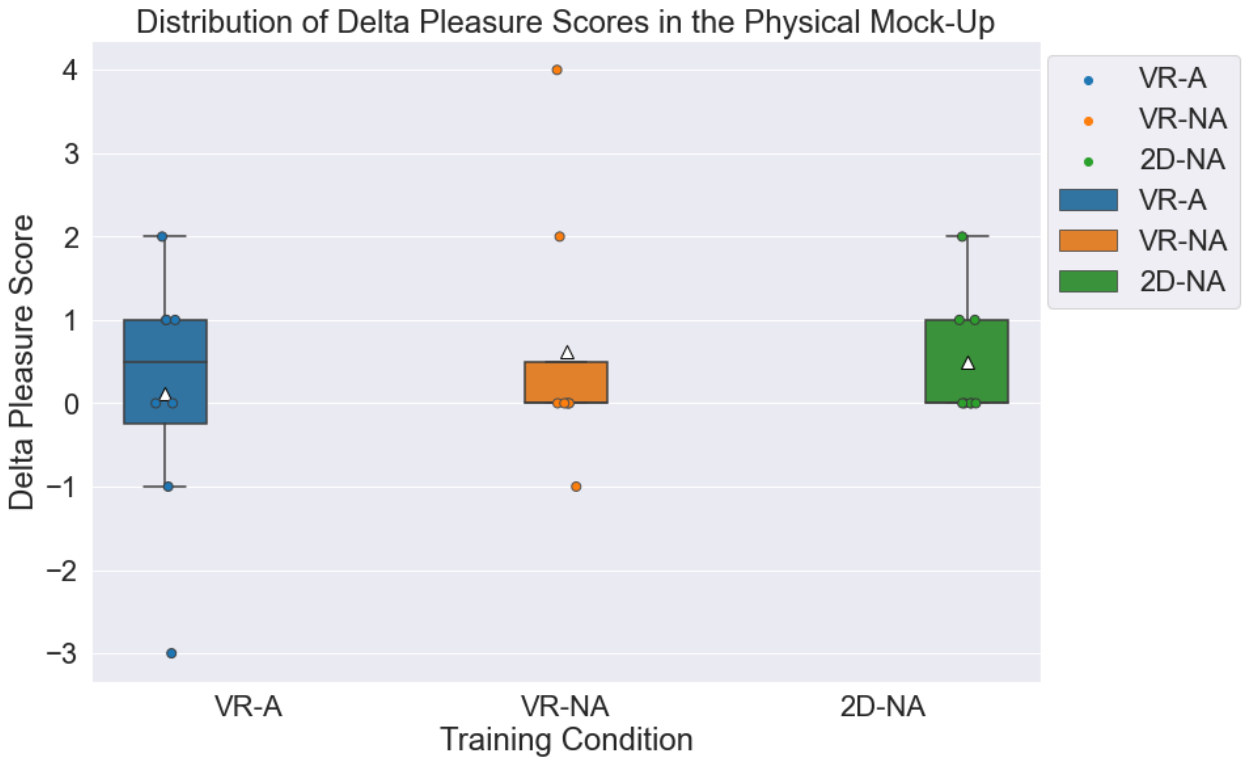

Figure 13. Distribution of delta pleasure scores partitioned by training condition (colors) in the physical mock-up. One datapoint represents the delta pleasure score for a given subject in the physical mock-up; each subject is assigned to a specific training condition. The white triangles represent the average delta pleasure score among subjects in a given training condition.

The VR-A training condition does not outperform the opposing training conditions in terms of higher changes in pleasure in the physical mock-up. The VR-A training condition has the lowest mean delta pleasure score, but each median delta value (one for each training condition) is the same. These differences in medians are clearly not significant.

3.3.5 Flow during Training

3.3.5.1 Flow Experience in Training

The distribution of flow experience scores across the 3 training sessions for each of the 3 training conditions is seen in Figure 14. A Mixed-Anova test was run on this data, with the factors being training condition, session, and the interaction between training condition and session. The dependent variable is flow experience. No significant results were found for the training condition factor ($p = 0.962$) or the interaction factor ($p = 0.087$). However, notice that the interaction factor is nearly significant. Despite not being meaningful on its own, significant results were produced for the session factor ($p = 0.000$).

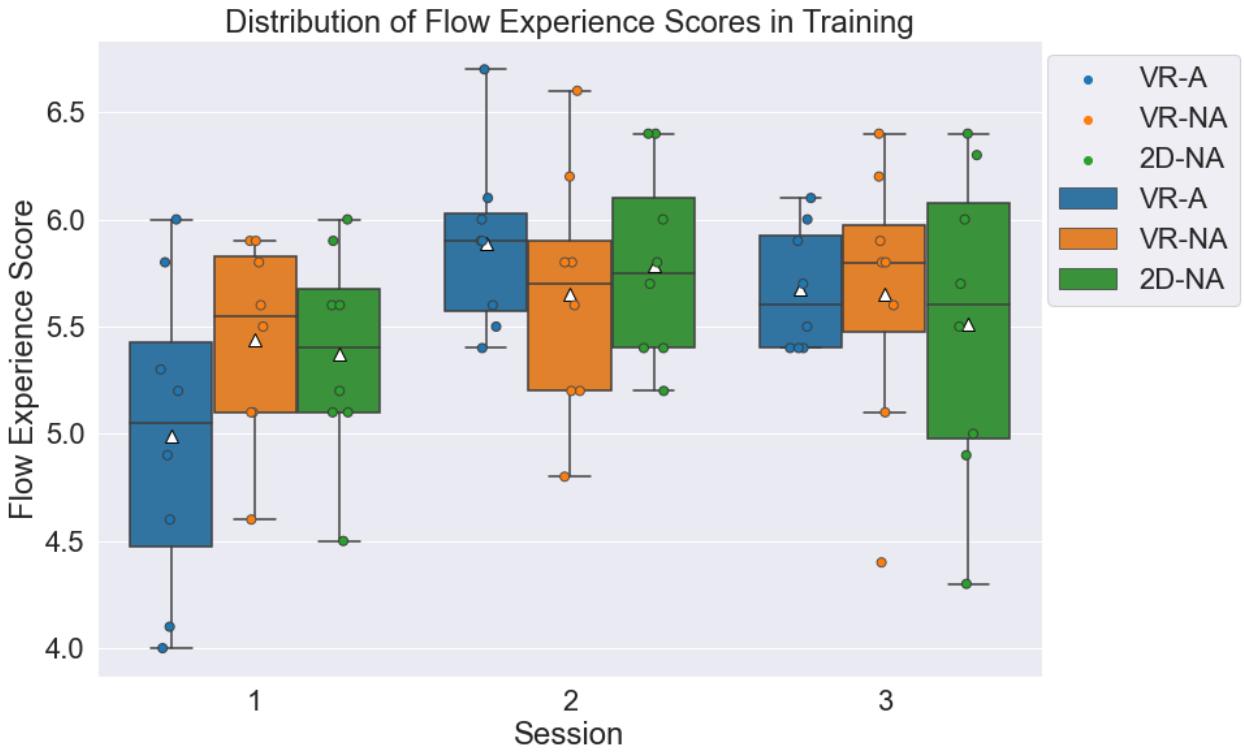

Figure 14. Distribution of flow experience scores for each session in training, partitioned by training condition (colors). One datapoint represents the flow experience score for a given subject in a given session; each subject is assigned to a specific training condition. The white triangles represent the average flow experience score among subjects in a given training condition and session.

Figure 14 shows that the VR-A training condition had a relatively large increase in flow experience from session 1 to 2 on average. Additionally, unlike the other training conditions, the VR-A training conditions had a sizable increase in flow experience from the beginning of training to the end of training. These results are nearly significant.

3.3.5.2 Perceived Task Importance in Training

The distribution of perceived task importance scores across the 3 training sessions for each of the 3 training conditions is seen in Figure 15. A Mixed-Anova test was run on this data, with the factors being training condition, session, and the interaction between training condition and session. The dependent variable is perceived task importance scores. No significant results were found for the training condition factor ($p = 0.099$), the interaction factor ($p = 0.955$), or the session factor ($p = 0.489$). Although, the interaction factor was nearly significant.

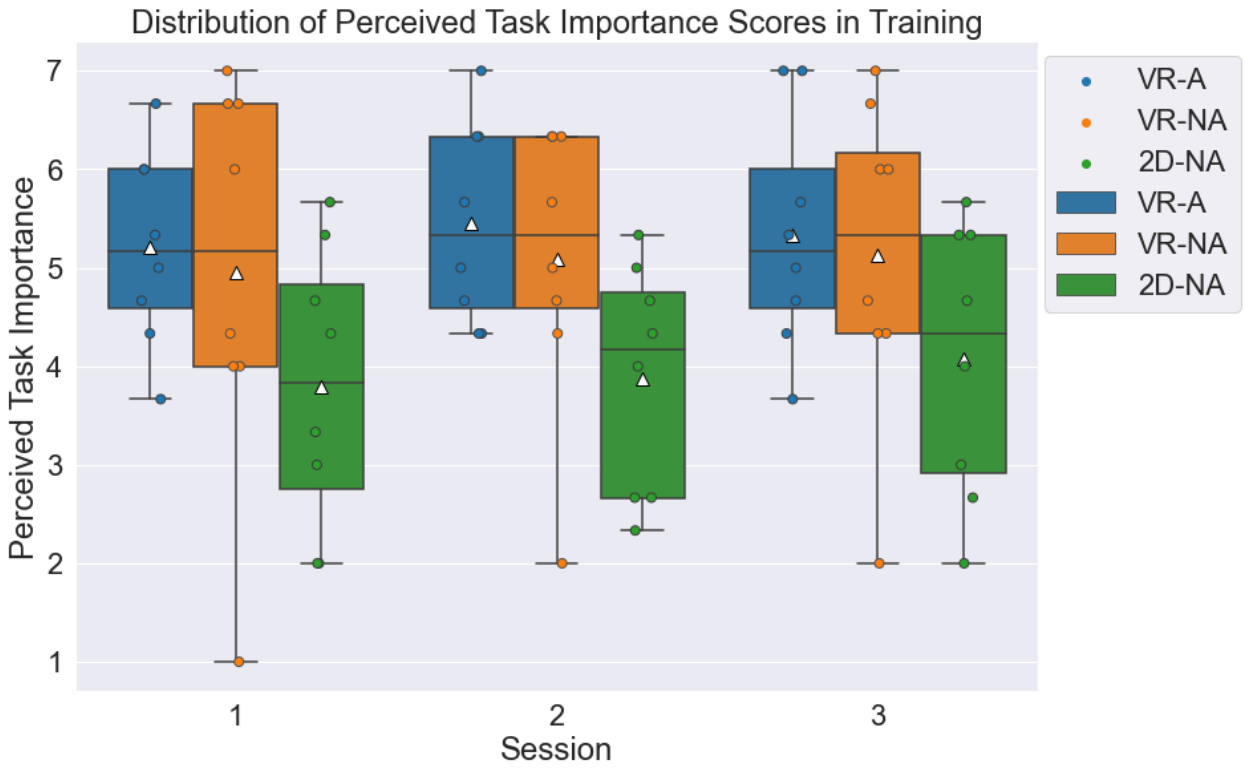

Figure 15. Distribution of perceived task importance scores for each session in training, partitioned by training condition (colors). One datapoint represents the perceived task importance score for a given subject in a given session; each subject is assigned to a specific training condition. The white triangles represent the average perceived task importance score among subjects in a given training condition and session.

Considering Figure 15, we see that the VR-A training condition had the highest perceived importance of the EDL tasks during all 3 sessions in training *on average*; conversely, the 2D-NA training condition had the lowest perceived importance of the EDL tasks for all 3 sessions in training *on average (and in terms of median values)*.

It appears that the training task environment (i.e. VR vs screens) plays a meaningful role in perceived task importance in training given the VR-NA training condition has high perceived task importance too; the VR-A and VR-NA training conditions have nearly the same median values of perceived task importance for all sessions in training. Again, results are not significantly different here, but absolutely interesting.

3.3.6 Flow in the Physical Mock-Up

3.3.6.1 Flow Experience in the Physical Mock-Up

The spread of flow experience scores in the physical mock-up for each training condition is seen in Figure 17. A Kruskal Wallis test on this data determined no significant difference in median flow experience scores across the training conditions in the physical mock-up ($p = 0.469$).

Figure 17. Distribution of flow experience scores partitioned by training condition (colors) in the physical mock-up. One datapoint represents the flow experience score for a given subject in the physical mock-up; each subject is assigned to a specific training condition. The white triangles represent the average flow experience score among subjects in a given training condition.

The VR-A training condition does not outperform the opposing training conditions in terms of flow experience scores in the physical mock-up. The 2D-NA training condition has the highest mean and median flow experience score in the physical mock-up. These differences in medians are not significant.

3.3.6.2 Perceived Task Importance in the Physical Mock-Up

The spread of perceived task importance scores in the physical mock-up for each training condition is seen in Figure 18. A Kruskal Wallis test run on this data determined no significant differences in median perceived importance of the EDL tasks across the training conditions in the physical mock-up ($p = 0.082$). However, note that the result was almost significant.

Figure 18. Distribution of perceived task importance scores partitioned by training condition (colors) in the physical mock-up. One datapoint represents the perceived task importance score for a given subject in the physical mock-up; each subject is assigned to a specific training condition. The white triangles represent the average perceived task importance score among subjects in a given training condition.

Although not significant (but nearly significant), looking at Figure 18, we can notice that the 2D-NA group has the lowest perceived task importance in the physical mock-up on average. The VR-NA training condition has the highest mean and median perceived task importance score in the physical mock-up, but not by much when compared to the VR-A training condition. Also we can recognize that the VR-A training condition has the smallest standard deviation in perceived task importance responses in the physical mock-up.

4 Discussion

Finally, let's interpret the boatload of results and consider them in a bigger picture setting.

4.1 Learning as Assessed in Training

Assuming the generalized logistic growth curve mixed effects model is correct, there is a noticeable effect of the training condition on integrated skill acquisition during training. Within the 30 training trials, the VR-A training condition was able to reach and maintain a higher integrated skill level then that of the VR-NA and 2D-NA training conditions. Thus, astronauts

trained with the VR-A training condition are associated with being more skillful when picking a landing site, piloting to a landing site, and touching down on the surface than opposing subjects by the time training is completed. It is critical in LDEMs that an astronaut is trained to the highest skill level before the mission begins.

However, not only is it consequential that astronauts are trained to the highest skill level by the end of training, it is also imperative that astronauts learn at the quickest rate in training to ensure a longer period of mastery. Put differently, learning at a faster rate saves more time for astronauts to master the tasks and skills already acquired; this is likely to result in less complications in the future. Surprisingly, the VR-NA training condition had the quickest rate of learning during training, but unsurprisingly, the VR-A training condition had the slowest. Given the number of difficulty levels (levels = 25) that the VR-A training condition experiences and the minimal amount of training, it isn't a perplexity that the VR-A training condition learns integrated skills at a slower rate. In fact, with an adaptive algorithm (i.e. independent 2-up-1-down locked), individuals can be locked at a low or high difficulty level, causing one's skill level to stay stagnant for longer periods of time. Conversely, non-adaptive training conditions are fixed at the same difficulty level, imposing consistency in task difficulty, allowing integrated skill level to increase at a faster rate; results likely reflect this. Additionally, it is highly plausible that the VR-A has the smallest slope as the VR-NA and 2D-NA training conditions never reach integrated skills above 1.5. Having to reach an integrated skill level 2.5, by default, takes longer than reaching a skill level of 1.5.

Finally, likewise with all human-subject experiments, subjects showed a high degree of variability in their integrated skill data. As well, there was a high degree of variability among the asymptote and scale parameter estimates at the training condition level. High subject variability, paired with high variability among parameter estimates at the training condition level, is strong evidence indicating a Mixed Effects Model was appropriate.

4.2 Skill and Performance in the Physical Mock-Up

In the event that an astronaut is given one chance to perform mission critical EDL tasks, the VR-A training condition has the highest integrated achieved skill. This means that astronauts trained with the VR-A training condition will have more successful missions by picking landing site selections closest to scientific sites of interest, having more accurate and precise piloting skills during flight, and by descending onto the surface safely.

Similarly, in the event that an astronaut is given 10 chances to perform mission critical EDL tasks in the true environment, the VR-A was again found to have the highest achieved mean integrated skill. With additional data spread over 10 trials, we can say with higher confidence that astronauts trained with the VR-A training condition will complete EDL tasks more successfully in a true LDEM mission. Recall that the VR-A training condition reached the

highest integrated skill level in training, showing good transfer of skills when transitioning to the real LDEM.

In regards to an all excellent performance for all 3 tasks when in the true LDEM environment, the VR-NA training condition was able to complete all tasks with excellent ratings the most often; this is critical for mission efficiency and safety. With minimal training sessions and an adaptive training algorithm, it is not a surprise that the VR-A training condition did not outperform the others on all excellent trials.

Finally, given the VR-A training condition had the highest achieved mean skill in the replicated true LEDM environment, it is not staggering that the VR-A training condition had the smallest number of poor performances. Overtly, as astronauts' lives are on the line in LDEMs, it is mandatory to establish a training condition that limits poor performances (specifically crashes) across all tasks.

Note that none of these results are significant, so accept them with caution. I thought they would be interesting to discuss anyways in a bigger picture setting.

4.3 Subjective Perceptions of the Training Conditions

Overall, the VR-A training condition rates their corresponding training condition the most usable in training. This means the VR-A training condition system is the most functable and operable compared to the other training conditions in training. A more usable training condition (or system) allows subjects to achieve tasks more dexterously and comfortably (Ref 15). Conversely, a less usable training system can result in significant losses, such as user motivation, user achievement, user resentment, user satisfaction, time, and money. When transitioning to the replicated true LDEM environment, notice that the 2D-NA training condition rated the system in the physical mock-up to be the most usable; this result is not a surprise as the system used in the physical mock-up most closely resembles the 2D-NA training condition (i.e. non-adaptive, no VR) as opposed to the other training conditions. It is reasonable that VR subjects were dismayed or overwhelmed by what felt like an entirely new system and setup in session 4. Interesting to think about is whether or not introducing VR subjects to the physical mock-up before training begins is helpful in establishing overall (or end goal) expectations, and thus improving usability in the physical mock-up for the VR subjects.

In terms of change in arousal from before and after a given session, the VR-A training condition resulted in the highest delta arousal score by the completion of training *on average*. According to the literature, a lower delta arousal score is associated with increased sleepiness from the beginning and end of the session, whereas a higher delta score is associated with increased physiological activity, feelings of being challenged, and heightened reactivity (Ref 13). When training astronauts for LDEMs, we wish to create training conditions that result in higher

changes in arousal during training. An adequate training procedure is geared towards challenging astronauts and preparing them for potentially high-stress situations in which they weren't expecting; this process allows subjects to gain technical skills, as well as interpersonal skills such as problem solving, decision making, and thinking on the spot. On the other hand, a low or negative delta arousal score in training means the subjects were not challenged enough and knew what to expect.

When subjects performed the 3 tasks in the physical mock-up session, the VR-A training condition experienced the lowest change in arousal (which was a change of 0), while the 2D-NA training condition had the highest change in arousal. It is likely that the VR-A training condition experienced the lowest change in arousal as subjects in this group were trained at higher difficulty levels, allowing subjects to feel more confident in what to expect and thus less challenged or stimulated. In fact, a delta arousal score of 0 in the physical mock-up (i.e. “true LDEM scenario”) is ideal as it corresponds to a perfect balance between sleepiness and being overly stimulated/hypersensitive/challenged. We don’t want astronauts to feel overwhelmed or anxious during the true LDEM. A delta arousal score of 0 in the physical mock-up is associated with proper training (e.g. knowing what to expect and having already been over the challenging obstacles).

In the matter of change in pleasure from before and after a given session, the 2D-NA training condition had consistently decreasing delta pleasure scores from the start to end of training. Additionally, important to discuss, is that the VR-A training condition never had negative median or mean change in pleasure scores. Evidently, a higher delta pleasure score is associated with increased satisfaction and enjoyment from the beginning to the end of the session, whereas a lower delta pleasure score is associated with increased displeasure, discontent, or sorrow (Ref 13). Satisfaction and enjoyment can lead to increased performance which is why we want to (although not absolutely required) employ a training condition that results in higher changes in pleasure when training astronauts for LDEMs. When subjects transitioned to the physical mock-up session, the VR-A training condition experienced the lowest change in pleasure, while the VR-NA training condition had the highest change in pleasure on average. The VR-A training condition likely experienced the lowest change in pleasure on average because subjects in this group were: 1) were able to train at higher difficulty levels, making VR-A subjects potentially more capable of performing the tasks in the physical mock-up and thus making them feel less accomplished for doing well, as well as 2) experienced changing difficulty levels in training, but a fixed difficulty level of 18 during the physical mock-up could make subjects in the VR-A group feel like they aren’t able to improve, resulting in less satisfaction. Note, however, that the VR-A training condition had the highest median value of delta pleasure scores in the physical mock-up.

Considering flow experience, the VR-A training condition resulted in the highest flow experience by the completion of training on average. In other words, the VR-A training condition started with the lowest flow experience in training, but ended with the highest flow experience by the end of training on average. It is not a surprise that the adaptive training condition began with the lowest flow experience in training as it takes some time for the adaptive training condition and the subject to find a good balance between the task being too hard and too easy (i.e. flow); in addition to the fact that many training algorithms and computer games do not use a difficulty adjustment algorithm. It is also not a surprise that the flow experience scores stay stagnant for the VR-NA and 2D-NA training conditions as their flow channel is essentially fixed with some biological variation due to experiencing a difficulty level of 12 only. Training conditions that result in higher flow experience are ones that find their task(s) at hand more intrinsically rewarding, focusing, creative, enjoyable, and immersive; additionally, they provide a better balance between boredom and frustration/anxiety (Ref 1). Note that by the end of training, the 2D-NA group has the least flow. When subjects performed the 3 tasks in the physical mock-up session, the VR-A training condition experienced the lowest flow experience, while the 2D-NA training condition had the highest flow experience on average. This is not a surprise as the VR-A training condition is switching from an adaptive trainer to a non-adaptive trainer as used in the physical mock-up (messing with consistency). Regardless, the flow experience score for the VR-A training condition is on the same scale as the 2D-NA flow experience score on average in the physical mock-up.

With reference to perceived task importance, the VR-A training condition had the highest perceived task importance for all training sessions on average. In other words, the VR-A training condition most strongly believed that each task was critical, more high stakes, and success on each task was imperative. Seemingly, in high stakes LDEMs, it is obligatory that astronauts believe task success is personal, mistakes must be avoided, and failing a task is a concern. When transitioning to the physical mock-up, the VR-NA training condition had the highest perceived task importance and the 2D-NA training condition had the lowest. This result is unexpected and should be dug into further in the future.

Note that some of these results are not significant, so accept them with caution. Also note that there are no significant results in which the VR-A training condition underperforms!

4.4 Limitations of the Analyses

It is critical to consider the limitations of the analyses performed, and thus the limitations of the results in the paper. For example, recall that the x-mid (inflection point) parameter is fixed at a value of 0.112 to help with convergence issues; the results of the analyses are still valid when acknowledging this limitation. Additionally, another limitation of this work is that the analyses performed rely on a small number of 24 subjects, n = 8 per training condition, due to time, money, and the number of willing participants. The many insignificant results likely reflect our

small sample size; increasing the sample size will increase the sensitivity of the tests and allow for more definitive conclusions. Apart from the small sample size, we also must acknowledge the limitation of the amount of training subjects underwent. Recall that subjects only undergo 3 training sessions, which is spread out across 6 days maximum. Not only is this inaccurate of how long true astronauts are trained for, but it also causes us to miss out on critical information regarding the VR-A training condition. For instance, the adaptive training condition is designed in a way that can hold subjects at low or high difficulty levels depending on their performance. Given that adaptive training conditions are not commonly used in video games or in high stakes training protocols, we do not expect the subjects to excel within the first 3 training sessions. In other words, we don't expect most subjects assigned to the VR-A training condition to reach harder difficulty levels, restricting us to truly seeing the benefits of adaptive training. It is inevitable that subjects trained at higher difficulty levels (i.e. difficulty levels not likely to actually take place in a true LDEM) will do better when difficulty levels are decreased (i.e. in the physical mock-up which resembles the true environment in which astronauts perform LDEM tasks). VR-A subjects need more time to reach those higher difficulty levels, and with minimal training, we are missing out on seeing this effect. Another limitation of these analyses involve knowing whether or not level 18 truly resembles the environment that astronauts will experience when picking, piloting to, and descending onto a selected landing site on Mars. Although one can never be certain, with thorough investigation of literature, we are confident that difficulty level 18 accurately resembles the environment that astronauts would experience when performing EDL tasks on Mars. Finally, the last limitation worthy of discussion is the fact that no power analyses have been conducted due to time constraints. This project should be extended in the future to address the power of each hypothesis test and/or the results from the mixed effects model. As well, this project could be extended to include more subjects, and optimize the VR-A training condition.

5 Concluding Thoughts

Overall, generalized logistic mixed effects modeling and appropriate hypothesis testing were performed to investigate the following overarching research questions: 1) how well do subjects learn across the 30 training trials, as a function of their training condition? 2) how did each training condition perform in the physical mock-up, what skill level did each training condition attain, and how does it compare across training conditions? 3) how does the subjective perceptions of each training condition differ by training condition in training and in the physical mock-up?, and 4) for each training condition, how does the subjects' subjective perceptions of their assigned training condition change as training progresses? Considerable results suggest that the VR-A training condition (i.e. the adaptive, immersive trainer) was able to reach a higher integrated skill level in training when compared to the opposing training conditions. When transitioning to the physical mock-up, no significant differences were found regarding achieved mean integrated skill, achieved integrated skill in the 1st trial, all excellent trials count, and

integrated poor performance count. Although, note that the VR-A training condition had the highest achieved mean integrated skill in all 10 trials, as well as in the 1st trial, and had the lowest poor performance counts in the physical mock-up. With respect to subjective interpretations of the different training conditions, results showed a significant difference in usability from the start to end of training for the VR-A and 2D-NA training conditions. Additionally, the VR-A training condition rates their corresponding training condition the most usable in training overall. A nearly significant interaction allowed us to infer that the VR-A training condition had a substantial increase in flow experience in training. Despite the insignificant results, results also showed convincing evidence to conclude that the VR-A training condition had incomparably higher perceived task importance in training then the other training conditions. In the physical mock-up, there were no significant differences in usability or flow experience between training conditions. However, the VR-NA training condition had relatively strong evidence suggesting higher subjective perceived task importance in the mock-up. Finally, there were no significant differences in change in affect (pleasure and arousal) in both training and the physical mock-up. With minimal training, implementation of a highly unused adaptive trainer, and small sample sizes, insignificant results for the VR-A training condition are not a surprise. Further work should be done to extend the project to include more subjects (or simulate more subject data) and power analyses. Not to mention, additional optimization testing should be done to enhance the VR-A training condition. Ultimately, with some further development and testing, immersive, adaptive training (VR-A training condition) should be accepted as a way to train astronauts for LDEMs.

See Reference 17 for all my code, assumption checks, and the datasets used in the analyses!

6 References

[1] Abuhamdeh, Sami. “Investigating the ‘Flow’ Experience: Key Conceptual and Operational Issues.” *Frontiers*, Frontiers, 13 Feb. 2020, https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00158/full.

[2] Brooke, John. “SUS: A Quick and Dirty Usability Scale.” *ResearchGate*, ResearchGate, Nov. 1995, https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale.

[3] Caldwell, Aaron R., et al. “Power Analysis with Superpower.” *Chapter 12 Violations of Assumptions*, 31 Mar. 2022, https://aaroncaldwell.us/SuperpowerBook/violations-of-assumptions.html.

[4] Caroli, A, et al. “The Dynamics of Alzheimer's Disease Biomarkers in the Alzheimer's Disease Neuroimaging Initiative Cohort.” *PMC PubMed Central*, U.S. National Library of Medicine, Aug. 2010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467365/.

[5] Csikszentmihalyi, Mihaly. “Boredom.” *Encyclopedia of Psychology, Vol. 1.*, 2000, pp. 442–444., https://doi.org/10.1037/10516-164.
[6] Engeser, Stefan, and Falko Rheinberg. “Flow, Performance and Moderators of Challenge-Skill Balance.” *SpringerLink*, Springer US, 9 Sept. 2008, https://link.springer.com/article/10.1007/s11031-008-9102-4.
[7] Gelman, Andrew, and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2009.
[8] “Getting Started with the Kruskal-Wallis Test.” *Research Data Services + Sciences*, University of Virginia Library, https://data.library.virginia.edu/getting-started-with-the-kruskal-wallis-test/.
[9] Kassambara, Peyton. “Mixed ANOVA in R.” *Comparing Multiple Means in R*, DataNovia, 29 Nov. 2019, https://www.datanovia.com/en/lessons/mixed-anova-in-r/.
[10] Lix, Lisa M., et al. “Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance ‘f’ Test.” *Review of Educational Research*, vol. 66, no. 4, 1996, p. 579., https://doi.org/10.2307/1170654.
[11] Mahr, Tristan. “Anatomy of a Logistic Growth Curve.” *Higher Order Functions*, Jekyll & Minimal Mistakes, 15 Feb. 2019, https://www.tjmahr.com/anatomy-of-a-logistic-growth-curve/.
[12] Murrar, Sohad, and Markus Brauer. “Mixed Model Analysis of Variance .” Edited by Bruce B. Frey, *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, Sage Reference, 26 Feb. 2018, https://psych.wisc.edu/Brauer/BrauerLab/wp-content/uploads/2014/04/Murrar-Brauer-2018-MM-ANOVA.pdf.
[13] Russell, James A., et al. “Affect Grid: A Single-Item Scale of Pleasure and Arousal.” *Journal of Personality and Social Psychology*, vol. 57, no. 3, 1989, pp. 493–502., https://doi.org/10.1037/0022-3514.57.3.493.
[14] “Self-Starting Nas Logistic Model.” *R Documentation*, Stanford University, https://web.stanford.edu/~rag/stat222/logiststart.pdf.
[15] Spencer, Donna. “What Is Usability?” *Step Two Designs*, 15 Sept. 2015, https://www.steptwo.com.au/papers/kmc_whatisusability/.
[16] “System Usability Scale (SUS).” *Usability.gov*, Department of Health and Human Services, 6 Sept. 2013, https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.
[17] Code and Datasets: Code and Data for Cumulative Experience Project
[18] Presentation Slides: Culminating Experience Presentation Slides
[19] Presentation Recording: CE_Presentation_Recording.mp4