AUTOMATED INDIVIDUALLY-ADAPTIVE ASTRONAUT TRAINING ALGORITHMS IN VIRTUAL REALITY FOR DEEP SPACE MISSIONS

by

Alessandro Verniani

B.S., University of California, Irvine, 2021

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado at Boulder in partial fulfillment

of the requirement for the degree of

Master of Science

Department of Aerospace Engineering Sciences

2023

Committee Members:

Allison P. Anderson (Chair)

Torin K. Clark

Daniel J. Szafir

Stephen K. Robinson

Verniani, Alessandro (M.S., Aerospace Engineering Sciences)

Automated Individually-Adaptive Astronaut Training Algorithms in Virtual Reality for Deep Space Missions

Thesis directed by Prof. Allison A. Anderson

Long-duration exploration missions (LDEM) pose a unique challenge for astronaut training. Astronauts may experience a degraded capcity to perform complex tasks due both to the time elapsed from initial ground training and to the neural decrements associated with spaceflight. This effect may be particularly pronounced for complex, mission-critical tasks such as maneuvering spacecraft during entry, descent, and landing (EDL). Since the time delays and crew constraints on deep space missions preclude facilitated, operator-mediated training, mitigating this risk requires a cost-effective, lightweight, and automated system for recurrent training. Virtual reality (VR), long-used as an immersive, easily-programmable, dynamic environment for training, is an ideal medium for training during LDEM.

To date, there is no literature investigating the effect of responsiveness, integration, and personalization on the efficacy of automated training algorithms. This study used a virtual simulator to train subjects to pilot and land a spacecraft on the surface of Mars and a physical mock-up of a spacecraft cockpit to put skills acquired during training to the test. The study assessed the effect of multiple training algorithms on skill acquisition, learning retention, progression of training difficulty, subtask performance, and skill transfer between the virtual and physical

environments. The training algorithms varied the threshold for difficulty progression (sensitivity), the effect of subtask performance on the difficulty progression of other subtasks (lockstep), and the use of fixed rather than adaptive difficulty progression.

The study found that highly responsive training algorithms leads to faster difficulty progressions and higher achieved difficulty in training but lower skill and performance in the cockpit environment. It also found that low levels of subtask integration which allow for discrete rather than unified subtask progressions leads to higher performance and achieved difficulty in training, and slightly better performance outcomes in the cockpit. Finally, the study found that personalized training leads to higher levels of skill and performance in both training and the cockpit compared to non-adaptive, fixed progression training.

Future work can build upon these results by analyzing the effect of responsiveness on the duration of the familiarization phase during training as a function of task complexity and expanding analysis on personalization to investigate the limiting effect of fixed training progression on top performing subjects. Future studies should investigate run-dependent shifts in PEST staircases, dynamic variable response paradigms which scale difficulty increments to subject performance, Bayesian methods to predict optimal challenge given both individual and aggregate data, subject-selected difficulty, and the incorporation of unobtrisvely-gathered psychophysiological data to estimate workload and challenge, closing the loop on characterizing and optimizing human performance in space.

DEDICATION

To the SpaceX Polaris Dawn crew,

For their kindness, enthusiastic support for research,

and fearless resolve for expanding the bounds of human spaceflight.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Allie Anderson for her support and mentorship; her brilliant insights and remarkable forebearance helped to make this work possible. I would also like to thank Dr. Torin Clark and Dr. Steve Robinson for their wisdom and support in this project.

I also thank Esther Putman for her steadfast advice, detailed insights, and unwavering humor, Sandra Tredinnick for her lightheand arted spirit instrumental advice in organizing the analysis, and Ellery Gavin for her remarkable acuity and willingness to assist in matters ranging from obstruse statistics to feline exposition.

I wish also to extend my appreciation to Wyatt Rees and Benjamin Peterson for their aid in developing and implementing the training algorithms, mangaging the database and server, and helping to test edge cases in user functionality at obscene hours of the night.

I would also like to greatly thank Sage Sherman, Cade McVey, Stella Cross, and Kiah May for their invaluable assistance in helping to test dozens of subjects; their alacrity in accommodating endless scheduling caprices and mishaps, and general excitement for helping this project come to fruition.

Finally, I would like to acknowledge the dispossessed Arapaho, Cheyenne, and Ute peoples on whose ancestral lands this research was conducted, and the continued importance of engaging with and learning from indigenous communities in Colorado.

CONTENTS

TABLES

Table

FIGURES

Figure

CHAPTER I

INTRODUCTION

1.1 MOTIVATION

Astronauts on long-duration exploration missions (LDEM) may experience a degraded capacity to perform complex tasks due both to the time elapsed from initial ground training (Arthur Jr. et al., 2009) and to the neural decrements associated with spaceflight (Eddy et al., 1998). This effect may be particularly pronounced for complex, mission-critical tasks (Childs and Spears, 1986) such as maneuvering spacecraft during entry, descent, and landing (EDL). Continued training for astronauts during LDEM would serve to attenuate skill attrition (Klostermann et al., 2022), stimulate cognitive task performance (Jiang et al., 2023; Holt, 2023), and even improve mental health outcomes (Carulli et al., 2019; Oluwafemi et al., 2011; Salomon et al., 2018). However, existing methods for facilitated astronaut training, including operator mediation and ad hoc difficulty modulation, are infeasible on deep space missions, where systems must be able to operate autonomously with minimal oversight from or dependence on Earth-based systems or operators (Wu and Vera, 2019; Doyle, 2003; Love and Harvey, 2014).

Mitigating the risk of degraded performance from ineffective or latent training requires a cost-effective, lightweight, and automated rather than facilitated method for training astronauts on LDEM. While facilitated training can be adapted to subject needs in real time by a human overseer or operator, we hypothesize that an autonomous training system must be able to respond to individual performance and

have comparable performance outcomes to be effective. Such a training system must meet the pragmatic constraints of spaceflight by being low-mass, cost-effective, compact, and requiring the least amount of operational overhead. Virtual reality (VR) is an immersive, low-cost, and programmable method for training that has been used effectively by NASA and other entities for more than three decades (Psotka, 1995). Its modularity, compactness, and growing adoption as a dynamic system makes it the ideal candidate for use as an immersive astronaut training testbed for deep space missions.

Developing personalized, individually-responsive automated training paradigms to facilitate learning in VR is crucial to developing a modular, easily-operable, Earth-independent system for crew training that counteracts skill degradation, maximizes retention, and leads to high performance outcomes in the spacecraft environment.

1.2 BACKGROUND

In recent years, space policy and funding directives from the United States government have clearly outlined human space exploration as a top national priority (National Space Policy of the United States of America, 2020). Such policy explicitly states that the United States "will lead the return of humans to the Moon for long-term exploration and utilization, followed by human missions to Mars and other destinations" (Presidential Policy Directive-4, 2017). This has culminated in ongoing bipartisan funding from the U.S. Congress to NASA's Artemis program, which seeks

to create a sustained human presence on the lunar surface (NASA, 2023). As NASA and its international partners push to develop the science and technology to enable deep space missions to the moon and Mars, it is important to advance human-centered autonomy to account for the challenges of spaceflight (Starek et al., 2015; Jonsson et al., 2007). This thesis therefore investigates the efficacy of adaptive, personalized, and integrated approaches to automating astronaut training for long-duration missions.

1.2.1 MOTOR SKILL DECAY

Future deep space missions beyond low-Earth orbit are projected to have durations ranging from several weeks to several years. For instance, NASA's Artemis III mission is anticipated to last up to four weeks (Creech et al., 2022; Smith et al., 2020), while a Mars mission would last, conservatively, between 1.5 to 2 years (Salotti and Heidmann, 2014; Herman et al., 2018), inclusive of the travel time between Earth and Mars, planetary surface operations, and a return trip to Earth given modern propulsion methods (Linck et al., 2019; Walberg, 2012; Sankaran et al., 2006). These durations, particularly for Mars missions, are well within the timescales at which complex task execution, and in particular tasks that require fine motor skills, are known to degrade through disuse. Studies of skill decay in pilots have shown for decades that flight skills decay rapidly and extensively after disuse (Arthur Jr. et al., 2009; Childs, Spears, and Prophet, 1983). Furthermore, skills which involve substantial cognitive, procedural, or accuracy-based components undergo greater and

more rapid decay over time than control-oriented skills (Childs and Spears, 1986; Hufford and Adams, 1961; Smith and Matheny, 1976). Skill loss in pilots is non-linear (Hendrickson et al., 2006), accelerates over a period of continued disuse (Svensson et al., 2013), and is particularly salient for manual control tasks (Casner et al., 2014). Even moderate lapses in skill proficiency can have outsized consequences in flight systems (Fanjoy and Keller, 2013). Moreover, the time-dependent degradation of motor skills is not unique to pilots. Surgical residents who primarily performed clinical research for two years, for example, were found to have significantly degraded fine psychomotor skills compared to residents who made regular use of psychomotor skills during surgery (Mohamadipanah et al., 2020). Thus, attrition of both fine and integrated motor skills ccurs in all people through disuse.

1.2.2 MICROGRAVITY-INDUCED CHANGES

Microgravity may exacerbate losses to motor skills and coordination beyond the attrition due to disuse. For instance, subjects displayed worse performance on instrument-control tasks in short-term microgravity (Steinberg et al., 2015), and medical professionals were found to apply more force and produce inferior surgical knots in microgravity induced by parabolic flight (Rafiq et al., 2006). Although acclimation can lead to partial motor skill recovery among astronauts, there are pronounced decrements to fine motor skills at gravitational transitions (Holden et al., 2022), including that which astronauts would experience during Mars entry when shifting from microgravity (0g) to Martian gravity (0.38g) (Cavagna, Willems, and

Heglund, 1998). Thus, the ability to train in the microgravity environment is an important way to maintain motor skills relevant to the spaceflight environment.

Furthermore, microgravity exposure leads to spaceflight-associated cognitive declines (Patel et al., 2020; Mamarella, 2020; Eddy et al., 1998), including in perceptual anticipation and spatial reasoning (Van Ombergen et al., 2017; Grabherr and Mast, 2009). Declines in cognitive task performance have been documented both in astronauts (Roberts et al., 2019; Schiflett, 2013) and subjects in head-down bed rest terrestrial analogs (Basner et al., 2021; Liu et al., 2012) and simulated microgravity (Yang and Shen, 2003). These effects persist in virtual reality (Jiang et al., 2023), suggesting that VR training when astronauts are experiencing cognitive decline may not attenuate the decline but rather serve to acclimate them to performing tasks under conditions for which few countermeasures exist.

Spaceflight has also been documented to cause neural decrements (Roy-O'Reilly, Mulavara, and Williams, 2021), including to the central nervous system (Newberg and Alavi, 1998; Clément and Ngo-Anh, 2013; Clément et al., 2020) and musculoskeletal system (Deschenes et al., 2002; Juhl IV et al., 2021). Moreover, microgravity has well-established effects on sensorimotor function (Clark, 2022; Clark et al., 2015; Clément, 2007) which require astronauts to relearn certain motor skills. These effects both reduce the efficacy of ground training and necessitate the use of continued training both to stymie further skill attrition and perhaps stimulate cognitive performance.

1.2.3 Recurrent Training for Skill Maintenance

Although some research on detecting skill decay is ongoing (Linde and Miller, 2019), the degradation of astronaut-relevant motor skills is subject to uncertainty and individual variability, necessitating a system to refresh training and practice complex functions recursively. The increasing level of autonomy in the navigation and control of spacecraft (Starek et al., 2015) is a concern as the amount of flight-relevant operational tasks that astronauts must perform decreases in scope and frequency (Markkula et al., 2018; Frank et al., 2013). For instance the presence of automation in aircraft was found to erode fine-motor flying skills in airline pilots (Haslbeck and Hoermann, 2016), and increasing automation in crewed spaceflight vehicles is likely to be a similar cause for skill disuse among astronauts.

An integrated theory of learning and forgetting suggests that both attrition and retention vary across three stages of learning, which are characterized as familiarization, consolidation, and tuning (Kim, Reuter, and Koubek, 2010). Practically speaking, different skills have varying risks of decay, modulated by how well they have been learned (Ritter et al., 2013), as seen in Figure 1.1. This also reinforces the idea that recurrent training, which increases the number of practice trial to which astronauts are exposed and reduces task completion time while increasing profiency, is an effective way to attenuate skill decay (Kluge and Frank, 2014). This substantiates the need for a low-mass, low-power, and low-volume system for astronauts to train frequently within the spacecraft itself.

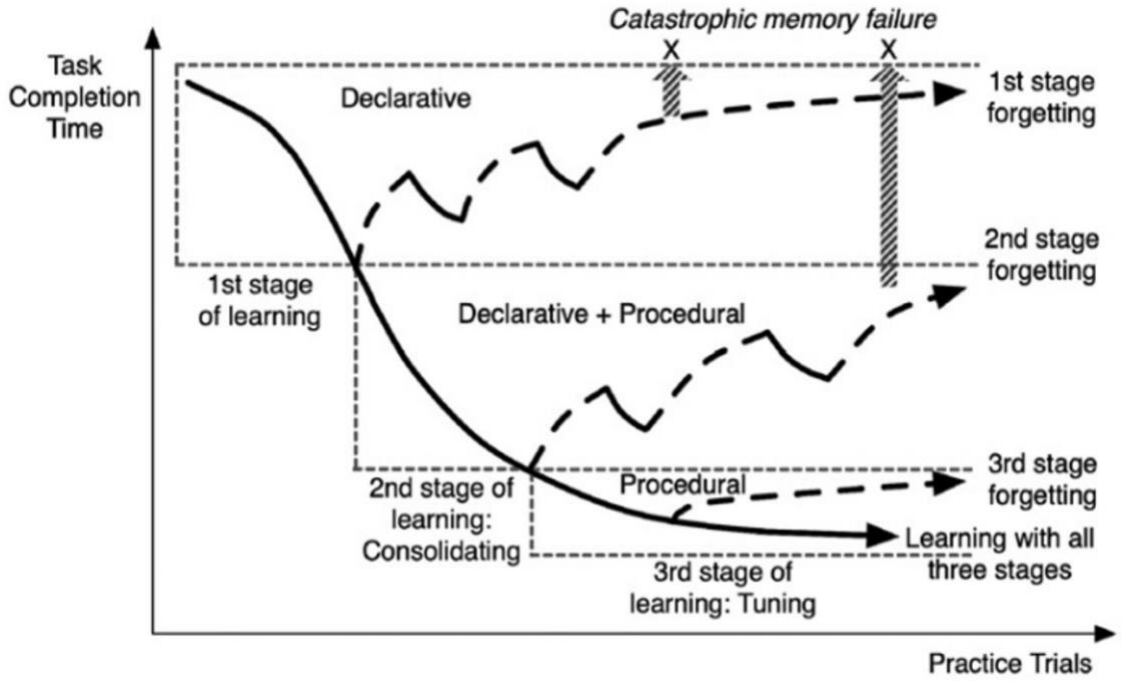


Figure 1.1: KRK Theory of Skill Learning and Retention for declarative and procedural learning (Kim, Reuter, and Koubek, 2010)

Such a system must also be capable of providing recurring training throughout the mission duration. Recurrent training is needed to maintain or enhance flight skills in pilots (Childs, Spears, and Prophet, 1983; Hollister et al., 1973), and even minor refresher interventions are effective at attenuating complex cognitive skill decay (Klostermann et al., 2022). A study on skill decay in non-performing surgeons found that cognitive training can improve performance, both alone and in combination with motor training (Kelc, Vogrin, and Kelc, 2020). Thus, the ability for astronauts to practice complex motor skills is paramount to minimizing the hazards associated with performance decrements from skill attrition. This requires providing immersive, high-fidelity, mission-like training throughout the LDEM duration.

1.2.4 ASTRONAUT TRAINING CHALLENGES

Astronauts typically receive a wealth of ground training before embarking on spaceflight missions. The Apollo program astronauts, for instance, trained for 18-24 months prior to lunar missions (Lim et al., 2010), while astronauts to the International Space Station (ISS) receive 6-12 months of training (Loehr et al., 2015). The crew for each of the 7 Apollo missions to the moon's surface were trained to use tools in an altered gravity environment, traverse the lunar surface with the rover, deploy and operate scientific instruments, take clear photographs, and to collect and document in situ samples of regolith and other material (Phinney, 2019; Messeri, 2014; Lofgren, Horz, and Eppler, 2011; El-Baz, 2011) for skills ranging from robotics operation and extravehicular activity (EVA) to ISS maintenance and emergency procedures, in addition to physical preparation (Sgobba et al., 2018; Marciacq and Bessone, 2009). Because astronaut training is necessarily complex and varied, spanning a large number of skills and knowledge of multiple systems, the risk of knowledge decay and skill attrition is particularly high.

Moreover, the time elapsed between ground training and mission-related task execution will be significantly higher on deep space missions, necessitating recurrent training during LDEM. To date, the execution of mission-related tasks, including navigation, piloting, and system operation (Lee, 1975; Murtazin and Petrov, 2012), has typically commenced within days (Donegan, 1965) or even hours (Seedhouse, 2016) of launch, including to vehicles and space stations in low-Earth orbit (LEO) and

the lunar surface (Butler, 1973; Pomeroy, 1973). For decades, astronauts have been trained to perform complex tasks in reduced gravity or microgravity, including to operate tools, perform vehicle maintenance, and pilot spacecraft. For instance, Gemini and Apollo astronauts were made to perform psychomotor tests during periods of weightlessness on the ASD zero-g aircraft in order to acclimate to changes in motion and behavior in microgravity (Mueller, 1963). Additionally, many astronauts first practice employing exercise-training protocols in microgravity by using neutral buoyancy facilities as analogs (Greenleaf et al., 1989). For deep space missions, accommodating the large number of skills required to perform novel operations, including surface EVAs or habitat repairs, is expected to require a more extensive training regimen of operational tasks (Thomas and Trevino, 1997; Sauro et al., 2023). Thus, future training systems must account for delays in the onset of skill use and associated skill attrition when preparing astronauts to perform complex operational tasks in altered gravity environments, something best accomplished through recurrent training.

Although the majority of astronaut training has been facilitated by operators and engineers, facilitated training is infeasible for deep space missions. The time delay for communication, which averages to be 2.56 seconds between Earth and the moon (Mishkin et al., 2007) and 5-20 minutes between Earth and Mars (Love and Reagan, 2013), is known to negatively impact performance, mood, and workload in subjects in analog missions who interface with a simulated ground crew (Diamond, 2015), and the delay makes space teleoperation infeasible (Sheridan, 1993).

Moreover, it is infeasible to bring dedicated trainers or operators on LDEM (Robertson et al., 2020), where there is a need for a crew composed of dedicated medics, engineers, or pilots (Saluja et al., 2008; Landon et al., 2017; Botella et al., 2016). Both of these limitations to facilitated training point to the need for an autonomous, Earth-independent system for recurrent astronaut training.

Finally, the mission complexity and associated hazards of deep space missions necessitates a dynamic, responsive, and programmable training system. To date, astronaut training has sought to prepare crew for emergencies, including by simulating subsystem and component malfunctions, requiring crew to run through off-nominal procedures, and engaging in simulated emergency responses, including medical events (Seedhouse, 2010; Strapazzon, 2014, Ewald, 2019). Typically, facilitated training is provided for the most probable and most consequential emergency scenarios to space vehicles or stations, such as collisions with micrometeoroids or debris, cabin fires, or failures in the ECLSS system (Escobar, Nabity, and Klaus, 2017; Jones, Hodgson, and Kliss, 2014). These are likely to be more unpredictable, and the number of hazards multiplied, during deep space missions, rendering it a challenge to train crew for all or most possible cases before the mission. A modular training system which can be remotely programmed ad hoc would be required to provide a slew of relevant training during transit and to respond to novel situational hazards during LDEM.

1.3 RESEARCH OBJECTIVES

The principal purpose of this thesis is to evaluate the efficacy of automated, individually–adaptive training algorithms for deep space missions. This is best described with the following three objectives:

Objective 1

To investigate the effect of training algorithm ***responsiveness*** *in learning and performance outcomes.*

Responsiveness refers to the sensitivity of an algorithm to an individual's performance levels. High responsiveness is characterized as fast or immediate upward/downward modulation of difficulty upon the detection of excellent subtask performance, while low responsiveness is characterized by an algorithm requiring, for example, multiple excellent performances before modulating difficulty upward. As discussed in Chapter 3, less responsive algorithms have higher performance thresholds for difficulty progression, and this conservatism reduces the probability of premature modulation. By contrast, highly responsive algorithms modulate subtask difficulty more easily, and thus more frequently, responding more sensitively to subject performance.

Hypothesis 1: Automated training algorithms with higher levels of responsiveness will lead to faster skill acquisition, increased learning retention, higher performance, and increased skill transfer between the virtual and physical environments compared to less responsive algorithms.

Objective 2

To investigate the effect of training algorithm ***integration*** *in learning and performance outcomes.*

Integration refers to the discretization of subtask difficulty modulation in training algorithms. Highly integrated algorithms require that the progression of difficulty among subtasks occurs in conjunction with one another, a condition hereafter referred to as "lockstep". This means that no subtask can become significantly more difficult than another, and that the difficulty of subtasks is unified, occurring in concert rather than progressing independently, leading to asymmetric skill acquisition (see Chapter 3.2). By contrast, algorithms with low levels of integration allow for discrete, independent modulation of difficulty across subtasks. This means that the progression of difficulty between subtasks can vary widely, according to subject performance at each subtask.

Hypothesis 2: Automated training algorithms with discrete rather than integrated subtask difficulty modulation will lead to faster skill acquisition, increased learning retention, higher performance, and increased skill transfer between the virtual and physical environments compared to highly integrated algorithms with lockstep.

Objective 3

To investigate the effect of training algorithm ***personalization*** *in learning and performance outcomes.*

Personalization refers to the individual adaptivity of training algorithms. A personalized algorithm responds to an individual's performance and modulates difficulty according to their needs. It therefore has a human-in-the-loop feedback system. Algorithms without a personalized response modulate difficulty according to a predefined progression. This progression does not vary in response to subject performance and is fixed. Such fixed progressions can be static, with difficulty never varying, linear, with difficulty increasing at a constant rate over time, or nonlinear. The fixed progression of an algorithm without personalization can be based on the median progression of difficulty among individuals who receive personalized training.

Hypothesis 3: Personalized, individually-adaptive automated training algorithms will lead to faster skill acquisition, increased learning retention, higher performance, and increased skill transfer between the virtual and physical environments compared to algorithms without individualized response.

1.4 THESIS OUTLINE

Chapter 2 of this thesis focuses on reviewing the literature on automated training and the use of virtual reality as an immersive medium.

Chapter 3 focuses on the experimental design. This includes screening and condition assignment and an overview of the virtual training simulator and spacecraft cockpit mock-up. The chapter also discusses performance grading for each

of the subtasks and algorithm implementation. Finally, it provides an overview of the statistical methods.

Chapter 4 summarizes the important results of the statistical analyses used to investigate the hypotheses.

Chapter 5 discusses the significance and implications of the results and addresses limitations and potential sources of error.

Chapter 6 summarizes the objectives and hypotheses and reestablishes the main results of the study. It also makes recommendations for modifying the experimental procedure and identifies future areas of study.

CHAPTER II

LITERATURE REVIEW

2.1 TRAINING IN VIRTUAL REALITY

Entities like NASA have used virtual reality (VR) for the last three decades to train astronauts (Psotka, 1995; Homan & Gott, 1996), including to perform operational tasks in neutral buoyancy (Sinnott et al., 2019; Everson et al., 2017), conduct simulated extravehicular-activity using hardware-in-the-loop simulations (Garcia, Schlueter, and Paddock, 2020), and to repair the Hubble Space Telescope (Loftin and Kenney, 1994). VR has also been used as a medium for training aircraft pilots (Dymora et al., 2020) and technicians (Vora et al., 2001 and 2002) to perform complex operational tasks, and as a training aid for manual spacecraft docking (Piechowski et al., 2020), including through the use of shared control and haptic guidance (Li, Patoglu, and O'Malley, 2009). Aside from operational skills, spatial disorientation in astronauts has been mitigated using VR for egress navigation training (Aoki, Oman, and Natapoff, 2007; Sinkjaer and Popović, 2009), and VR was used for orientation and postural training in a simulated spacecraft cabin (Zhu et al., 2015). Moreover, VR is effective at imparting complex skills used by surgeons in the operating room environment (Seymour et al., 2002; Aïm et al., 2016). Thus, VR is a robust system for operational, sensorimotor, and even spatial training, all three of which are important components for astronaut training.

Immersive training in VR is effective at stress inoculation on simulated astronaut tasks (Finseth et al., 2021) and at imparting complex, operationally relevant skills (Thurman and Mattoon, 1994; Ng et al., 2019) and is becoming increasingly more widespread. The transfer and equivalence to real world tasks has been demonstrated extensively (Kozak et al., 1992; Kenyon and Afenya, 1995; Rose et al., 2010; Moskaliuk, Bertram, and Cress, 2012; Hamblin, 2005; Park et al., 2007), including with automated scenario generation (Zook et al., 2012). Skill retention in both minimally and maximally immersive VR training systems (desktop vs. head-mounted display (HMC), respectively) is high for procedural skills (Farr et al., 2022), and highest for subjects who used HMDs when training to gain complex military medical skills (Siu et al., 2016). Skill acquisition is highest among those who train in VR, especially if used in concert with physical and/or haptics-mediated controls (Butt, Kardong-Edgren, and Ellertson, 2018). Given that ground-based flight simulators for astronauts and pilots are cumbersome, hardware intensive, and require facilitation by operators and trainers, VR has the potential to be the lightweight, programmable, cost-effective, and easily-operable alternative (Gupta et al., 2008) required for use on deep space missions.

2.2 FLOW THEORY OF LEARNING

To use VR as an immersive framework within which to implement a variety of training algorithms, it is necessary to characterize both the desired result and existing precedent for training paradigms. According to one theory of learning, optimal learning occurs when participants are engrossed in an activity, entering a

flow state characterized by a sense of temporal dilation and a cessation of self-awareness (Csíkszentmihályi, 1990). Thus, an optimal training model is one where subjects are continuously engrossed in the task.

The flow theory of learning has been used to model post-secondary student engagement with learning material (Shernoff et al., 2003), to analyze acquisition of spatial system knowledge (Smith, 2005), and to study the effect of learning through online or virtual systems (Liu, Liao, and Pratt, 2009; Almeida and Buzády, 2019; Huang, Backman, and Backman, 2010; Cheng, 2020). In this learning model, maintaining flow requires that a task be sufficiently challenging to stimulate learning (Liu, Liao, and Peng, 2005), but not so challenging so as to be overwhelming or so simple so as to cause inattention (Oliveira dos Santos et al., 2018). The narrow conditions require to maintain a state of flow can be represented as a channel between non-optimal combinations of challenge vs. expertise, as seen in Figure 2.1.

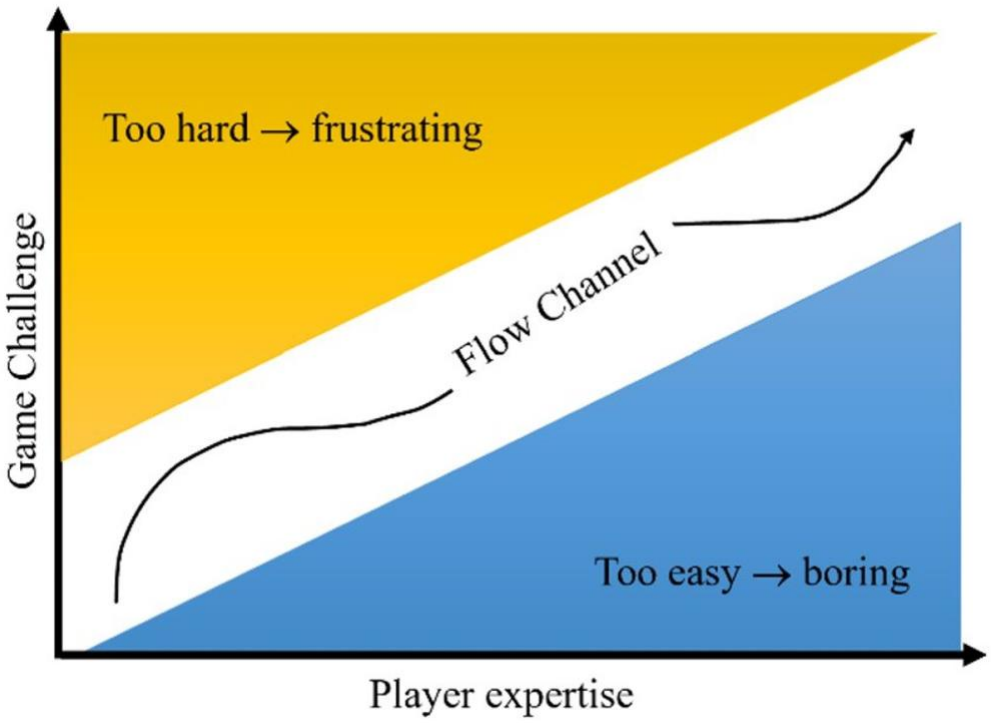


Figure 2.1: Flow channel as a function of task challenge and player expertise (Putman et al., 2022)

When applied to tasks requiring motor skills as an element, research suggests that practice is the most important factor for the "relatively permanent" improvement in skill performance (Carveth and Adams, 1964), that feedback plays a central role in reinforcement (Annett, 1969; Anderson, Magill, and Sekiya, 1999) and both perceptual-motor skill learning (Fitts, 1964; Marteniuk, 1976) and coordination/control (Newell, 1981; Salmoni, Schmidt, and Walter, 1984; Schmidt et al., 1999). Although acquired skill increases with practice, the challenge point framework suggests that training efficiency is increased by modulating difficulty to account for the level of performer. Thus, the complexity of the task and the

environment in regulating the learning potential during practice work in tandem, and automated adjustment of these components can enhance motor learning applied to a variety of skills when "optimal challenge" is met, including for rehabilitation (Descarreaux, Passmore, and Cantin, 2010; Onla-or and Winstein, 2008) and simulation-based surgical practice (Gofton, 2006).

Moreover, because the ratio of task challenge to expertise is dynamic (Choi, Kim, & Kim, 2007), the ideal training algorithm is able to respond to performance markers indicating that a subject has left the flow channel and modulate difficulty to re-attain optimality to balance challenge and learning. Dynamic difficulty adjustment (DDA) is a method of modifying a game or training regimen's features, behaviors, scenarios, or difficulty in real-time depending on player skill to maintain an optimal level of challenge or flow (Zohaib, 2018; Hunicke, 2005). In computer games, DDA has been used through real-time anxiety-based affective feedback (Liu et al., 2009; Xue et al., 2017) and through the use of AI to estimate player skill level (Silva, Silva, and Chaimowicz, 2015; Missura, 2015). When applied to training, DDA has been used to estimate skill level with heuristic value averages (Demediuk et al., 2018), with reinforcement learning (Lopes and Lopes, 2023), and with meta-learning algorithms using deep learning on small data sets (Moon and Seo, 2020).

Although each of these methods strives to maintain a flow state in users to maintain high levels of motivation and challenge, there is a gap in the literature concerning the use of subject performance as the sole input for dynamic difficulty adjustment in adaptive training.

2.3 ADAPTIVE TRAINING

The simplest training method is that of fixed difficulty, which has been found to result in higher improvement of performance compared to a linear fixed progression (Orvis et al., 2008). However, dynamic difficulty adjustment increases task engagement (Xue et al., 2017; Missura et al., 2009; Hunicke et al., 2009), is more easily usable (Benyon, 1993), and improves the experience and reported stimulation among subjects (Sampayo-Vargas et al., 2013; Constant et al., 2019; Lang et al., 2018). Moreover, previous research demonstrates that training outcomes are improved when practice is designed so that the task difficulty is appropriately matched to a performer's skill (Guadagnoli and Lee, 2010) and when there is variability in training conditions (Schmidt, 1975). Therefore, dynamically changing or modulating difficulty as a function of some predefined rule, or algorithm, is more effective than both high and low levels of unchanging difficulty.

One type of adaptive algorithm is an adaptive staircase modeled after the Parameter Estimation by Sequential Testing (PEST) method in signal detection theory. This requires a number of consecutive positive signal detections before reducing the salience of the signal (Taylor, 1967) and where a higher threshold minimizes false positive detections (Pollack, 1968). These kinds of algorithms employ the same PEST principle by requiring a certain number of satisfactory performances during training before increasing the difficulty, increasing the probability that the subject is able to perform well at increased difficulty and minimizing the risk of premature difficulty modulation (Levitt, 1971). A low threshold (e.g. changing

difficulty after only one successful performance) might lead to increasing the difficulty before the subject is ready, a premature modulation which would correspond to a false positive detection in PEST. However, a too-high threshold (e.g. changing difficulty after 3 successful performances) may lead to subject fatigue and a departure from the flow channel due to boredom at a stagnant difficulty level (Leek, 2001).

A common adaptive staircase that is frequently used is Two-Up/1-Down (2↑,1↓), where the user must perform well twice to increase difficulty, but must perform poorly only once for the difficulty to be decreased. Adaptive staircases were found to be more effective than both high and low fixed difficulty (Gabay, Karni, and Banai, 2017) and the Two-Up/1-Down (2↑,1↓) variant has been used for rehabilitation training in virtual environments (Grimm, Naros, & Gharabaghi, 2016). The One-Up/1-Down 1↑,1↓ staircase was used *de facto* in a variety of studies, including for neurorehabilitation (Cameirão et al., 2010), balance and gait training (Kumar et al., 2018; Koenig et al., 2011), training of fine motor movements (Saurav et al., 2018; Dhiman et al., 2016), and haptics-mediated attentional lengthening (Yang et al., 2016). Although both variants have been used for a variety of purposes across training modalities, there is no literature investigating the effect of a training algorithm's responsiveness to performance on the rate of skill acquisition, progression through training, or on skill transfer and performance outcomes.

Moreover, past investigations of training which involve multiple components typically design experiments such that esubjects are trained to proficiency on one

task at a time before progressing to new ones, including studies on military training (Gagne, 1962) and those investigating procedural learning with virtual collaborators (Rickel and Johnson, 1999 and 2010). However, an aspect of interest is training which involves multiple subtasks in parallel to accomplish the greater, or composite, task. There is no literature investigating the effect of subtask integration on the rate of skill acquisition, progression through training, or on skill transfer and performance outcomes.

Studies of motor skill training in virtual environments (VE) showed that subjects who trained virtually under a 1↑,1↓ paradigm displayed a significant improvement in performance compared to subjects who trained under a fixed progression paradigm, both virtually and physically (Gray, 2017). Moreover, subjects who trained adaptively in VE were found to display higher performance in a physical environment and, when reevaluated after 1 month, were found to retain higher performance compared to subjects who trained under fixed progression. However, the study was limited to the use of a projector screen rather than a head-mounted display (HMD) and focused purely on motor skills. There is no literature examining the acquisition and retention of complex task learning relevant to human spaceflight, namely tasks that have components of both motor learning and strategy and decision making.

Furthermore, a study on adaptive training in virtual reality for military medical skills used an Adaptive Control of Thought/Rational (ACT-R) cognitive architecture to model learning and forgetting in order to recognize skill deficiencies

in performance and adapt the training schedule accordingly (Siu et al., 2016). However, this system relied on kinematics and electromyography (EMG) to estimate individual cognitive, perceptual, and psychomotor states and workload, and was thus a system of psychophysiological adaptivity.

Literature on minimally invasive, performance-based adaptivity modulated by algorithms is scant, and there is no literature investigating the effect of unified versus discrete modulation of subtask difficulty in automated training algorithms, nor a rigorous examination or comparison of staircase threshold sensitivity on learning and performance outcomes. Furthermore, although dynamic difficulty adjustment has been explored, there is a need to better understand the efficacy of individually-adaptive, personalized paradigms and to demonstrate the feasibility of virtual reality as a medium for automated astronaut training on deep space missions.

In addition, previous studies of adaptive training used immersive VR for simple procedural tasks (Sampayo-Vargas et al., 2013; Constant and Levieux, 2019; Spiel et al., 2017), and where adaptive training is applied to complex operational tasks, it is typically done physically (Gray, 2017; Plass et al., 2019). Therefore, there is a need to investigate the efficacy of adaptive training in immersive VR on complex operational tasks. Moreover, theoretical and empirical adaptive training systems research has focused on aptitude-treatment interactions, macro and micro interactions, and two-step approaches to optimize engagement (Raybourn, 2007), but questions remain pertaining to how individual difference variables affect those

chosen for adaptation and the relative effectiveness of different adaptive training approaches (Landsberg et al., 2012).

This research addresses gaps in the literature surrounding adaptive training in immersive VR for complex operational tasks, the effect of unified versus discrete modulation of subtask difficulty in performance outcomes, and the effect of staircase threshold sensitivity on skill acquisition and performance. Furthermore, although dynamic difficulty adjustment has been explored, this study addresses a gap concerning the efficacy of using subject performance data as the primary input for a feedback mechanism, or algorithm, use to adapt difficulty for training. This research therefore focuses on investigating the efficacy of individually-adaptive, personalized training paradigms using performance metrics rather than obtrusive physiological measures and to demonstrate the feasibility of virtual reality as a medium for automated astronaut training on deep space missions.

CHAPTER III

METHODS

3.1 EXPERIMENTAL DESIGN

3.1.1 SCREENING AND CONDITION ASSIGNMENT

This experimental design was approved by the Institutional Review Board (IRB) at the University of Colorado, Boulder, under protocol #21-0349. A total of 48 subjects (24M/24F, ages 18-54, avg. 23.82 years) in good general health were recruited for participation in the study. Subjects were prescreened and excluded from the study if they scored above the 90th percentile on the Motion Sickness Susceptibility Questionnaire (Reason, 1968; Golding, 1998) to avoid the potential for motion sickness during VR training in highly susceptible individuals. Subjects were excluded if they reported having color blindness or vision uncorrectable to 20/20 to avoid confounds surrounding variability in the perception of the primary flight displays and their indicators. Subjects were also excluded if they reported consuming alcohol 6 or fewer hours prior to the study. Subjects completed a demographic survey, which included questions about prior piloting and flight experience and prior use of VR systems, and a reaction time test. These tests were designed to allow us to account for individual variability in statistical analyses of training and performance outcomes.

The experimental data collection was completed over 4 days, or sessions. During the first 3 sessions, subjects were trained to perform an entry, descent, and landing (EDL) task in virtual reality. Sessions were spaced 18-48 hours apart from

one another, and each session contained 10 training trials for a total of 30 trials. The difficulty of each subtask was modulated depending on the algorithm to which they were assigned. Subjects were randomly assigned to one of four training conditions: Two-Up/One-Down with Lockstep (2↑,1↓L), Two-Up/One-Down Unlocked (2↑,1↓UL), One-Up/One-Down with Lockstep (1↑,1↓L) (Locked), and Median Fixed Progression (MFP). Each of these algorithms is described in detail in section 3.2. For the final session, the subject performed For the final session, the subject performed the EDL task in the Aerospace Research Simulator (AReS), shown in Figure X. Difficulty of the task in the simulator was fixed at a level for which no subjects had trained.

For each session, subjects upon arrival reported their total hours of sleep from the previous night to account for individual variability in restedness in statistical analysis. Subjects also completed an Affect Grid (Russell & Mendelsohn, 1989; Killgore, 1998) before and after each session to provide information on induced changes in emotional state as a result of the training and testing sessions. At the conclusion of each trial, a modified Bedford Work Scale (BWS) survey (Roscoe & Ellis, 1990; Casner & Gore, 2010) was presented to subjects to measure cognitive workload. Following each session, subjects completed the System Usability Scale (SUS) (Peres, Pham, & Phillips, 2013; Vlachogianni & Teslios, 2020) and Flow Short Scale (FSS) (Yoshia et al., 2013) surveys to provide self-reported information on degree of task challenge and ease of system use. In accordance with this experimental design, the following training groups were formed, as shown in Table 3.1:

Group	Training Algorithm	Subtask Modulation	Adaptivity	Sex	Total	Age
1	2↑,1↓	Locked	Adaptive	4M/4F	8	22-32 (25.38)
2	2↑,1↓	Unlocked	Adaptive	4M/4F	8	18-35 (22.38)
3	1↑,1↓	Locked	Adaptive	4M/4F	8	18-54 (27.13)
4	Median Fixed Progression	Locked	Non-Adaptive	4M/4F	8	18-25 (20.38)

Table 3.1: Summary of subject distribution across training groups

3.1.2 VIRTUAL TRAINING SIMULATOR

A training simulator was developed for EDL of a spacecraft on Mars. It was designed to emulate the Lunar Landing Training Vehicle (LLTV), a physical demo vehicle used by NASA to train Apollo astronauts to throttle an array of maneuvering thrusters to land on the lunar surface (Hatch, Pennington, and Cobb, 1967), considered the gold standard for training astronauts to maneuver and land spacecraft (Engle, 2012). While a mock-up vehicle replicates real flight dynamics, operational controls and interfaces, and evokes a realistic stress response by imparting the sensations of motion, they are expensive, can be extremely dangerous, require staff support, and cannot be easily scaled or modified (Brady and Paschall, 2010; NASA DFRC, 2004). A virtual simulator, by contrast, can replicate flight dynamics, operational controls and interfaces, and evoke stress responses while removing the dangers, cost, and operational complexity intrinsic to past trainers. Moreover, unlike mock-up vehicles which have an all-or-nothing binary approach to performance, they

can adjust difficulty. Most pertinently, virtual simulators can modulate difficulty autonomously, allowing for recurrent, unfacilitated astronaut training during LDEM. A detailed description of the virtual EDL simulator and its development can be found in Putman et al., 2022.

The simulator was designed to train subjects through three subtasks: 1) landing site selection, where the user selects a landing site located centrally between a variable number of randomly distributed sites of scientific interest (SSI) within a topological map of Mars terrain using the cursor on a joystick, 2) piloting, where the user must manually control the spacecraft's pitch and roll to navigate to the landing site location using a joystick and a guidance cue on the primary flight display despite simulated wind perturbations, and 3) landing burn, where the user must use a hand-thruster to control the descent velocity given a limited amount of propellant. The associated displays for the three subtasks are shown in Figures 3.1 and 3.2:

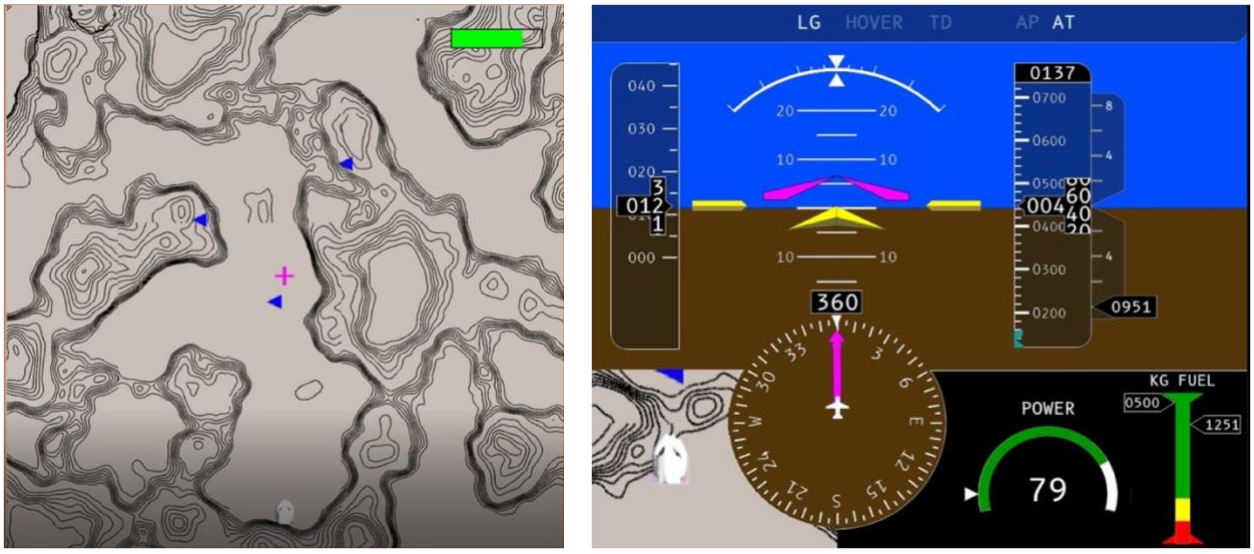


Figure 3.1: (Left) Topological map on the secondary flight display during the Landing Site Selection (LS) subtask; (Right) Primary flight display with guidance cue, altimeter, velocity meter, fuel gauges, flight vector indicator, and mini-map for the Manual Control (MC) subtask

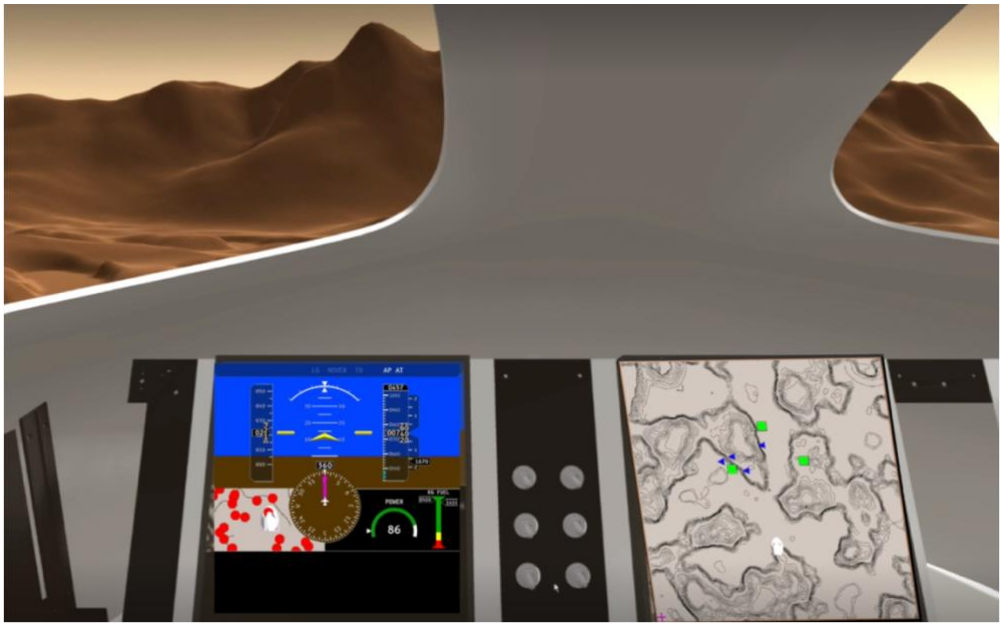


Figure 3.2: View of the virtual cockpit with primary and secondary flight displays and generated Martian landscape during the Terminal Descent (TD) subtask

Each of the subtasks had 24 possible levels of difficulty (1 – 25) with level 18, the fixed difficulty of the AReS cockpit, being skipped during training to ensure a novel difficulty for all subjects. A head-mounted display (HMD, HTC Vive Pro) was used to project the simulated interior of the AReS spacecraft cockpit mock-up to subjects during training (Figure 3.3). The virtual displays and cockpit environment were designed to emulate those of AReS, the physical cockpit mock-up. Subjects used a physical joystick and hand-thruster to perform tasks (Figure 3.4), and all physical inputs to both were recorded in a server in addition to performance data.

Figure 3.3: Subject wearing head-mounted display during virtual training

Figure 3.4: Hand-thruster (right) and joystick (left) used during EDL subtasks

3.1.3 Cockpit Mock-Up Testing

Following completion of the three training sessions, subjects performed ten trials in the Aerospace Research Simulator (AReS) spacecraft cockpit mock-up located in the University of Colorado, Boulder's Aerospace Engineering Sciences building to assess skill transfer from the virtual to an analogous, high-fidelity physical environment. The AReS mock-up is shown in Figure 3.5, and a view of the cockpit interior is shown in Figure 3.6. Each of the subtasks was fixed at a difficulty of level 18 across trials regardless of subject performance.

Figure 3.5: View of the AReS cockpit mock-up in the Bioastronautics High Bay at the University of Colorado, Boulder, with external monitors and controls visible

Figure 3.6: Subject performing EDL subtasks in AReS cockpit mock-up

Subjects in each of the four training conditions went through the same 3 session training paradigm and 1 session cockpit test. The initial difficulty level for all three subtasks during training was 12, while the difficulty level for all subtasks in the cockpit was fixed at 18 through all 10 sessions. Figure 3.7 displays the experimental design graphically.

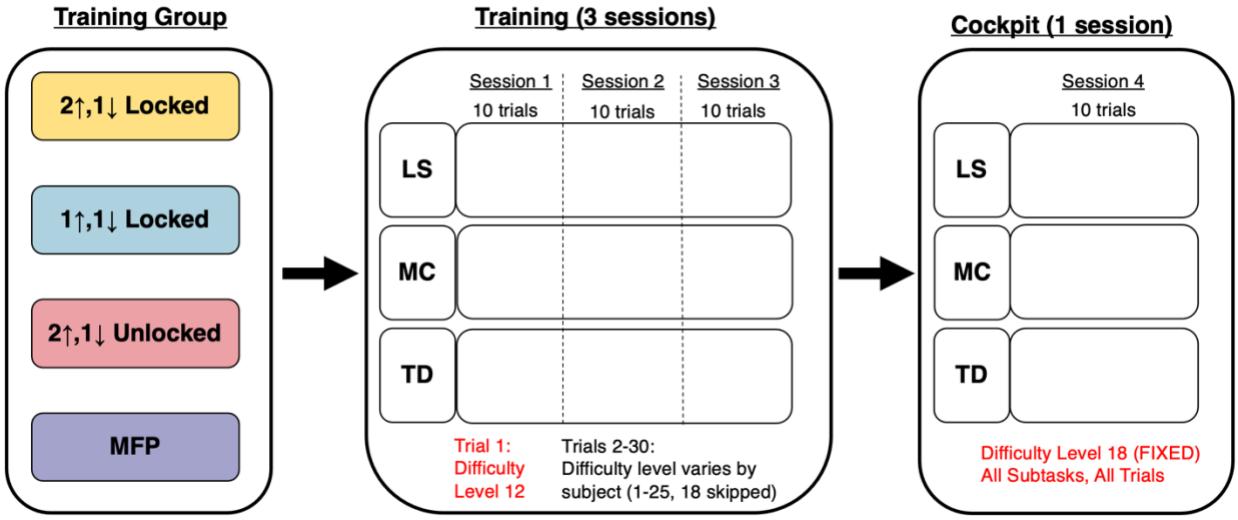


Figure 3.7: Graphic of experimental design, including training and cockpit sessions

3.1.4 Performance Grading

The inability for subjects to perceive how changes in difficulty are calculated or executed renders automated training algorithms more effective (Andrade et al., 2005). Moreover, simplified performance feedback in the form of qualitative grading rather than returning numerical values is more readily intuitive, improves intrinsic motivation, and increases skill acquisition (Vollmeyer & Rheinberg, 2005; Wilson et

al., 2017). Thus, a trivariate grading system was developed that scored performance as either excellent, adequate, or poor. These results were displayed to subjects at the conclusion of each trial for each subtask with corresponding green, yellow, and red color schema, respectively, to facilitate comprehension, as seen in Figure 3.8.

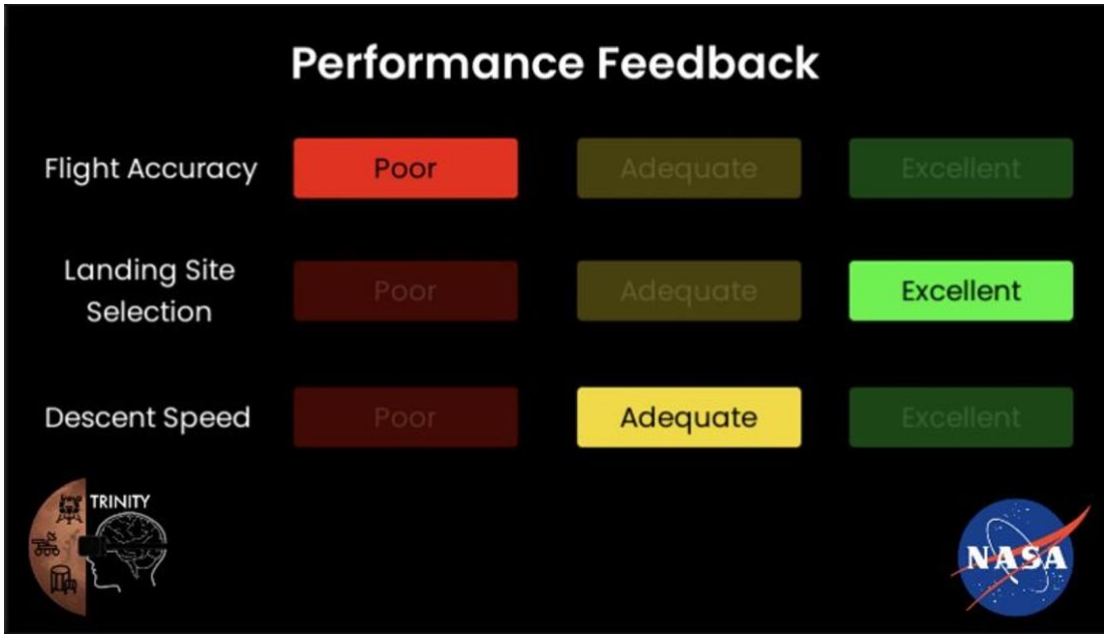


Figure 3.8: Post-trial performance feedback screen for EDL subtasks

Since subject performance falls on a spectrum, thresholds to demarcate excellent and poor performance were developed by means of pilot testing. These thresholds were dependent upon subtask difficulty and become increasingly stringent at higher difficulty levels, as described in the following sections.

I. LANDING SITE SELECTION

Randomly generated Martian surface features include denser topological lines at higher difficulty levels. The number of sites of scientific interest (SSI) increases at higher difficulty levels. The location closest to the calculated SSI centroid which is under an 10% terrain steepness threshold is selected as the ideal location by the computer. The distance of the user-selected site compared to the ideal location is used to score performance. At low difficulty levels, three possible landing sites are automatically displayed, and the subject must choose the most ideal site. For all other difficulty levels, selection occurs freely over the displayed map. Selection of a site that is distant from the SSI centroid or that is located on terrain exceeding 15% steepness is graded as poor, and the latter is specifically demarcated as a crash if the steepness exceeds 20%. Subjects are given 8 seconds to select a landing site, with a visual timer present on the upper right side of the flight display. If subjects fail to select a landing site before the 8 seconds elapse, a site is automatically selected to enable task continuation, but the landing site selection subtask is recorded as a crash.

II. MANUAL CONTROL

A navigation guidance cue shown on the primary flight display (PFD) follows a flight path calculated as a function of the selected landing site location. The subject executes pitch and roll commands to align a triangle, representing instantaneous spacecraft orientation, with the guidance gue. Subject joystick inputs are used to determine deviation from the ideal pitch and roll commands by means of a root-mean-square deviation (RMSD). Higher RMSD values lead to poor performance grading,

and the threshold becomes more stringent at higher difficulty levels. Wind perturbations randomly applied to the spacecraft cause deviations from the guidance cue which require correction in both the pitch and roll axes. Wind perturbations are first introduced at difficulty level 7, and the frequency and amplitude, or severity, of such perturbations increase at higher difficulty levels. In this phase of flight, the rate of descent is controlled autonomously.

Difficulty in this subtask is adjusted by changing the manual control requirements (pitch only vs. pitch and roll), the amount of fuel allotted for piloting, the magnitude, frequency, and directionality of wind perturbations on the spacecraft, and the amount of time with the guidance cue disabled where subjects were required to pilot without it. Subjects used a flight display with information on ground speed, altitude, vertical descent rate, a miniature version of the topographic map, and a vector of spacecraft velocity to aid them. If piloting commands continually and greatly differ from the computed flight path and guidance cue, the subject will burn through a finite amount of propellant used for the piloting subtask without reaching the desired landing site destination. Such a trial leads to a poor performance that is specifically demarcated as a crash.

III. TERMINAL DESCENT

In the final subtask, subjects must use a hand-thruster to modulate the thrust of a descent engine to descend in altitude and touch-down at a velocity lower than 120 ft/min. The amount of propellant for the descent engine diminishes at higher

difficulty levels, requiring more aggressive thrust modulations. At the highest difficulty levels, the ideal descent profile (which optimizes fuel use) is one which cuts thrust to initiate free-fall before applying maximal thrust in the seconds before vehicle descent speed exceeds a threshold at which the vehicle can no longer be decelerated before impact, modulated to reach zero velocity at the instant where the spacecraft reaches zero altitude. The selection of a landing site on complex terrain creates uncertainty in the final landing altitude, requiring trial and error in the range of 0-20 ft in which touchdown may occur. Touchdown at a velocity exceeding 200 ft/min is graded as a poor performance and is specifically demarcated as a crash. This may occur as a result of poor thrust modulation, or by consuming all available fuel at an appreciable altitude.

3.2 ALGORITHM IMPLEMENTATION

Four different training conditions were developed. These can be broadly divided into two groups, adaptive and non-adaptive. Non-adaptive algorithms hold difficulty across subtasks fixed at predetermined levels irrespective of performance. By contrast, adaptive algorithms alter subtask difficulty as a function of subject performance across a range of disparate but interconnected subtasks. The core facet of adaptivity is the use of human performance to close the feedback loop of automated difficulty modulation.

3.2.1 ADAPTIVE ALGORITHMS

I. Two-Up/One-Down (2↑,1↓)

The first adaptive progression takes the form of Two-Up/One-Down (2↑,1↓), a fixed linear response where difficulty is quantized and can both ascend and descend by linear increments of one. As mentioned in Chapter 2, this staircase is modeled on the PEST method for signal detection, whereby the strength of a signal is diminished after successive correct detections of a stimulus. A higher number of required correct detections increases fidelity, with diminishing effect. Thus, when applied to training paradigms, subjects in the 2↑,1↓ staircase are required to manifest excellent performance on a subtask twice at the same level of difficulty and in succession before that difficulty is modulated up by one level. Conversely, subjects who perform poorly just once on a subtask will have the algorithm modulate the difficulty down by one level for that subtask. The staircase is fixed throughout the training, and the step-sizes are fixed at one.

II. One-Up/One-Down (1↑,1↓)

Another variant of adaptive progression is the One-Up/One-Down (1↑,1↓) staircase, a fixed linear response where difficulty is again quantized and can both ascend and descend by linear increments of one. In 1↑,1↓, the threshold for upward progression is a single excellent performance on a subtask, rendering the variant more sensitive to subject performance. The threshold for downward progression is a single poor performance, as in the 2↑,1↓ variant. The staircase is fixed throughout the training, and the step-sizes are fixed at one. A comparison of the two progressions is displayed in Figure 3.9, and a visualization of the way in which adaptive staircases

can modulate difficulty to remain within the hypothesized flow channel is shown in Figure 3.10 for varying levels of responsiveness.

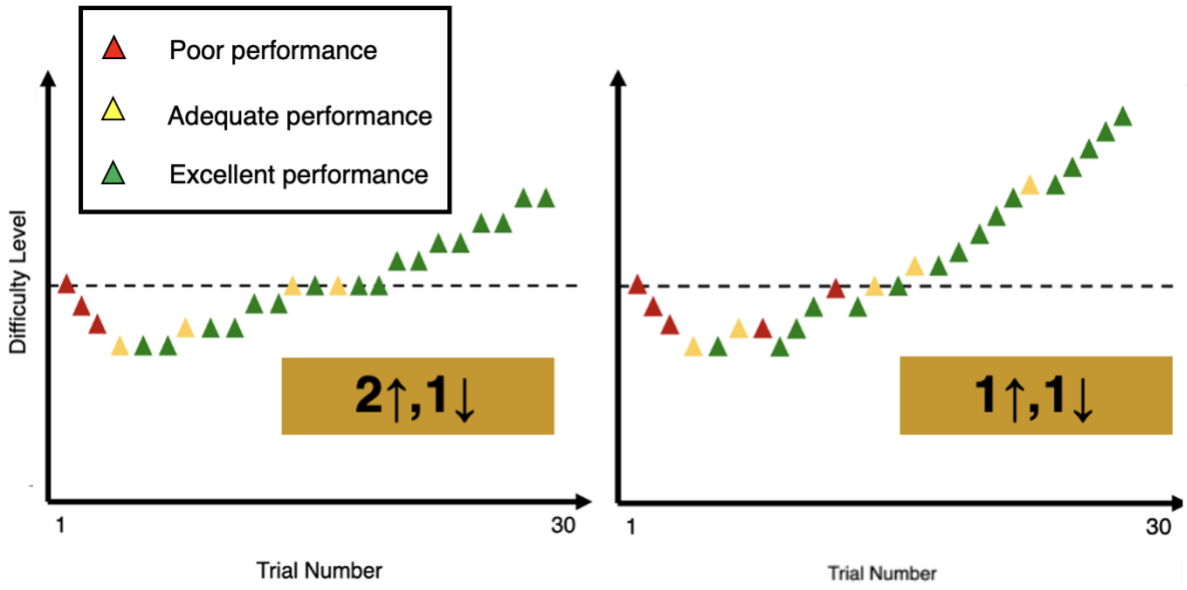


Figure 3.9: Example of difficulty progression for 2↑,1↓ (left) and 1↑,1↓ (right)

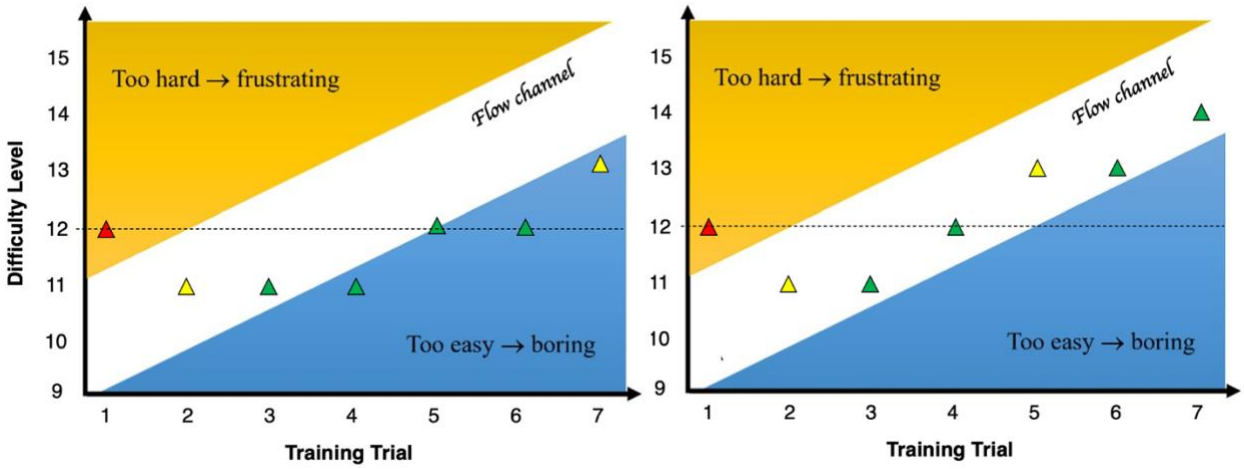


Figure 3.10: Staircase progressions for 2↑,1↓ (left) and 1↑,1↓ (right) overlaid on an example flow channel, with higher responsiveness better able to maintain flow

III. Lockstep

Lockstep describes the inhibition of upward modulation on one or more subtasks by poor subject performance on at least one subtask. This serves to prevent asymmetric learning by requiring that subtask modulation occurs within +/- 1 level of synchrony. In training paradigms with an adaptive staircase, once the difficulty for one of the subtasks is decreased to two or more levels below the other subtasks, lockstep is triggered, preventing the 2↑,1↓ and 1↑,1↓ algorithms from applying their staircases nominally except to decrease difficulty after poor performance. If downward modulation occurs on a subtask which is not driving lockstep, difficulty is free to return to the prior level under the requirements of the staircase (e.g. two successive excellent performances) but may not exceed the level of difficulty at the time at which lockstep was first triggered.

By contrast, an unlocked staircase allows for discrete, mutually-independent modulation of subtask difficulty. It assumes that although subtasks are sequential and thematically interconnected, they require disparate skills and are likely to incur varying levels of propensity between subjects. When the 2↑,1↓ and 1↑,1↓ algorithms are unlocked, there are three staircases functionally operating in parallel, modulating difficulty according to subject performance for individual subtasks. This allows for asymmetric progression on the basis that subjects will learn more effectively at varying levels of challenge across subtasks.

3.2.2 NON-ADAPTIVE ALGORITHMS

I. Median Fixed Progression

Median Fixed Progression (MFP) is a non-adaptive, fixed progression based on the median difficulty level across subtasks incurred by subjects in a baseline condition. The composite was formed using data from subjects trained with the 2↑,1↓ algorithm and with lockstep enabled (2↑,1↓ Locked). MFP mimics the progression characteristics of adaptivity without responding to individual subject performance. It serves to isolate the effect of adaptivity on subjects with performance data, and thus training needs, which differ from the average, either because of exceptional ability or unique difficulty in skill acquisition. The MFP condition captures the initial decline in difficulty across subtasks as subjects familiarize themselves with subtasks and associated controls, as well as the eventual and gradual increase in difficulty as subjects become familiar with controls and begin honing particular motor skills.

3.3 STATISTICAL METHODS

This research investigates the efficacy of different training algorithm features using training in virtual reality (VR) by altering task difficulty as a function of subject performance across a range of disparate but interconnected subtasks. It is hypothesized in this study that personalized training algorithms which adapt task difficulty to subject performance and which possess high levels of responsiveness and integration have improved outcomes in skill acquisition during training, increased skill transfer between the virtual and physical environments, and improved final performance in a physical cockpit mock-up. The three features of study are visualized in Figure 3.7:

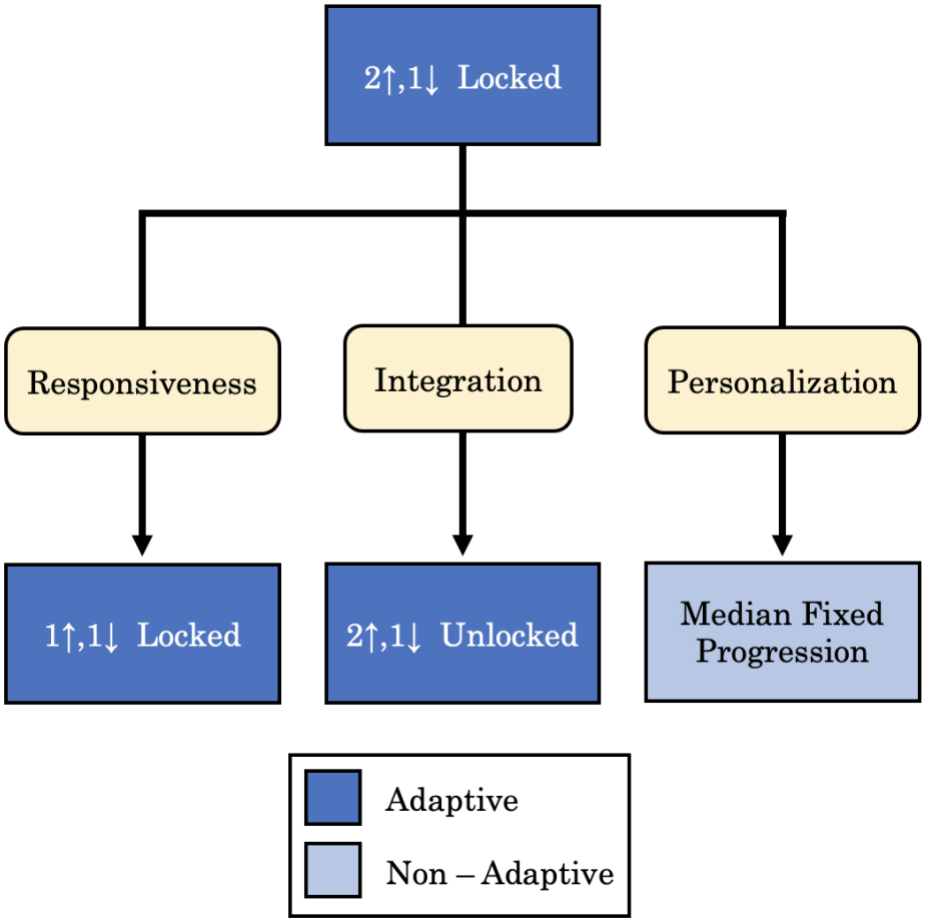


Figure 3.11: Schematic of training algorithms and associated variables of interest

To test the hypotheses surrounding the effect of responsiveness, integration, and personalization on outcomes in both training and the physical cockpit environment, a range of statistical tests and associated post-hoc tests were established to investigate each of the variables of interest in both the training phase, as listed in Table 3.1, and in the AReS cockpit mock-up, as listed in Table 3.2:

<table>
<tr><th>Training Trial</th><th colspan="2">Dependent Variable</th><th colspan="2">Independent Variable(s)</th><th>Main Test</th><th>Post-Hoc Test</th></tr>
<tr><td rowspan="4">1-30</td><td colspan="2" rowspan="2">Difficulty Level
(1 – 25)</td><td>Between:

Training Algorithm*</td><td>Within:

Session
(1, 2, 3)</td><td rowspan="2">Mixed-Effects ANOVA</td><td rowspan="2">Dunnett’s Test</td></tr>
<tr><td rowspan="2">Skill</td><td rowspan="2">All
(-1 to +1)</td><td>Between:

Training Algorithm</td><td>Within:

Session
(1, 2, 3)</td></tr>
<tr><td>Low
(-1 to 0)</td><td colspan="2">Training Algorithm</td><td>Welch’s ANOVA</td><td>Tukey’s Range Test</td></tr>
<tr><td rowspan="3">Performance</td><td>3 Excellent</td><td colspan="2" rowspan="3">Training Algorithm</td><td rowspan="3">Kruskal-Wallis H-Test</td><td rowspan="3">Dunn’s Test</td></tr>
<tr><td># of Crashes</td></tr>
<tr><td>Total
(-1, 0, 1)</td></tr>
<tr><td rowspan="5">30</td><td colspan="2">Difficulty Attained
(1 – 25)</td><td colspan="2" rowspan="5">Training Algorithm</td><td rowspan="5">Kruskal-Wallis H-Test</td><td rowspan="5">Dunn’s Test</td></tr>
<tr><td colspan="2">Skill Attained
(-1 to + 1)</td></tr>
<tr><td rowspan="3">Performance</td><td>3 Excellent</td></tr>
<tr><td># of Crashes</td></tr>
<tr><td>Total
(-1, 0, 1)</td></tr>
</table>

*Included: 2↑,1↓ Locked, 1↑,1↓ Locked, 2↑,1↓ Unlocked; Excluded: MFP

Table 3.2: Statistical methods for evaluating training outcomes

<table>
<tr><th>Cockpit Trial</th><th colspan="2">Dependent Variable</th><th colspan="2">Independent Variables(s)</th><th>Main Test</th><th>Post-Hoc Test</th></tr>
<tr><td rowspan="4">1</td><td rowspan="3">Performance</td><td>3 Excellent</td><td colspan="2" rowspan="4">Training Algorithm</td><td rowspan="4">Kruskal-Wallis H-Test</td><td rowspan="4">Dunn’s Test</td></tr>
<tr><td># of Crashes</td></tr>
<tr><td>Total
(-1, 0, 1)</td></tr>
<tr><td colspan="2">Skill
(-1 to +1)</td></tr>
<tr><td rowspan="5">1-10</td><td rowspan="3">Performance</td><td>3 Excellent</td><td colspan="2" rowspan="3">Training Algorithm</td><td rowspan="3">Kruskal-Wallis H-Test</td><td rowspan="3">Dunn’s Test</td></tr>
<tr><td># of Crashes</td></tr>
<tr><td>Total
(-1, 0, 1)</td></tr>
<tr><td rowspan="2">Skill</td><td>All
(-1 to +1)</td><td>Between:

Training Algorithm</td><td>Within:

Trial
(1 – 10)</td><td>Mixed-Effects ANOVA</td><td>Dunnett’s Test</td></tr>
<tr><td>Low
(-1 to 0)</td><td colspan="2">Training Algorithm</td><td>Welch’s ANOVA</td><td>Tukey’s Range Test</td></tr>
</table>

Table 3.3: Statistical methods for evaluating AReS cockpit outcomes

Here, skill refers to subject performance normalized by subtask difficulty and is a continuous range between -1 and +1. A skill of +1 indicates perfect performance at the highest difficulty, a skill of 0 indicates adequate performance at a medium difficulty level, and a skill of -1 indicates poor performance, or a crash, at the lowest difficulty levels, with scores varying between these markers. Low skill refers to any

negative integer score and was assessed separately as a key metric for identifying disparities in trained or attained skill, in addition to tests across all skill grades.

A mixed-effects analysis of variance omnibus test was used to test for differences between training groups (fixed effect) and within training session or trial (random effect) on a range of variables, including trained difficulty level, skill, and performance. Mixed-design ANOVA was selected because the dependent variables were continuous repeated measures across trials in both training and the AReS cockpit mock-up. For each dependent variable, a different ANOVA was run for each of the 3 subtasks and for the 3 subtasks summed togther into an integrated measure. The mixed-effects ANOVA was used to identify significant differences between groups, within sessions or trials, and to identify significant interactions. Moreover, unlike other analyses of variance, two-way mixed-effects ANOVAs have been determined to be robust against outliers and normality (Schmider et al., 2010; Milligan, Wong, and Thompson, 1987; Mair and Wilcox, 2020), making it ideal for subject data with, for instance, a high incidences of crashes or a large number of difficulties at level 12, from which all subjects began. Dunnett's test was used post-hoc as a multiple comparison procedure to isolate training groups with significant differences using 2↑,1↓ Locked as the control condition.

The residuals of each ANOVA were used to check that parametric assumptions were met. Grubbs' test was used to identify outliers, the Shapiro-Wilk test and Q-Q plots were used to assess normality, and Levene's test was used to assess homogeneity of variance. For difficulty level, a continuous variable ranging from 1 –

25, residuals for each of the three were normally distributed, and outliers were sufficiently few that they were not removed from the training data, which was taken only from adaptive training groups (see Appendix B). For skill level, a continuous variable ranging from 0 – 1, residuals for MC were normally distributed, but both LS and TD had outliers which caused the distribution to violate normality and homogeneity of variance (see Appendix B). Since the preponderance of outliers were the result of subjects who had crashed during LS or TD, trials containing a crash were removed from the training data and analyzed separately.

A Kruskal-Wallis H test by ranks was used as a non-parametric method to test for differences between training groups on categorical dependent variables such as performance (graded as -1, 0, 1), number of trials where the scores were "Excellent" for all three subtasks, and number of trials with a crash recorded. When significant differences were found, Dunn's test for non-parametric pairwise comparisons was used post-hoc to isolate the training groups with significant differences in the dependent variable for both a single trial or sum of all trials (Dinno, 2015).

Finally, Welch's ANOVA for unequal variances was used to determine significant differences between training groups on continuous metrics such as low skill (-1 to 0), since groups had unequal numbers of trials where subjects were given skill grades beneath a certain value. Tukey's Range Test was used post-hoc to identify which particular training groups had significant differences in the dependent variable across all trials.

CHAPTER IV

RESULTS

4.1 Training Results

4.1.1 Difficulty

A two-way mixed-effects ANOVA was conducted, with adaptive training algorithm (2↑,1↓ L vs. 1↑,1↓ L vs. 2↑,1↓ UL) as the between-subject independent variable, training session (1, 2, or 3) as the within-subject independent variable, and subject as the blocking factor. On the LS subtask, there was a significant main effect of training algorithm on differences in difficulty, $F(2, 21) = 5.065$, $p < 0.05$, $\eta 2 = 0.325$. A post-hoc Dunnett's test with the baseline training algorithm (2↑,1↓ L) as the control showed that subjects in the 1↑,1↓ L training group had significantly higher difficulty progressions in LS across training sessions compared to subjects in the 2↑,1↓ L training group ($p < 0.05$), and subjects in the 2↑,1↓ UL training group also had significantly higher difficulty progressions in LS compared to 2↑,1↓ L ($p < 0.001$). The mean difficulty progressions are shown in Figure 4.1, and the distributions of difficulty across subtasks are plotted in Figure 4.2.

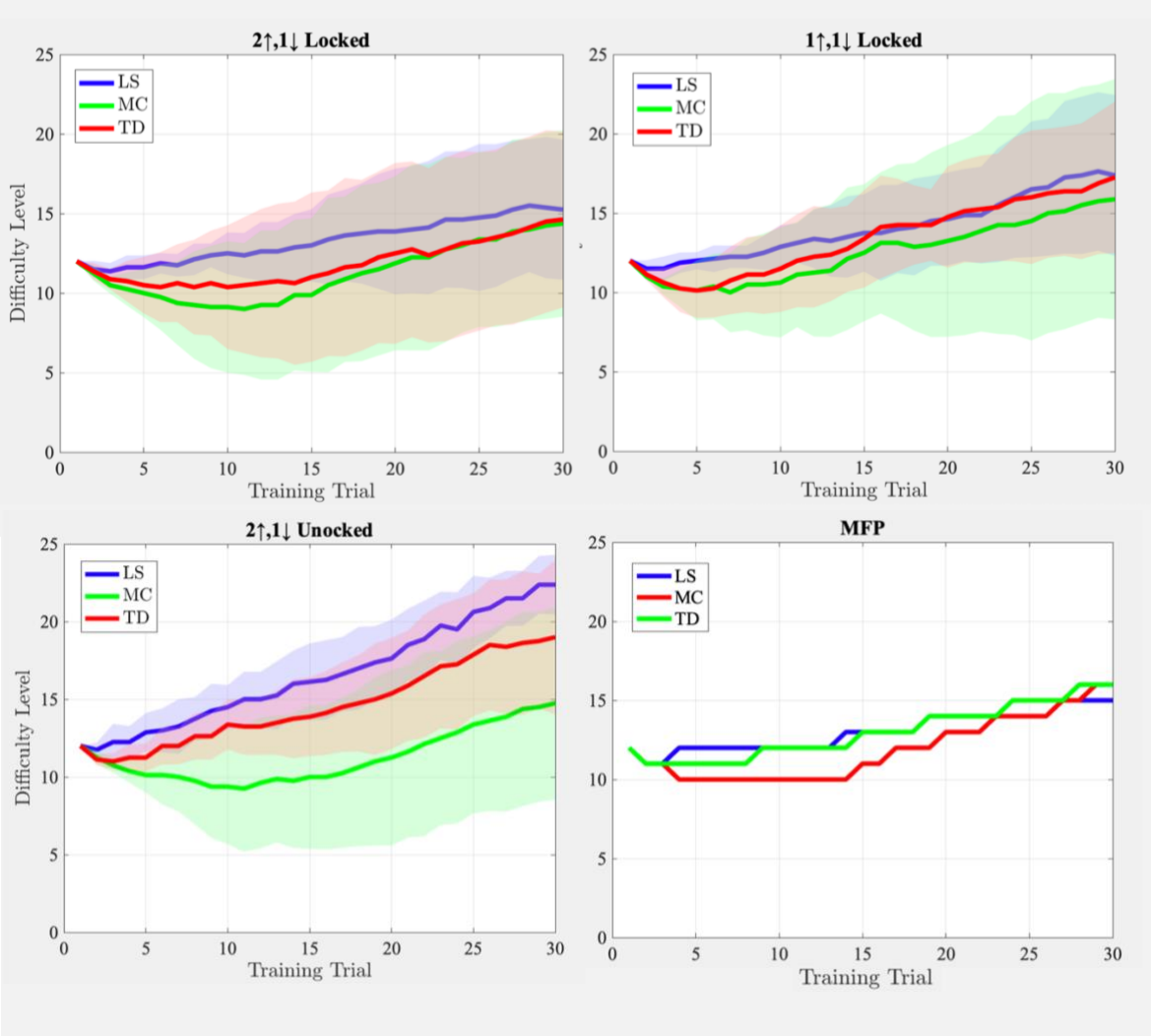


Figure 4.1: Difficulty progressions on EDL subtasks across training algorithms (mean: solid line, 95% CI: shaded region)

There was also a significant main effect of training session on the LS subtask, $F(2, 42) = 50.130$, $p < 0.001$, $\eta 2 = 0.705$, and a significant interaction between training algorithm and session, $F(4, 42) = 3.846$, $p < 0.01$, $\eta 2 = 0.268$. Additionally, there was a significant main effect of training session for the MC subtask, $F(2, 42) = 13.712$, $p < 0.001$, $\eta 2 = 0.395$, for the TD subtask, $F(2, 42) = 35.376$, $p < 0.001$, $\eta 2 = 0.627$, and for the integrated average of all subtasks, $F(2, 42) = 34.179$, $p < 0.001$, $\eta 2 = 0.619$.

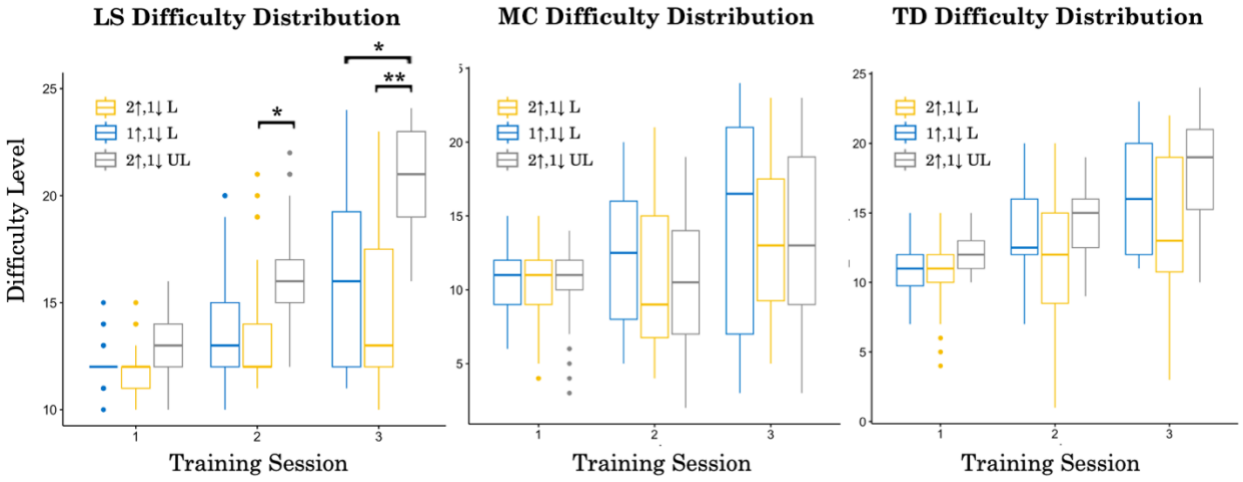


Figure 4.2: Difficulty distributions among adaptive training groups across subtasks

A Kruskal-Wallis *H*-Test by ranks found a significant difference between adaptive training groups on difficulty attained on the 30th trial in LS, $\chi 2(2) = 9.031$, $p < 0.05$. A post-hoc Dunn's test showed that subjects in the 2↑,1↓ UL training group attained significantly higher difficulty levels in LS than subjects in the 2↑,1↓ L group ($p < 0.01$). The difference in attained difficulty can be seen in Figure 4.3.

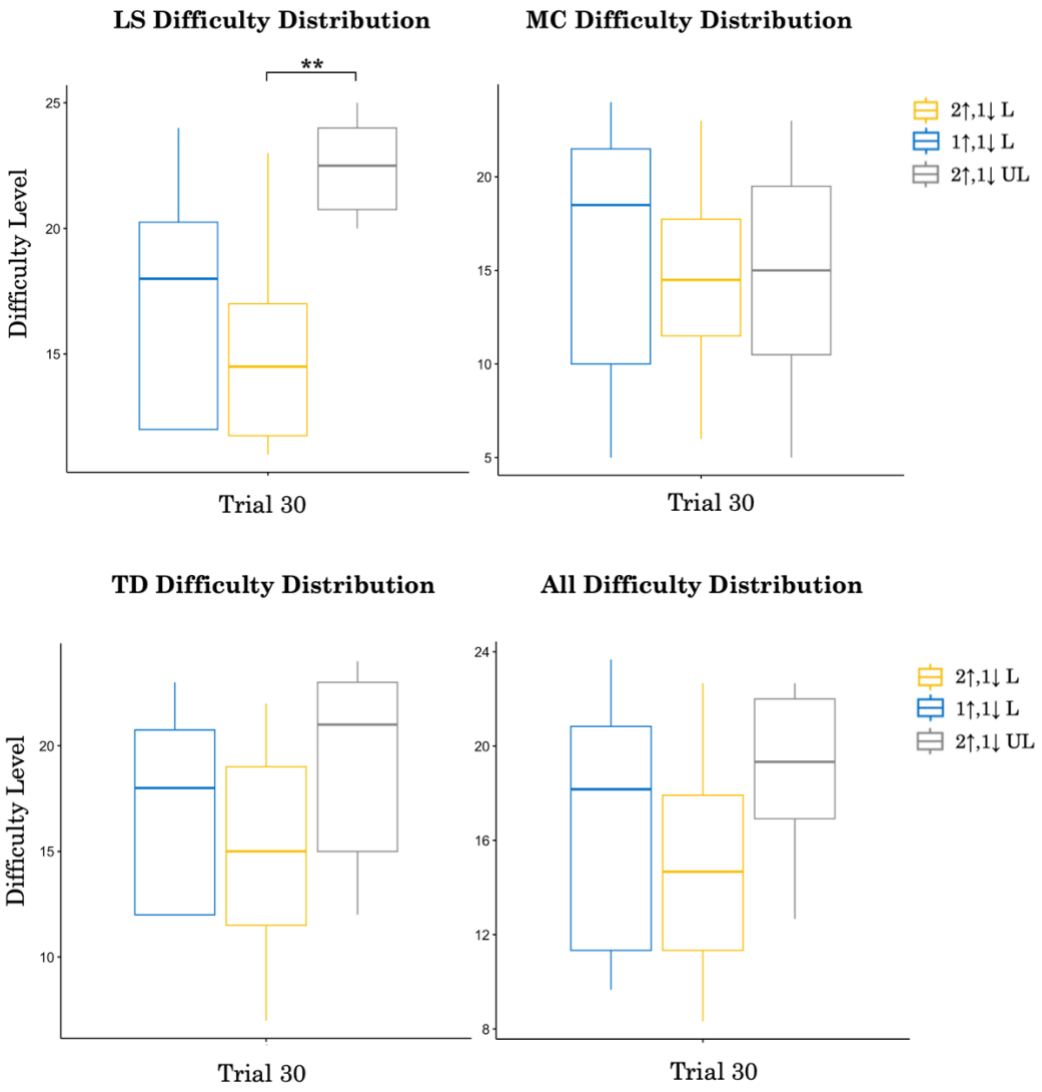


Figure 4.3: Attained difficulty distribtutions among adaptive training groups

The results of the mixed-effects ANOVA for differences in difficulty progression between adaptive training groups and across the 3 training sessions for each subtask are tabulated below in Table 4.1 along with associated post-hoc Dunnett's test results.

Subtask	Source	F	p	η2	p (Post-hoc)
LS	Group	F(2,21) = 5.065	0.016 *	0.325	p = 0.0288 * (B vs. 2↑,1↓ UL) p < 0.0001 *** (B vs. 1↑,1↓ L)
	Session	F(2,42) = 50.13	<0.0001 ***	0.705	
	Interaction	F(4,42) = 3.846	0.009 **	0.268	
MC	Group	-	NS	-	-
	Session	F(2,42) = 13.71	<0.0001 ***	0.395	-
	Interaction	-	NS	-	-
TD	Group	-	NS	-	-
	Session	F(2,42) = 35.38	<0.0001 ***	0.627	-
	Interaction	-	NS	-	-
All	Group	-	NS	-	-
	Session	F(2,42) = 34.18	<0.0001 ***	0.619	-
	Interaction	-	NS	-	-

Blocked subjects, NS: Not significant, B: Baseline, *p < 0.5, **p < 0.01, ***p < 0.001

Table 4.1: Results of mixed-effects ANOVA on difficulty progression (sessions 1-3)

The results of the Kruskal-Wallis tests for differences in attained difficulty on the 30th training trial between adaptive training groups for each subtask are tabulated below in Table 4.2 along with results of the post-hoc Dunn's test.

Subtask	$\chi 2$	p	p (Post-hoc)
LS	$\chi 2(2) = 9.0312$	0.011 *	p = 0.01 ** (B vs. 2↑,1↓ UL)
MC	-	NS	-
TD	-	NS	-
All	-	NS	-

Table 4.2: Results of Kruskal-Wallis on attained difficulty (training trial 30)

4.1.2 Skill

Welch's ANOVA for unequal sample sizes found a significant difference between all training groups on the number of subjects who demonstrated low (-1 to 0) skill across all training trials on the LS subtask, $F(3, 76.332) = 8.688$, $p < 0.001$, $\eta 2 = 0.151$. Tukey's post-hoc test showed that subjects in MFP had significantly higher rates of low skill on LS across training trials ($p < 0.01$, 95% CI [0.0883, 0.5299]). The distributions of subtask skill for all training groups is plotted in Figure 4.4.

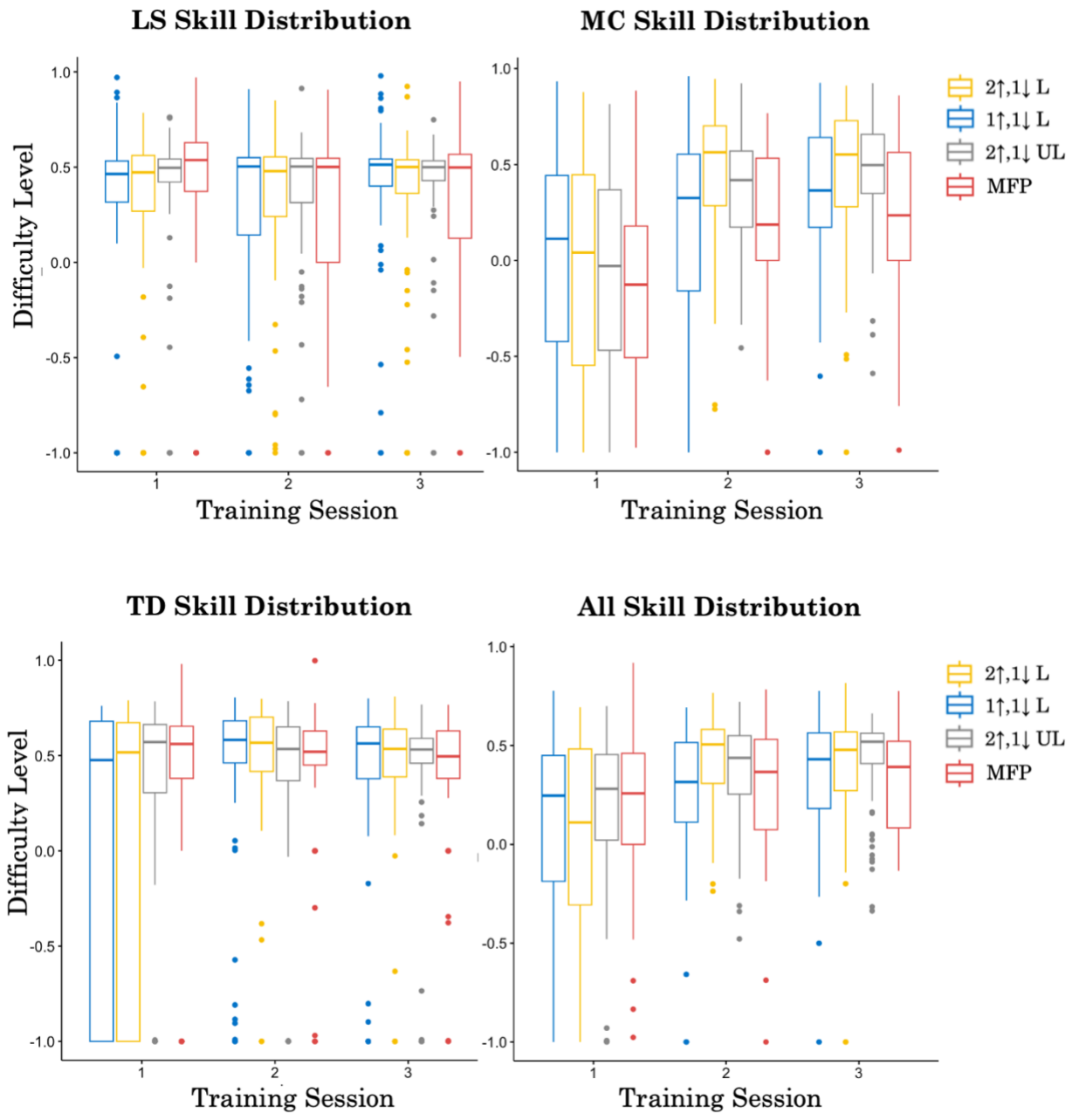


Figure 4.4: Skill distributions across all training groups and subtasks

The test also found a significant difference between all training groups on the number of subjects who demonstrated low skill across all training trials on the MC subtask, $F(3, 140.67) = 3.676$, $p < 0.05$, $\eta 2 = 0.039$. Tukey's post-hoc test showed that

subjects in MFP had significantly higher rates of low skill on MC across training trials ($p < 0.01$, 95% CI [0.035, 0.2993]).

Finally, Welch's ANOVA also found a significant difference between all training groups on the number of subjects who demonstrated low skill across all training trials on the TD subtask, $F(3, 84.6) = 12.069$, $p < 0.001$, $\eta 2 = 0.271$. Tukey's post-hoc test showed that subjects in MFP had significantly higher rates of low skill on TD across training trials ($p < 0.001$, 95% CI [0.2413, 0.6144]).

When taking into account only the 30th (final) training trial, a Kruskal-Wallis *H*-Test by ranks found a difference between all training groups on attained skill in MC which approached significance, $\chi 2(3) = 7.377$, $p = 0.061$. Moreover, a Kruskal-Wallis *H*-Test by ranks found a difference on the integrated average of skill across subtasks, which also approached significance, $\chi 2(3) = 7.119$, $p = 0.068$.

The results of the mixed-effects ANOVA for differences in total skill between training groups and across the 3 training sessions are tabulated below in Table 4.3 for each subtask.

Subtask	**Source**	**F**	**p**	**η2**	**p (Post-hoc)**
LS	Group	-	NS	-	-
	Session	-	NS	-	-
	Interaction	-	NS	-	-
MC	Group	-	NS	-	-

	Session	F(2,56) = 53.42	<0.0001 ***	0.395	-
	Interaction	-	NS	-	-
TD	Group	-	NS	-	-
	Session	F(2,56) = 8.715	0.001 ***	0.627	-
	Interaction	-	NS	-	-
All	Group	-	NS	-	-
	Session	F(2,56) = 26.44	<0.0001 ***	0.619	-
	Interaction	-	NS	-	-

Table 4.3: Results of mixed-effects ANOVA on total skill (training sessions 1-3)

The results of Welch's ANOVA for differences in low skill between training groups for all subtasks on the 3 training sessions are tabulated in Table 4.4.

Comparison	**F**	**p**	**η2**	**p (Post-hoc)**
B vs. 2↑,1↓ UL	F(3,84.6) = 12.069	<0.0001 ***	0.271	NS
B vs. 1↑,1↓ L				NS
B vs. MFP				p < 0.001 ***

Table 4.4: Results of Welch's ANOVA on low skill (training sessions 1-3)

The results of the Kruskal-Wallis tests for differences in attained skill on the 30th training trial between adaptive training groups for each subtask are tabulated below in Table 4.4 along with results of the post-hoc Dunn's test.

Subtask	χ2	p	p (Post-hoc)
LS	-	NS	-
MC	$\chi2(3) = 7.3778$	0.0608	-
TD	-	NS	-
All	-	NS	-

Table 4.5: Results of Kruskal-Wallis on attained skill (training trial 30)

4.1.3 Performance

A Kruskal-Wallis *H*-Test by ranks across all 30 training trials found a significant difference between training groups on performance in TD, $\chi2(3) = 11.885$, $p < 0.01$. However, a post-hoc Dunn's test showed that the significant differences between groups did not belong to pairwise comparisons against the baseline condition. Another Kruskal-Wallis *H*-Test by ranks across training trials found a difference between training groups on performance in LS which approached significance, $\chi2(3) = 6.423$, $p = 0.093$. Performance grades across all training trials are shown in Figure 4.5.

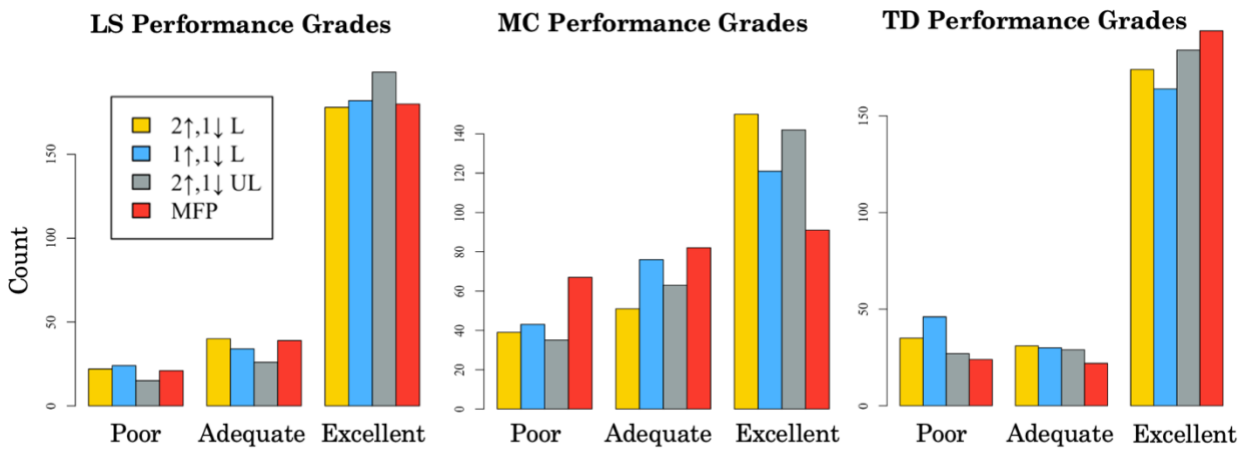


Figure 4.5: Performance grades across subtasks for all training trials

For the 30th training trial in particular, a Kruskal-Wallis *H*-Test by ranks found a significant difference between training groups on attained performance on the integrated median of all subtasks, $\chi 2(3) = 9.874$, $p < 0.05$. A post-hoc Dunn's test showed that subjects in the MFP training group attained significantly lower median performance across all subtasks than subjects in the 2↑,1↓ L group ($p < 0.05$). Counts of trials with crashes recorded for any subtask are plotted in Figure 4.6.

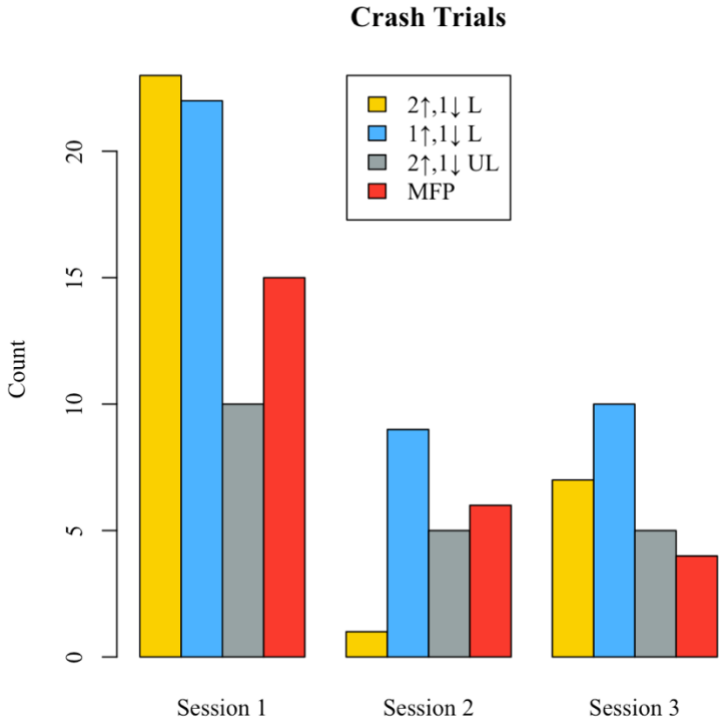


Figure 4.6: Crash trial count across groups and sessions during training

A Kruskal-Wallis *H*-Test by ranks across all 30 training trials found a significant difference between training groups on the number of crashes across all subtasks, $\chi 2(3) = 9.393$, $p < 0.05$. However, a post-hoc Dunn's test showed that the significant differences between groups did not belong to pairwise comparisons against the baseline condition. Moreover, a Kruskal-Wallis *H*-Test by ranks across all 30 training trials found a significant difference between training groups on the number of trials on which subjects scored excellent for all subtasks, $\chi 2(3) = 20.731$, $p < 0.001$. However, a post-hoc Dunn's test showed that the significant differences between groups did not belong to pairwise comparisons against the baseline condition. Counts of trials with excellent performance in all three subtasks are shown in Figure 4.7.

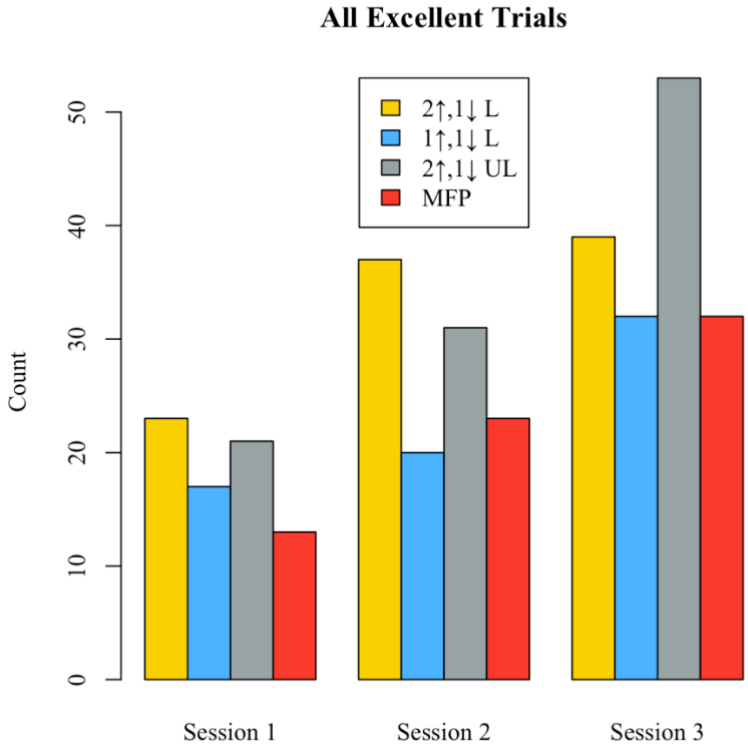


Figure 4.7: All excellent trial count across groups and sessions during training

Finally, a Kruskal-Wallis *H*-Test by ranks across the 30th training trial found a significant difference between training groups on the number of trials on which subjects scored excellent for all subtasks, $\chi 2(3) = 9.118$, $p < 0.05$. However, a post-hoc Dunn's test showed that the significant differences between groups did not belong to pairwise comparisons against the baseline condition.

The results of the Kruskal-Wallis tests for differences in performance across training sessions between training groups for each subtask are tabulated below in Table 4.6 along with results of the post-hoc Dunn's test.

Subtask	Range	χ2	p	p (Post-hoc)
LS	Sessions 1-3	$\chi2(3) = 6.4233$	0.0927	-
MC	Sessions 1-3	-	NS	-
TD	Sessions 1-3	$\chi2(3) = 11.8847$	0.0078 *	-
All	Sessions 1-3	-	NS	-
LS	Trial 30	-	NS	-
MC	Trial 30	$\chi2(3) = 9.4935$	0.0234 *	-
TD	Trial 30	-	NS	-
All	Trial 30	$\chi2(3) = 9.8741$	0.0197 *	p = 0.0403 * (B vs. MFP)

Table 4.6: Results of Kruskal-Wallis on total performance during training

The results of the Kruskal-Wallis tests for differences in the number of triple excellent (3E) trials attained on the 30th training trial between training groups for each subtask are tabulated below in Table 4.7 along with results of the post-hoc Dunn's test.

Subtask	Range	χ2	p	p (Post-hoc)
LS	Trial 30	$\chi2(3) = 6.4233$	0.0927	-
MC	Trial 30	-	-	-
TD	Trial 30	$\chi2(3) = 11.8847$	0.0078 *	-

All	Trial 30	-	-	-

Table 4.7: Results of Kruskal-Wallis on number of attained triple excellent performances (training trial 30)

4.2 Cockpit Results

4.2.1 Performance

A Kruskal-Wallis *H*-Test by ranks across all 10 cockpit trials found a significant difference between training groups on performance in TD, $\chi 2(3) = 9.319$, $p < 0.05$. A post-hoc Dunn's test showed that the significant differences between groups did not belong to pairwise comparisons against the baseline condition. No other group comparisons of total performance, crashes, and all excellent trials for both 1st cockpit trial and full cockpit session attained significance. The number of crashes in both the 1st cockpit trial and full cockpit session are shown in Figure 4.8, and the number of all excellent performance scores for both the 1st cockpit trial and full cockpit session are shown in Figure 4.9.

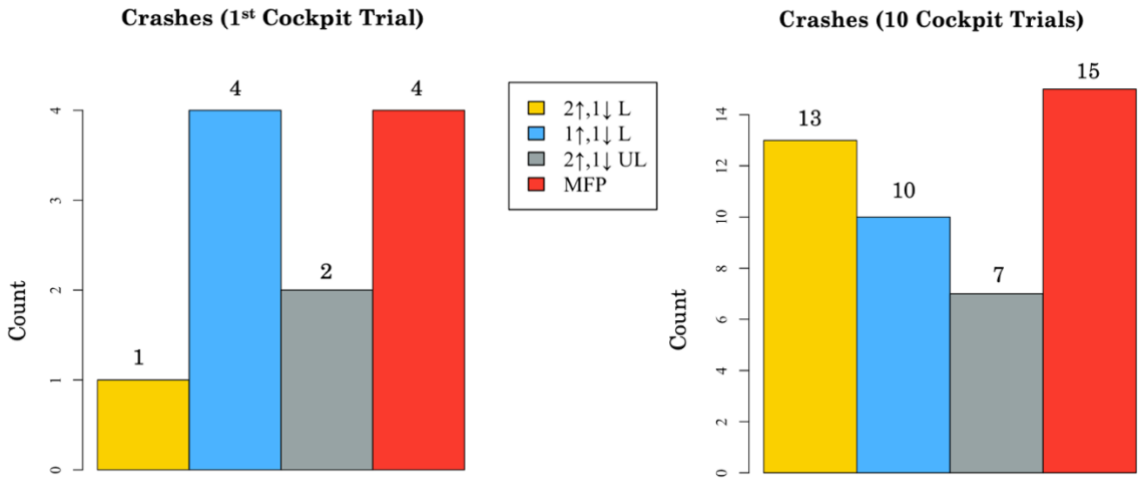


Figure 4.8: Crashes in the cockpit (left: 1st trial, right: all 10 trials) between groups

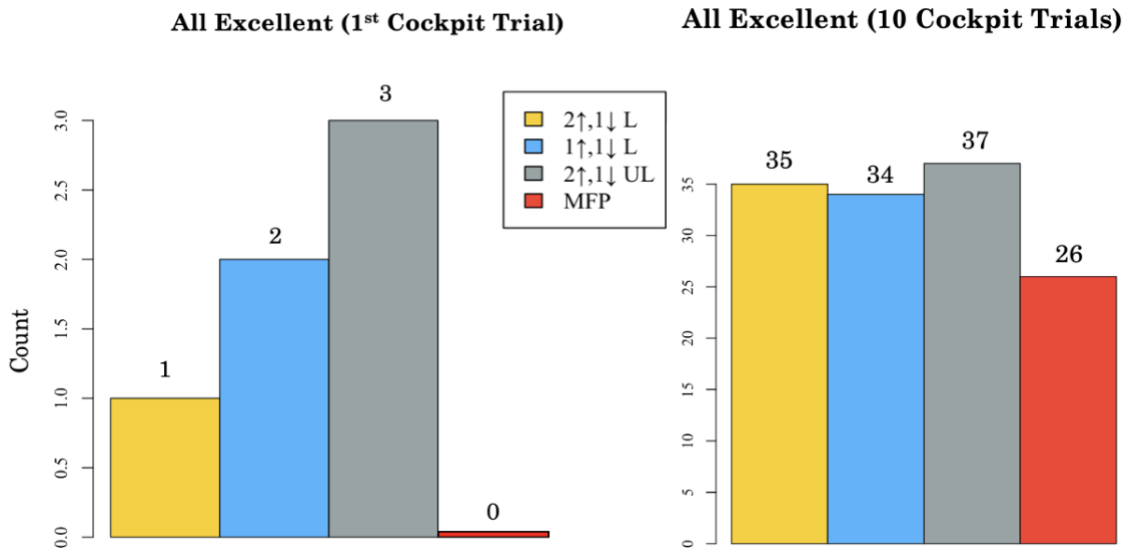
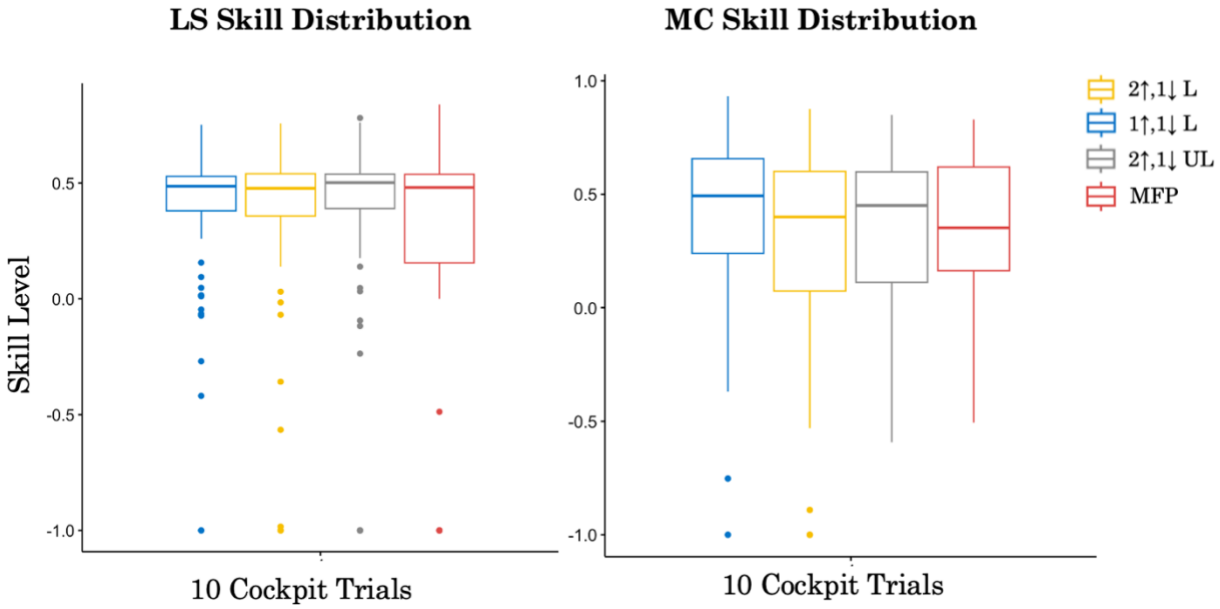


Figure 4.9: All excellent scores in the cockpit (left: 1st trial, right: all 10 trials)

4.2.2 Skill

Welch's ANOVA for unequal sample sizes found a significant difference between all training groups on the number of subjects who demonstrated low (-1 to

0) skill across all cockpit trials on the TD subtask, $F(3, 32.813) = 4.141$, $p < 0.05$, $\eta 2 = 0.192$. Tukey's post-hoc test showed that subjects in MFP had significantly higher rates of low skill on TD across trials in the cockpit mock-up ($p < 0.05$, 95% CI [0.0494, 0.7056]). Moreover, Welch's ANOVA for unequal sample sizes found a significant difference between all training groups on the number of subjects who demonstrated low skill across all cockpit trials and across all subtasks, $F(3, 24.434) = 4.141$, $p < 0.05$, $\eta 2 = 0.123$. Tukey's post-hoc test showed that subjects in MFP had significantly higher rates of low skill on TD across cockpit trials ($p < 0.05$, 95% CI [0.0251, 0.4888]). The skill distributions for subjects in each group across cockpit trials are shown in Figure 4.10.



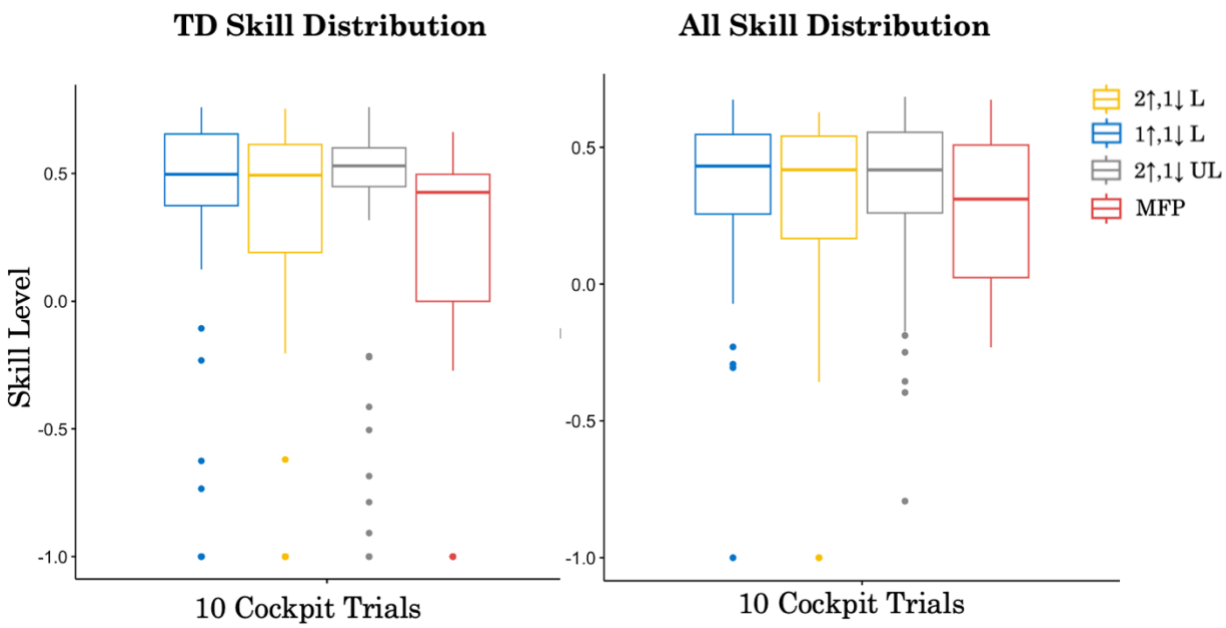


Figure 4.10: Skill distribution across training groups for all cockpit trials

The results of Welch's ANOVA for differences in low skill between training groups for all subtasks on the 10 AReS cockpit trials are tabulated in Table 4.8.

Subtask	F	p	η2	p (Post-hoc)
LS	-	NS	-	-
MC	-	NS	-	-
TD	F(3, 32.813) = 4.141	0.0135 *	0.192	p < 0.05 * (B vs. MFP)
All	F(3, 24.434) = 4.141	0.0169 *	0.123	p < 0.05 * (B vs. MFP)

Table 4.8: Results of Welch's ANOVA on low skill (cockpit trials 1-10)

CHAPTER V

DISCUSSION

5.1 RESPONSIVENESS

The difficulty progressions between 1↑,1↓L and 2↑,1↓L varied by subtask, with LS showing differences in progression and significantly higher difficulty levels in 1↑,1↓L across the training trials. The subtask progressions converge by trial 15 for 1↑,1↓L and continue in concert, while a noticeable rift between subtasks persists in the baseline 2↑,1↓L condition throughout trials, narrowing only at the end. This suggests that increased sensitivity accelerates the rate of convergence. Between subtasks, the largest differences in difficulty progression are in the MC and TD subtasks, where the average difference is nearly 3 levels for the final 2 training sessions for TD.

Moreover, we observe familiarization dips in all subjects, but the dips occur earlier (trial 5-7 in 1↑,1↓ vs. trial 11-12 for 2↑,1) and are slightly less pronounced in 1↑,1↓ (level 9) compared to 2↑,1↓ (level 10). The difference in minima is small, but the difference in the training trial at which the minima occurs is, predictably, a factor of two apart. This suggests that 1↑,1↓ is allowing subjects to exit the familiarization dip faster and more steadily than the less sensitive baseline condition.

Although this difference in staircase sensitivity is hard-coded into the algorithm, it is interesting to note that oscillations in the 1↑,1↓ progression from potential "false positive" premature modulations of difficulty would stymie the tendency for progressions to diverge, but this is not seen to occur. However,

heightened sensitivity is accompanied by a detectable increase in inter-trial variance and motility, as seen in the distribution of difficulty being far larger across subtasks and training sessions than for the baseline. This result is to be expected, since a higher level of responsiveness more closely follows the variability in subject performance across trials, and allows for double the possible difference in staircase between subjects over the thirty training trials.

However, higher responsiveness was accompanied by a lower number of all excellent performances across all training trials, and a higher number of crashes across all training trials compared to the baseline. Moreover, the total number of excellent performances was slightly lower on the MC and TD subtasks, which data suggests are the more difficult, and thus limiting, training factors. The average skill of 1↑,1↓ subjects was lower in MC, the most difficult subtask, and on the average of all subtasks for both the second and third training session compared to 2↑,1↓. These relatively poorer outcomes in skill and performance in training suggests a non-trivial inimical effect of premature modulation and variability on training efficacy.

Further, it is interesting to note that although subjects in the 1↑,1↓ condition trained, on average, nearly at the AReS cockpit's level 18 fixed difficulty, increased sensitivity had mixed results in the mockup. For instance, on the first trial, a quarter of subjects attained all excellent ratings, double the number of subjects who did so in the 2↑,1↓ condition. However, half of subjects in 1↑,1↓ crashed! This rate is four times higher than subjects in 2↑,1↓ and comparable only to subjects who trained in MFP, a non-adaptive training algorithm. Despite this, there were 25% fewer crashes among

1↑,1↓ subjects compared to 2↑,1↓ subjects through all cockpit trials. Given the preponderance of poor performance in the first trial, however, this may simply suggest that subjects in the 1↑,1↓ group were more amenable to the physical environment and acclimated faster than those in the baseline, who were nonetheless better preared. Over the ten cockpit trials, subjects in the two conditions converged to similar performance and skill levels, indicating that both eventually acclimated to the physical mock-up conditions during the course of the session.

The mixed results displayed by 1↑,1↓ subjects on difficulty progression, skill, and performance, both in training and in the mockup, suggests that responsiveness is a double-edged sword. On one hand, higher responsiveness exposes subjects to higher subtask difficulties, a difference which was significant for an easier subtask like LS, suggesting that high responsiveness is ideal for reaching adequately challenging difficulty levels faster. However, for complex, difficult functions such as MC or TD, the increased volatility of a highly responsive system, which more frequently modulates difficulty prematurely, and the tendency to spend fewer trials at each difficulty level may both lead to poorer skill transfer and performance.

5.2 INTEGRATION

The unlocked algorithm shows clearly discrete staircases, implying that subjects likely learn the 3 subtasks independently. Removing lockstep seems to allow subjects to remain in a different flow channel for each subtaskThe removal of lockstep had a clear effect on difficulty progression, with landing site selection being

unimpeded from shooting to an average difficulty of 22 in the 2↑,1↓UL condition compared to approximately 15 in the baseline condition. This suggests that this subtask was easier than the rest, such that it was consistently subjected to lockstep in the baseline condition. Similarly to SS, TD shows an almost immediate divergence in average difficulty level, and the separation widens over the training trials between the two conditions.

However, it appears that subjects had almost identical progressions in both of the 2U1D conditions. This result is expected since MC is the limiting subtask; in 2↑,1↓L, MC imposes lockstep on other subtasks in baseline, so it is not itself affected by it, while in 2↑,1↓UL, MC is totally independent of other subtasks by design. Thus, we would expect the variability and progression to be the same since they are under the same PEST staircase. Their similarity suggests that the 2 conditions did not have inherently better/worse subjects, validating the randomness in condition assignment and further validating results from comparisons.

Further, it is interesting to note that, on average, subjects in the 2↑,1↓unlocked condition trained at or past the AReS cockpit's level 18 fixed difficulty, and the average was much higher than that of the baseline condition for non-limiting subtasks. This suggests that integration allowed subtask difficulty to progress to a seemingly natural level of challenge unimpeded by other subtasks.

The unlocked group had slightly higher performance on the LS and TD subtasks across training, indicating that independent progressions more optimally modulated difficulty according to subject training needs. Interestingly, the

distribution of skill was tighter for subjects in the unlocked condition than those in the locked condition for all subtasks, and the distribution was tightest among all training algorithms. This supports the notion that high levels of integration create more variability in skill level for subtasks affected by lockstep, whereas this effect dissipates almost entirely when progression between subtasks is left to vary.

5.3 PERSONALIZATION

Subjects in the median fixed progression condition initially mimicked the performance and skill acquisition of their counterparts in the 2↑,1↓L condition, with slightly less achievement throughout trials, but there is a marked divergence in skill acquisition near the middle of the training trials, after which skill begins to decline. Moreover, the significant difference in the number of trials in which subjects were graded as having low skill in MFP versus 2↑,1↓L, a difference that held for each subtask across training trials, indicates that the average subject was forced to perform subtasks at a higher challenge than an adaptive staircase would have provided them, and their performance and skill in training suffered as a result.

Notably, the difference persisted and remained significant by the end of training, as seen in the significantly higher rates of low skill on the piloting subtask, MC, compared to the baseline condition, and on the average measure of skill over all subtasks. These differences indicate that personalization is a crucial factor for automated training, and that mimicry alone is a non-optimal method for training.

Although cockpit trials were not significantly different from other groups, subjects in MFP nonetheless displayed interesting patterns of performance and skill compared to the baseline. For instance, on the first trial in the cockpit, a potent proxy for whether training adequately prepared subjects to perform EDL, none of the subjects in MFP attained all excellent across subtasks, while between 1 and 3 subjects did for each of the adaptive groups. That subsequent skill and performance also did not achieve significance suggests that subjects were able to learn in AReS itself. Moreover, MFP had the highest number of crashes of all training groups and was the only algorithm which exceeded the baseline in total number of crashes in the cockpit, though by a small number. However, half of all subjects crashed on the first trial, a rate four times higher than in the baseline group. The relative rates at which subjects were unable to perform the tasks nominally suggests that the lack of personalization in training did not adequately prepare them for the rigors of the physical mock-up.

CHAPTER VI

CONCLUSION

This research investigated the effects of sensitivity, integration, and personalization on the efficacy of automated, individually-adaptive astronaut training algorithms in virtual reality for long-duration exploration missions. The study found that high sensitivity for difficulty progression leads to higher achieved difficulty in training, that discrete rather than unified ("locked") modulation of subtask progression leads to higher achieved difficulty in training, and that personalized training leads to higher levels of skill acquisition and performance than non-adaptive, fixed progression training. Sensitivity, integration, and personalization may not have significant effects on skill transfer and cockpit performance given sufficient training time.

This work addresses the literature gap examining the acquisition and retention of complex task learning relevant to human spaceflight, namely tasks that have components of both motor learning and strategy and decision making. It also addressed the effect of unified versus discrete modulation of subtask difficulty in automated training algorithms and provided a rigorous comparison of staircase threshold sensitivity on learning and performance outcomes. Furthermore, although dynamic difficulty adjustment had been explored the efficacy of individua, this work provided more data on the feasibility of individually-adaptive, personalized training paradigms and the use of virtual reality as a medium for automated astronaut training on deep space missions.

6.1 LIMITATIONS

Analyses in training were limited to considering training sessions rather than individual trials, which made for more robust statistical results but reduced the granularity of analysis to the aggregate of ten trials. Further, the subtasks were not equally challenging for all subjects, introducing variability and affecting the *de facto* form and function of lockstep, which engaged for generally difficult subtasks and may have stifled progression and flow in easier subtasks. Crashes due to high skill levels, for instance on the terminal descent subtask whereby subjects tested the limits of vehicle control by removing all thrust and applying maximal thrust at the last possible opportunity, were not differentiable from crashes due to low skill.

Additionally, analysis of personalization was limited by an asymmetry in measurable variables. Although it is hypothesized that low-performing subjects who train on a progression fixed to the median of eight subjects will be forced to encounter difficulties higher than they are prepared for and that performance metrics will suffer as a result, the converse is also true but less apparent: high-performing subjects will be anchored, or limited, to the median, and will fail to reach their full training potential within thirty trials. Since it is difficult to quantitatively assess potential and deviations from expected performance, this facet of MFP was not captured in analysis. Such subjects, however, may have diluted the poor performance across subtasks of subjects with more considerable training needs, reducing the general effect.

Although subjects were recruited at random and were evently split by sex, the majority of subjects were in their early twenties and ethnic background was relatively homogeneous, factors which may affect the generalizability of the results applied to older and/or minority populations. Moreover, the challenges inherent to human subject testing limited the sample size to 8 subjects for each condition, which combined with inter-subject variability reduced statistical resolution and obscured potentially significant differences in performance in both the virtual environment and AReS cockpit mock-up. Finally, individual variability in intrinsic motivation, restfulness, and confidence in approaching tasks may create noise in the data that could not be easily corrected through facilitation given the autonomous nature of the training.

6.2 FUTURE WORK

Future analysis will incorporate surveys taken by subjects during testing, including a flow survey, workload questionnaire, system usability scale (SUS) survey, and affect grids to investigate differences in reported measures of flow and experience between groups and study any overlap with performance data. Further work could investigate differences in the magnitude and minima of difficulty decline during the familiarization period to further assess the effect of responsiveness. Assessing results within trial rather than session may provide better results on the interaction between training group and training progression. Additionally, skill transfer between sessions

can be compared across groups by analyzing first trial performance at the onset of each new session.

This work made primary use of modified PEST staircases, altering the progression threshold, presence of lockstep, and adaptivity paradigm. The results suggest that reduced sensitivity may be beneficial during familiarization, when subjects are prone to sudden changes in performance, but may prove stifling during nominal progression. Thus, investigating the efficacy of shifting between staircase sensitivities as a function of consecutive runs of excellent performance may aid in understanding the optimal inflection point between PEST staircases and complement the data comparing sensitivity of the two most common staircases.

Additionally, each of the training conditions in this work modulated the difficulty in fixed increments of one; developing a system of dynamic rather than fixed linear response would allow for investigation of more optimal reconverge into a flow channel for subjects. For instance, an algorithm which reacts to a crash by decreasing difficulty by multiple levels may accelerate the familiarization process by descending to the needed amount before allowing subjects to progress. Such a method would be more sensitive to the "second derivative" of performance progress, and studying the difference in performance outcomes that such increased sensitivity might have would be a valuable insight.

Moreover, each of the algorithms in this work used a predefined paradigm, such as a PEST staircase, to take performance as an input in a closed-loop manner. Although these were individually-adaptive, future work should investigate the use of

data-driven Bayesian models to predict the probability of failure given a certain difficulty level to identify the optimal training difficulty. Such a model could weight both an individual's performance profile, and compare it to performance data from other subjects to identify patterns and outliers faster than a fixed linear response staircase like 2↑,1↓L. Moreover, a Bayesian training algorithm may provide a good benchmark for assessing deficiencies in other algorithms by identifying cases where the algorithm modulated difficulty differently than the statistical model predicted would be ideal for a subject, allowing for closer investigation of responsiveness, integration, and personalization.

Another fascinating area of future work might be to allow users to self-select difficulty levels during each trial for each subtask. Assuming that subjects eventually choose difficulties that most closely follow their training aptitude and needs, the selected progressions could be compared to those modulated by autonomous algorithms to measure how closely, or optimally, automated, individually-adaptive algorithms are able to predict a subject's location within, or departure from, a flow channel and return them to it. Such work may be complicated by differences in risk tolerance and is sensitive to variability in intrinsic motivation.

Finally, and perhaps the most exciting, would be to incorporate psychophysiological monitoring, particularly non-invasive methods for determining cognitive workload and stress, to predict challenge using more variables than simply performance. Such a system may be more robust to differences in motivation and could detect lax demeanors and heightened stress alike. The ability to autonomously

modulate a training system to more closely fit the needs of crew is an important area for future development of deep space missions.

Bibliography

Aannett, J. (1969). *Feedback and human behaviour: The effects of knowledge of results, incentives and reinforcement on learning and performance*. Penguin Books.

Abbas, Z. A., & North, J. S. (2018). Good-vs. poor-trial feedback in motor learning: The role of self-efficacy and intrinsic motivation across levels of task difficulty. *Learning and Instruction*, *55*, 105–112. https://doi.org/10.1016/j.learninstruc.2017.09.009

Adamovich, S. V., Fluet, G. G., Tunik, E., & Merians, A. S. (2009). Sensorimotor training in virtual reality: A review. *NeuroRehabilitation*, *25*(1), 29–44. https://doi.org/10.3233/nre-2009-0497

Aïm, F., Lonjon, G., Hannouche, D., & Nizard, R. (2016). Effectiveness of virtual reality training in orthopaedic surgery. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, *32*(1), 224–232. https://doi.org/10.1016/j.arthro.2015.07.023

Almeida, F., & Buzády, Z. (2019). Learning entrepreneurship in higher education through flow theory and FLIGBY game. *International Journal of Virtual and Personal Learning Environments*, *9*(1), 1–15. https://doi.org/10.4018/ijvple.2019010101

Andrade, G., Ramalho, G., Santana, H., and Corruble, V. Extending Reinforcement Learning to Provide Dynamic Game Balancing. Edinburgh, United Kingdom, 2005

Anderson, D. I., Magill, R. A., & Sekiya, H. (2001). Motor learning as a function of KR schedule and characteristics of task-intrinsic feedback. *Journal of Motor Behavior*, *33*(1), 59–66. https://doi.org/10.1080/00222890109601903

Aoki, H., Oman, C. M., & Natapoff, A. (2007). Virtual-Reality-Based 3D navigation training for emergency egress from spacecraft. *Aviation, Space, and Environmental Medicine*, *78*(10), 774–783.

Arthur Jr., W., Bennett Jr., W., Stanush, P. L., & McNelly, T. L. (1998a). Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance*, *11*(1), 57–101. https://doi.org/10.1207/s15327043hup1101_3

Arthur Jr., W., Bennett Jr., W., Stanush, P. L., & McNelly, T. L. (1998b). Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance*, *11*(1), 57–101. https://doi.org/10.1207/s15327043hup1101_3

Basner, M., Stahn, A. C., Nasrini, J., Dinges, D. F., Moore, T. M., Gur, R. C., Mühl, C., Macias, B. R., & Laurie, S. S. (2021). Effects of head-down tilt bed rest plus elevated CO2 on cognitive performance. *Journal of Applied Physiology*, *130*(4), 1235–1246. https://doi.org/10.1152/japplphysiol.00865.2020

Beck, L. A. (1992). CsikszentMihalyi, Mihaly. (1990). Flow: The psychology of optimal experience. *Journal of Leisure Research*, *24*(1), 93–94. https://doi.org/10.1080/00222216.1992.11969876

Benyon, D. (1993). Adaptive systems: A solution to usability problems. *User Modeling and User-Adapted Interaction*, *3*(1), 65–87. https://doi.org/10.1007/BF01099425

Botella, C., Baños, R. M., Etchemendy, E., García-Palacios, A., & Alcañiz, M. (2016). Psychological countermeasures in manned space missions: "EARTH" system for the Mars-500 project. *Computers in Human Behavior*, *55*, 898–908. https://doi.org/10.1016/j.chb.2015.10.010

Brady, T., & Paschall, S. (2010a, March). The challenge of safe lunar landing. *2010 IEEE Aerospace Conference*. http://dx.doi.org/10.1109/aero.2010.5447029

Brady, T., & Paschall, S. (2010b, March). The challenge of safe lunar landing. *2010 IEEE Aerospace Conference*. http://dx.doi.org/10.1109/aero.2010.5447029

Butler, G. V. (1973). Skylab. *NASA NTRS*. https://ntrs.nasa.gov/api/citations/19730005104/downloads/19730005104.pdf

Butt, A. L., Kardong-Edgren, S., & Ellertson, A. (2018). Using game-based virtual reality with haptics for skill acquisition. *Clinical Simulation in Nursing*, *16*, 25–32. https://doi.org/10.1016/j.ecns.2017.09.010

Cameirão, Badia, Oller, & Verschure. (2010). Neurorehabilitation using the virtual reality based Rehabilitation Gaming System: Methodology, design, psychometrics, usability and validation. *Journal of NeuroEngineering and Rehabilitation*, *7*(1), 1–14. https://doi.org/10.1186/1743-0003-7-48

Carulli, M., Bordegoni, M., Bernecich, F., Spadoni, E., & Bolzan, P. (2019, August 18). A multisensory virtual reality system for astronauts' entertainment and relaxation. *Volume 1: 39th Computers and Information in Engineering Conference*. http://dx.doi.org/10.1115/detc2019-97836

Carveth, J. W., & Adams, J. A. (1964). Effects of Practice on Dial Reading. *University of Illinois*.

Casner, S. M., Geven, R. W., Recker, M. P., & Schooler, J. W. (2014). The Retention of Manual Flying Skills in the Automated Cockpit. *Human Factors*, *56*(8), 1506–1516. https://doi.org/10.1177/0018720814535628

Cavagna, Willems, & Heglund. (n.d.). Walking on Mars. *Nature*, *393*(6686), 636–636. https://doi.org/10.1038/31374

Cheng, Y.-M. (2020). Investigating medical professionals' continuance intention of the cloud-based e-learning system: An extension of expectation–confirmation model with flow theory. *Journal of Enterprise Information Management*, *34*(4), 1169–1202. https://doi.org/10.1108/JEIM-12-2019-0401

Childs, J. M., & Spears, W. D. (1986). Flight-Skill decay and recurrent training. *Perceptual and Motor Skills*, *62*(1), 235–242. https://doi.org/10.2466/pms.1986.62.1.235

Childs, J. M., Spears, W. D., & Prophet, W. W. (1983). *Private pilot flight skill retention 8, 16, and 24 months following certification*. Defense Technical Information Center; Embry-Riddle Aeronautical University. https://apps.dtic.mil/sti/citations/ADA133400

Choi, D. H., Kim, J., & Kim, S. H. (2007). ERP training with a web-based electronic learning system: The flow theory perspective. *International Journal of Human-Computer Studies*, *65*(3), 223–243. https://doi.org/10.1016/j.ijhcs.2006.10.002

Clark, T. K. (2022). *Handbook of Space Pharmaceuticals: Effects of Spaceflight on the Vestibular System* (1st ed.). Springer Nature. (Original work published 2022)

Clark, T. K., Newman, M. C., Oman, C. M., Merfeld, D. M., & Young, L. R. (2015). Modeling human perception of orientation in altered gravity. *Frontiers in Systems Neuroscience*, *9*. https://doi.org/10.3389/fnsys.2015.00068

Clément, G. (2007). Using your head: Cognition and sensorimotor functions in microgravity. *Gravitational and Space Biology*, *20*(2), 65–78.

Clément, G., & Ngo-Anh, J. T. (2012). Space physiology II: Adaptation of the central nervous system to space flight—past, current, and future studies. *European Journal of Applied Physiology*, *113*(7), 1655–1672. https://doi.org/10.1007/s00421-012-2509-3

Clément, G. R., Boyle, R. D., George, K. A., Nelson, G. A., Reschke, M. F., Williams, T. J., & Paloski, W. H. (2020). Challenges to the central nervous system during human spaceflight missions to Mars. *Journal of Neurophysiology*, *123*(5), 2037–2063. https://doi.org/10.1152/jn.00476.2019

Constant, T., & Levieux, G. (2019, May 2). Dynamic difficulty adjustment impact on players' confidence. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. http://dx.doi.org/10.1145/3290605.3300693

Creech, S., Guidi, J., & Elburn, D. (2022). Artemis: An overview of NASA's activities to return humans to the moon. *IEEE Xplore*, *50100*. https://ieeexplore.ieee.org/abstract/document/9843277

Csikszentmihalyi, Mihaly. (1990). Flow: The Psychology of Optimal Experience." *Journal of Leisure Research*, 24(1), pp. 93–94

Demediuk, S., Tamassia, M., Raffe, W. L., Zambetta, F., Mueller, F. "Floyd," & Li, X. (2018, January 29). Measuring player skill using dynamic difficulty

adjustment. *Proceedings of the Australasian Computer Science Week Multiconference*. http://dx.doi.org/10.1145/3167918.3167939

Descarreaux, M., Passmore, S. R., & Cantin, V. (2010). Head movement kinematics during rapid aiming task performance in healthy and neck-pain participants: The importance of optimal task difficulty. *Manual Therapy*, *15*(5), 445–450. https://doi.org/10.1016/j.math.2010.02.009

Deschenes, M. R., Giles, J. A., McCoy, R. W., Volek, J. S., Gomez, A. L., & Kraemer, W. J. (2002). Neural factors account for strength decrements observed after short-term muscle unloading. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *282*(2), R578–R583. https://doi.org/10.1152/ajpregu.00386.2001

Dhiman, A., Solanki, D., Bhasin, A., Bhise, A., Das, A., and Lahiri, U. Design of Adaptive Haptic-Enabled Virtual Reality Based System for Upper Limb Movement Disorders: A Usability Study. Presented at the 2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob), 2016

Diamond, M. J. (2015). *Will communication delays impact mission controllers? An investigation of mood, performance, and workload during analog missions* [Master's Thesis]. University of North Dakota.

Donegan, J. J. (1967). *Apollo Mission Profile* (pp. 135–158).

Doyle, R. (2003). *Autonomy needs and trends in deep space exploration*. Defense Technical Information Center (Jet Propulsion Laboratory, California Institute of Technology. https://apps.dtic.mil/sti/citations/ADA485027

Dymora, P., Kowal, B., Mazurek, M., & Romana, S. (2021). The effects of Virtual Reality technology application in the aircraft pilot training process. *IOP Conference Series: Materials Science and Engineering*, *1024*(1), 012099. https://doi.org/10.1088/1757-899x/1024/1/012099

Eddy, D., Schiflett, S., Schlegel, R., & Shehab, R. (1998). Cognitive performance aboard the life and microgravity spacelab. *Acta Astronautica*, *43*(3–6), 193–210. https://doi.org/10.1016/S0094-5765(98)00154-4

El-Baz, F. (2011). Training Apollo astronauts in lunar orbital observations and photography. *Analogs for Planetary Exploration: Geological Society of America Special Paper*, *483*, 49–65.

Elburn, D. (2022, November 15). *Artemis III: NASA's first human mission to the Lunar South Pole*. NASA. https://www.nasa.gov/feature/artemis-iii

Engle, M. (2004, June 19). Operational considerations for manned lunar landing missions - Lessons learned from Apollo. *Space 2004 Conference and Exhibit*. http://dx.doi.org/10.2514/6.2004-6081

Escobar, C., Nabity, J., & Klaus, D. (2017, July 16). *Defining ECLSS robustness for deep space exploration*. Texas Tech University Libraries. http://hdl.handle.net/2346/73061

Everson, T., McDermott, C., Kain, A., Fernandez, C., Horan, B., Everson, T., McDermott, C., Kain, A., Fernandez, C., & Horan, B. (2017, February 9). Astronaut Training using Virtual Reality in a Neutrally Buoyant Environment. *6th Engineering, Science and Technology Conference - Panama (ESTEC)*. https://ridda2.utp.ac.pa/handle/123456789/4340

Fanjoy, R. O. (2013). Flight skill proficiency issues in instrument approach accidents. *Journal of Aviation Technology and Engineering*, *3*(1). Purdue University Press. https://doi.org/https://doi.org/10.7771/2159-6670.1069

Farr, A., Leon Pietschmann, Zürcher, P., & Bohné, T. (2023). Skill retention after desktop and head-mounted-display virtual reality training. *Experimental Results*, *4*. https://doi.org/10.1017/exp.2022.28

Finseth, T., Dorneich, M. C., Keren, & Franke. (2021). The effectiveness of adaptive training for stress inoculation in a simulated astronaut task. *Proceedings of the 2021 HFES 65th International Annual Meeting*.

Fitts, P. M. (1964). Perceptual-Motor Skill Learning11This chapter is based in part on research supported by the U. S. Air Force, Office of Scientific Research, under Contract No. AF 49 (638)-449. In *Categories of Human Learning* (pp. 243–285). Elsevier. http://dx.doi.org/10.1016/b978-1-4832-3145-7.50016-9

Frank, J., Sprikovska, L., McCann, R., Wang, L., Pohlkamp, K., & Morin, L. (2013). *Autonomous mission operations*. IEEE Xplore.

Gabay, Y., Karni, A., & Banai, K. (2017). The perceptual learning of time-compressed speech: A comparison of training protocols with different levels of difficulty. *PLOS ONE*, *12*(5), e0176488. https://doi.org/10.1371/journal.pone.0176488

Gagne, R. M. (1962). Military training and principles of learning. *American Psychologist*, *17*(2), 83–91. https://doi.org/10.1037/h0048613

Garcia, A. D., Schlueter, J., & Paddock, E. (2020, January 5). Training Astronauts using Hardware-in-the-Loop Simulations and Virtual Reality. *AIAA Scitech 2020 Forum*. http://dx.doi.org/10.2514/6.2020-0167

Golding, J. F. (1998). Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain Research Bulletin*, *47*(5), 507–516. https://doi.org/10.1016/s0361-9230(98)00091-4

Grabherr, L., & Mast, F. W. (2009). Effects of microgravity on cognition: The case of mental imagery - IOS Press. *Journal of Vestibular Research*, *20*(1–2), 53–60. https://doi.org/10.3233/VES-2010-0364

Gray, R. (2017). Transfer of training from virtual to real baseball batting. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02183

Greenleaf, J. E., Bulbulian, R., Bernauer, E. M., Haskell, W. L., & Moore, T. (1989). Exercise-training protocols for astronauts in microgravity. *Journal of Applied Physiology*, *67*(6), 2191–2204. https://doi.org/10.1152/jappl.1989.67.6.2191

Grimm, F., Naros, G., & Gharabaghi, A. (2016). Closed-Loop task difficulty adaptation during virtual reality reach-to-grasp training assisted with an exoskeleton for stroke rehabilitation. *Frontiers in Neuroscience*, *10*. https://doi.org/10.3389/fnins.2016.00518

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior*, *36*(2), 212–224. https://doi.org/10.3200/jmbr.36.2.212-224

Gupta, S. K., Anand, D. K., Brough, J. E., Schwartz, M., & Kavetsky, R. A. (2008). *Training in Virtual Environments: A Safe, Cost-Effective, and Engaging Approach to Training*. University of Maryland, College Park.

Hamblin, C. J. (2005). *Transfer of training from virtual reality environments* [Doctoral Thesis]. Wichita State University.

Haslbeck, A., & Hoermann, H.-J. (2016). Flying the needles. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(4), 533–545. https://doi.org/10.1177/0018720816640394

Hatch, H. G., Pennington, J. E., & Cobb, J. B. (1967). *Dynamic simulation of lunar module docking with Apollo command module in lunar orbit*. NASA Langley Research Center.

Hendrickson, S. M. L., Goldsmith, T. E., & Johnson, P. J. (2006). Retention of airline pilots' knowledge and skill. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(17), 1973–1976. https://doi.org/10.1177/154193120605001755

Herman, D. A., Tofil, T. A., Santiago, W., Kamhawi, H., Polk, J. E., Snyder, J. S., Hofer, R. R., Picha, F. Q., Jackson, J., & Allen, M. (2018). *Overview of the development and mission application of the advanced electric propulsion system (AEPS)*. NASA Technical Reports Server (NTRS). https://ntrs.nasa.gov/citations/20180001297

Holden, K. (2022). Effects of long-duration microgravity and gravitational transitions on fine motor skills - Kritina Holden, Maya Greene, E. Vincent, Anikó Sándor, Shelby Thompson, Alan Feiveson, Brandin Munson, 2022. *Human Factors*. https://doi.org/10.1177_00187208221084486

Hollister, LaPointe, Oman, & Tole. (1973). *Identifying and determining skill degradations of private and commercial pilots*. Defense Technical Information Center; MIT Measurement Systems Lab. https://apps.dtic.mil/sti/citations/AD0771101

Holt, S. (2023). Virtual reality, augmented reality and mixed reality: For astronaut mental health; and space tourism, education and outreach. *Acta Astronautica*, *203*, 436–446. https://doi.org/10.1016/j.actaastro.2022.12.016

Homan, D., & Gott, C. (1996, July 29). An integrated EVA/RMS virtual reality simulation, including force feedback for astronaut training. *Flight Simulation Technologies Conference*. http://dx.doi.org/10.2514/6.1996-3498

Huang, Y.-C., Backman, S. J., & Backman, K. F. (2010). Student attitude toward virtual learning in second life: A flow theory approach. *Journal of Teaching in Travel & Tourism*, *10*(4), 312–334. https://doi.org/10.1080/15313220.2010.525425

Hunicke, R. (2005, June 15). The case for dynamic difficulty adjustment in games. *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*. http://dx.doi.org/10.1145/1178477.1178573

Jacob, & Teuteberg. (2017, January 1). *Game-Based Learning, Serious Games, Business Games und Gamification –Lernförderliche Anwendungsszenarien, gewonnene Erkenntnisse und Handlungsempfehlungen*. Springer Fachmedien Wiesbaden. https://link.springer.com/chapter/10.1007/978-3-658-16742-4_8

Jiang, A., Gong, Y., Yao, X., Foing, B., Allen, R., Westland, S., Hemingray, C., & Zhu, Y. (2023a). Short-term virtual reality simulation of the effects of space station colour and microgravity and lunar gravity on cognitive task performance and emotion. *Building and Environment*, *227*, 109789. https://doi.org/10.1016/j.buildenv.2022.109789

Jiang, A., Gong, Y., Yao, X., Foing, B., Allen, R., Westland, S., Hemingray, C., & Zhu, Y. (2023b). Short-term virtual reality simulation of the effects of space station colour and microgravity and lunar gravity on cognitive task performance and emotion. *Building and Environment*, *227*, 109789. https://doi.org/10.1016/j.buildenv.2022.109789

Jones, H. W., Hodgson, E. W., & Kliss, M. H. (2014, July 13). *Life support for Deep Space and Mars*. Texas Tech University Libraries. http://hdl.handle.net/2346/59729

Jonsson, A., Morris, R. A., & Pedersen, L. (2007). Autonomy in space: Current capabilities and future challenge. *AI Magazine*, *28*(4), 27–27. https://doi.org/10.1609/aimag.v28i4.2066

Juhl, Buettmann, Friedman, DeNapoli, Hoppock, & Donahue. (2021). Update on the effects of microgravity on the musculoskeletal system. *Npj Microgravity*, *7*(1), 1–15. https://doi.org/10.1038/s41526-021-00158-4

Kelc, R., Vogrin, M., & Kelc, J. (2020). Cognitive training for the prevention of skill decay in temporarily non-performing orthopedic surgeons. *Acta Orthopaedica*, *91*(5), 523–526. https://doi.org/10.1080/17453674.2020.1771520

Kenyon, R. V., & Afenya, M. B. (1995). Training in virtual and real environments. *Annals of Biomedical Engineering*, *23*(4), 445–455. https://doi.org/10.1007/bf02584444

Killgore, W. D. S. (1998). The affect grid: a moderately valid, nonspecific measure of pleasure and arousal. *Pyschological Reports*, *83*, 639–642.

Kim, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science*, *14*(1), 22–37. https://doi.org/10.1080/1464536x.2011.573008

Klostermann, M., Conein, S., Felkl, T., & Kluge, A. (2022). Factors influencing attenuating skill decay in high-risk industries: A scoping review. *Safety*, *8*(2). https://doi.org/10.3390/safety8020022

Kluge, A., & Frank, B. (2014). Counteracting skill decay: Four refresher interventions and their effect on skill and knowledge retention in a simulated process control task. *Ergonomics*, *57*(2), 175–190. https://doi.org/10.1080/00140139.2013.869357

Koenig, A., Novak, D., Omlin, X., Pulfer, M., Perreault, E., Zimmerli, L., Mihelj, M., & Riener, R. (2011). Real-Time closed-loop control of cognitive load in

neurological patients during robot-assisted gait training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *19*(4), 453–464. https://doi.org/10.1109/tnsre.2011.2160460

Kumar, D., González, A., Das, A., Dutta, A., Fraisse, P., Hayashibe, M., & Lahiri, U. (2018). Virtual reality-based center of mass-assisted personalized balance training system. *Frontiers in Bioengineering and Biotechnology*, *5*. https://doi.org/10.3389/fbioe.2017.00085

Landon, L. B., Rokholt, C., Slack, K. J., & Pecena, Y. (2017). Selecting astronauts for long-duration exploration missions: Considerations for team performance and functioning. *REACH*, *5*, 33–56. https://doi.org/10.1016/j.reach.2017.03.002

Landsberg, C. R., Astwood, R. S., Jr., Van Buskirk, W. L., Townsend, L. N., Steinhauser, N. B., & Mercado, A. D. (2012). Review of adaptive training system techniques. *Military Psychology*, *24*(2), 96–113. https://doi.org/10.1080/08995605.2012.672903

Lang, Y., Wei, L., Xu, F., Zhao, Y., and Yu, L.-F. Synthesizing Personalized Training Programs for Improving Driving Habits via Virtual Reality. Presented at the 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2018

Lee, C. M. (1975). Mission profile of the Apollo-Soyuz Test Project - NASA Technical Reports Server (NTRS). *AIAA Student Journal*, *13*.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*(8), 1279–1292. https://doi.org/10.3758/BF03194543

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

Li, Y., Patoglu, V., & O'Malley, M. K. (2009). Negative efficacy of fixed gain error reducing shared control for training in virtual environments. *ACM Transactions on Applied Perception*, *6*(1), 1–21. https://doi.org/10.1145/1462055.1462058

Lim, D. S. S., Warman, G. L., Gernhardt, M. L., McKay, C. P., Fong, T., Marinova, M. M., Davila, A. F., Andersen, D., Brady, A. L., Cardman, Z., Cowie, B., Delaney, M. D., Fairen, A. G., Forrest, A. L., Heaton, J., Laval, B. E., Arnold, R., Nuytten, P., Osinski, G., … Williams, D. (2010). Scientific field training for human planetary exploration. *Planetary and Space Science*, *58*(6), 920–930. https://doi.org/10.1016/j.pss.2010.02.014

Linck, E., Crane, K. W., Zuckerman, B. L., Corbin, B. A., Myers, R. M., Williams, S. R., Carioscia, S. A., Garcia, R., & Lal, B. (2019). *Evaluation of a human mission to Mars by 2033*. Institute for Defense Analyses. https://apps.dtic.mil/sti/citations/AD1122304

Linde, & Miller. (2019). Applications of future technologies to detect skill decay and improve procedural performance. *Military Medicine*, *184*(Supplement_1), 72–77. https://doi.org/10.1093/milmed/usy385

Liu, C., Agrawal, P., Sarkar, N., & Chen, S. (2009). Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *International Journal of Human-Computer Interaction*, *25*(6), 506–529. https://doi.org/10.1080/10447310902963944

Liu, Q., Zhou, R., Chen, S., & Tan, C. (2012). Effects of head-down bed rest on the executive functions and emotional response. *PLOS ONE*, *7*(12). https://doi.org/10.1371/journal.pone.0052160

Liu, S.-H., Liao, H.-L., & Peng, C.-J. (2005). Applying the technology acceptance model and flow theory to online e-learning users' acceptance behavior. *Issues In Information Systems*. https://doi.org/10.48009/2_iis_2005_175-181

Liu, S.-H., Liao, H.-L., & Pratt, J. A. (2009). Impact of media richness and flow on e-learning technology acceptance. *Computers & Education*, *52*(3), 599–607. https://doi.org/10.1016/j.compedu.2008.11.002

Loehr, J. A., Guillams, M. E., Petersen, N., Hirsch, N., Kawashima, S., & Ohshima, H. (2015). Physical training for long-duration spaceflight. *Aerospace Medicine and Human Performance*, *85*(10), 14–23. https://doi.org/https://doi.org/10.3357/AMHP.EC03.2015

Lofgren, G. E., Horz, F., & Eppler, D. (2011). Geologic field training of the Apollo astronauts and implications for future manned exploration. *The Geological Society of America*, *483*, 33–48.

Loftin, R. B., & Kenney, P. J. (1994). Virtual environments in training: NASA's Hubble Space Telescope mission. *16th Interservice/Industry Training Systems & Education Conference*.

Lopes, & Lopes. (2022, January 1). *A Review of Dynamic Difficulty Adjustment Methods for Serious Games*. Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-23236-7_11

Love, S. G., & Harvey, R. P. (2014). Crew autonomy for deep space exploration: Lessons from the Antarctic Search for Meteorites. *Acta Astronautica*, *94*(1), 83–92. https://doi.org/10.1016/j.actaastro.2013.08.001

Love, S. G., & Reagan, M. L. (2013). Delayed voice communication. *Acta Astronautica*, *91*, 89–95. https://doi.org/10.1016/j.actaastro.2013.05.003

Mair, & Wilcox. (2019). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, *52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

Mammarella, N. (2020). The effect of microgravity-like conditions on high-level cognition: A review. *Frontiers in Astronomy and Space Sciences*, *7*. https://doi.org/10.3389/fspas.2020.00006

Marciacq, J.-B., & Bessone, L. (2009). Crew training safety. In *Safety Design for Space Systems* (pp. 745–815). Elsevier. http://dx.doi.org/10.1016/b978-0-7506-8580-1.00025-7

Markkula, G., Romano, R., Madigan, R., Fox, C. W., Giles, O. T., & Merat, N. (2018). Models of human decision-making as tools for estimating and optimizing impacts of vehicle automation. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(37), 153–163. https://doi.org/10.1177/0361198118792131

Marteniuk, R. G. (1976). Cognitive information processes in motor short-term memory and movement production. In *Motor Control* (pp. 175–186). Elsevier. http://dx.doi.org/10.1016/b978-0-12-665950-4.50012-2

Messeri, L. (2014). Earth as Analog: The Disciplinary Debate and Astronaut Training that Took Geology to the Moon. *Astropolitics*, *12*(2–3), 196–209. https://doi.org/10.1080/14777622.2014.964131

Milligan, G. W., Wong, D. S., & Thompson, P. A. (1987). Robustness properties of nonorthogonal analysis of variance. *Psychological Bulletin*, *101*(3), 464–470. https://doi.org/10.1037/0033-2909.101.3.464

Mishkin, A., Lee, Y., Korth, D., & LeBlanc, T. (2007). Human-Robotic missions to the Moon and Mars: Operations design implications. *2007 IEEE Aerospace Conference*. http://dx.doi.org/10.1109/aero.2007.352960

Missura, O. (2015). *Dynamic difficulty adjustment* [Doctoral Dissertation]. Rheinische Friedrich-Wilhelms-Universität Bonn.

Mohamadipanah, H., Perrone, K., Peterson, K., Garren, M., Parthiban, C., Sunkara, A., Zinn, M., & Pugh, C. (2020). Can virtual reality be used to track skills decay during the research years? *Journal of Surgical Research*, *247*, 150–155. https://doi.org/10.1016/j.jss.2019.10.030

Moon, H.-S., & Seo, J. (2020, October 20). Dynamic difficulty adjustment via fast user adaptation. *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*. http://dx.doi.org/10.1145/3379350.3418578

Moskaliuk, J., Bertram, J., & Cress, U. (2013). Training in virtual environments: Putting theory into practice. *Ergonomics*, *56*(2), 195–204. https://doi.org/10.1080/00140139.2012.745623

Mueller, D. D. (1963). *Zero gravity indoctrination for the Gemini/Apollo astronauts*. Defense Technical Information Center. https://apps.dtic.mil/sti/citations/AD0402786

Murtazin, R., & Petrov, N. (2012). Short profile for the human spacecraft Soyuz-TMA rendezvous mission to the ISS. *Acta Astronautica*, *77*, 77–82. https://doi.org/10.1016/j.actaastro.2012.03.019

Newberg, A. B., & Alavi, A. (1998). Changes in the central nervous system during long-duration space flight: Implications for neuro-imaging. *Advances in Space Research*, *22*(2), 185–196. https://doi.org/10.1016/S0273-1177(98)80010-0

Newell, K. M. (1985). Coordination, control and skill. In *Advances in Psychology* (pp. 295–317). Elsevier. http://dx.doi.org/10.1016/s0166-4115(08)62541-8

Ng, Y.-L., Ma, F., Ho, F. K., Ip, P., & Fu, K. (2019). Effectiveness of virtual and augmented reality-enhanced exercise on physical activity, psychological outcomes, and physical performance: A systematic review and meta-analysis of randomized controlled trials. *Computers in Human Behavior*, *99*, 278–291. https://doi.org/10.1016/j.chb.2019.05.026

Oluwafemi, F. A., Abdelbaki, R., Lai, J. C.-Y., Mora-Almanza, J. G., & Afolayan, E. M. (2021). A review of astronaut mental health in manned missions: Potential interventions for cognitive and mental health challenges. *Life Sciences in Space Research*, *28*, 26–31. https://doi.org/10.1016/j.lssr.2020.12.002

Ombergen, V., Demertzi, Tomilovskaya, Jeurissen, Sijbers, Kozlovskaya, Parizel, de Heyning, V., Sunaert, Laureys, & Wuyts. (2017). The effect of spaceflight and microgravity on the human brain. *Journal of Neurology*, *264*(1), 18–22. https://doi.org/10.1007/s00415-017-8427-x

Onla-or, S., & Winstein, C. J. (2007). Determining the optimal challenge point for motor skill learning in adults with moderately severe parkinson's disease. *Neurorehabilitation and Neural Repair*, *22*(4), 385–395. https://doi.org/10.1177/1545968307313508

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human Behavior*, *24*(5), 2415–2433. https://doi.org/10.1016/j.chb.2008.02.016

Park, J., MacRae, H., Musselman, L. J., Rossos, P., Hamstra, S. J., Wolman, S., & Reznick, R. K. (2007). Randomized controlled trial of virtual reality simulator training: Transfer to live patients. *The American Journal of Surgery*, *194*(2), 205–211. https://doi.org/10.1016/j.amjsurg.2006.11.032

Patel, Brunstetter, Tarver, Whitmire, Zwart, Smith, & Huff. (2020). Red risks for a journey to the red planet: The highest priority human health risks for a mission to Mars. *Npj Microgravity*, *6*(1), 1–13. https://doi.org/10.1038/s41526-020-00124-6

Peres, S. C., Pham, T., & Phillips, R. (2013). Validation of the system usability scale (SUS). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 192–196. https://doi.org/10.1177/1541931213571043

Phinney, W. C. (2015). *Science training history of the apollo astronauts*. NASA Technical Reports Server (NTRS); NASA. https://ntrs.nasa.gov/citations/20190026783

Piechowski, S., Pustowalow, W., Arz, M., Rittweger, J., Mulder, E., Wolf, O. T., Johannes, B., & Jordan, J. (2020). Virtual reality as training aid for manual spacecraft docking. *Acta Astronautica*, *177*, 731–736. https://doi.org/10.1016/j.actaastro.2020.08.017

Plass, J. L., Homer, B. D., Pawar, S., Brenner, C., & MacNamara, A. P. (2019). The effect of adaptive difficulty adjustment on the effectiveness of a game to develop executive function skills for learners of different ages. *Cognitive Development*, *49*, 56–67. https://doi.org/10.1016/j.cogdev.2018.11.006

Pollack, I. (1968). Methodological examination of the PEST (Parameter Estimation by Sequential Testing) procedure. *Perception & Psychophysics*, *3*(4B).

Pomeroy, W. (1973). Skylab Press Kit. *NASA*.

Psotka, J. (1995a). Immersive training systems: Virtual reality and education and training. *Instructional Science*, *23*(5), 405–431. https://doi.org/10.1007/BF00896880

Psotka, J. (1995b). Immersive training systems: Virtual reality and education and training. *Instructional Science*, *23*(5–6), 405–431. https://doi.org/10.1007/bf00896880

Rafiq, A., Hummel, R., Lavrentyev, V., Derry, W., Williams, D., & Merrell, R. C. (2006). Microgravity effects on fine motor skills: Tying surgical knots during parabolic flight. *Aviation, Space, and Environmental Medicine*, *77*(8), 852–856.

Raybourn, E. M. (2007). Applying simulation experience design methods to creating serious game-based adaptive training systems. *Interacting with Computers*, *19*(2), 206–214. https://doi.org/10.1016/j.intcom.2006.08.001

Rickel, J., & Johnson, W. L. (1999a). Virtual humans for team training in virtual reality. *Proceedings of the Ninth World Conference on AI in Education*, 578–585.

Rickel, J., & Johnson, W. L. (1999b). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, *13*(4–5), 343–382. https://doi.org/10.1080/088395199117315

Ritter, F. E., Yeh, K.-C., Cohen, M. A., Weyhrauch, P., Kim, J. W., & Hobbs, J. N. (2013). Declarative to procedural tutors: a family of cognitive architecture-based tutors. *Proceedings of the 22nd Annual Conference on Behavior Representation in Modeling & Simulation*.

Roberts, D. R., Asemani, D., Nietert, P. J., Eckert, M. A., Inglesby, D. C., Bloomberg, J. J., George, M. S., & Brown, T. R. (2019). Prolonged microgravity affects human brain structure and function. *American Journal of Neuroradiology*, *40*(11), 1878–1885. https://doi.org/10.3174/ajnr.A6249

Robertson, J. M., Dias, R. D., Gupta, A., Marshburn, T., Lipsitz, S. R., Pozner, C. N., Doyle, T. E., Smink, D. S., Musson, D. M., & Yule, S. (2020). Medical event management for future deep space exploration missions to Mars. *Journal of Surgical Research*, *246*, 305–314. https://doi.org/10.1016/j.jss.2019.09.065

Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., & Penn, P. R. (2000). Training in virtual environments: Transfer to real world tasks and

equivalence to real task training. *Ergonomics*, *43*(4), 494–511. https://doi.org/10.1080/001401300184378

Roy-O'Reilly, Mulavara, & Williams. (2021). A review of alterations to the brain during spaceflight and the potential relevance to crew in long-duration space exploration. *Npj Microgravity*, *7*(1), 1–9. https://doi.org/10.1038/s41526-021-00133-z

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502. https://doi.org/10.1037/0022-3514.57.3.493

Salamon, N., Grimm, J. M., Horack, J. M., & Newton, E. K. (2018). Application of virtual reality for crew mental health in extended-duration space missions. *Acta Astronautica*, *146*, 117–122. https://doi.org/10.1016/j.actaastro.2018.02.034

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin*, *95*(3), 355–386. https://doi.org/10.1037/0033-2909.95.3.355

Salotti, J.-M., & Heidmann, R. (2014). Roadmap to a human Mars mission. *Acta Astronautica*, *104*(2), 558–564. https://doi.org/10.1016/j.actaastro.2014.06.038

Saluja, I. S., Williams, D. R., Woodard, D., Kaczorowski, J., Douglas, B., Scarpa, P. J., & Comtois, J.-M. (2008). Survey of astronaut opinions on medical crewmembers for a mission to Mars. *Acta Astronautica*, *63*(5–6), 586–593. https://doi.org/10.1016/j.actaastro.2008.05.002

Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education*, *69*, 452–462. https://doi.org/10.1016/j.compedu.2013.07.004

Sankaran, K., Cassady, L., Kodys, A. D., & Choueiri, E. Y. (2004). A survey of propulsion options for cargo and piloted missions to Mars. *Annals of the New York Academy of Sciences*, *1017*(1), 450–467. https://doi.org/10.1196/annals.1311.027

Santos, W. O. dos, Bittencourt, I. I., Isotani, S., Dermeval, D., Marques, L. B., & Silveira, I. F. (2018). Flow Theory to Promote Learning in Educational Systems: Is it Really Relevant? *Revista Brasileira de Informática Na Educação*, *26*(02). https://doi.org/10.5753/rbie.2018.26.02.29

Saurav, K., Dash, A., Solanki, D., and Lahiri, U. Design of a VR-Based Upper Limb Gross Motor and Fine Motor Task Platform for Post-Stroke Survivors. Presented at the 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 2018.

Sauro, F., Payler, S. J., Massironi, M., Pozzobon, R., Hiesinger, H., Mangold, N., Cockell, C. S., Frias, J. M., Kullerud, K., Turchi, L., Drozdovskiy, I., & Bessone, L. (2023). Training astronauts for scientific exploration on planetary surfaces: The ESA PANGAEA programme. *Acta Astronautica*, *204*, 222–238. https://doi.org/10.1016/j.actaastro.2022.12.034

Schiflett, S. G. (1992). Microgravity effects on standardized cognitive performance measures - NASA Technical Reports Server (NTRS). *NASA. Johnson Space Center, 5th Annual Workshop on Space Operations Applications and Research (SOAR 1991), Volume 2*. NASA.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology*, *6*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, *82*(4), 225–260. https://doi.org/10.1037/h0076770

Schmidt, R. A., Lee, T. D., Winstein, C. J., Wulf, G., & Zelaznik, H. Z. (1982). *Motor Learning and Control: A Behavioral Emphasis* (6th ed.). Human Kinetics.

Seedhouse, E. (n.d.). *Prepare for Launch* (pp. 127–172). Springer and Praxis Publishing. (Original work published 2010)

Seedhouse, E. (2016). Dragon at the International Space Station. In *SpaceX's Dragon: America's Next Generation Spacecraft* (pp. 45–62). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-21515-0_4

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual reality training improves operating room performance. *Annals of Surgery*, *236*(4), 458–464. https://doi.org/10.1097/00000658-200210000-00008

Sgobba, T., Landon, L. B., Marciacq, J.-B., Groen, E., Tikhonov, N., & Torchia, F. (2018). *Space Safety and Human Performance* (pp. 721–793). Butterworth Heinemann. https://www.sciencedirect.com/science/article/pii/B9780081018699000169?via%3Dihub (Original work published 2018)

Sheridan, T. B. (1993). Space teleoperation through time delay: Review and prognosis. *IEEE Transactions on Robotics and Automation*, *9*(5), 592–606. https://doi.org/10.1109/70.258052

Shernoff, D. J., Csikszentmihalyi, M., Shneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*, *18*(2), 158–176. https://doi.org/10.1521/scpq.18.2.158.21860

Silva, M. P., do Nascimento Silva, V., & Chaimowicz, L. (2015, November). Dynamic difficulty adjustment through an adaptive AI. *2015 14th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*. http://dx.doi.org/10.1109/sbgames.2015.16

Sinnott, C., Liu, J., Matera, C., Halow, S., Jones, A., Moroz, M., Mulligan, J., Crognale, M., Folmer, E., & MacNeilage, P. (2019, November 12). Underwater virtual reality system for neutral buoyancy training: Development and evaluation. *25th ACM Symposium on Virtual Reality Software and Technology*. http://dx.doi.org/10.1145/3359996.3364272

Siu, K.-C., Best, B. J., Kim, J. W., Oleynikov, D., & Ritter, F. E. (2016). Adaptive virtual reality training to optimize military medical skills acquisition and retention. *Military Medicine*, *181*(5S), 214–220. https://doi.org/10.7205/milmed-d-15-00164

Smith, M., Craig, D., Herrman, N., Mahoney, E., Krezel, J., McIntyre, N., & Goodliff, K. (2020). The artemis program: An overview of NASA's activities to return humans to the moon. *IEEE Xplore*. https://ieeexplore.ieee.org/abstract/document/9172323

Starek, Açıkmeşe, Nesnas, & Pavone. (2016a, January 1). *Spacecraft autonomy challenges for next-generation space missions*. Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-662-47694-9_1

Starek, Açıkmeşe, Nesnas, & Pavone. (2016b, January 1). *Spacecraft autonomy challenges for next-generation space missions*. Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-662-47694-9_1

Steinberg, F., Kalicinski, M., Dalecki, M., & Bock, O. (2015). Human performance in a realistic instrument-control task during short-term microgravity. *PLOS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0128992

Strapazzon, G., Pilo, L., Bessone, L., & Barratt, M. R. (2014). CAVES as an environment for astronaut training. *Wilderness & Environmental Medicine*, *25*(2), 244–245. https://doi.org/10.1016/j.wem.2013.12.003

Svensson, E., Angelborg-Thanderz, M., Borgvall, J., & Castor, M. (2013). *Skill decay, reacquisition training, and transfer studies in the swedish air force: A retrospective review* (1st ed.). Routledge. (Original work published 2013)

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782–787. https://doi.org/10.1121/1.1910407

Thomas, G. A., & Trevino, L. A. (1997, July 1). Extravehicular activity metabolic profile development based on Apollo, Skylab, and Shuttle missions. *SAE Technical Paper Series*. http://dx.doi.org/10.4271/972502

Thurman, R. A., & Mattoon, J. S. (1994). Virtual reality: Toward fundamental improvements in simulation-based training. *Educational Technology*, *34*(8), 56–64. https://doi.org/10.2307/44428231

Vlachogianni, P., & Tselios, N. (2021). Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review.

Journal of Research on Technology in Education, *54*(3), 392–409. https://doi.org/10.1080/15391523.2020.1867938

Vora, J., Nair, S., Gramopadhye, A. K., Duchowski, A. T., Melloy, B. J., & Kanki, B. (2002). Using virtual reality technology for aircraft visual inspection training: Presence and comparison studies. *Applied Ergonomics*, *33*(6), 559–570. https://doi.org/10.1016/s0003-6870(02)00039-x

Vora, J., Nair, S., Gramopadhye, A. K., Melloy, B. J., Medlin, E., Duchowski, A. T., & Kanki, B. G. (2001). Using virtual reality technology to improve aircraft inspection performance: Presence and performance measurement studies. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(27), 1867–1871. https://doi.org/10.1177/154193120104502703

Walberg, G. (1993). How shall we go to Mars? A review of mission scenarios. *Journal of Spacecraft and Rockets*, *30*(2), 129–139. https://doi.org/10.2514/3.11521

White House. (2017a). *National Space Policy of the United States of America*. White House Archives. https://trumpwhitehouse.archives.gov/wp-content/uploads/2020/12/National-Space-Policy.pdf

White House. (2017b). *Presidential memorandum on reinvigorating america's human space exploration program – The White House*. White House Archives. https://trumpwhitehouse.archives.gov/presidential-actions/presidential-memorandum-reinvigorating-americas-human-space-exploration-program/

Wu, S.-C., & Vera, A. H. (2019). *Supporting Crew Autonomy in Deep Space Exploration: Preliminary Onboard Capability Requirements and Proposed Research Questions. Technical Report of the Autonomous Crew Operations Technical Interchange Meeting*. NASA Technical Reports Server (NTRS). https://ntrs.nasa.gov/citations/20190032086

Xue, S., Wu, M., Kolen, J., Aghdaie, N., & Zaman, K. A. (2017). Dynamic difficulty adjustment for maximized engagement in digital games. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. http://dx.doi.org/10.1145/3041021.3054170

Yang, J. J., & Shen, Z. (2003). Effects of microgravity on human cognitive function in space flight. *Space Medicine & Medical Engineering*, *16*(6), 463–467.

Yoshida, K., Asakawa, K., Yamauchi, T., Sakuraba, S., Sawamura, D., Murakami, Y., & Sakai, S. (2013). The flow state scale for occupational tasks: Development, reliability, and validity. *Hong Kong Journal of Occupational Therapy*, *23*(2), 54–61. https://doi.org/10.1016/j.hkjot.2013.09.002

Zhu, X., Liu, Y., Zhou, B., An, M., Chen, X., Hu, F., & Jiang, G. (2015). Design of a virtual training system and application of an evaluation scheme for orientation in a spacecraft cabin. In *Lecture Notes in Electrical Engineering* (pp. 653–661). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-662-48224-7_78

Ziv, A., Wolpe, P. R., Small, S. D., & Glick, S. (2006). Simulation-Based medical education: An ethical imperative. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, *1*(4), 252–256. https://doi.org/10.1097/01.sih.0000242724.08501.63

Zohaib, M. (2018). Dynamic difficulty adjustment (DDA) in computer games: A review. *Advances in Human-Computer Interaction*, *2018*. https://doi.org/https://doi.org/10.1155/2018/5681652

Zook, A., Lee-Urban, S., Riedl, M. O., Holden, H. K., Sottilare, R. A., & Brawner, K. W. (2012, May 29). Automated scenario generation. *Proceedings of the International Conference on the Foundations of Digital Games*. http://dx.doi.org/10.1145/2282338.2282371

Appendix A: ADDITIONAL VISUALIZATIONS AND RESULTS

Figure A.1: Average difficulty progressions for EDL subtasks during training across adaptive training algorithms

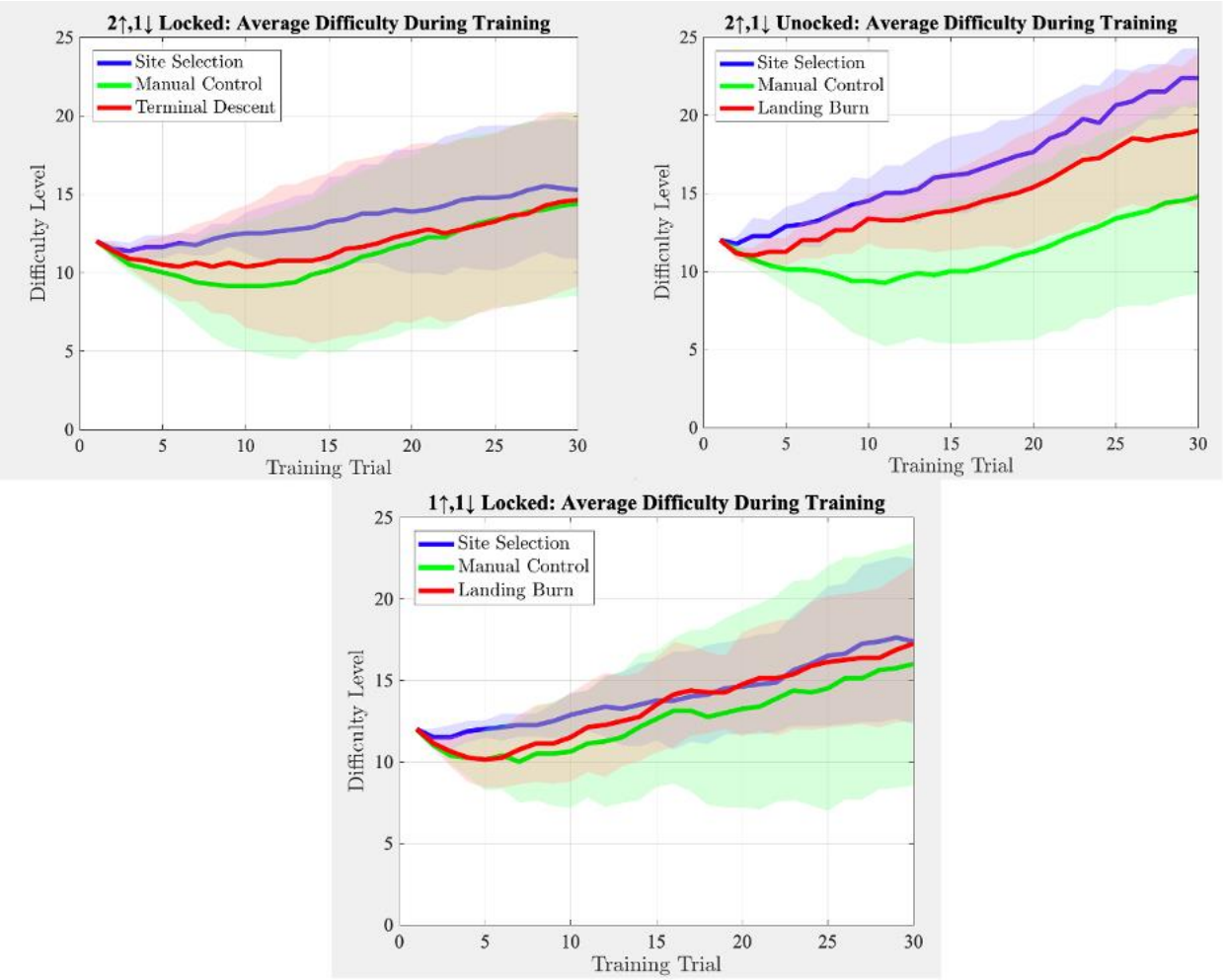


Figure A.2: Average performance on three subtasks during VR training (integration)

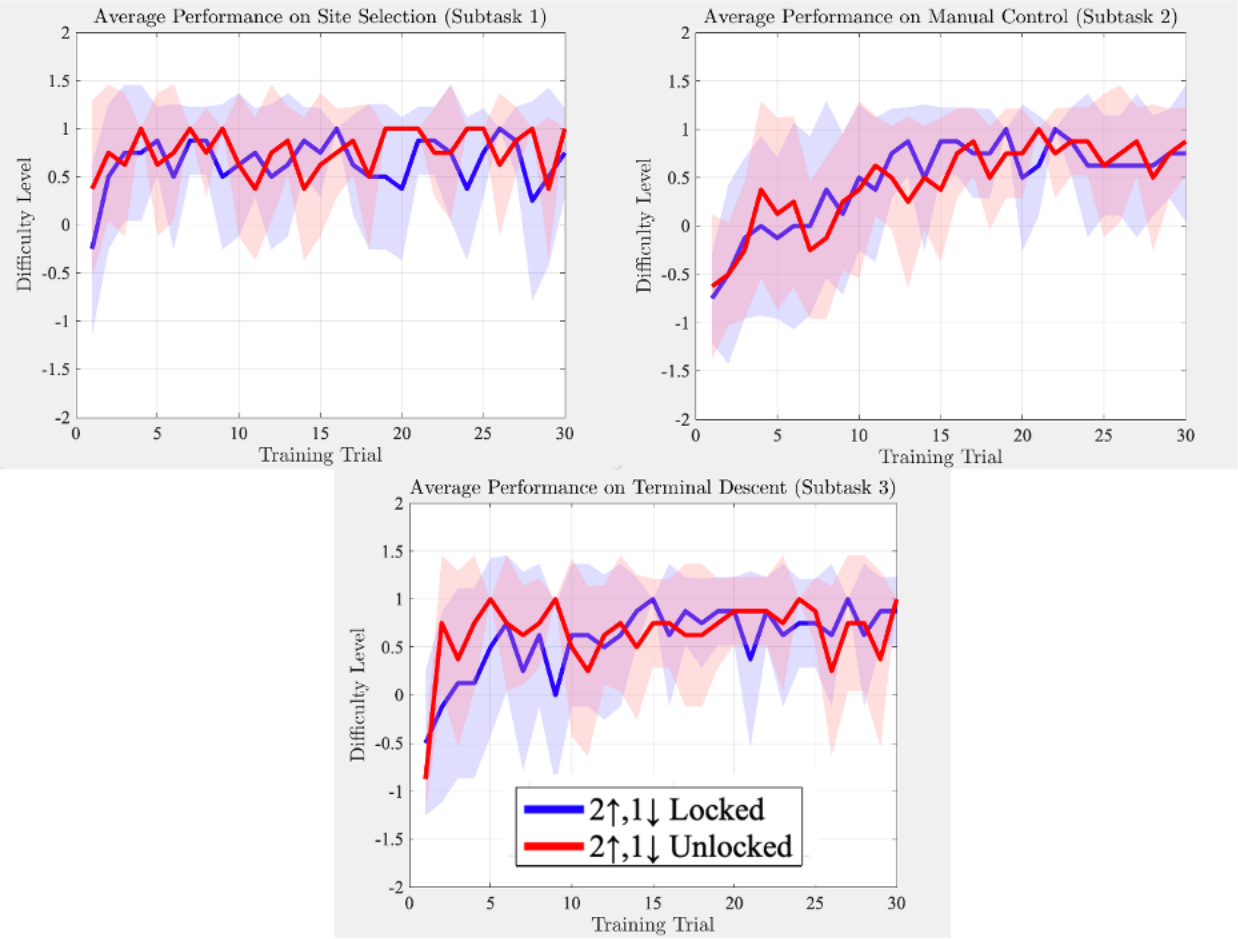


Figure A.3: Average performance across three subtasks during VR training (responsiveness)

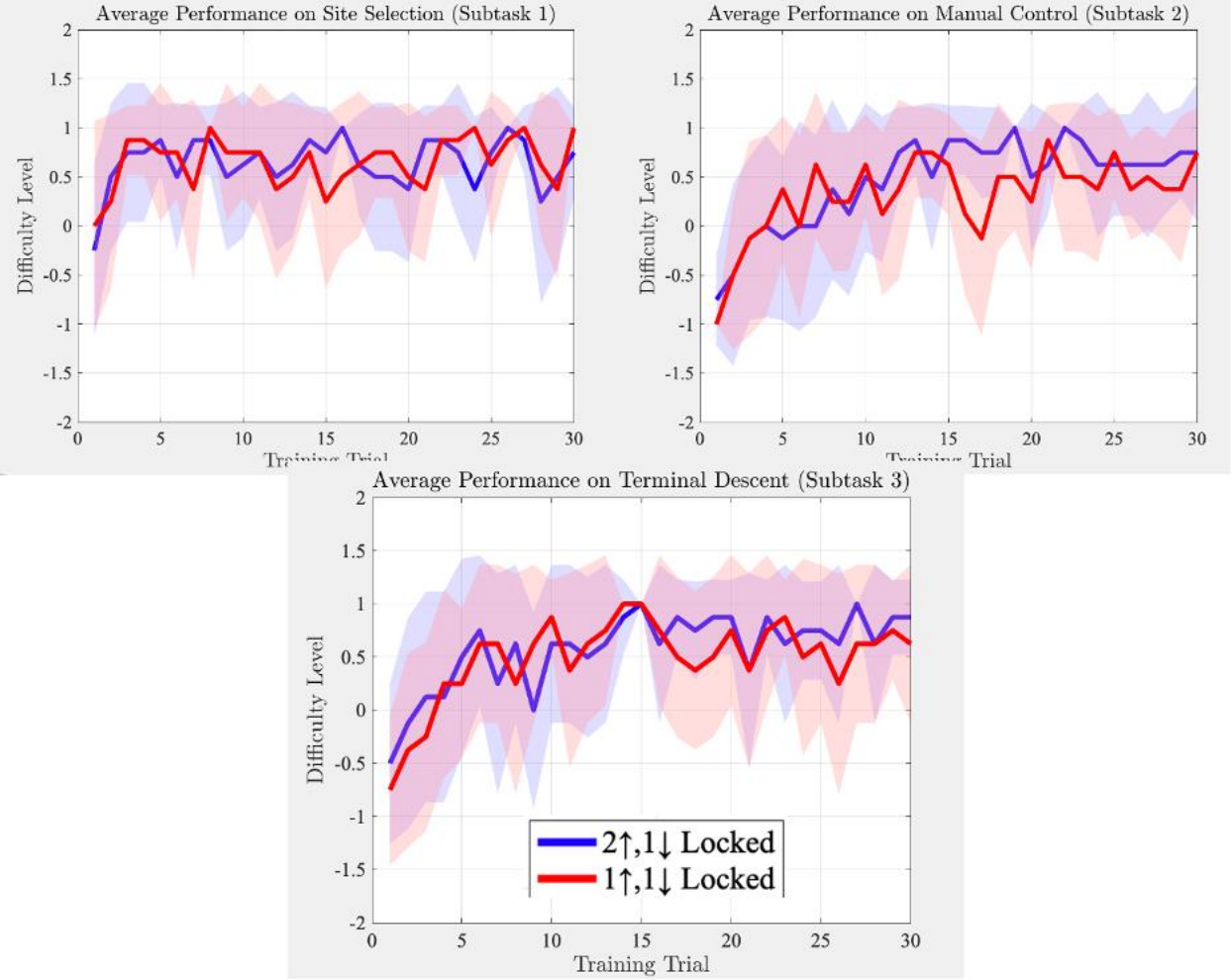
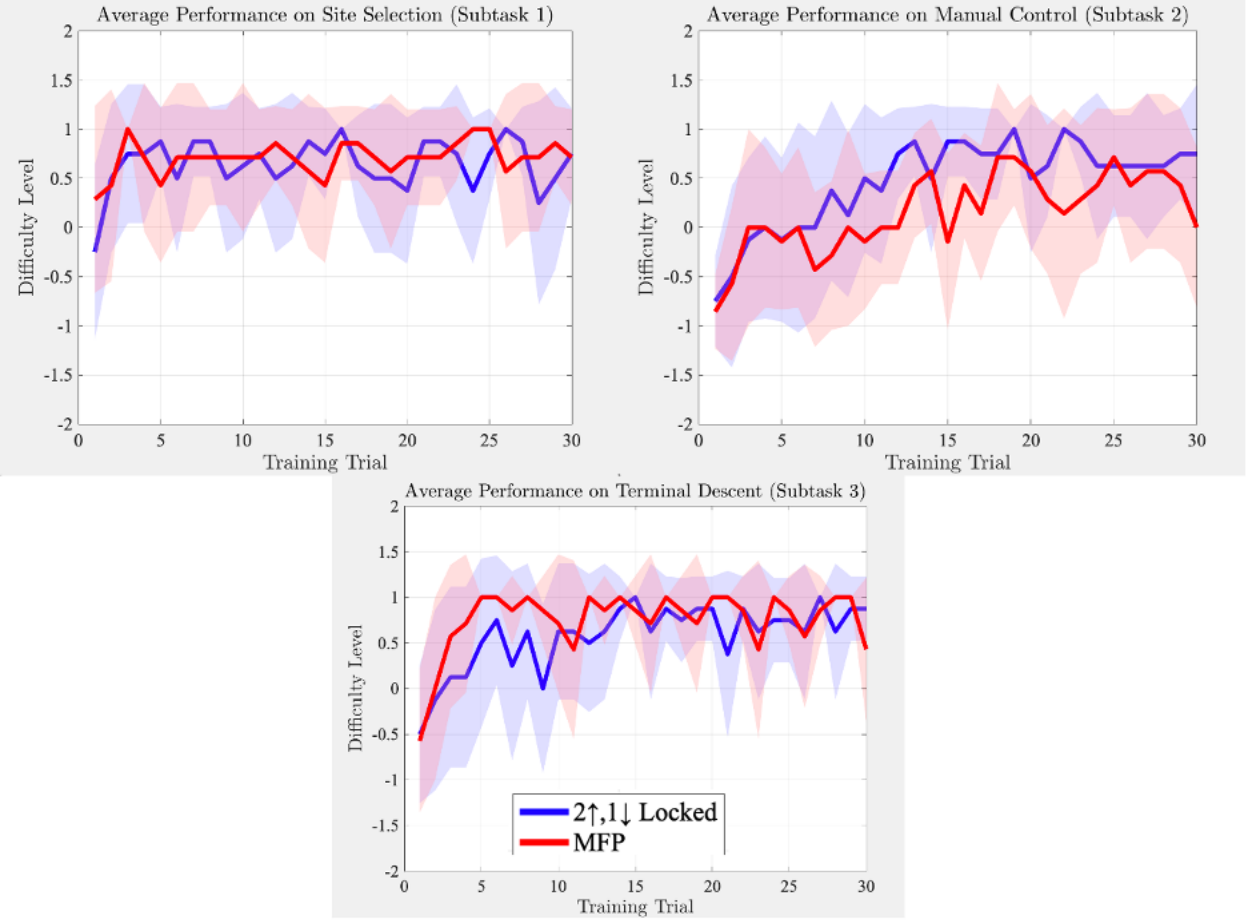


Figure A.4: Average performance on three subtasks during VR training (personalization)



Appendix B: ASSUMPTION CHECKS FOR PARAMETRIC TESTS

Figure B.1: Residuals from Mixed-Effects ANOVA on LS Difficulty Data (All 30 Training Trials)

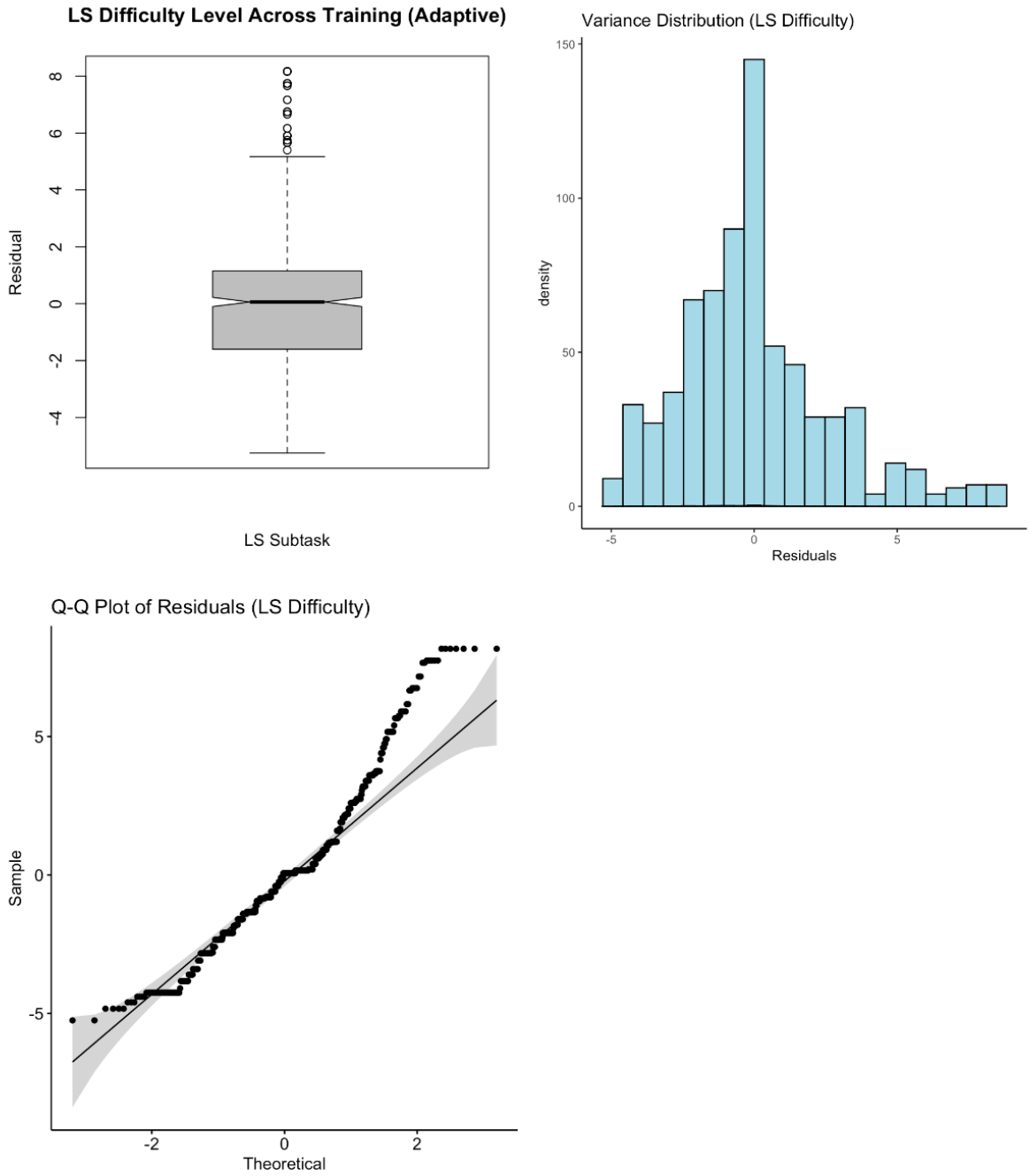


Figure B.2: Residuals from Mixed-Effects ANOVA on MC Difficulty Data (All 30 Training Trials)

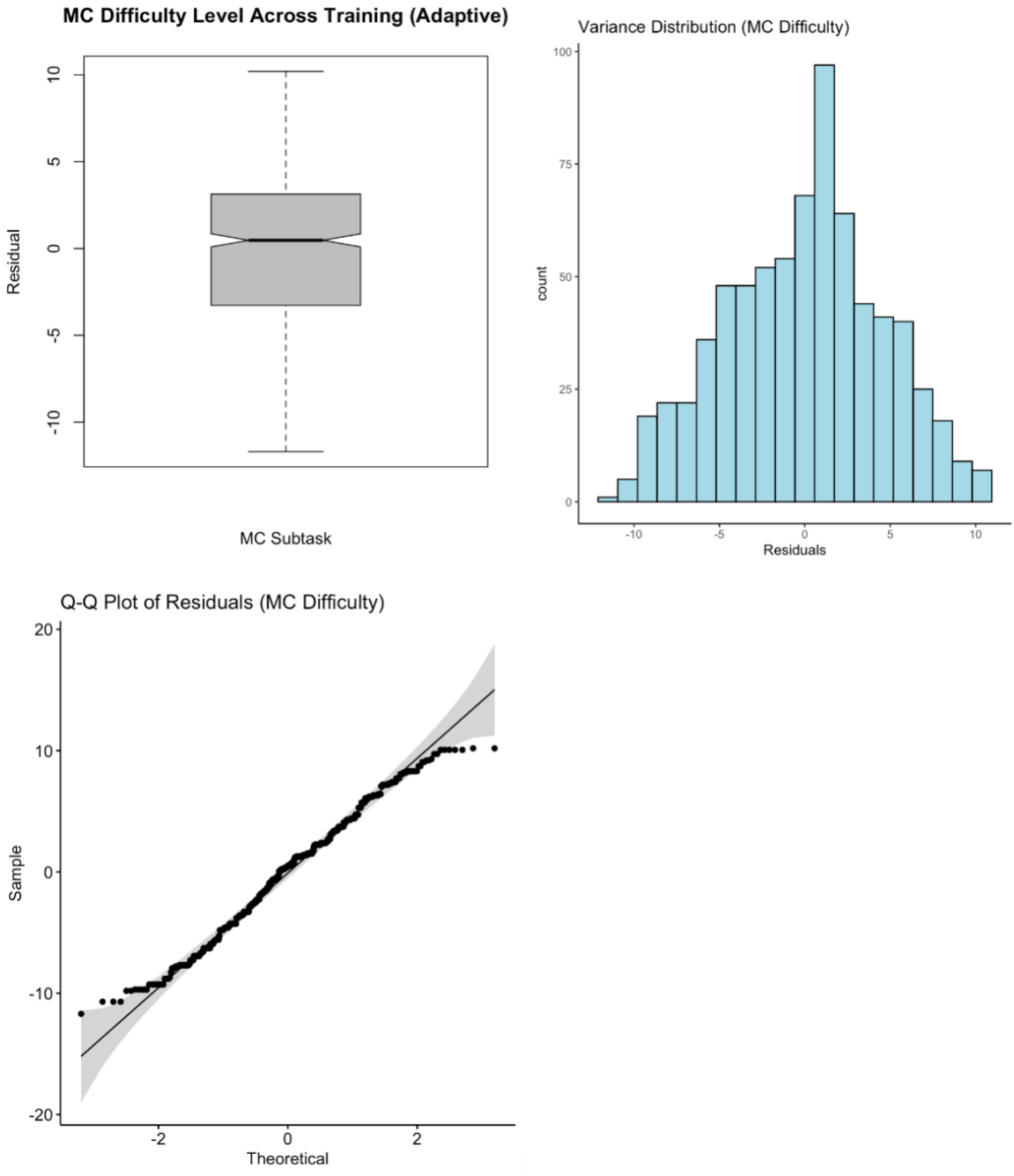


Figure B.3: Residuals from Mixed-Effects ANOVA on TD Difficulty Data (All 30 Training Trials)

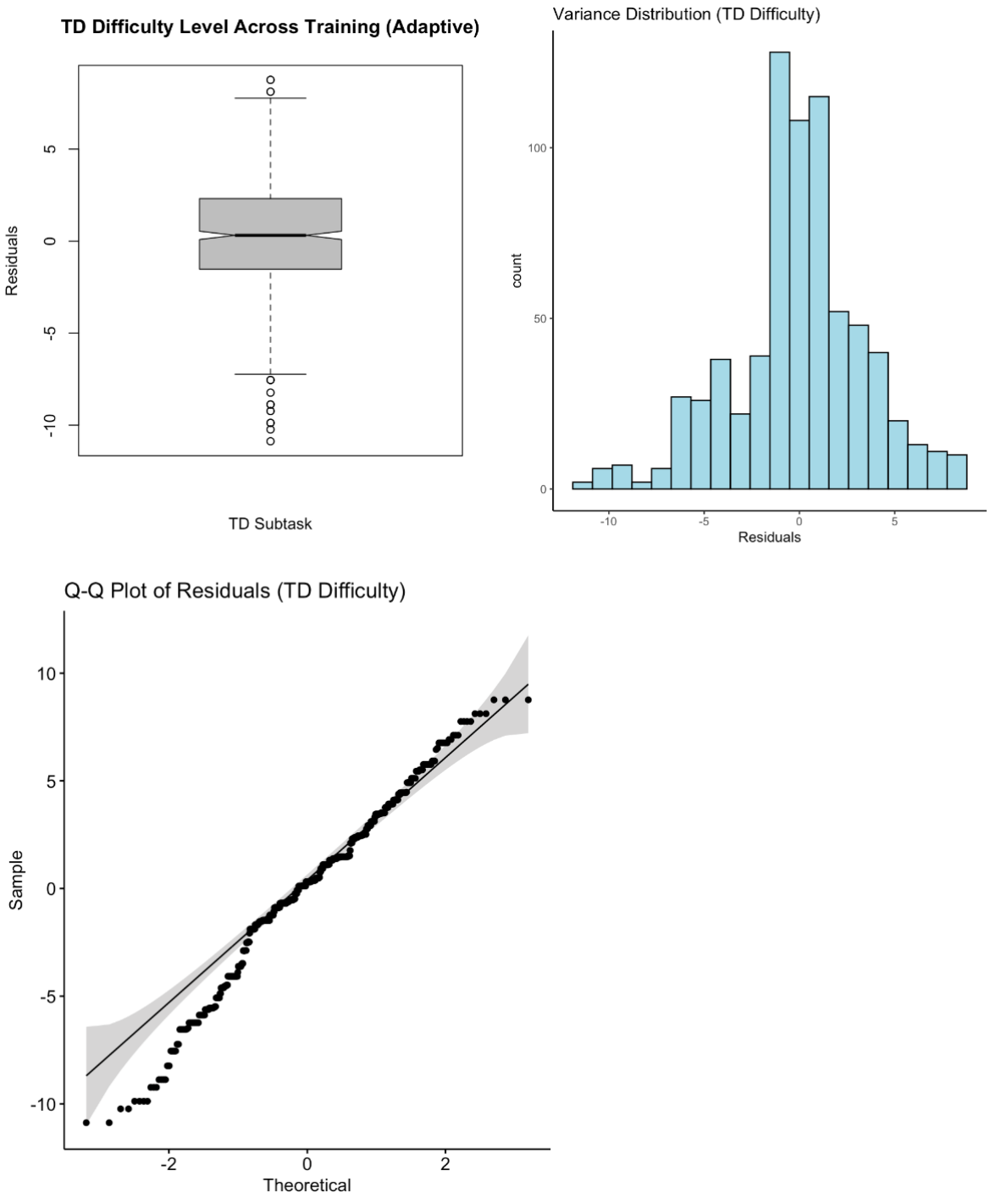


Figure B.4: Residuals from Mixed-Effects ANOVA on LS Skill Data (All 30 Training Trials)

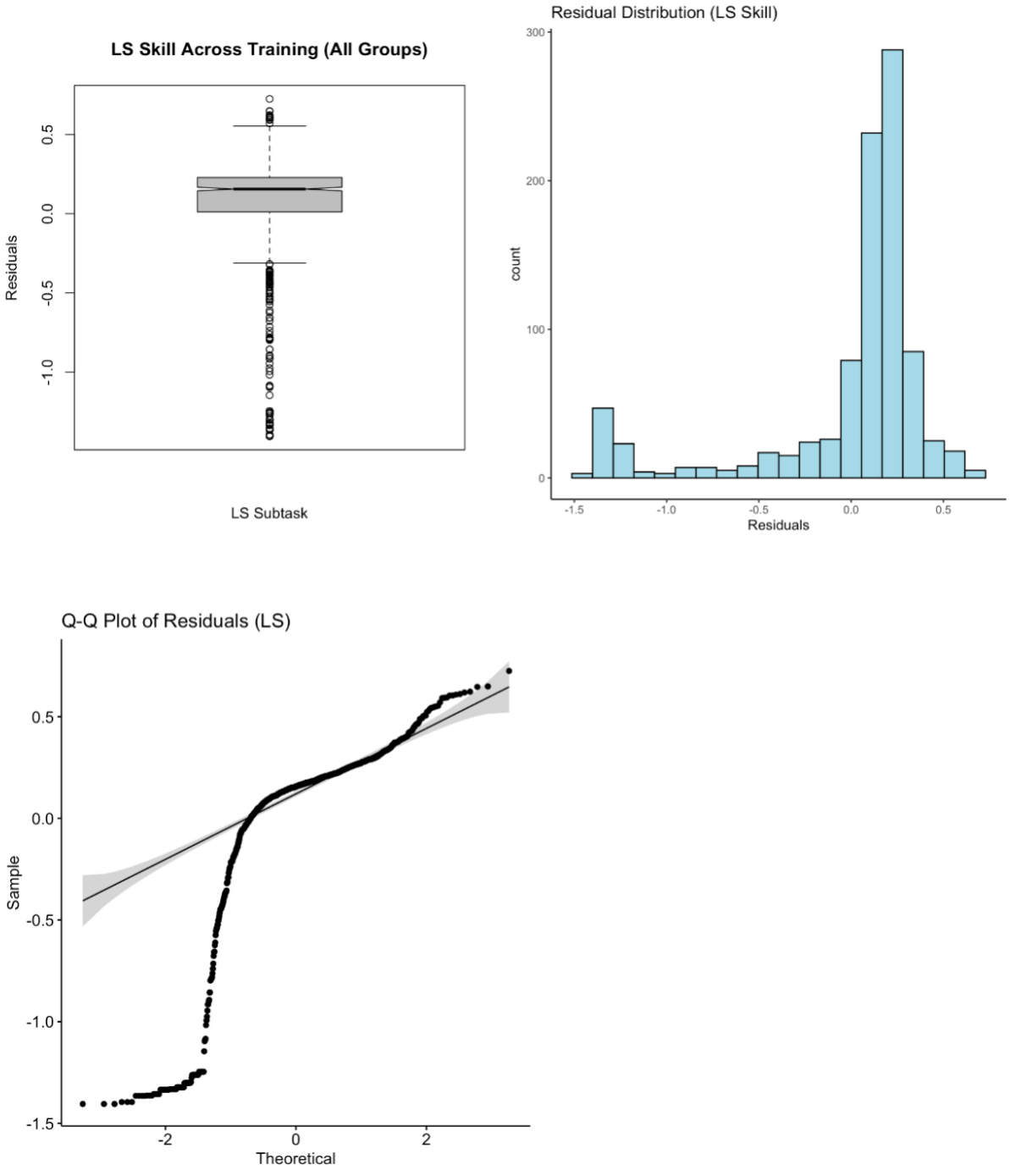


Figure B.5: Residuals from Mixed-Effects ANOVA on MC Skill Data (All 30 Training Trials)

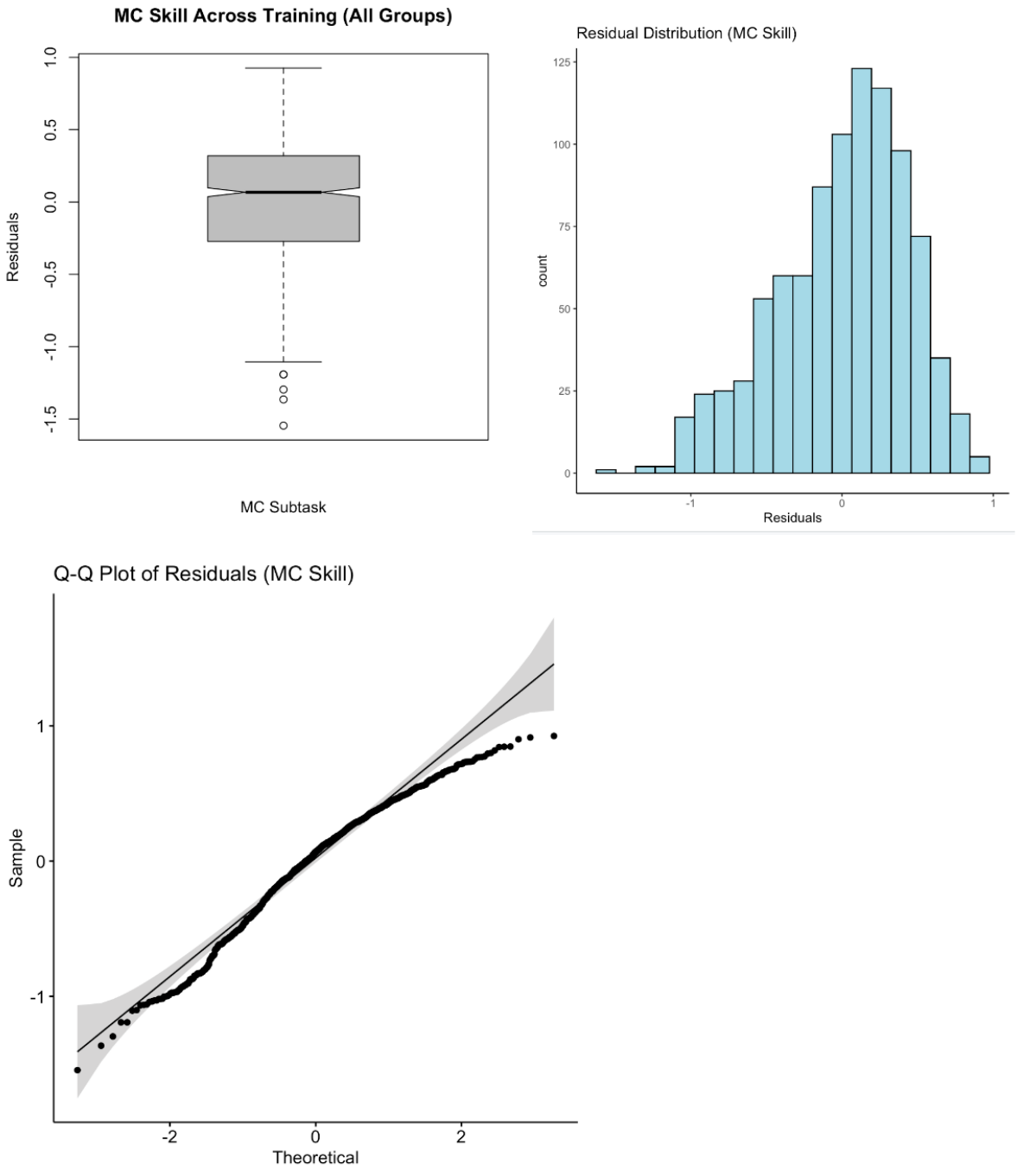


Figure B.6: Residuals from Mixed-Effects ANOVA on TD Skill Data (All 30 Training Trials)

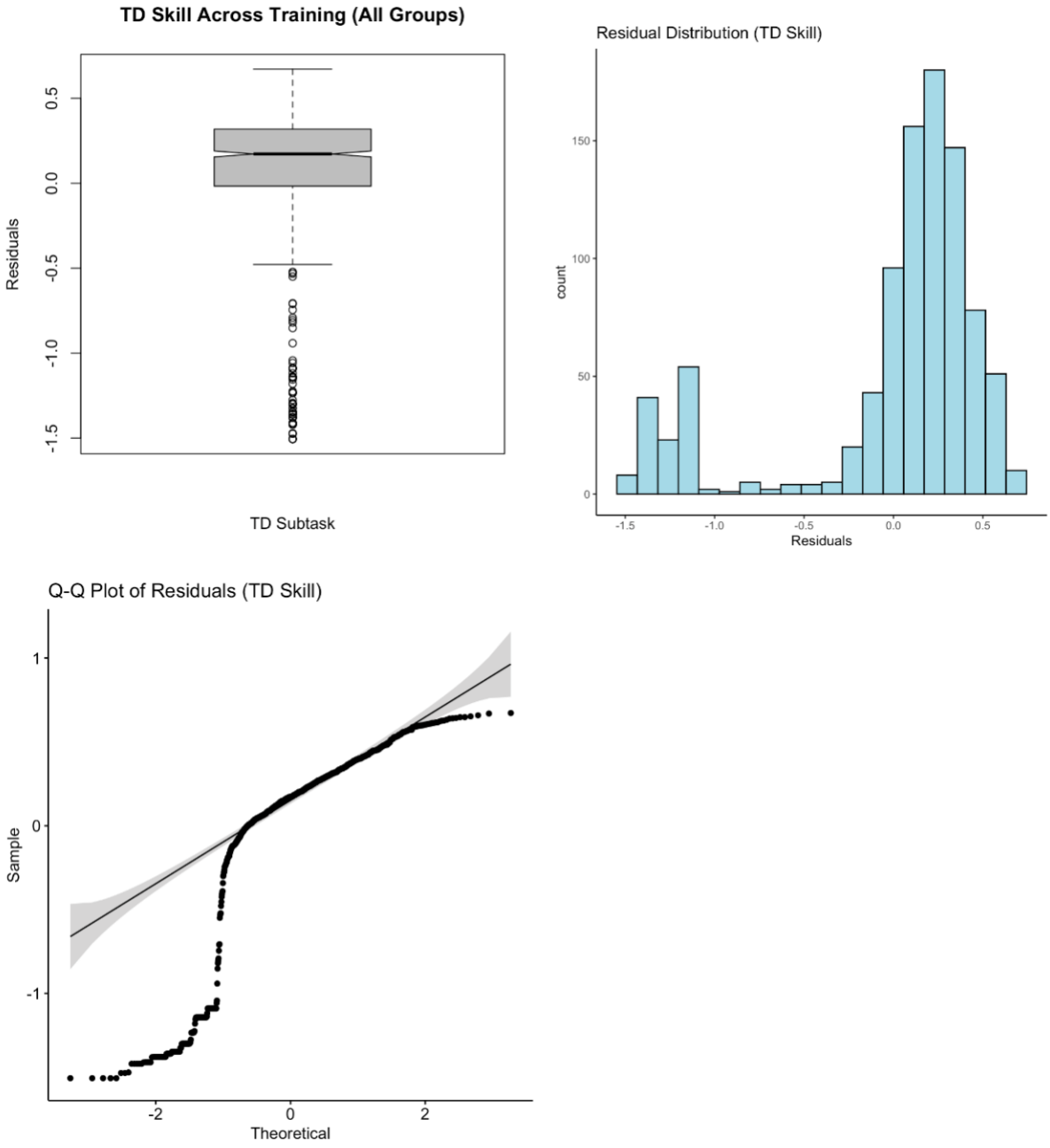


Figure B.7: Residuals from Mixed-Effects ANOVA on LS Skill Data (All 10 Cockpit Trials)

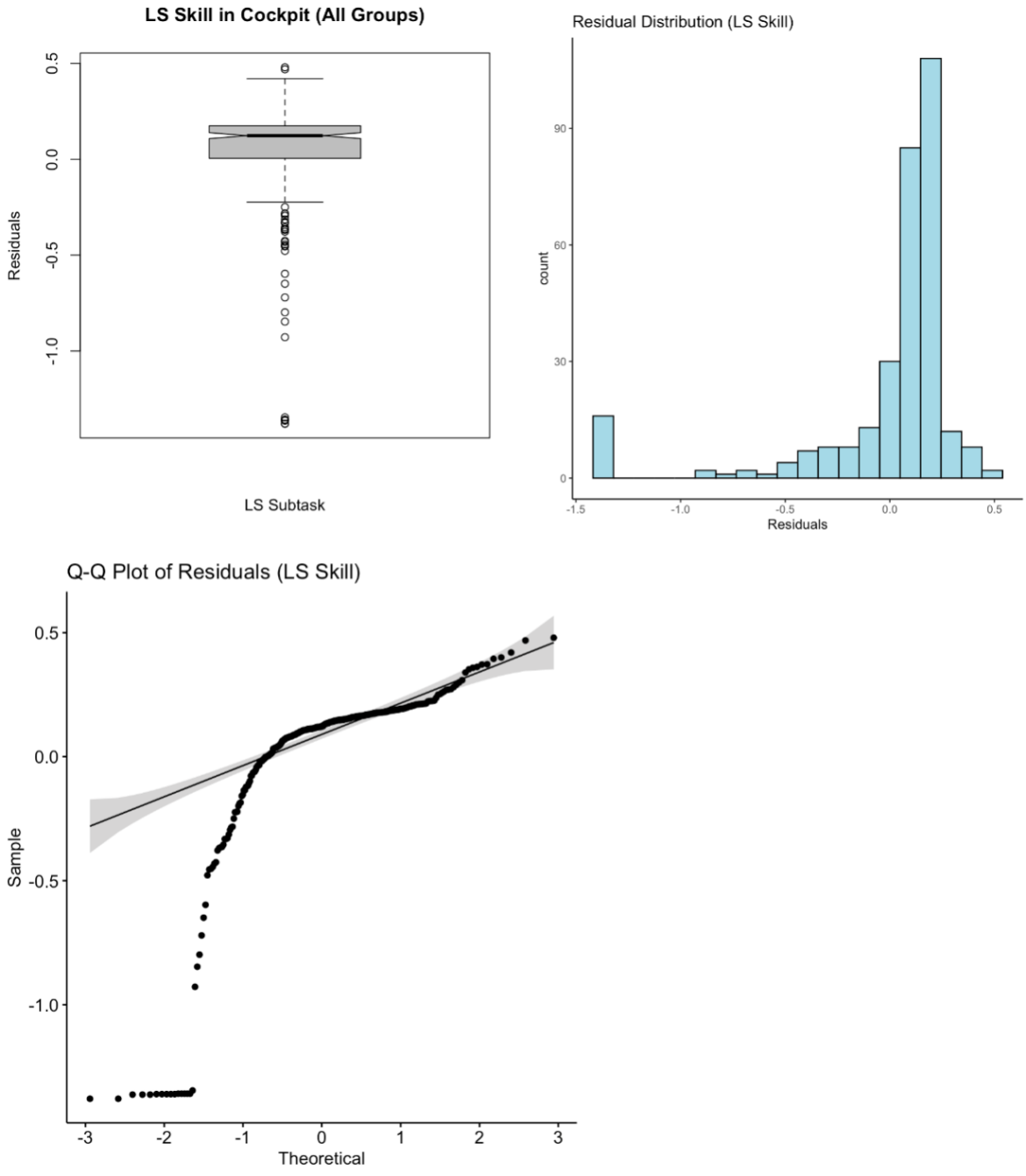


Figure B.8: Residuals from Mixed-Effects ANOVA on MC Skill Data (All 10 Cockpit Trials)

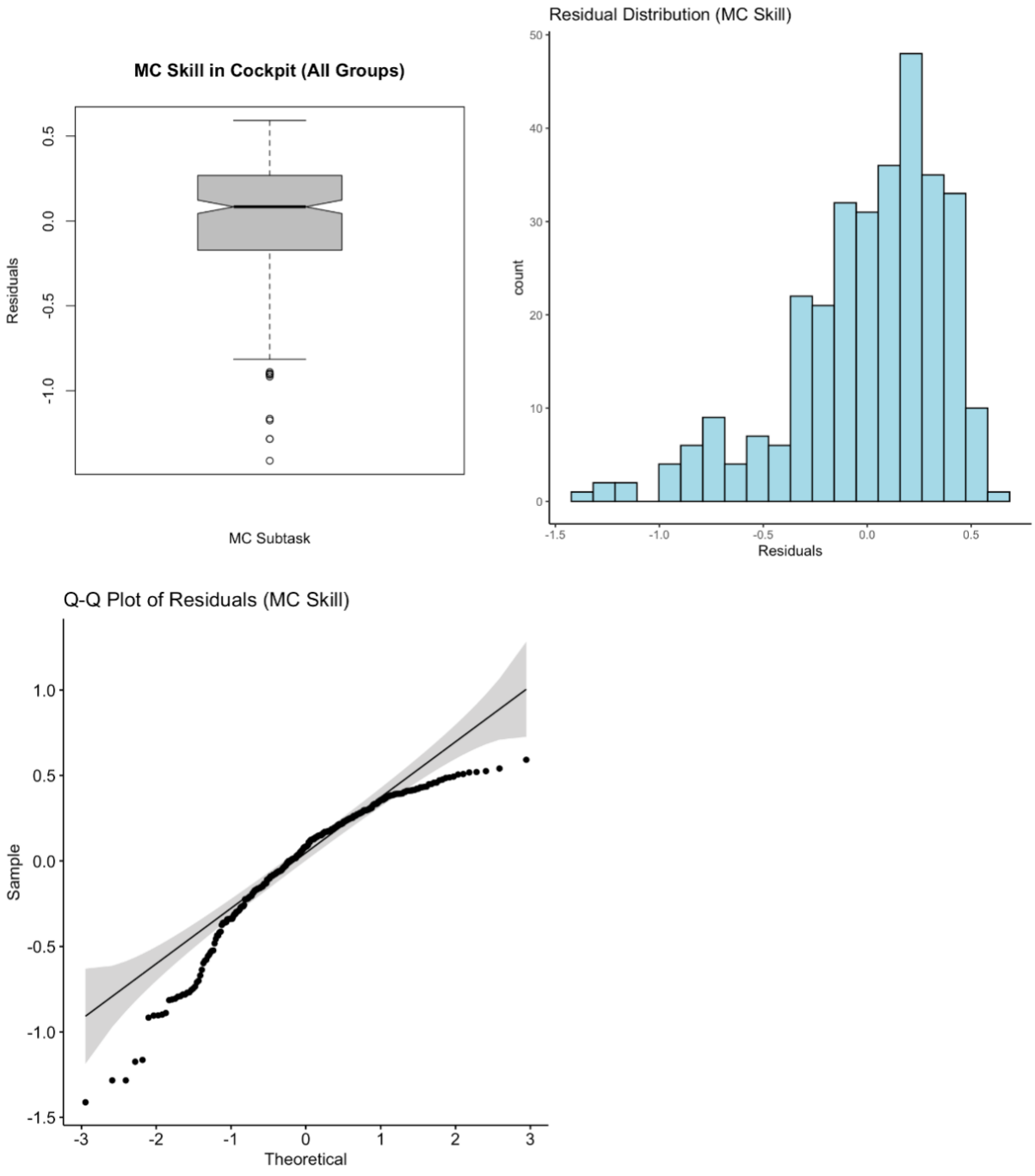


Figure B.9: Residuals from Mixed-Effects ANOVA on TD Skill Data (All 10 Cockpit Trials)