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

Presented by Sandra Tredinnick

Culminating Experience Project
Spring 2023

Presentation Overview

Project Background/Motivation

- ➢ Training Astronauts for Long Duration Exploration Missions
- ➢ Understanding the difference between Virtual Reality (VR) and 2D screen-based training
- ➢ Understanding the difference between adaptive training and non-adaptive training

Overall research goals

Narrowed down research questions of interest

Methods / Analyses

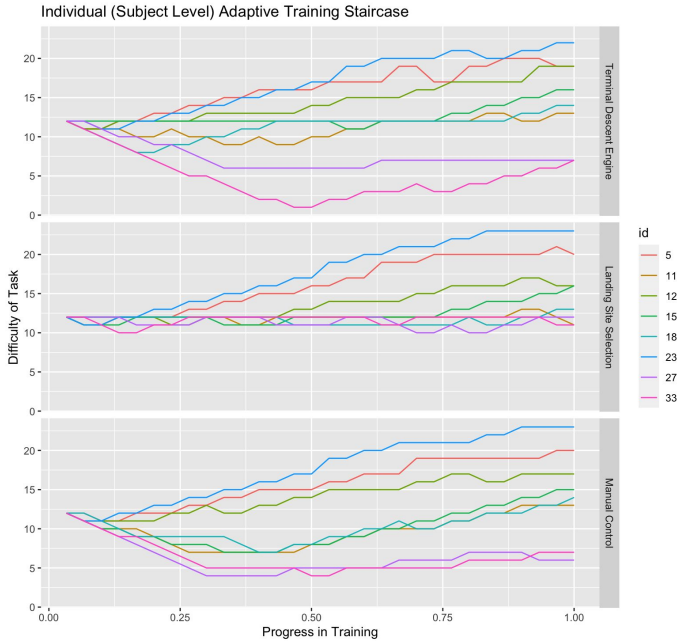
Results

Discussion + Conclusions

With advancing technology and the increasing push for manned space missions, how should astronauts be trained for Entry, Descent, and Landing tasks on long duration exploration missions?

Via virtual reality or normal 2D screens (i.e. computer monitors)?

Using difficulty progressive training protocols (i.e. adaptive)?



Virtual Reality in Training

Computer-generated simulation of a 3D environment

The Entry, Descent, and Landing environment was constructed via Unity Game Engine. Use of the VR headset allows for full immersion into this environment.

Allows for extended periods of training post deployment

Allows astronauts to mentally and physically prepare for missions in a safe environment

Most accurate representation of the true environment besides empirical reality

2D Screens in Training

A true 3D environment simulated onto a 2D monitor/display

Individuals complete training in front of a desktop monitor that displays the Entry, Descent, and Landing simulated environment

Larger, more high mass hardware, making it a less portable training option → cant bring large equipment on missions

Most commonly used and accepted training environment (baseline trainer)

More complex system when dealing with hardware issues

Adaptive Training

Changing difficulty level [to complete a task or activity] based on performance

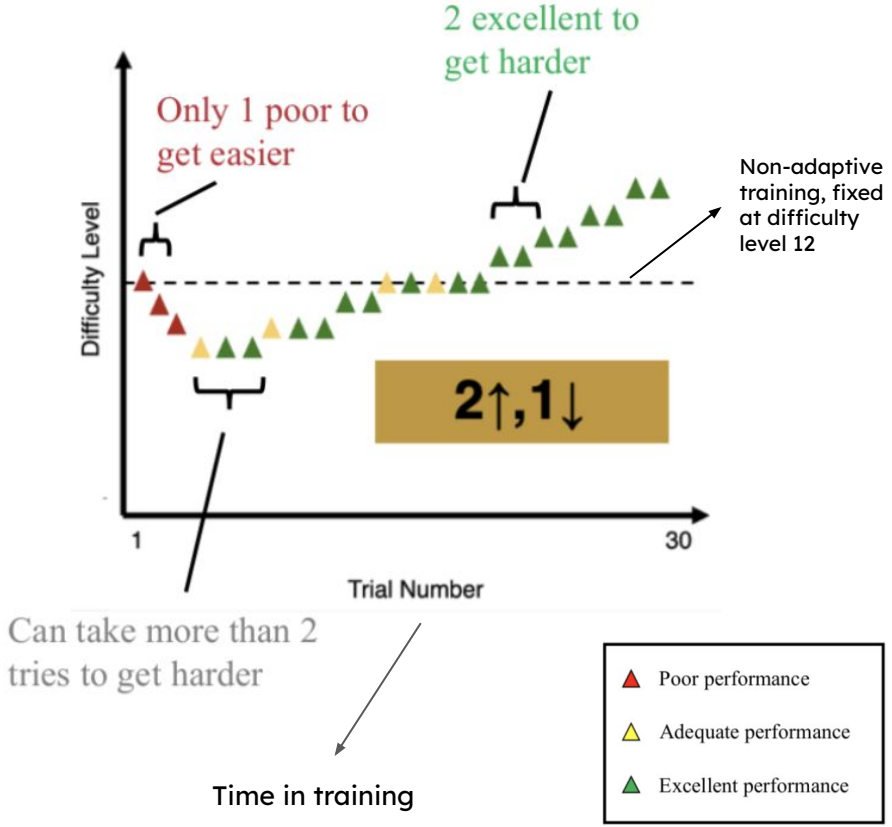
Personalized training system

Found to improve outcomes and increase task engagement

➢ Constant et al., 2019; Sampayo-Vargas et al., 2013

This research utilizes a 2-up-1-down (2↑1↓) locked adaptive training algorithm → 2 consecutive "excellent" performances in training results in a step up in difficulty, whereas a single "poor" performance results in a step down in difficulty. Difficulty levels for the Entry, Descent, and Landing tasks can't be more than one difficulty level away from each other to move-up in difficulty

Let's visualize the (2↑1↓) locked adaptive training algorithm used in this research…



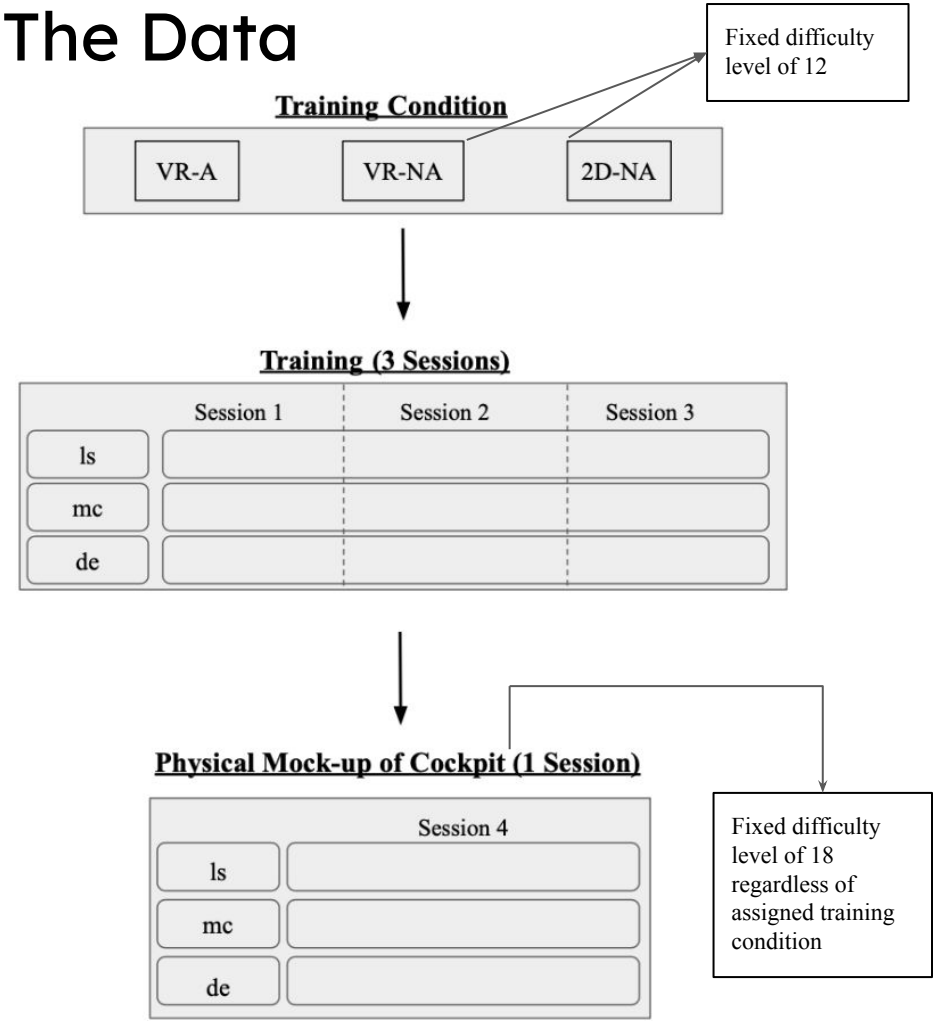
Overall Research Goals

To understand the influence of adaptive, immersive virtual reality training on astronauts' ability to perform Entry, Descent, and Landing tasks on long duration exploration missions.

To understand the impacts of adaptive, immersive training, 3 training conditions were constructed: 1) Adaptive VR (VR-A), 2) Non-adaptive VR (VR-NA), and 3) Non-adaptive screens (2D-NA). In addition, 3 Entry, Descent, and Landing (EDL) tasks were created.

* The VR-A training condition is the adaptive, immersive trainer

The Data



n = 24 subjects (8 subjects per training condition)

Subjects assigned to 1 of the following training conditions: 1) adaptive VR (VR-A), 2) non-adaptive VR (VR-NA), and 3) non-adaptive screens (2D-NA)

Participants completed 3 training sessions using their assigned training condition and 1 session in the physical mock-up

10 trials in each session, where subjects were scored on their skill level on 3 Entry, Descent, and Landing tasks:

- Landing Site Selection (ls), Manual Control (mc), and Terminal Descent Engine (de)

3 surveys administered: System Usability, Flow Short Scale, and Affect Grid

Overview: Scoring Metrics - Skill and Performance

For each subject, raw skill metrics were recorded for every task and trial. Raw skill metrics were then converted into performance and difficulty controlled skill metrics.

Difficulty Controlled Skill metric: [regardless of the task] range from 0-1

Performance metrics: binned skill values, where a value of (-1) is considered poor performance, a value of (0) is considered adequate performance, and a value of (1) is considered excellent performance

Integrated skill and performance metrics were also constructed, which are the summed skills or performances achieved across the entry, descent, and landing tasks for a given subject and trial. Note that integrated skill values range from 0-3, while performance metrics vary across analyses.

A progress in training variable was also constructed which is the given trial divided by 30

* See my write-up for further information on scoring metrics

A Few Remarks on Survey Data (Subjective Perceptions)

The **Flow Short Scale survey** is measured according to 7 attributes, namely: flow experience, perceived task importance, demands, skills, perceived fit of demands and skills, fluency of performance, and absorption. My research/project focuses on *flow experience and perceived task importance only.*

The **Affect Grid survey** is measured according to 2 attributes, specifically arousal and pleasure. *My research analyzes both attributes.*

There is only one dimension of the **System Usability survey**.

The System Usability and Flow Short Scale surveys are administered after each session, while the Affect Grid survey is measured pre and post every session. For easier interpretation, I use delta scores (e.g. post - pre) when investigating change in affect.

For reference, the hardware used in *training* is as follows…

Hand-thruster (left)

Joystick (right)

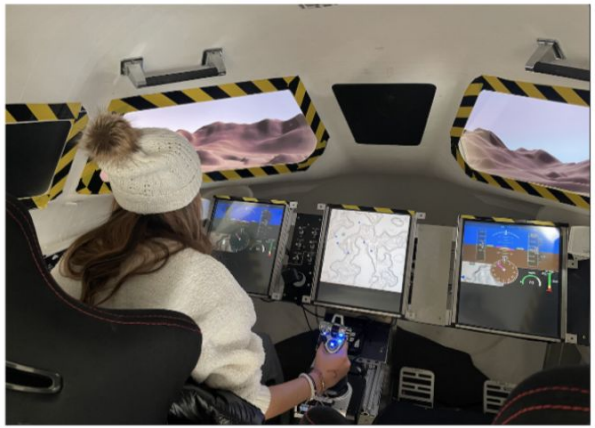
Head-mounted display (HMD)

HTC Vive Pro

VR subjects train with the VR headset (right figure), all other subjects train with a normal computer desk monitor. All subjects use the same hand-thruster and joystick controls.

For reference, the hardware used in *the physical mock-up* is as follows…

AReS: Aerospace Research Simulator



Cockpit interior with physical displays

Hand-thruster (left)

Joystick (right)

* All subjects, regardless of their assigned training condition, use the same hardware, system, and difficulty level in the physical mock-up

Research Questions

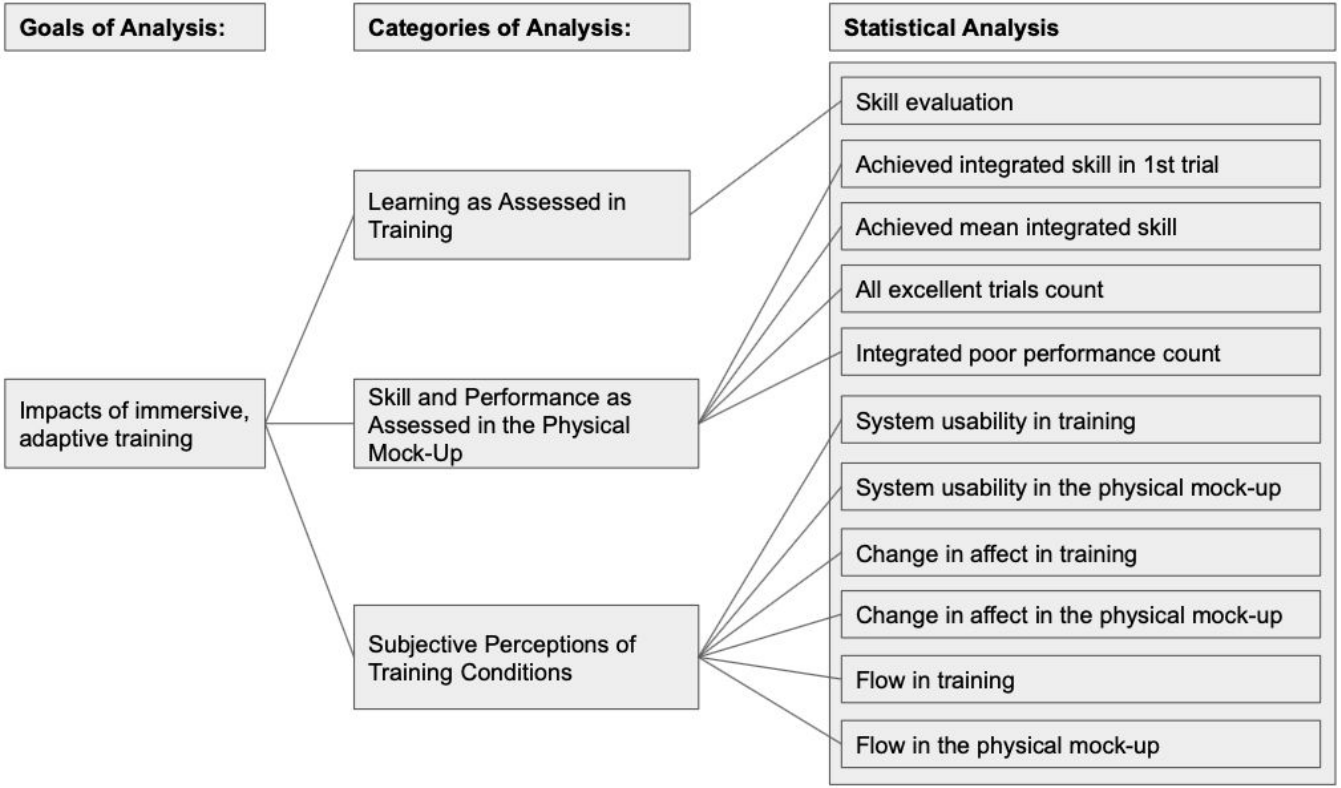
1) How well do subjects learn Entry, Descent, and Landing (EDL) skills in training as a function of their training condition?

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 do the subjective perceptions of each training condition differ by training condition in training and in the physical mock-up?

4) For each training condition, how does the subjects' subjective perceptions of their assigned training condition change as training progresses?

For complexity reasons, my research questions were partitioned into 3 overarching categories of statistical analyses, consisting of several analyses in each category…



To perform each analysis, one of the following statistical methods was implemented (as appropriate)...

Nonlinear Mixed Effects Modeling - Generalized Logistic Growth Curve *

- ➢ Used to assess learning in training
- ➢ Varies its scale and asymptote parameters by subject and training conditions (i.e. 3 level model). Xmid is fixed.
- ➢ Predicts integrated skill as a function of progress in training

Mixed-Anova *

- ➢ Used to assess system usability, flow, and change in affect for each training condition during the training phase
- ➢ Factors are always ‘Session’, ‘Training condition’, and their interaction

Kruskal-Wallis *

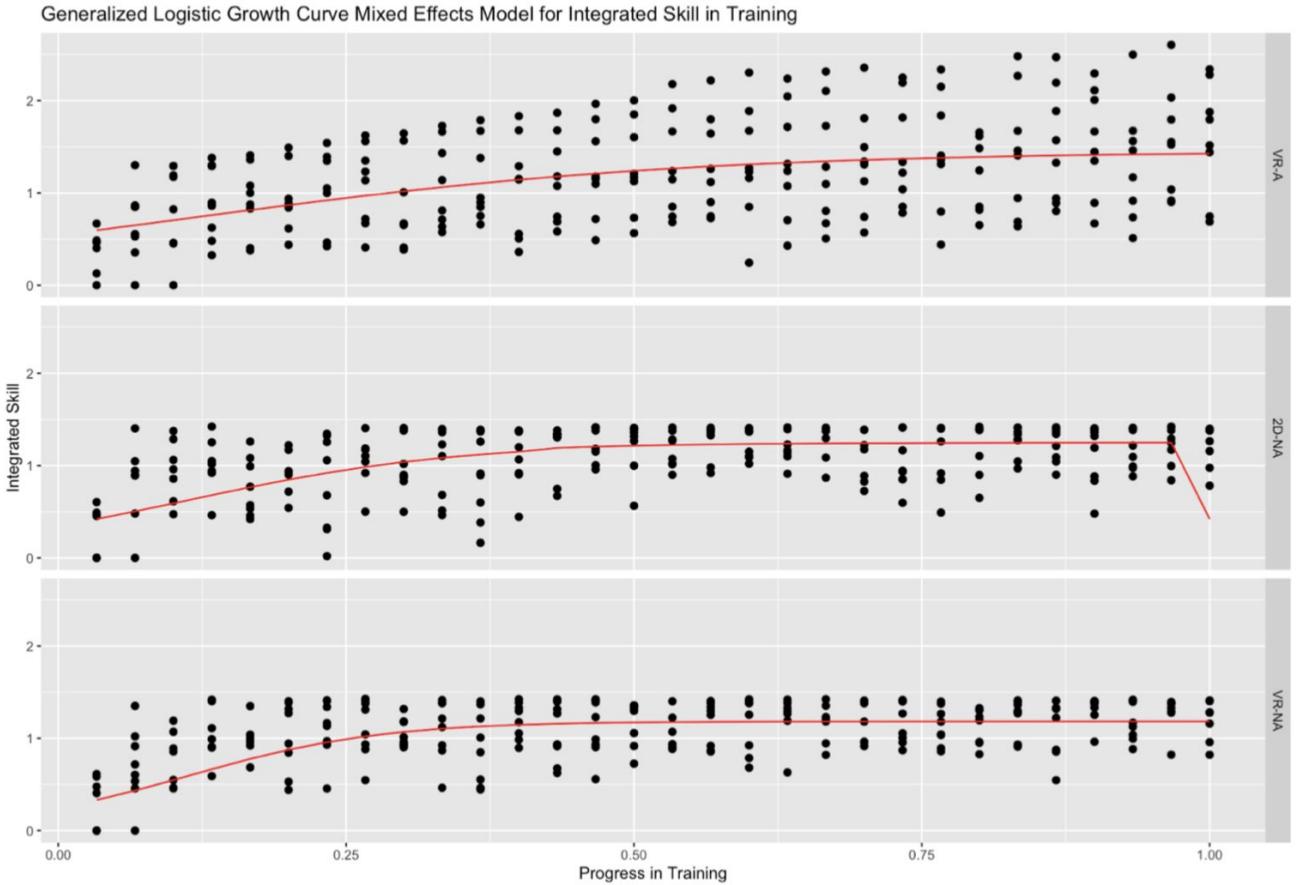
- ➢ Employed to assess skill and performance in the physical mock-up, as well as system usability, flow, and change in affect in the physical mock-up for each training condition

*Follow-up post hoc tests (pairwise t-tests or Dunn’s) were performed following significant results

*All hypothesis tests were conducted with a global α level of 0.05

Results

I Learning as Assessed in Training



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.

My model infers that the VR-A training condition had the largest asymptote and scale parameter estimates

Standard Deviations of the Parameter Estimates at the Training Condition Level (J)
$\hat{\sigma}_{\text{Asymptote,J}} = 0.122$
$\hat{\sigma}_{\text{Scale,J}} = 0.062$

I Learning as Assessed in Training Continued…

Estimated parameter coefficients for the generalized logistic growth curve mixed effects model at the training condition level

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.

Ⅱ Skill and Performance as Assessed in the Physical Mock-up

The integrated achieved skill in the 1st trial, as well as the achieved mean skill across all 10 trials in the physical mock-up showed no significant differences between training conditions, $p = 0.887$ and $p = 0.928$ respectively.

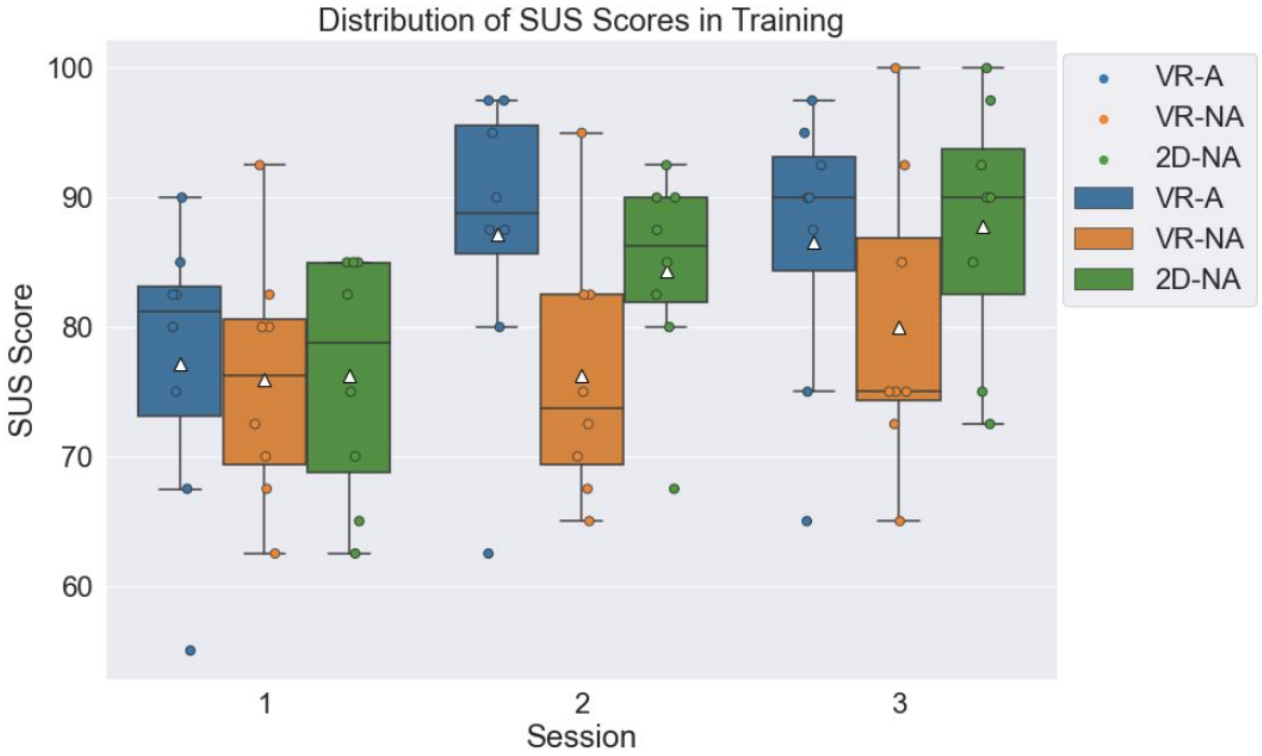
Similarly, the integrated poor performance counts and all excellent trials counts showed no significant differences among training conditions, $p = 0.841$ and $p = 0.778$ respectively.

⇒ Although, the VR-A training condition had the highest integrated achieved skill in the 1st trial and across all 10 trials in the physical mock-up, as well as had the lowest number of integrated poor performance counts in the physical mock-up

Ⅲ Results: Subjective Perceptions of Training Conditions

System Usability in Training

$p_{\text{training condition}} = 0.379$
$p_{\text{interaction}} = 0.044$



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.

Significant difference in system usability for the interaction factor

↓

There is a significant difference in system usability from the start to end of training for the VR-A and 2D-NA training conditions on average

Significant difference in system usability from session 1 to session 2 for the VR-A training condition on average → rapid and immediate increase in usability

Flow in Training: Flow Experience

$p_{\text{training condition}} = 0.962$
$p_{\text{interaction}} = 0.087$

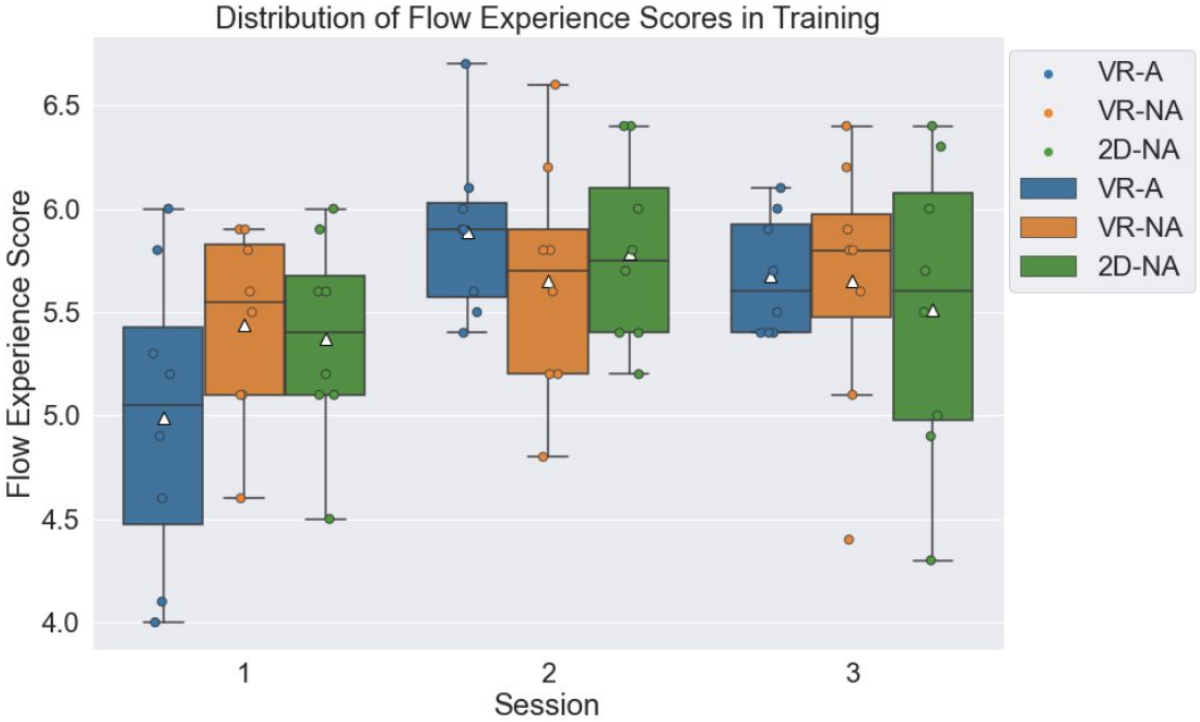


Figure 14. Distribution of flow experience scores for each session in training, partitioned by training condition (colors). One datapoint represents the flow experience score for a given subject in a given session; each subject is assigned to a specific training condition. The white triangles represent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

Nearly significant interaction revealed that there is a substantial increase in flow experience from session 1 to 2 for the VR-A training condition on average, with decreasing standard deviation

Flow experience stays relatively stagnant for the VR-NA and 2D-NA training conditions throughout training

Flow in Training: Perceived Task Importance

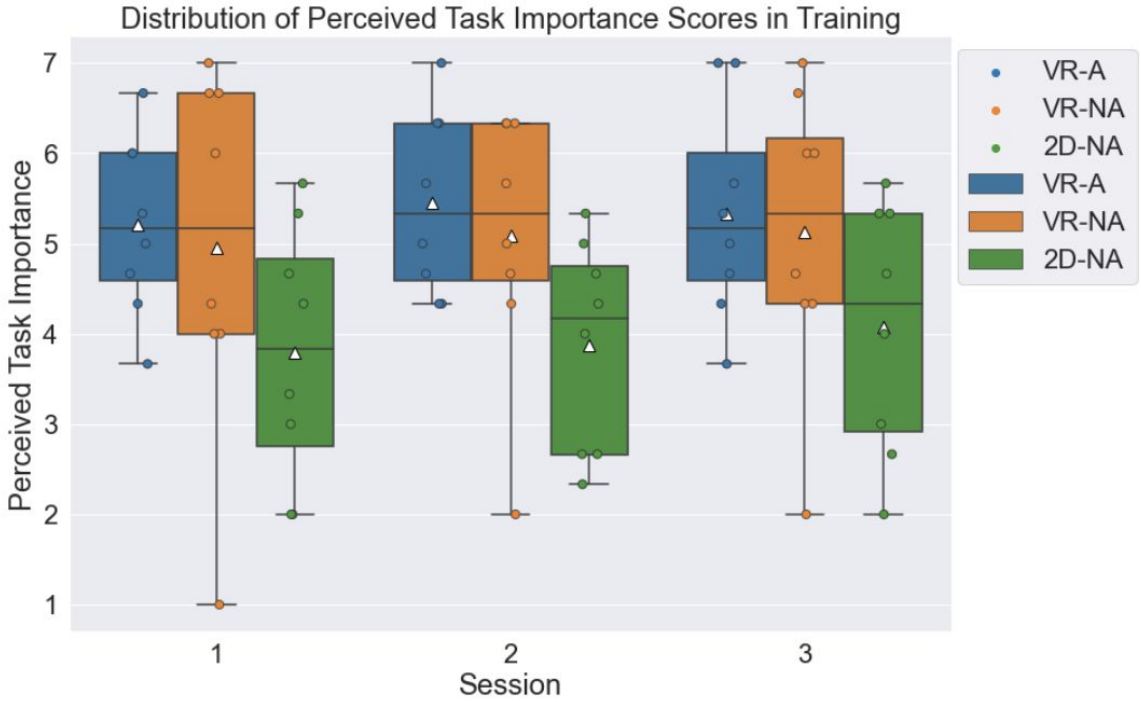


Figure 15. Distribution of perceived task importance scores for each session in training, partitioned by training condition (colors). One datapoint represents the perceived task importance score for a given subject in a given session; each subject is assigned to a specific training condition. The white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

The VR-A training condition has the highest perceived task importance in all 3 training sessions on average

The 2D-NA training condition has the lowest perceived task importance for all sessions in training on average

$p_{\text{training condition}} = 0.099$

Flow in the Physical Mock-Up: Perceived Task Importance

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.

The VR-NA training condition has a nearly significant difference in the median value of perceived task importance than the 2D-NA training condition in the physical mock-up

Notice the 2D-NA training condition has an appreciably lower perceived task importance score then the VR training conditions on average

$p_{\text{training condition}}$ =

Insignificant Results

No significant differences between training conditions for flow experience, system usability, and change in affect (in terms of both arousal and pleasure) scores in the physical mock-up

No significant differences for change in affect, in terms of both arousal and pleasure, in training

Discussion + Conclusions

I Learning as Assessed in Training

There is a noticeable effect of the training condition on integrated skill acquisition during training

➢ The VR-A training condition reached the highest integrated skill level by the end of training, while also learning integrated skills at the lowest rate

Given that long duration exploration missions are high-stakes, it is critical to employ a training condition that trains astronauts to the highest integrated skill level by the end of training. Learning skills at a quicker rate is valuable, but not absolutely necessary

High degree of variability in asymptote and scale parameter estimates at both the training condition and subject level → Mixed Effects modeling is appropriate!

Ⅱ Skill and Performance as Assessed in the Physical Mock-up

No significant results, however…

Astronauts trained with the VR-A training condition are able to pick landing sites closest to scientific sites of interest, more accurately and precisely pilot to a selected landing site, and descend onto the surface the safest when given only one chance to perform each task on a true long duration exploration mission. This also holds when astronauts are given 10 chances to perform each task on a true long duration exploration mission.

Astronauts trained with the VR-A training condition had the lowest number of integrated poor performances when transitioning to the true long duration exploration mission environment. With crew and astronaut lives on the line, poor performances (especially crashes) must be limited

Ⅲ Subjective Perceptions of Training Conditions

The VR-A training condition allows astronauts to achieve tasks more dexterous and comfortably, with higher user motivation, achievement, and satisfaction in training

- ➢ It is not a surprise that the 2D-NA training condition rated the system in the physical mock-up the most usable as the system used in the physical mock-up most closely resembles the 2D-NA system in training

It takes longer for astronauts trained with the VR-A condition to get into a good state of flow experience in training; however, after initial training, astronauts are able to reach the highest level of flow experience by the completion of training. This is not shocking as it takes a few trials for the adaptive trainer to correct for the personalization of each astronaut's skill level.

- ➢ The VR-A training condition had the lowest flow experience in the physical mock-up, likely caused by the transition to a non-adaptive, non-virtually immersive system in the mock-up.

More than other training conditions, astronauts trained with the VR-A condition believed each entry, descent, and landing task is critical, high-stakes, and success on missions is imperative in training. When transitioning to the physical mock-up, the 2D-NA training condition had a substantially lower perceived task importance regarding the 3 tasks then the other training conditions.

** The many insignificant results are not bewildering given the small sample size and the minimal amount of training*

Ⅲ Subjective Perceptions of Training Conditions Continued…

Astronauts trained with the VR-A condition felt more challenged, and were more physiologically active, as well as, had the highest heightened reactivity in training. Conversely, astronauts trained with the conflicting training conditions had higher levels of sleepiness and boredom in training. In the physical mock-up, the VR-A training condition had no change in arousal, while the other training conditions had higher changes. A non-zero change in arousal in the physical mock-up is associated with not being as prepared to handle the tasks (i.e. didn't know what to expect) in the true long duration exploration mission.

Each training condition had similar changes in pleasure scores in both training and in the physical mock-up. The 2D-NA training condition had consistently decreasing pleasure scores as training progressed; there could be a correlation between VR training and pleasure. We wish to construct a training condition that does not yield negative pleasure scores, but not critical. Decrease in pleasure after performing a task isn't bad if they were unsuccessful (shouldn't be content after crashing).

** The many insignificant results are not bewildering given the small sample size and the minimal amount of training*

Limitations of this Research

Unstableness of the multilevel generalized logistic growth curve model at the end of training for the 2D-NA training condition

Analyses relied on a small sample size (n = 24 subjects, 8 per training condition)

Minimal amount of training

⇒ washing out the true effects of adaptive, immersive training

Is difficulty level 18 the most accurate representation of the true environment astronauts experience when performing entry, descent, and landing tasks on long duration exploration missions?

**Power analyses, optimization techniques, and additional subject collection/training should be carried out!

THANK YOU!!!

Questions? Email me!

Sandra Tredinnick
satr3210@colorado.edu

University of Colorado Boulder
Bioastronautics

LISA
Laboratory for Interdisciplinary Statistical Analysis

NASA

TRINITY

Acknowledgements

This research was funded by NASA under Grant No. 80NSSC21K1140 and supported by the University of Colorado Boulder, as well as UC Davis.

Thank you to all the school participants, students, professors, and staff who participated in the study. A special thanks to Dr. Allison Anderson (Principal Investigator), Ellery Galvin, Esther Putnam, Alessandro Verniani, and Dr. Eric Vance (Research Advisor) for their contributions to this project and making it possible.

University of Colorado Boulder
Bioastronautics

LISA
Laboratory for Interdisciplinary Statistical Analysis

NASA

TRINITY

References

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.

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.

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.

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/.

Constant, Thomas, and Guillaume Levieux. “Dynamic Difficulty Adjustment Impact on Players' Confidence.” *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, https://doi.org/10.1145/3290605.3300693.

Csikszentmihalyi, Mihaly. “Boredom.” *Encyclopedia of Psychology, Vol. 1.*, 2000, pp. 442–444., https://doi.org/10.1037/10516-164.

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.

Gelman, Andrew, and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2009.

“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/.

References

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/.

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.

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/.

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.

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.

Sampayo-Vargas, Sandra, et al. “The Effectiveness of Adaptive Difficulty Adjustments on Students' Motivation and Learning in an Educational Computer Game.” *Computers & Education*, vol. 69, 2013, pp. 452–462., https://doi.org/10.1016/j.compedu.2013.07.004.

“Self-Starting Nls Logistic Model.” *R Documentation*, Stanford University, https://web.stanford.edu/~rag/stat222/logistart.pdf.

References

Spencer, Donna. “What Is Usability?” *Step Two Designs*, 15 Sept. 2015, https://www.steptwo.com.au/papers/kmc_whatisusability/.

“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.

Walk, Wolfgang, et al. “Design, Dynamics, Experience (DDE): An Advancement of the MDA Framework for Game Design.” *SpringerLink*, Springer International Publishing, 1 Jan. 1970, https://link.springer.com/chapter/10.1007/978-3-319-53088-8_3.