Concurrent Bandwidth Feedback for Complex Manual Control Tasks

By

John A. Karasinski

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Mechanical and Aerospace Engineering

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Stephen K. Robinson, Chair

Ron A. Hess

Zhaodan Kong

Committee in Charge

2020

Contents

List of Figures

List of Tables

Abstract

Concurrent Bandwidth Feedback for Complex Manual Control Tasks

Augmented feedback has been demonstrated to be an effective technique for improving human subject performance for a variety of simple and low-dimensional tasks, but more complex and realistic tasks are rarely explored in the literature. Instead of simply providing additional guidance, augmented feedback alerts operators to critical features of a task that they may not otherwise be aware of. However, past research has revealed a significant caution — that many forms of augmented feedback lead to the guidance hypothesis, which manifests as decreased performance when the feedback is removed.

The research presented here explores a novel approach to a specific type of augmented feedback, called concurrent bandwidth feedback, and how it might be effectively applied to avoid causing the guidance hypothesis when training for complex tasks. Concurrent bandwidth feedback is provided to an operator in real-time, during task execution, when a specific signal deviates outside of an acceptable range of values. This "Instructor Model" of feedback highlights display elements when urgent attention is needed from the operator and, consequently, appears less frequently as operator performance improves. Through the analysis of four human-in-the-loop experiments, we establish that subjects exposed to concurrent bandwidth feedback immediately and significantly improve task performance. By varying the functional task complexity in an aircraft flight task, we demonstrate that the improvements in operator performance increase with task complexity. Investigation of immediate and 24-skill retention shows that concurrent bandwidth feedback does not result in the guidance hypothesis. While increased human performance usually results in higher levels of cognitive demand, subjective workload measurements and objective secondary task performance indicate that our subjects did not experience additional cognitive load when using our novel concurrent bandwidth feedback.

Control theory-based modeling techniques are explored to further understand and predict the effects of concurrent bandwidth feedback. To explain the increased task performance, the Structural Model of the human pilot was extended by introducing a new control block that models the concurrent bandwidth feedback received by the operator. Using data from

our aircraft flight task and the Structural Model, we show that exposure to the concurrent bandwidth feedback results in increased error sensing and gain compensation, raising the resultant crossover frequency and ultimately improving pilot performance.

In all of our studies, subjects exposed to concurrent bandwidth feedback had higher levels of performance, did not experience increased workload, and did not suffer from the guidance hypothesis, indicating that concurrent bandwidth feedback can be used as an effective training technique for complex manual control tasks.

Acknowledgments

Thank you to my fellow Human/Robotics/Vehicle Integration and Performance Lab members, who elevated this work with their high standards. Thank you to Professor Robinson, who provided the inspiration for this project and has guided me throughout the whole process. Thank you to my committee members, Professor Hess and Professor Kong, who provided support in shaping the research. Thank you to Richard Joyce and Sarah O'Meara, who were always willing to discuss research over coffee. Thank you to everyone that took the time to read and edit early drafts of this manuscript. Thank you to the subjects who volunteered their time and made this research possible.

Thank you to the San Jose State University Research Foundation, who provided financial support. Thank you to the Link Foundation, who selected me for the Advanced Training and Simulation Fellowship and provided financial support. Thank you to NASA Ames Research Center's Human Systems Integration Division for having me as a Pathways Intern, providing financial support, and the knowledge and expertise of your wonderful staff.

Thank you to my family, without whom I would, quite literally, not be here.

Chapter 1

Introduction

1.1 Motivation

We aim to improve performance and decrease learning times for novice operators of highly complex motor control tasks without increasing cognitive workload. We are specifically interested in modeling and improving human performance in flight tasks, which generally require extensive training to master. The Federal Aviation Administration (FAA), for instance, requires a minimum of 1,500 hours as a pilot to captain a U.S. airline (Federal Aviation Administration, 2013). Being able to decrease this training time could lead to significant cost savings, and the predictive ability provided by modeling human performance could allow for the safer operation of aircraft.

Motor control tasks consist of a wide variety of skills such as playing tuba, pole vaulting, or flying an aircraft. An individual's performance in any of these skills can change dramatically as they transition from a novice to an expert through training. We are interested in measuring and modeling this change in performance as it changes over the course of the training process. We are also interested in developing methods to improve this performance without increasing human workload.

Humans rely on several kinds of feedback during training to improve their performance in motor control tasks. Feedback can be largely grouped into two types: internal, or intrinsic feedback, and external, or extrinsic feedback. Intrinsic feedback is anything a person can infer using their own senses: the feel of the valves of a tuba as it is played, the sense of balance mid-jump, or the sound an aircraft engine makes during a climb. Extrinsic feedback, conversely, is

provided by an external source, often in the form of an expert instructor. Extrinsic feedback comes in many forms and has a long history of improving performance in a large variety of motor control tasks.

We will focus on a specific type of extrinsic feedback known as concurrent bandwidth feedback (CBF), which is a combination of two forms of feedback. Concurrent feedback is provided in real-time, as an operator is completing a task. Bandwidth feedback is provided only when a parameter deviates outside a designated range or bandwidth. Concurrent bandwidth feedback is, therefore, feedback provided to an operator in real-time when a signal deviates out of a predefined range. This type of feedback has been shown to improve performance in many simple motor control tasks, but has not been investigated in complex, high degree of freedom tasks. We are interested in measuring, modeling, and predicting the effects of concurrent bandwidth feedback (CBF) on human performance in complex manual control tasks.

It is important to note that this feedback should be thought of as qualitative feedback, not as an additional form of quantitative guidance. We are not interested in adding additional displays or gauges to control interfaces, but would prefer to modify existing indicators, during training, to better inform an operator as to how well they are performing a task. Despite extensive evidence demonstrating the effectiveness of this feedback, the mechanism by which it improves performance has yet to be explained or integrated into human performance models. In this work, we will attempt to explain why concurrent bandwidth feedback is effective in enhancing learning and integrate this explanation into a model.

1.2 Background

1.2.1 Augmented Feedback

The concept of feedback for complex engineering systems was popularized when closed-loop control systems were first developed and has since been redefined many times (Wierner, 1948). In the context of the current research, a convenient definition of feedback comes from Ramaprasad, "[f]eedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way". The aforementioned "gap" is the error between the current and desired values and can be

conveyed to the operator of a system in a variety of ways. As mentioned previously, feedback can be broadly classified into two types: intrinsic feedback, which is generated from within the context of the action itself, and extrinsic feedback, which is given from an external source (Laurillard, 2002).

Extrinsic feedback, which is also known as augmented feedback, has been extensively studied in the field of motor learning (Sigrist et al., 2013). In their 2013 review, Sigrist et al. write "[i]t is generally accepted that augmented feedback, provided by a human expert or a technical display, effectively enhances motor learning." There are a variety of forms of augmented feedback that can be further classified by how, when, and what form the feedback is provided. Each of these choices can significantly impact the effectiveness of augmented feedback, and the summary of Sigrist et al.'s review is available in Figure 1.1. Concurrent, or real-time, feedback is displayed to the operator while the task is being executed while terminal feedback, in contrast, is displayed after the task is complete. Bandwidth feedback is only displayed to the operator when some parameter is inside (on-track feedback) or outside (off-track feedback) of an acceptable, predefined tolerance limit.

Experimentation with bandwidth feedback traces its origins to Thorndike's 1927 line-

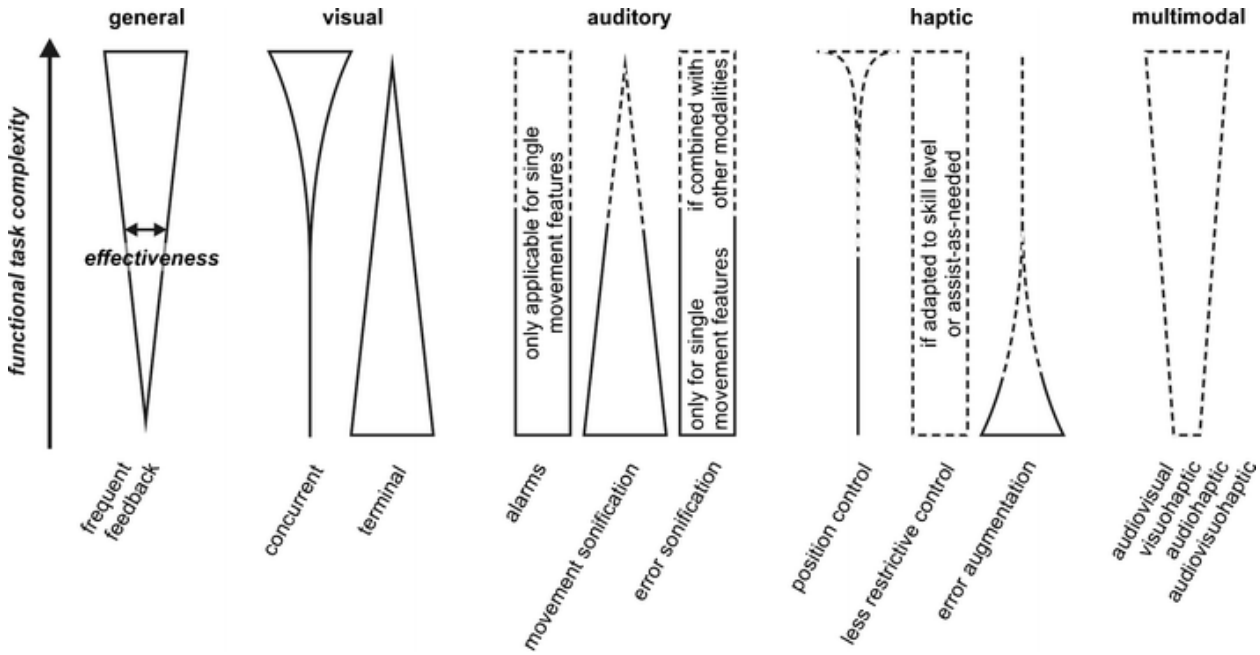


Figure 1.1: The effectiveness of different types of feedback as a function of functional task complexity (Sigrist et al., 2013). Broader shapes indicate that the feedback is more effective.

drawing experiment. In his experiment, subjects were seated and blindfolded at a table and asked to draw lines of 3, 4, 5, or 6 inches. The experiment was divided into two groups of subjects: one group received no verbal feedback, while the other group was told "right" if they were within 1/8th of an inch of the desired length for the 3 inch line, or 1/4 of an inch for the other three line lengths, and "wrong" if they were outside this bandwidth. Subjects that received the verbal bandwidth feedback improved from an initial median "right" percentage of 13% to a final "right" percentage of 54% after several training sessions. The feedback was then removed after these training trials, during which time subjects dropped to a median percentage of 26%, similar the performance of subjects who never received feedback. Subjects in Thorndike's experiments are thought to have become dependent on the verbal feedback (extrinsic feedback) rather than on their visual or proprioceptive sense (intrinsic feedback) to such an extent that they could no longer perform at the previously-demonstrated high levels with the verbal feedback removed. Similar effects were observed by Kinkade in his one-dimensional compensatory tracking task. In this study, artificial noise was injected into the reference indicator for half of the subjects and half of the subjects were provided audio augmented feedback (Kinkade, 1963). He found that augmented feedback always improved performance but that subjects with the artificial noise could not retain this benefit when the feedback was removed. In the case of Kinkade's work, the artificial noise led subjects to become dependent on the augmented feedback, and they were no longer able to use the intrinsic cues in the task. This is consistent with the guidance hypothesis (which was not formalized for another fifty years after Thorndike's experiment was concluded), which states that consistent feedback during the acquisition phase of learning leads to a dependency on that feedback (Salmoni et al., 1984). Due to this effect, for training to be useful it is important to evaluate augmented feedback in retention to test if subjects are dependent on the extrinsic (augmented) feedback rather than the intrinsic feedback provided by the task. Care must be taken to design augmented feedback such that it does not block or overshadow the task intrinsic feedback.

Payne and Hauty performed one of the first concurrent bandwidth feedback studies in 1955. In their study, subjects completed a multidimensional pursuit test which required them to scan four simulated aircraft instruments and counter their drift by adjusting simulated

aircraft controls. Subjects were placed into one of three feedback groups: a control level, where no feedback was provided, a second level, which included a single peripheral visual signal when a deviation in one of the displays occurred but did not specify which instrument, and a third level, which provided individual indicators for each of the four instruments and noted the locus of the deviation. They found a very significant effect between the different feedback groups, with the control group performing the worst, the second level performing better, and the third level performing better still. Subjects completed the test every hour for a four-hour period. Performance dropped across all three groups as time elapsed and subjects fatigued, but, despite this fatigue, the performance of the subjects in level three was superior at the end of this period compared to that of the subjects in the control group at the beginning of the experiment. They concluded by stating that "the increment is a positive function of the specificity of the information supplied, it can be ascribed largely to the directive properties of the cues, i.e., the cues impose a more efficient temporal and spatial organization upon [the subject's] scanning behavior" (Payne and Hauty, 1955).

Gordon and Gottlieb performed a rotary pursuit study investigating the effects of on-track and off-track concurrent bandwidth feedback in 1967. Subjects in their study were placed into one of three groups: control, on-track feedback, and off-track feedback. The subjects in the bandwidth feedback groups had to track a 0.75 inch by 0.75 inch target with 0.187 inch rigid stylus tip. Subjects used the stylus to track the target on a screen, and the position of the stylus and target were recorded. For subjects in the on-track feedback group, a light bulb was illuminated when they were on target, and for subjects in the off-track group, the light bulb was illuminated when they were not on the target. While both the on-track and off-track groups performed better than the control group, the off-track group performance was slightly superior. This finding was consistent with the results Williams and Briggs found in a similar task. Additionally, subjects in the feedback groups completed several trials at the end of the experiment without feedback and did not experience the loss of performance which is often seen due to the guidance hypothesis. This indicates that subjects were able to use the feedback to better learn the task and were not completely dependent on the feedback. Subjects used their own intrinsic feedback to learn the task and were able to take advantage of the concurrent bandwidth feedback to better learn and perform the task

without becoming dependent on the external feedback.

Cote et al. performed two studies to investigate the effects of visual augmented feedback and task difficulty in a two-dimensional pursuit tracking task. They presented their augmented feedback as additional, continuous displays of task difficulty and performance accuracy in the form of a bar graph. The performance accuracy augmented feedback was updated in real-time as subjects completed the two-dimensional tracking task, effectively displaying the radial offset between the cursor and target. In the first of their two experiments, trainees were split into two groups: fixed-difficulty training and adaptive training, where the difficulty of the task gradually increased with trainee skill level. In each of these groups, half of the subjects received augmented feedback while the other half did not. In their second experiment, subjects completed adaptive training followed by an adaptive transfer task, and subjects were again split such that they did or did not receive feedback in each part of the experiment. In their analysis of the results, they concluded that "visually presented augmented feedback in the form of bar graphs in a two-dimensional pursuit tracking task does not aid tracking performance in training nor transfer" (Cote et al., 1978). We hypothesize that subjects either

1. did not find feedback useful and ignored it, or

2. found the feedback useful but over-used it such that they did not attend to the actual task

The first hypothesis is proposed because the additional information provided by the bar charts did not provide the subjects with novel information that they could not directly observe from the primary display. Translating the error from the visual, two-dimensional information to a one-dimensional bar graph provides less information than one would observe from the primary display. The perceived workload associated with using this feedback may have been so large that subjects decided to ignore it. The second hypothesis is that subjects found the feedback so useful that they no longer looked at the actual task, similar to what was observed in our early experiments with augmented feedback (Karasinski, 2016). This seems less likely for this case, as the two-dimensional information would not be fully captured by the augmented feedback provided in this study, and subjects would have needed to refer

back to the primary display when large errors were presented in the augmented feedback display. Cote et al. further surmised that "the motor skill task used in these studies was too complex to permit the subjects to use the feedback cues provided," suggesting that our first hypothesis is more likely.

Lintern has published a number of studies which involved providing supplemental visual cues (augmented feedback) to subjects, who were flight-naive or otherwise in the very early stages of training, to learn to land an airplane (Lintern, 1980; Lintern et al., 1990, 1997). Early experiments in 1980 showed that an "adaptive augmented feedback group," who were provided additional command guidance cues when deviating from a specified performance envelope, successfully outperformed a control group with only a basic guidance display during retention trials. The augmented feedback also showed statistically significant improvement in actual aircraft landings conducted with a small number of subjects in a light aircraft. In 1990, Lintern et al. expanded this work with subjects taking part in a university training program. In their study, subjects in augmented feedback groups again performed better than a paired student in a control group. They were able to reduce the effects of individual flight instructors by pairing subjects in the augmented feedback groups and their control group counterpart with the same instructor. They showed that the adaptive guidance augmentation provided significant performance benefits, and that subjects that trained in their simulator for approximately two hours could reduce their actual flight training time by one and a half hours without any degradation in performance. Finally, Lintern et al. (1997) investigated the interaction effect between several variables, including augmented guidance and the quality of scene fidelity. They showed that augmented guidance again increased flight performance but that this was only the case for low scene fidelity. This is likely because the "quality of natural feedback will interact with effectiveness of augmented feedback [such] that augmented guidance would be more useful when the natural feedback was impoverished" (Lintern et al., 1997). In other words, subjects used the augmented feedback less frequently when the visual display of the task was of higher fidelity. This effect is likely to be highly dependent on the way in which the augmented feedback is presented and represents an interesting opportunity for future work.

A follow-on study to Lintern's work with augmented feedback and aircraft landing train-

ing was completed by Huet et al. (2009). They split their subjects into three groups: a control, which received no feedback, a self-controlled, augmented feedback group, and a yoked feedback group. Subjects in the augmented feedback groups were provided with Precision Approach Path Indicators (PAPIs) in the form of four lights that the subjects were trained to interpret. Subjects in the self-controlled group could request a PAPI at any time by pressing a button on the controller, after which the PAPI remained on the screen for two seconds. The yoked feedback group participants were each paired with one member of the self-controlled group and were shown the PAPIs at the same time. Their results showed that subjects in the self-controlled augmented feedback group performed the best, followed by the yoked and the control groups. While they theorized that subjects in the self-control group would gradually reduce the number of PAPI requests through training, this was not observed, and subjects tended to request the same number of PAPIs throughout the entire experiment. Despite this, the performance benefits seen in both the augmented feedback groups persisted in retention trials, suggesting that the guidance hypothesis did not hold in this case.

Recent work in the field has taken advantage of modern computing and sensory-signaling technology to produce higher fidelity simulations. In 2011, de Groot et al. investigated the effects of concurrent bandwidth feedback on learning a lane-keeping task in a driving simulator. Similar to Gordon and Gottlieb, they investigated the effects of on-track and off-track feedback compared to a control group. Instead of using a visual indicator, however, de Groot et al. used haptic feedback in the form of a vibrating chair for their feedback groups. They found that on-target and off-target groups had better lane-keeping performance than the control group, and that, similar to Gordon and Gottlieb and Williams and Briggs, the off-target group performed best. Retention trials, however, showed that much of this performance improvement was lost when the feedback was removed, which was in accordance with the guidance hypothesis. Still, some minor performance improvement was retained by the off-target group, which the authors partially attributed to the onset advantage (Fischer and Miller, 2008). The onset advantage "suggests that the sudden onset of a stimulus is a more powerful perceptual event than a stimulus offset, facilitating low-level perceptual processing and resulting in faster reaction times" (de Groot et al., 2011). This effect could

explain a repeated finding that off-track feedback is superior to on-track feedback, even if the effect is generally small. de Groot et al. also measured response time to a secondary task as an estimate of workload but found no differences across groups.

1.2.2 SAFER Experiment

In order to determine the effects of feedback for training more complex tasks such may be found in human spaceflight, we designed a series of investigations into concurrent bandwidth feedback in a four degree of freedom Simplified Aid for EVA Rescue (SAFER) task (Karasinski, 2016; Karasinski et al., 2016, 2017). SAFER is a small, propulsive jet pack worn during U.S. spacewalks for self-rescue in the event of detachment (Vassigh et al., 1998). Subjects were tasked with flying a SAFER simulation to perform a virtual inspection of the International Space Station's (ISS) solar arrays. Subjects were responsible for controlling their three-dimensional position and roll of their jet pack while pitch and yaw were automatically controlled by an autopilot.

Subjects were initially placed 40 feet away from the solar array and were asked to close to 30 feet and hold this distance for the remainder of the task. They could gauge their distance from the solar array using the indicator on the guidance display and the out-the-helmet display. Subjects were then asked to inspect four waypoints on the solar array and were given a guidance display for navigation to the waypoints.

Two vertically arranged displays in the simulator were available to complete the task (see Figure 1.2). The primary display contained an out-the-window view of the solar array and, depending on which group the subject was in, one of the guidance displays (see Figure 1.3). The secondary display, located directly below the primary display, portrayed information about the subject's current mode, remaining fuel, and a "comm" light (see Figure 1.4). This communication or "comm" light on the flight display was used as the secondary task to measure the subject's workload during each trial. This light was a colored concentric circle on the secondary display and changed from a teal color to a blue or a green color every 5 to 7 seconds (at a pseudorandom interval), and the subjects responded by pressing the corresponding button on the joystick. This secondary task was displayed on a separate screen from the flight tasks, and the change in color could not be easily distinguished from the subjects' peripheral vision, requiring the subjects to establish a visual scan pattern that

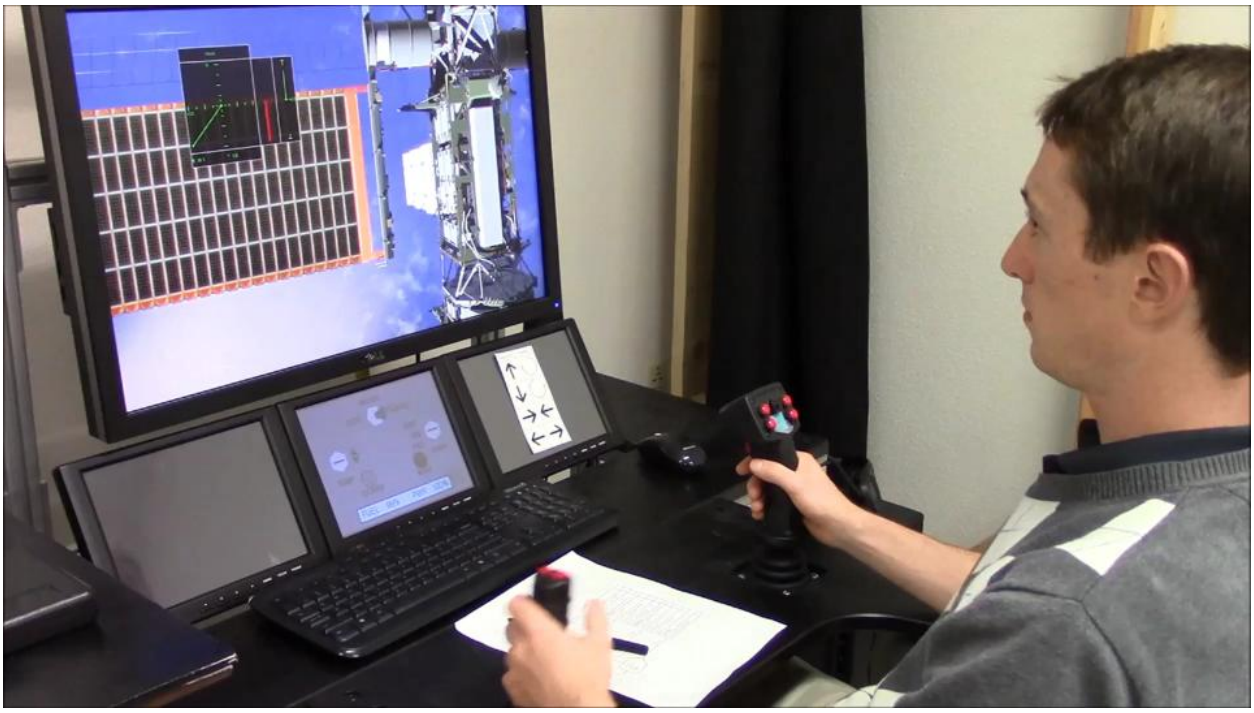


Figure 1.2: A subject from the Simplified Aid for EVA Rescue (SAFER) experiment seated in the fixed-base simulator (Karasinski, 2016).

Name	Display Max	Sig. Figs	Feedback
Control	10 (ft)	1	No
Precise	5 (ft)	2	No
Feedback	10 (ft)	1	Yes

Table 1.1: Experimental differences between subject groups.

included sequential attention to both primary and secondary displays.

In our experiment, subjects were placed into one of three groups: a control, a precise guidance group, and a concurrent bandwidth feedback (CBF) group. Subjects in the precise guidance group were given an extra significant figure in their guidance display and their analog display was scaled twice as large as the control's (but had half of the maximum value) of their flight parameters. Subjects in the CBF group had two display elements that would change from a green to a red color when the subject's performance was outside a predefined range. The two elements represented the distance from the solar array and the roll of the jet pack, and subjects were informed that monitoring and minimizing the error in

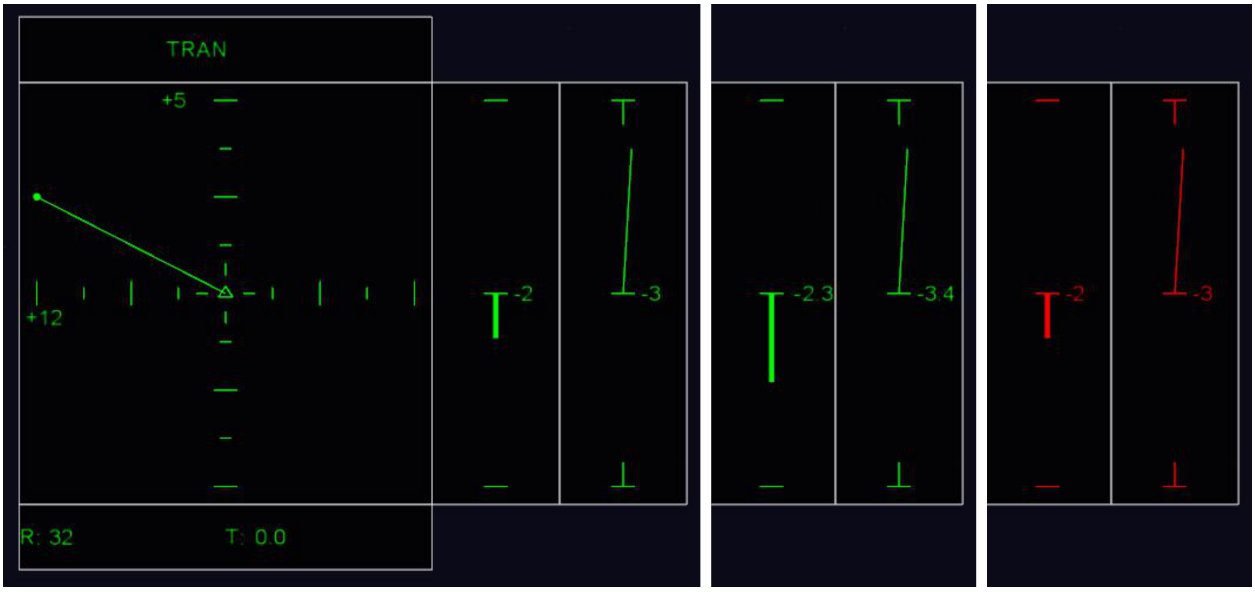


Figure 1.3: Simplified Aid for EVA Rescue (SAFER) guidance and feedback display. The left side of the figure shows the interface presented to the Control, the middle shows the change to the interface for the Precise group, and the right shows the Feedback group's interface when it is activated, while all three groups are in the same state.

these variables was their primary task. The summary of the differences between the three groups are presented in Table 1.1, and Figure 1.3 highlights the interface differences.

Performance was measured as mean absolute distance error (MAE), and results across trials are shown in Figure 1.5a. Both treatment groups performed better than the control group, with the CBF group performing the best and having the least error. The treatments had a very different effect on workload than performance, however, and subjects in the precision group reported significantly higher workload than the control group, while subjects in the CBF group reported significantly less workload than the control group (see Figure 1.5b).

The concurrent bandwidth feedback also had the added benefit of significantly reducing the amount of time required to train the subjects to their maximum skill level. Subjects with the CBF performed better on their first trial than subjects in the control group did on their last, which was after approximately two hours spent training on the task.

In summary, subjects in the precise and feedback groups had better performance in both the distance and roll tasks than subjects in the control group, and subjects in the feedback group had an initial performance that was superior to the final performance observed by subjects in the control group. The magnitude of this effect is surprising and appears to

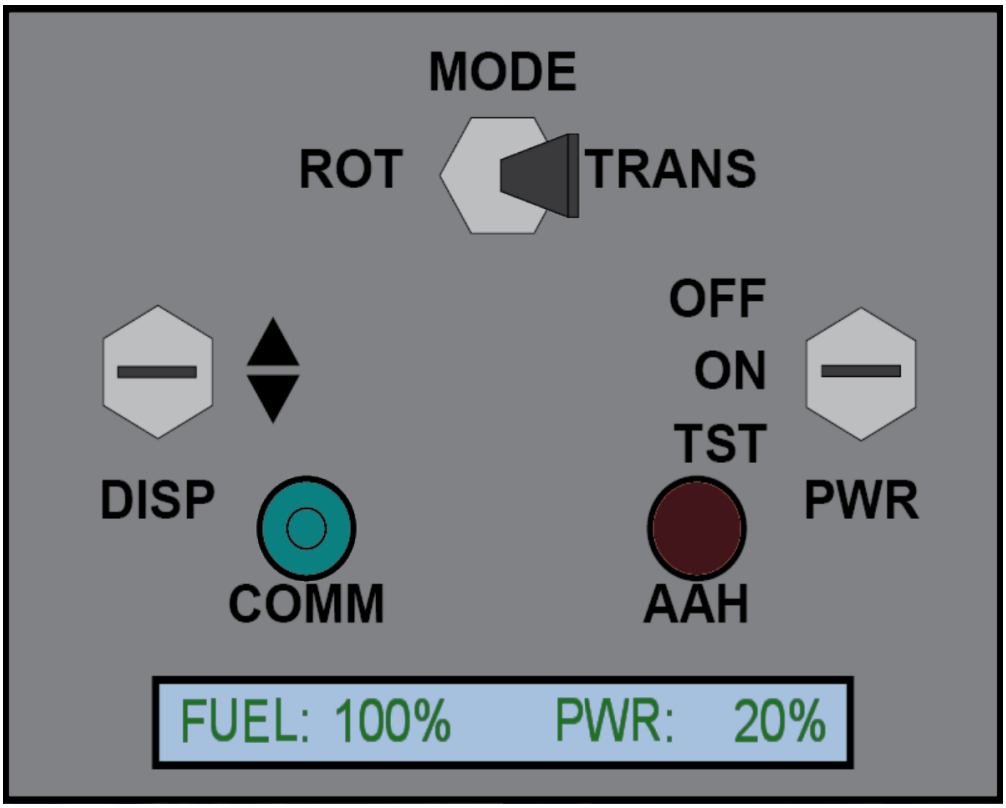
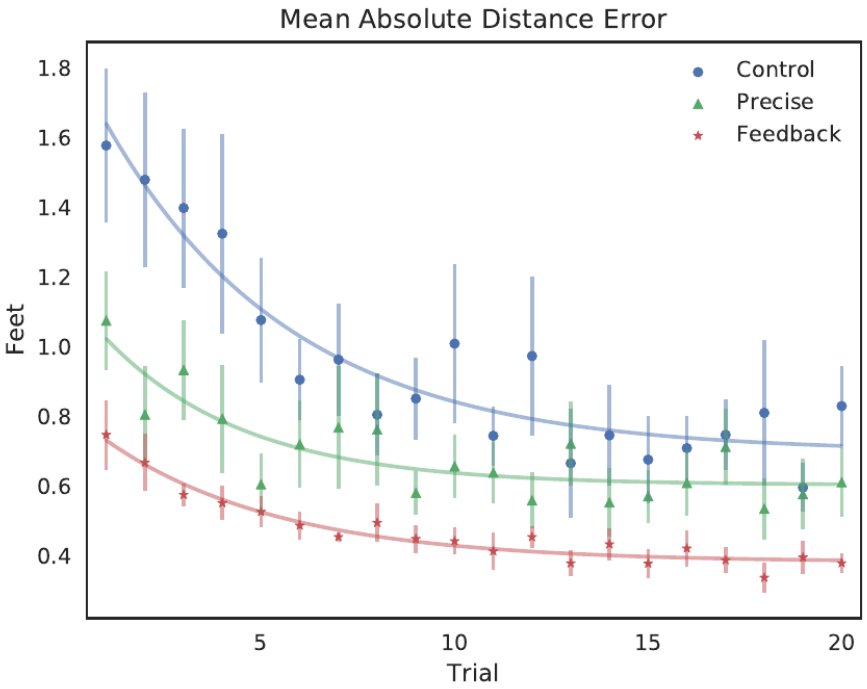


Figure 1.4: Simplified Aid for EVA Rescue (SAFER) secondary display. The current flight mode, remaining fuel and power, and "comm" light are all presented.

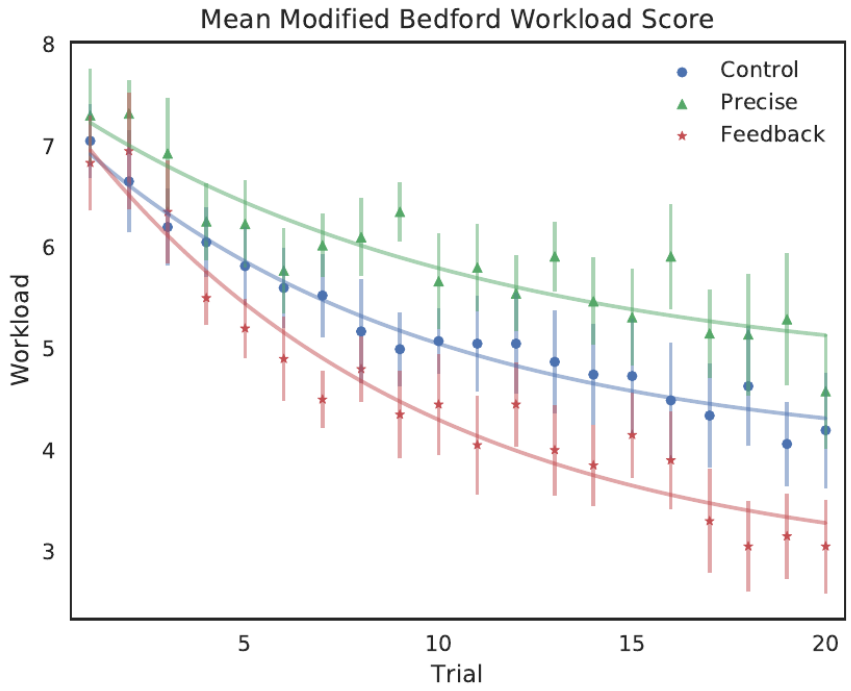
confirm that concurrent bandwidth feedback can greatly accelerate the learning of the task. While subjective workload was not significantly different among the groups, the subjects with feedback also reported the lowest workload. While presenting more precise guidance information can also improve subject performance, we showed that it does so at the cost of higher workload and longer task completion times compared to the control group. Concurrent bandwidth feedback was effective at improving subject performance and resulted in subjects reporting lower levels of subjective workload, making it the superior option for training.

1.2.3 Concurrent Bandwidth Feedback

Through our experimental work with augmented feedback, we have found concurrent bandwidth feedback to be an effective technique for enhancing performance without increasing workload. It is important to carefully consider all aspects of the task when deciding what, when, and how to provide augmented feedback. While we performed an initial assessment of different whats, whens, and hows when designing the augmented feedback technique pre-



(a) Mean absolute distance error.



(b) Mean subjective workload rating.

Figure 1.5: Subjects with concurrent bandwidth feedback (CBF) performed the best (a) and reported the lowest workload (b). Errors are the standard error of the mean (Karasinski, 2016).

sented in this dissertation, we neither completed an exhaustive study, nor did we complete extensive experimental comparisons. Instead, when designing this feedback, we considered the role of an expert instructor in training a complex skill, such as learning to fly an airplane. In this instructor model, the trainee has a basic understanding of the task, but the expert instructor can inform them on aspects of the task that they may not be aware of or that they do not have sufficient cognitive margin to attend to. By pointing at elements on a display or providing minimal verbal feedback, a skilled instructor can call attention to opportunities to improve task execution in real-time with minimal distraction to the student. The instructor chooses some bandwidth of acceptable performance for each aspect of the task, and their role in providing feedback gradually fades away as the trainee becomes more experienced with the task. Figure 1.6, adapted from Hodges and Williams, presents a schematic of the primary what, when, and how variables that need to be considered when providing augmented feedback. The choices involved in choosing the what, when, and how to provide feedback are presented here.

Deciding *what* information to provide to subjects is not trivial, especially when the task contains a large number of relevant variables. Under the paradigm of the instructor model, we have chosen to highlight elements that are provided in the guidance display rather than add on additional augmented feedback elements. Sigrist et al. states that "feedback in complex tasks should be prescriptive–that is, feedback should inform the learning on how to correct the error–rather than descriptive (i.e., information about occurrence of an error)", even though we have found that purely descriptive feedback can be effective for training complex flight tasks without increasing trainee workload. Many of the tasks considered by Sigrist et al. were targeting specific, short movements or related to sports tasks which do not normally include guidance displays like those found in aerospace tasks. As these guidance displays are prevalent in the aerospace domain, and our subjects were trained in how to use them to accomplish their tasks, it may be that prescriptive feedback is not necessary in this case. The subjects are already aware of their errors, and the descriptive feedback provided by the concurrent bandwidth feedback simply provides a description of what the most temporally relevant error is. It has also been suggested that the motor task is relatively simple for flight and driving tasks and that most of the difficulty associated with

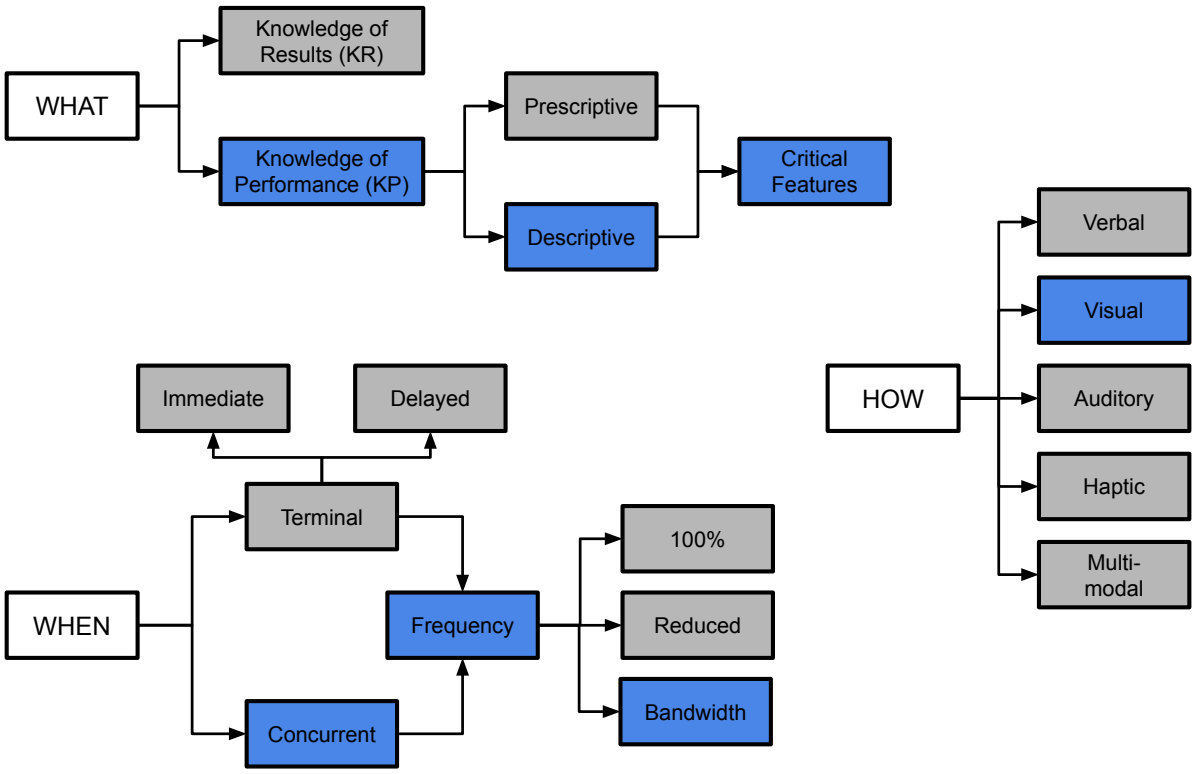


Figure 1.6: A schematic of the primary what, when, and how variables that need to be considered when providing augmented feedback, adapted from (Hodges and Williams, 2020). Blue boxes identify the primary feedback considered in this dissertation.

the task comes from the high levels of cognitive demand (Todorov et al., 1997). Under this consideration, simply knowing which error to focus on at any given time may be sufficient for improving performance and could be responsible for the reduction in workload we observed in our previous SAFER study. In any case, the type of skill and task is certainly "a critical factor in determining the effectiveness and the appropriateness of the corrective feedback types" (Tzetzis et al., 2008).

Choosing *when* to provide feedback can be one of the most import aspects of augmented feedback. Providing feedback terminally, or when the task is over, may be too late to provide a performance benefit, while providing feedback concurrently, in real-time as the task is being executed, can overwhelm the trainee. Sigrist et al. note that "it seems that the more complex the task, the more the trainee can profit from concurrent feedback." When considering the instructor model approach, we chose to focus on concurrent feedback. It is well established

that the frequency of feedback should decrease with increased skill level (Wulf et al., 1998; Wulf and Shea, 2002; Guadagnoli and Lee, 2004; Timmermans et al., 2009). It has also been shown that fading feedback, which appears less frequently over time, is beneficial for both terminal and concurrent feedback (Crowell and Davis, 2011; Kovacs and Shea, 2011). By reducing the frequency of feedback, subjects learn to develop their internal models of the task and identify their own task errors, reducing their dependency on the feedback. The role of the instructor should similarly decrease over time as trainees become more proficient. This conclusion led us to the concept of presenting concurrent feedback when performance variables deviated outside of an acceptable bandwidth. Bandwidth feedback, however, is not without its own difficulties. Several authors have noted that "[b]andwidth feedback has been shown to be effective; however, setting the error threshold is not trivial" (Timmermans et al., 2009; Ribeiro et al., 2011; Sigrist et al., 2013). This issue has usually been associated with terminal bandwidth feedback, where it is challenging to decide what types of descriptive summary statistics are appropriate and what bandwidths are sufficient. By instead choosing to present the feedback concurrently we can use an operational limit or chosen bandwidth of acceptable performance, which can make the choice of setting the error threshold simpler.

The *how* of presenting feedback has largely focused on which modality (or modalities) to use. Under the paradigm of the instructor model, we considered the verbal and visual modalities. Many aerospace tasks, however, involve significant secondary tasks in the form of managing interactions with air traffic control or communicating over other voice loops. To avoid interference and competition with these tasks, the concurrent bandwidth we developed was exclusively visual, involving color changes on the guidance display. One advantage of focusing on the visual modality is that this is where a majority of research in augmented feedback has focused, allowing us to compare between other research studies more easily.

In summary, concurrent bandwidth feedback is provided to an operator in real-time when a signal deviates out of a predefined range. In this work, the feedback is always displayed visually and changes the color of an already existing element of the display. Figure 1.7 illustrates the process of taking the signal from a task relevant feature and applying this technique. For example:

1. A task relevant signal is identified. Figure 1.7a shows an example signal, the pitch

disturbance from the aircraft flight task presented in Chapter 5.

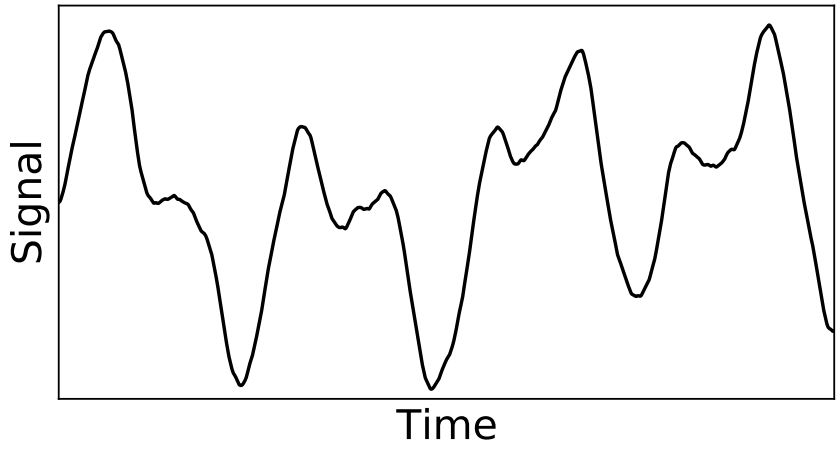
2. An operationally relevant performance bandwidth is identified. Figure 1.7b shows an example bandwidth of acceptable performance.

3. Visual feedback is applied to an existing part of the display and is presented to the operator concurrently. Figure 1.7c shows an example of feedback that would be presented to an operator. In this example, an element of the display would change from green (acceptable performance) to red (unacceptable performance) when the operator deviates outside of the acceptable pitch range.

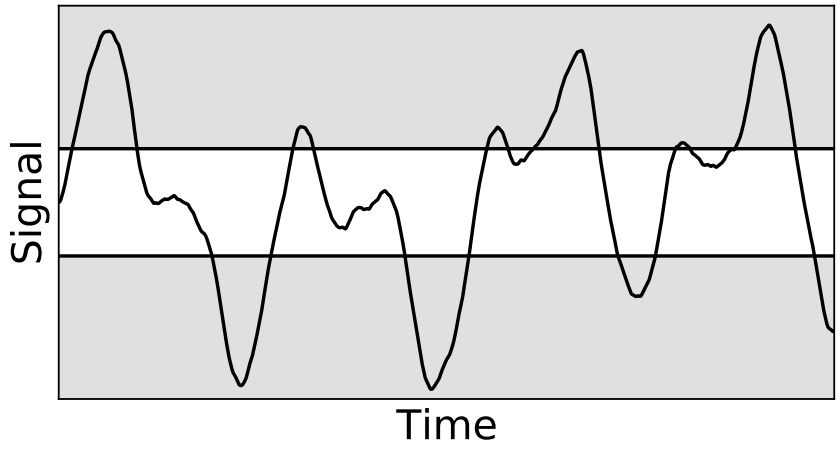
1.2.4 Workload

Improved performance, through some kind of feedback or other technique, often comes at the cost of increased workload, which can lead to a loss of the ability to maintain high performance over a sustained time frame due to fatigue (Karasinski, 2016). While improving performance is the most common motivator for supplying augmented feedback, the workload of subjects must also be considered. If performance benefits come at the cost of increased or unsustainable workload levels, for instance, the cost of augmented feedback may be too high. For this reason, we are interested in measuring the workload associated with participants to evaluate if including concurrent bandwidth feedback has a net positive effect.

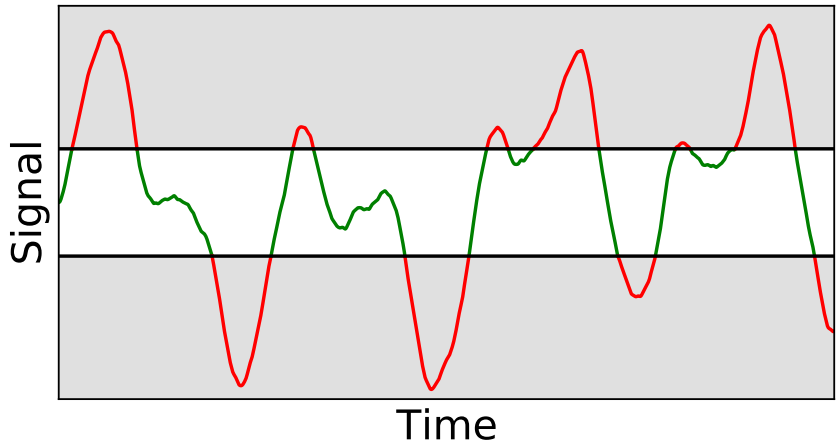
In the human factors community, one of the challenges associated with measuring workload has been the inability to agree on a single definition of the term. As Hart notes in her 2006 report, "[t]he many definitions [of workload] that exist in the psychological literature are a testament to the complexity of the construct". Workload has been defined by Hart and Staveland as "the perceived relationship between the amount of mental processing capability or resources and the amount required by the task." Hart later provided an excellent definition of workload that is broadly acceptable for many types of tasks, "[w]orkload is a term that represents the cost of accomplishing mission requirements for the human operator". More specific definitions are available when focusing in the aerospace domain. After sending a questionnaire to some 350 military and airline pilots, Ellis and Roscoe analyzed the results and proposed that "[p]ilot workload is the integrated mental and physical effort required to



(a) The signal of a task critical feature.



(b) The signal with an operationally relevant bandwidth.



(c) The signal with concurrent bandwidth feedback applied.

Figure 1.7: The implementation of concurrent bandwidth feedback to the signal of a task critical feature.

satisfy the perceived demands of a specified flight task". Regardless of the precise definition, is generally agreed that having a low workload indicates that it would be easy to complete additional tasks, while having a high workload suggests that it would be difficult.

While there are several objective techniques that attempt to capture workload, the most commonly used techniques are subjective in nature (Hart and Staveland, 1988). In the domain of aircraft flight tasks, the two most commonly used subjective workload measurements are the Modified Bedford Workload Scale and the NASA Task Load Index (NASA-TLX) (Roscoe and Ellis, 1990; Hart and Staveland, 1988; Hart, 2006). Both techniques are used in the studies presented in this dissertation and generally show good agreement for the types of tasks that we are interested in.

The Modified Bedford Workload Scale is a 10-point scale that asks operators to rate their spare time to attend to additional tasks. While taking the survey, operators ask themselves a series of questions to determine if the workload for the task was satisfactory, tolerable, possible, or impossible. The scale ranges from a score of 1 (a "piece of cake") to a 10 (adequate performance was impossible), see Figure A.1. The benefits of this scale include that the resulting workload score is a single number which can be easily analyzed statistically, that it is extremely quick to complete, and that it is easy and intuitive for inexperienced subjects to complete. Some of the negative features of the scale are that it only considers the amount of spare cognitive ability to attend to additional tasks and that there can be large inter-rater variability. Despite this, the scale is widely used for measuring the workload associated with aircraft flight tasks and has significant heritage in the aerospace domain.

The NASA Task Load Index (NASA-TLX) is one of the best-known and commonly used subjective workload measures. The NASA-TLX has been in use for thirty years and has been validated over a large variety of tasks (Hart, 2006). The NASA-TLX is a multidimensional rating scale which uses the magnitude and ranking of six subscales to produce an overall estimate of subjective workload (Hart and Staveland, 1988). The six subscales are: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration, see Figure A.2. Each of these scales is rated on a 0 (Very Low) to 100 (Very High) scale, with the exception of Performance, which is rated from 0 (Perfect) to 100 (Failure). After marking a value for each of these subscales, subjects then make fifteen pairwise weightings,

allowing them to rate each pair of subscales based on its perceived contribution to their overall workload. A final, overall workload score is computed by multiplying each subscale's score by the number of times it was chosen in the pairwise weightings, adding these values, and dividing by fifteen. As certain subscales may be more or less important than others, depending on the task being evaluated, researchers can omit subscales or simply not compute the overall score.

In addition to subjective measures of workload, there are a variety of techniques which aim to estimate objective workload. One of the most common objective measurement techniques is the secondary task, which requires subjects to complete the primary task then use any spare cognitive margin to respond to an additional task (Gawron, 2008). Secondary tasks can provide a measure more sensitive to differences in workload and performance than a single task alone and allow for a common measure between experimental conditions (Slocum et al., 1971). Care must be taken, however, to ensure that the secondary task does not intrude upon primary task performance (Williges and Wierwille, 1979). It is important that the secondary task requires a relatively small percentage of the overall cognitive margin compared to the main task, otherwise the distinction between primary and secondary task begins to blur. In our previous studies, we have used a multiple-choice reaction time task as an objective workload measurement. In this multiple-choice secondary task, subjects are presented with several different stimuli, each of which requires a different response (Lysaght et al., 1989). A subject's objective workload can then be inferred by either the percentage of secondary tasks which were correctly responded to within a given time, the number of secondary tasks which were correctly responded to in a trial, or both. We have previously found this type of task to be correlated with subjective workload scales in the aforementioned SAFER task (Karasinski et al., 2017). This secondary task approach does not work well for simple tasks, as completing it begins to compete for attention with the primary task but works well for more complex tasks and so is a valid technique for our interests.

1.2.5 Pilot Modeling

In addition to popularizing the concept of feedback, the creation of control theory in the early 1940s also provided the tools required for the mathematical modeling of the human pilot. At that time, new weapons were being created for World War II that could only be

used effectively with trained operators working with a machine. While it was thought that a human could be viewed as a unique kind of servomechanism in the control feedback loop, it was still unclear what factors affected human performance. Early work by Tustin and others extended the control theory framework and applied these theories to actual human operators (Tustin, 1944). Significant attention was focused on "attempt[ing] to find the laws of relationship of movement and error. In particular, it was hoped that this relationship [would] be approximately linear and so permit well developed theory of 'linear servomechanisms' to be applied to manual control in the same way as it applies to automatic following" (Tustin, 1944). This would allow for the prediction of human performance and the ability to predict the limits of human control.

These early works were summarized in McRuer and Krendel's report, "Dynamic Response of Human Operators." This work evaluated measurements for single-input/single-output (SISO) manual control systems and developed predictive models consistent with this data. Indeed, McRuer and Krendel wrote, "[i]t is possible, without doing violence to the data, to obtain describing functions which are generally applicable to the results of the many diverse

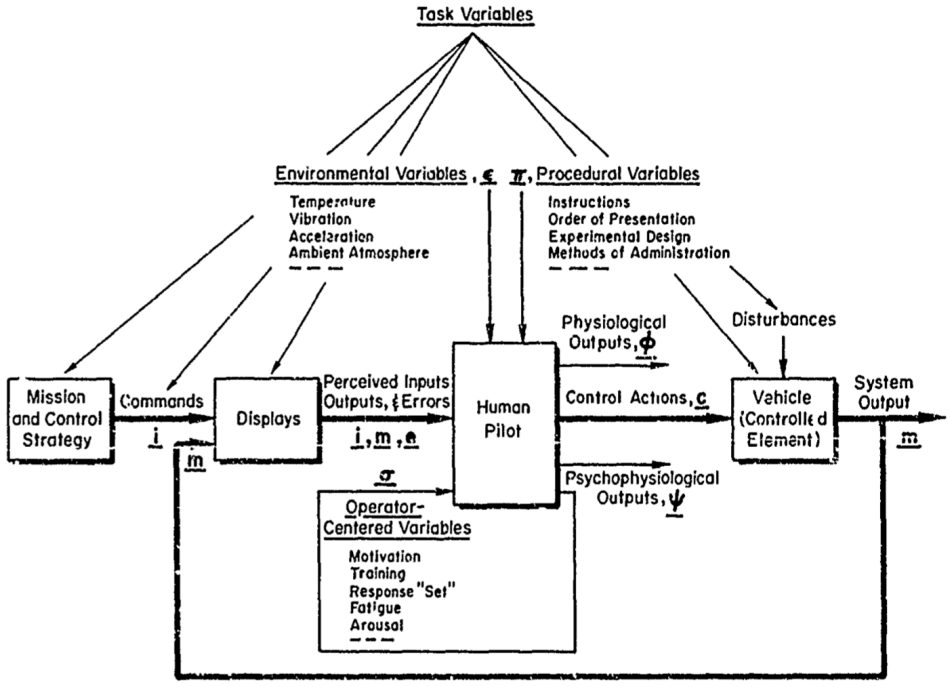


Figure 1.8: Variables affecting the pilot/vehicle system, from McRuer and Krendel (1974).

experiments." The report concludes by describing a hypothetical transfer function of the human operator which includes a time delay, a neuromuscular lag, and a gain. McRuer's early model of the complete pilot/vehicle system is presented in Figure 1.8. McRuer revisited these results in 1974 after two decades of supporting engineering and experimental psychology experiments and was able to further generalize these results to a wide variety of system dynamics. In his study, McRuer completed a detailed analysis which included the human response to proportional, rate/velocity, and acceleration type controlled element dynamics (see Table 1.2). The result of this report was the now famous "crossover model," which relates the operator and controlled element transfer characteristics by the equation

$$Y_c(jw)Y_p(jw) = \frac{w_c e^{-jw\tau_e}}{jw} \tag{1.1}$$

where Y_c is the controlled element transfer function, Y_p is the approximate human operator transfer function, w_c is the crossover frequency, and τ_e is the effective time delay of the pilot. The crossover model is so named as it allows for linear behavior at approximately -20 dB/decade slope in the region of the crossover frequency. The approximate human operator response to several controlled element transfer functions and their combined open-loop transfer function are presented in Table 1.3. Modeling the human pilot with the crossover enabled a more complete view of the pilot/vehicle system and allowed for human factors recommendations towards the design of new vehicles. Even today, the crossover model is used as the standard for describing pilot/vehicle systems at the crossover frequency (McRuer and Krendel, 1965, 1974; Xu et al., 2017).

The continued demand for human pilot models for use in informing vehicle design, as well as predicting, preventing, and explaining accidents, has led to a variety of more complex pilot models since the creation of the crossover model. A recent review by Xu et al. in 2017 surveyed the state of the art in human pilot modeling and grouped existing models into three classes of models based on: control theory, human physiology, and intelligence techniques (Xu et al., 2017). Classical models based on control theory include the McRuer crossover model and optimal control models by Kleinman et al. developed in the early 1970s (Kleinman et al., 1970; Baron et al., 1970). Models based on human physiology were developed to understand human pilot perception and control behavior and include the Structural Model (Hess, 1980,

Controlled Element Form	Aerospace Control	Automobile Control
K_c	Attitude control with ACAH system	Speed control
$\frac{K_c}{s}$	Attitude control with a rate command system	Heading control at low to moderate speeds
$\frac{K_c}{s^2}$	Attitude control of a spacecraft with damper off	Longitudinal position control

Table 1.2: Example Applications of Idealized Controlled Element Forms, adapted from McRuer and Krendel (1974).

Controlled Element Transfer Function Y_c	Approximate Human Operator Transfer Function Y_p	Open-Loop Transfer Function Y_cY_p
K_c	$\frac{K_p e^{-\tau_1 s}}{s}$	$\frac{w_c e^{-\tau_e s}}{s}$
$\frac{K_c}{s}$	$K_p e^{-\tau_2 s}$	$\frac{w_c e^{-\tau_e s}}{s}$
$\frac{K_c}{s^2}$	$K_p s e^{-\tau_3 s}$	$\frac{w_c e^{-\tau_e s}}{s}$

Table 1.3: Summary of Human Operator Approximate Characteristics, adapted from McRuer and Krendel (1974).

1990, 1997), Hosman's descriptive model (Hosman, 1996; Hosman and Stassen, 1999), and the biodynamic model (Griffin, 2001). Recent intelligence models take advantage of techniques including fuzzy control and neural networks (Zaychik et al., 2006; Gestwa and Bauschat, 2003). Of these three overarching sets of models, the models based on human physiology are of the greatest interest here due to their potential to include the effects of feedback.

While the crossover model was very successful in predicting pilot behavior, it did not attempt "to describe the underlying structure which contributes to human pilot dynamics (Hess, 1980)." For this reason, the Structural Model is of interest due to the incorporation of multiple sensory channels and models of visual acuity and the time-varying human pilot (Hess, 2009). The Structural Model includes the effects of the neuromuscular system,

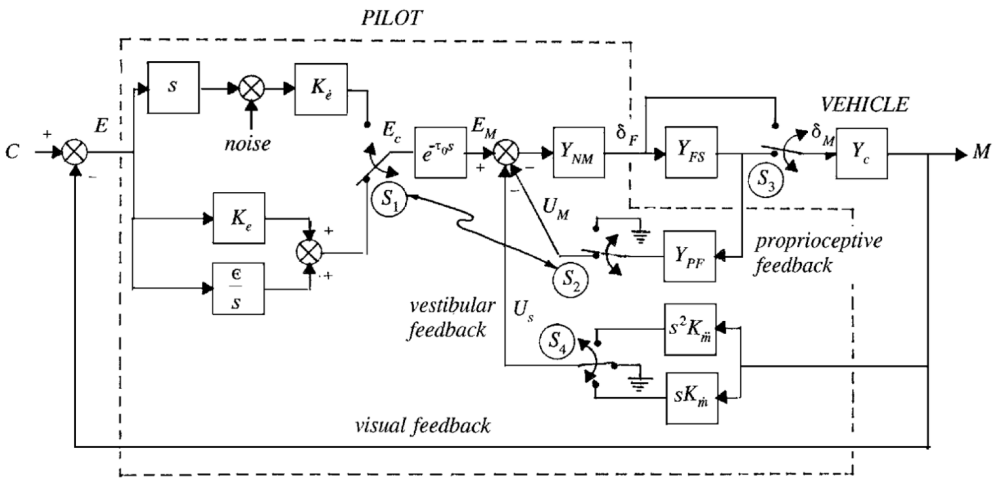


Figure 1.9: The Structural Model of the Human Pilot, from Hess (1997).

the force-feel characteristics of the input device, and the contributions of proprioceptive, vestibular, and visual feedback, see Figure 1.9. One of the key strengths of the Structural Model is the relatively small number of free parameters that need to be set to predict pilot performance. The model has been used in predicting and evaluating handling qualities and pilot-induced oscillation rating levels for helicopters, the Boeing 747, the Lockheed C-5A, and twin ducted-fan aircraft (Hess and Joyce, 2013; Andreea-Irina and Achim, 2014; Grant et al., 2015). Hess has also used the model to investigate how pilot control characteristics change with time due to flight anomalies, changing flight dynamics, and sudden increases in task demand (Hess, 2009, 2016). The results of this model have been compared to the results of a human-in-the-loop simulation for a well-trained subject and showed good comparison (Hess, 2016). Recent work from Bachelder et al. has included modifications to the Structural Model to link pilot performance and workload, and to enable the modeling of pulsive pilot behavior (Bachelder et al., 2017, 2018).

1.2.6 Summary

Augmented feedback has been used in a large variety of motor control tasks and has generally been found to improve performance. Until recently, however, only simple tasks such as physical movements or low-dimensional pursuit tasks have been investigated. More recent works, including the lane-keeping task by de Groot et al. and our previous work with the SAFER task, have indicated that concurrent bandwidth feedback can also be quite

effective for complex tasks (Karasinski et al., 2017). Unlike simple tasks, in which the guidance hypothesis dominates when feedback is removed, there is some evidence that concurrent bandwidth feedback can be removed after training complex tasks without a loss of performance. The decrease in required learning time, improved maximum performance, and decreased workload seen in the SAFER task show that concurrent bandwidth feedback may prove to be most useful very early in training when subjects are first exposed to complex, highly dynamic tasks. As concurrent bandwidth feedback can improve performance for complex tasks without an increase in workload, it may prove a useful technique for training other complex manual control tasks.

There has been considerable improvement in the field of pilot modeling since McRuer's crossover model, especially with models that incorporate human physiology. The Structural Model, in particular, has been very effective in predicting pilot performance, handling qualities, pilot-induced oscillation rating levels, and workload for a variety of system dynamics. None of these pilot models, however, are able to include the effects of concurrent bandwidth feedback. The performance improving potential of this feedback make this a compelling feature to be incorporated into a pilot model.

1.3 Research Questions

We set out to accomplish two aims with the research included in this dissertation. These aims build on each other, starting with a compensatory tracking task, extending to surface electromyography and aircraft flight tasks, and finishing with a theoretical model. This dissertation contains the results of human-in-the-loop subject testing experiments which were designed to understand the effects of concurrent bandwidth feedback. Using the data from these experiments, the Structural Model was modified to integrate the effects of this feedback into a human performance model. To investigate the two aims outlined below, we completed four experiments and the development of a model.

Aim One Investigate the effects of concurrent bandwidth feedback on human performance and workload in complex manual control tasks.

Aim Two Extend the Structural Model of the human pilot to include the effects of concurrent bandwidth feedback.

There are several research questions that we are interested in answering by completing these aims, which include:

1. Can concurrent bandwidth feedback (CBF) improve human performance in complex manual control tasks?
   (a) Can CBF reduce the required training time to peak performance?
   (b) Can CBF be removed after reaching peak performance without reducing subject performance (i.e., does the guidance hypothesis not hold)?
   (c) Can performance be increased without increasing workload?
2. Can we develop a model of human performance that includes the effects of concurrent bandwidth feedback?
   (a) Can we use this model to estimate operational limits?

1.4 Summary

We have introduced our goals and motivation for the design and use of concurrent bandwidth feedback. The background for the experimental work has been described and the open research questions that we explore in this work have been summarized. In the following chapters, we first present a systematic assessment of current and upcoming human automation/robotic integration technologies and research topics (Chapter 2), then report on four experiments involving augmented feedback (Chapters 3-6), propose a theoretical model which explains the observed effects of the feedback (Chapter 7), and summarize our findings and proposed future work (Chapter 8).

Chapter 2

Trade Study

The primary motivation of this dissertation was to help advance the human exploration of space. This chapter investigates the current state of upcoming automation/robotic technologies and opportunities for advancing the state of the art through directed research. The data gathered through the literature review presented here, combined with interviews from ten subject matter experts across industry and academia, was combined in a trade study to produce a prioritized list of research topics that must be addressed to further space exploration. This trade study helped to focus this research on the leading research topic which was identified–improving training for human automation/robotic systems and tasks. Portions of this chapter were originally compiled for the report, "Enabling Technologies for Deep-Space Human Spaceflight: Human-Automation and Robotic Systems Trade Analysis" by John Karasinski, Sherrie Holder, and Stephen Robinson, which was submitted to NASA in August 2019.

2.1 Executive Summary

This investigation focused on a systematic assessment of current and upcoming human automation/robotic (HAR) integration, or HARI, technologies and research topics. Analysis was focused on research and technology that address critical gaps in spaceflight-relevant HARI knowledge, and prioritizing the research required for successful human performance and HAR integration. This is essential for NASA's Human Factors and Behavioral Performance Element to understand the critical human-automation/robotic integration design

challenges for future space exploration. A multi-dimensional trade analysis was performed to objectively score HARI research topics and specific technologies resulting in recommended research priorities for NASA investment. A series of factors informing overall return on investment potential were used in weighted analysis of each technology. Factors included characteristics such as TRL and applicability to relevant spaceflight tasks. While these factors for assessment pertained directly to HARI technologies, research topics were assessed through direct relationships with those technologies (Figure 2.1).

To understand the HARI trade space, we reviewed relevant literature across the past ten years and interviewed ten subject matter experts (SMEs) across industry and academia to investigate the current state of HARI technology, challenges facing development, upcoming automation/robotic technologies across a wide range of fields, and opportunities for advancing the state of the art through directed research. Based on the gathered information, we derived a list of HARI research topics essential to addressing HARI development challenges and advancing the state of the art, as well as a list of specific HARI technologies with application toward HAR tasks common to either long duration deep space exploration (orbital) missions, space surface exploration missions, or both. An initial set of factors for assessment

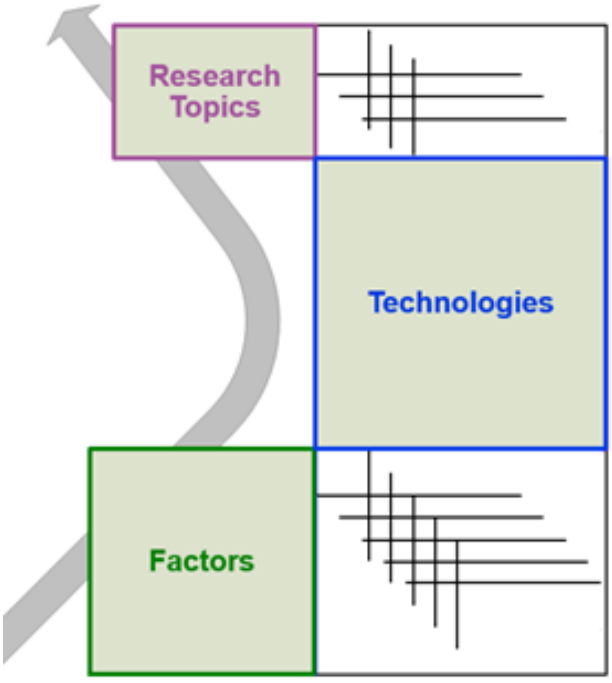


Figure 2.1: Top-level trade study approach used in the HARI analysis.

of technologies was developed. These factors were characteristics of technologies that effect the potential impact of development on NASA missions, or overall return on research investment. Factors included HAR task applicability, task (capability) enabling, potential to reduce or introduce risk, and Technology Readiness Level (TRL), among others.

Factors were provided to a group of NASA HARI stakeholders from NASA Ames Research Center and NASA Johnson Spaceflight Center. They were asked to review eight factors and rank them from most important to least important for consideration of HARI technology investment potential. NASA stakeholders were informed that these factors would be weighted and used to conduct a trade study designed to help NASA prioritize which technologies and, consequently, which HARI research topics, should be pursued in support of future long duration exploration missions. After gathering their input, the stakeholder's scores of the factors were averaged and ranked. The final ranks were used in our trade analysis.

A multi-dimensional trade analysis was performed to objectively assess HARI research topics and specific technologies. The factors for assessment were traded directly with HARI technologies, while research topics were assessed through direct relationships with those technologies. Technologies were first assessed against each factor in a series of individual one-dimensional trade analyses (each trading technologies against one factor). The results of these factor-level trades were normalized and used to score technologies, taking into account the relative factor weights, in the factor-to-technology dimension of the larger analysis (Figure 2.1). The research-topic-to-technology dimension was assessed based on relationships between the two. Research topics and technologies were defined as related if a given technology supports the research topic such that its development would fundamentally drive investigation of that topic. The scores for each related technology for a given research topic were summed to achieve the total score for that topic.

The top-ranking research topics were: (1) Improving training for HAR systems and tasks, (2) Establishing appropriate trust in automation/robotics systems, and (3) Understanding human intent. The top-ranking technologies identified from the trade study were: (1) Machine Learning, (2) Autonomous obstacle detection/imaging, (3) Robotic/human information interfaces, and (4) Artificial Intelligence. These results reflect the surveyed background literature and the information gathered from our SMEs. These top-ranking research topics

were driven by their associated highly scoring technologies, while the top-ranking technologies have seen enormous advancements in research interest and development over the past few years, and all offer a large benefit to the tasks required by NASA on future missions. The top-ranking technologies all benefited from high marks across all factors.

Based on the trade analysis performed, it is recommended that NASA prioritize research investment in the topics of improving training for HAR systems and tasks, establishing appropriate trust in autonomous/robotic systems, and understanding human intent. These top-ranked research topics can be traced to trends of broad task applicability, high potential for risk reduction, low potential for risk reduction, and are areas whose study supports the advancement of research in lower-ranked topics as well. Investigation of these research topics will provide a fundamental foundation for addressing challenges that face implementation of HARI technology solutions in future exploration missions.

2.2 Introduction

Technological advancements in automation and robotics necessitate appropriate integration between these systems and their human operators. To date, there has not been a systematic evaluation of the HARI design challenges for human spaceflight critical to current and upcoming automation/robotic technologies. Industries like transportation, air traffic management, and defense are investing significant time and effort to investigate and solve the many design challenges involved in human-automation/robotic integration. NASA's Human Factors and Behavioral Performance Element needs to understand the critical human-automation/robotic integration design challenges for future space exploration. A survey of the upcoming research topics and technologies which can be applied to NASA from a range of industries and domains is needed in order to reduce the risks associated with human spaceflight.

Therefore, an assessment of the upcoming technologies and open research challenges critical to effective human and automation/robotic integration (HARI) systems across industries and domains is essential to inform the design and development of safe, efficient future systems. The objective of the current project is to conduct a systematic assessment of the space-relevant HARI automation/robotic technologies in order to prioritize necessary

research required for successful human performance and HAR integration. This includes identification of aspects that influence the relative importance of technology for spaceflight, or factors, for assessing prioritization of HARI related research and technologies.

2.3 Project Background

The overall objective of this study was to investigate HARI technologies on the horizon with the potential to support critical HAR tasks and use trade analysis to assess these technologies against critical factors for investment in order to determine recommendations for research and development. The project was designed with two Phases, with Phase 1 focused on gathering background information and identification of specific technologies, and Phase 2 focused on trade analysis. Tasks were originally proposed for each Phase as shown in Tables 2.1 and 2.2. This project largely followed the original two-phase plan, with background research informing the design of a trade study aimed to provide recommendations of technologies/research to pursue. However, specific tasks were redirected, in coordination with the Human Automation / Robotics Integration (HARI) Discipline Scientist (DS) (NASA Civil Servant at Ames), as we gathered the background information in Phase 1 and learned more about the trade space.

In exploring HARI technologies and risks and challenges facing development, as described in the Phase 1 tasks, through literature review and interviews with Subject Matter Experts (SMEs), it became evident that technology implementations as described in Task 1.3 would vary widely due to dependence on specific mission design, even when constrained to a specific HAR task. It would not be possible or practical to capture the space of all possible specific technology implementations at such a detailed level. The primary goal of this project was to explore a trade space of HARI solutions or directions for research, not to trade on mission designs. Rather than explore a subset of implementations whose applicability to a HAR task would be limited mission to mission, we chose to raise the level of the technology/research trade space and explore broader solutions to HARI challenges as they apply to HAR tasks common to the scope of long duration orbital and planetary surface exploration missions. For example, exploring the potential of Augmented Reality/Virtual Reality (AR/VR) technology in general as applied to HAR tasks, as opposed to a specific implementation of AR/VR to

train for surface operations that assumes a human-robot team makeup (a mission design decision).

In reviewing a draft of the report described in Task 1.5, it became apparent that the HARI solutions identified fell into two categories. While some were technologies which support or enable HAR tasks, others were research topics related to those technologies whose study will fundamentally drive future HARI capabilities and directly address HARI challenges. Given the importance of the research topics identified for addressing HARI risk, the trade analysis plan was directed to capture both technologies and research topics (and the relationship between them). Additionally, with each technology and research topic applicable to a range of HAR tasks as described above, rather than having a separate analysis for each task, HAR tasks became a critical factor for comparison across the trade space, providing a more complete evaluation between technologies. Although the process outlined in the Phase 2 tasks was followed, the focus of the trade analysis was shifted to reflect the nature of the trade space. This shift allowed the study to produce relevant recommendations on closing HARI risk as intended.

2.4 Background Research

To begin the assessment of space-relevant HARI critical factors, we first completed a comprehensive literature review of the field of human and automation/robotics interaction. Background literature primarily focused on survey papers from the past ten years, but also included prominent papers from noted authors in the field. Primary research was also gathered from discussions with subject matter experts in human factors and human-robot interaction related fields. Findings and lessons learned from this investigation are provided in this report.

2.4.1 Literature Review

We completed a review of human factors and automation/robotics integration survey papers published over the past decade, with an increased focus on the past five years. Non-survey papers from highly cited and established experts were also added to this review to provide additional insights. When reading these papers, care was taken to note recurrent topics and technologies that received specific focus, were forecast to generate additional

Phase 1	
Task	Task Description
1.1	Work with the Human Automation / Robotics Integration (HARI) Discipline Scientist (DS) (NASA Civil Servant at Ames) to understand the HAR tasks and the key mission architecture design constraints that will heavily influence future HAR system design.
1.2	Identify and consider current and near-term future (expected to be operational within the next 10 years) HAR classes of technologies and capabilities that are relevant to the required HAR tasks.
1.3	Identify possible technological implementations, i.e., automation and robotic systems, to accomplish these HAR tasks, taking into account the mission architecture design constraints in the Concept of Operations.
1.4	Determine the associated HARI design and research challenges associated with each HAR technological solution.
1.5	Coordinate with the HARI DS and develop a report describing the potential technologies and associated risks and challenges of each.

Table 2.1: Tasks initially proposed for Phase 1 of the HARI Trade Analysis.

interest in the near future or were otherwise noted as requiring greater study.

As a result of this literature review, major themes in human and automation/robotic integration technology development and research were identified, see Table 2.3.

Each of these topics is briefly discussed below, referencing their fundamental papers when possible, as well as their forecasts from the previously reviewed articles.

Machine Learning

Machine learning (ML) is among the most commonly mentioned topics which authors forecast as being essential to the future of human-robotic interaction (Wang et al., 2018). Machine learning has enabled significant benefits in a variety of automation/robotics systems but has also given rise to the need for explainable systems and has raised additional questions

Phase 2	
Task	Task Description
2.1	Develop draft criteria and associated weighting for the trade space evaluation of the suitability of each class of technology for each relevant mission task. Among potential decision criteria are: technology readiness, safety-criticality, crew-time savings, unique capability, and minimum frequency of interaction between the human and automation/ robotic system.
2.2	Work with the HARI DS and other NASA stakeholders to arrive at a consensus for criteria and relative weighting, by participating in a series of virtual meetings, an on-site workshop, or technical interchange meetings as organized and implemented by NASA and KBRwyle.
2.3	Develop an analysis method for applying the DS-agreed upon criteria to evaluate each technology for each DRM task, be it computationally modeled, empirically data driven or based on subject-matter expertise.
2.4	Complete the trade analysis using the selected analysis method and criteria, and develop recommendations for the most likely automation/ robotic implementations; identify the corresponding human integration design challenges associated with developing each HAR system.
2.5	Develop and deliver a final report on the findings, which will include recommendations for each of the HAR tasks.

Table 2.2: Tasks initially proposed for Phase 2 of the HARI Trade Analysis.

	Machine Learning	Flexible/ Adaptive/ Adaptable Automation	Networked Multi-robot Systems, Swarms	Trust
Admoni and Scassellati (2017)				x
Ahmad et al. (2017)	x	x		
Chen and Barnes (2014)	x	x	x	
Endsley (2017b)	x	x		x
Guiochet et al. (2017)				x
Kehoe et al. (2015)	x		x	
Kolling et al. (2016)		x	x	
Liu and Wang (2018)	x			
Losey et al. (2018)	x	x		x
Lu et al. (2016)		x		
Ososky et al. (2013)			x	x
Parasuraman and Wickens (2008)		x		
Phillips et al. (2016)				x
Rautaray and Agrawal (2015)	x			
Schaefer et al. (2016)				x
Sheridan (2016)	x			x
Vagia et al. (2016)		x		
Wang et al. (2018)	x		x	
Zamora et al. (2017)	x			

Table 2.3: Table of the key papers reviewed, and the topics discussed in each.

about trust. While machine learning techniques may be effective, they are rarely easily explainable, and operators often have difficulty understanding exactly why a system behaves as it does. Additionally, as these systems have become more sophisticated, they have can now continuously learn and update their behavior, making it challenging for operators to maintain both system understanding and appropriate levels of trust (Chen and Barnes, 2014).

Machine learning techniques such as hidden Markov models, Gaussian mixture models, and radial basis function neural networks, though usually requiring a supervised training phase, have been shown to be very effective in predicting human intent in the context of physical human-robotic interaction (Losey et al., 2018). In reviewing which machine learning algorithms are currently being used, Zamora et al. found that neural networks accounted for an overwhelming majority, but that both supervised and unsupervised algorithms were about equally common (Zamora et al., 2017). ML is essential to the fields of vision-based hand gesture recognition and non-visual gesture recognition, without which gesture recognition devices would be impossible (Rautaray and Agrawal, 2015; He, 2018; Liu and Wang, 2018). As computer technology continues to rapidly advance, the ability to detect, track, and classify gestures in real-time has enabled this technology to be implemented in manufacturing and other industrial plants. Liu et al. specifically note a need to combine different ML algorithms to improve efficiency, and state that deep learning techniques are now enabling non-wearable sensors (Liu and Wang, 2018). ML has also been used to vary the personality and behavior of adaptive social robots (Ahmad et al., 2017).

In her 2017 paper, Endsley noted the research needs for the next thirty years of designing and building fully autonomous systems (Endsley, 2017b). Several of these needs specifically concern machine learning techniques, including validating autonomy software, learning system consistency and transparency. There are currently no effective techniques for validating autonomy software, as "traditional methods fail to address the complexities of learning systems. Exhaustive testing of rules and potential system states will not be possible and understanding boundary conditions will be difficult" (Endsley, 2017b). Validating machine learning solutions is currently an active area of research. There is concern about consistency in learning systems, as different systems will learn using different techniques and provide different levels of feedback about how their automation has changed based off

new data. Endsley notes the lack of transparency in learning systems as a unique challenge, saying "[t]he actual logic and lessons 'learned' by neural networks and deep learning software are typically opaque not only to the human operator but also to software developers who may not fully understand how the system will behave in all circumstances" (Endsley, 2017b). These problems are exemplified by Sheridan, who notes that "[i]t is becoming clear that many complex traffic situations are exceedingly difficult for computer vision and artificial intelligence to 'understand' and that many accidents are avoided by social interaction between drivers, such as mutual eye contact, hand signals, and so on. Understanding the social aspects of driving in traffic, as well as the degree to which cars can be safely automated, demands much further research."

Flexible, Adaptive, or Adaptable Automation

Flexible, adaptive, and adaptable automation are widely praised in the literature for their ability to provide dynamic levels of automation. The flexibility to provide different sets of automated features during different mission phases, for instance, is an effective requirement for many modern tasks. One such example is the autopilot software used in modern transport aircraft, which includes multiple modes of automation for takeoff, cruise, and landing. Chen and Barnes define flexible automation as "systems that invoke various levels of automation depending on the operator's state, critical events in the environment, or algorithms related to specialized problem sets." Chen and Barnes and others have subdivided flexible automation into subtypes, based on the involvement of humans in the decision making process: adaptive automation—where tasks are assigned using conditions established before a mission, adjustable automation—where the human decides when to invoke automation, and mixed-initiative systems—where both the human and the system jointly decide how to allocate tasks (Chen and Barnes, 2014; Beer et al., 2014). These dynamic changes in the role of the human in the human-automation interaction are meant to "either increase the robot's level of autonomy at the expense of the human's authority, or, conversely, increase the human's control over the shared cooperative activity at the expense of the robot's autonomy" (Losey et al., 2018). These systems help maintain overall performance while attempting to reduce workload and maintain situational awareness for their human operators (Kaber et al., 2006).

Among these three automation technology areas, adaptive automation has seen the most

research, and many authors have used it in empirical studies (Vagia et al., 2016). The primary difficulty with adaptive automation lies in "thorny human factors issue of [function] allocation...which has been met with marginal success" (Vagia et al., 2016). Optimal assignment of tasks between the operator and the system is difficult as it requires excellent understanding of the performance of the operator and the system's response to the operator. It also requires the operator to be fully aware of the functional allocation at all times, otherwise mode confusion may occur.

Flexible automation can react to dynamic changes in the environment, and researchers have been able to include real-time sensor data of human physiological states to bring the operator's workload and situational awareness into the loop. Monitoring the human allows the system to automatically take over tasks when workload is high, and has been used to send control back to the human when the system notes that they have become complacent or as an attempt to increase situational awareness (Lu et al., 2016). This type of automation is already present in self-driving vehicles on the road today—self-driving vehicles require that drivers have their hands on the wheel even when in self-driving/lane-keeping modes. While this flexible automation is often effective in common and well understood systems such as driving, there is some concern that flexible automation may prove detrimental in complex and potentially unpredictable systems such as robotic swarms (Kolling et al., 2016).

While adaptive and adaptable automation has been the subject of many experiments over the past few decades, the question of who should be in charge of setting the level of automation remains an open question in need of further study, though mixed-initiative systems may provide the best of both worlds (Parasuraman and Wickens, 2008; Chen and Barnes, 2014). Chen and Barnes conclude their review by noting that "[m]ixed-initiative architectures take advantage of the synergy between the more sophisticated worldview of an experienced human as well as the agent's logical precision and more rapid latencies." This architecture is inherently complex and difficult to study as individual differences such as age, expertise, and trust have large effects when interacting with these systems (Schaefer et al., 2016). Further research is recommended into different types of flexible automation, especially when dealing with very complex systems.

Networked Multi-robot Systems and Swarms

Human automation/robotics interaction has traditionally focused on a single robotic system, but the miniaturization of computer technology has made swarm or multi-robot systems an increasingly viable option. The ability for swarms to dynamically reconfigure themselves in response to changing environmental variables and task demands, however, can lead to complex requirements on the human operator. There remain important questions to be answered in the realm of human systems integration with swarms, especially regarding human supervisory control (Kolling et al., 2016). There is a specific concern with monitoring human workload and situational awareness as the number of robots increases. Depending on the number and ability of robots and the type of tasks being performed, it is possible to quickly overburden the swarm operator, especially when the operator is required to negotiate swarm-swarm interactions. Kolling et al.'s 2016 review breaks the cognitive complexity of the human-robot system into three complexities: robots performing independent activities, with complexity $O(n)$, which allows more robots to be controlled simply by adding more operators in a linear manner; robots interacting with other robots fully autonomously, with complexity $O(1)$, which allows for a fixed number of robots to control any number of robots; and the case where robot-robot interaction must be controlled by an operator, with complexity $O(> n)$, as the dependencies between robots results in more demand faster than the number of robots grows. See Figure 2.2 for a graphical illustration of control complexity under each of these conditions.

Ongoing research into human-swarm interaction and multi-robot systems has primarily focused on coordinated swarm control, changing swarm topology, and describing the state of the swarm in a more understandable way (Wang et al., 2018). The development and design of human-swarm interfaces for multi-robot collaboration and, particularly, unmanned aerial vehicle teams is another important set of ongoing research. The ability for swarms to multitask and the requirement for the human operator to quickly task switch have been shown to cause detrimental effects on overall system performance (Chen and Barnes, 2014). High workload phases have been shown to be most sensitive to interruptions from tasks switching, suggesting that task switching should be avoided during these phases unless absolutely necessary (Norman and Draper, 1986). Issues relating to multitasking, task switching, and the

loss of situational awareness can be mitigated with properly designed human-swarm interfaces. Chen and Barnes outlined several of the prominent issues in user interface design and offered solutions in the form of guidelines. They identified six issues ranging from "maintaining operator's ultimate decision authority" to "visualization and training techniques enhance human-agent collaboration", and presented guidelines based on the findings of their review.

The concept of robots and automation systems that rely on externally networked support has also been explored by researchers (Kehoe et al., 2015). New topics of research using "the cloud" or otherwise networked robotics include big data, cloud computing, collective robot learning, and human computation, see Figure 2.3. Other key technologies which can be enhanced with networked robotic systems include human-robot collaboration technology, autonomous navigation technology under non-structured environments, multi-agent robot systems (swarms), and emotion recognition (Wang et al., 2018). Issues associated with the rise in cloud technology include the need for techniques to consider time varying latency and quality of service, system security from remote intrusion, privacy concerns, and big data cleaning and filtering techniques. Currently, cloud computing can be described as a framework with consists of three levels: Infrastructure as a Service (IaaS), where bare oper-

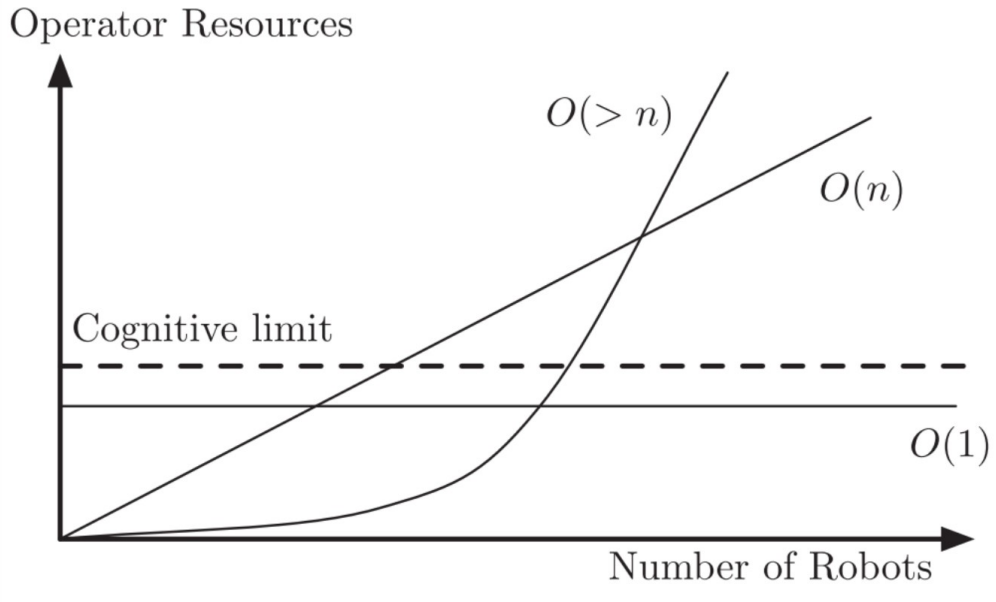


Figure 2.2: Graphical illustration of the concept of control complexity in a human–multirobot system (Kolling et al., 2016).

ating systems are available; Platform as a Service (PaaS), where more structure is provided, including access to application frameworks, databases, and programming languages; and Software as a Service (SaaS), where software is made available online rather than as a local service (Kehoe et al., 2015).

Trust

Human trust has numerous definitions but for our purposes can be defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004). Trust has become an increasing topic of research as robotics increasingly moves out of traditional settings such as manufacturing and into more common-place locations such as the office and the home. Trust has a large impact on the physical safety of people operating around robots, as improper trust can lead a person to inadvertently place themselves in harm's way. Schaefer et al. (2016) define trust as a three-dimensional expression of a relational property:

1. An individual's overall, long-term propensity to trust in general
2. A transient, momentary trust response to immediate ambient conditions
3. How 1) and 2) evolve over time

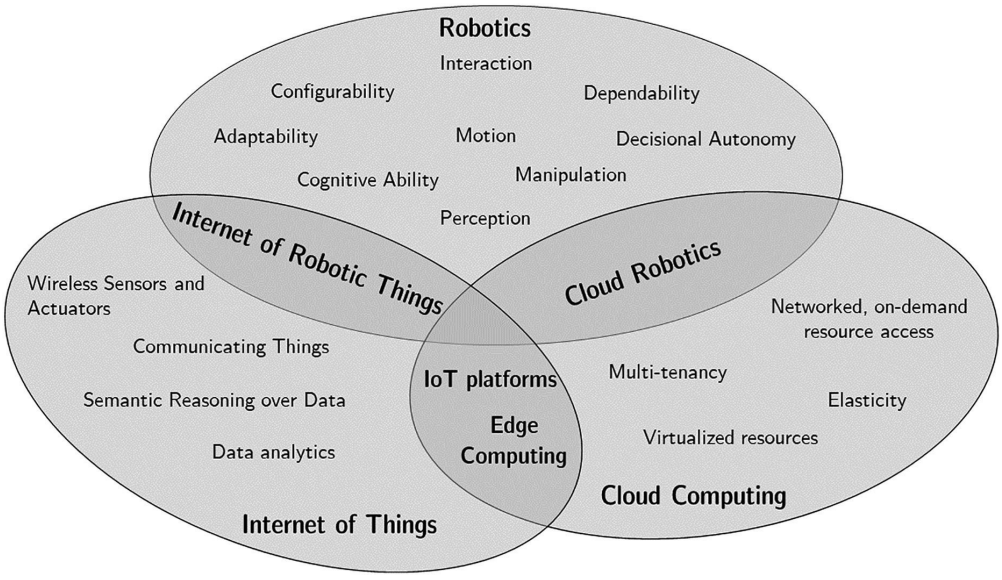


Figure 2.3: Combining Robotics, the Internet of Things, and Cloud computing has resulted in many new possibilities such as Cloud Robotics (Simoens et al., 2018).

Their meta-analysis found strong effects between human-robot interaction and analyzed the factors that determine trust. A robot or robotic system's ability to garner trust relies on several factors. See Figure 2.4 for Schaefer et al.'s conceptual organization of influencing the development of trust. Trust is commonly assessed using surveys which attempt to measure the individual factors that establish trust. These scales attempt to measure individual elements of trust, asking about the operator's assessment of the automation's competence, predictability, and dependability, among other factors. Of these measurement techniques, two of the most commonly used scales are the "Checklist for Trust between People and Automation" (Jian et al., 2000) and versions of Muir and Moray's subjective rating scales, though research into real-time techniques is ongoing (Seppelt and Lee, 2019). Appropriate trust is important when shared control between human-robot teams is essential, as the human is more likely to arbitrate additional tasks to the robot when this trust is established (Losey et al., 2018).

Ososky et al. made several propositions regarding human trust of robotics, including:

- Humans are easily influenced by superficial characteristics of robots
- Human subjective assessment of trust in robots ultimately determines the use of robotic systems

They noted that robot characteristics had the strongest influence on trust in human-robot

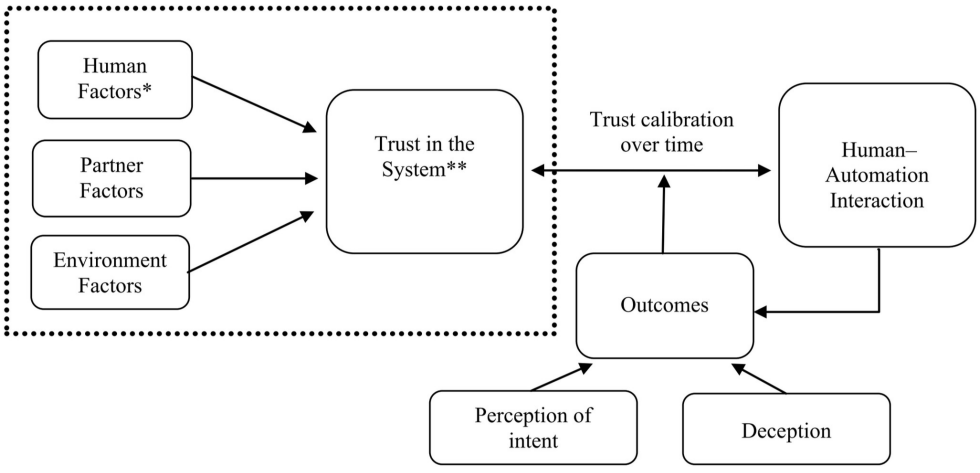


Figure 2.4: A conceptual organization of trust influences highlighting trust development (Schaefer et al., 2016).

teams, which included factors such as reliability, transparency, and anthropomorphic qualities. One example of this is de Visser et al.'s 2016 study, which found that anthropomorphic automation associated greater trust resilience. de Visser et al. concluded their study by suggesting that designers incorporate these features into future robots as a deliberate design choice to garner greater trust. Even very simple additions such as high levels of mutual gaze have been shown to increase trust, while gaze aversions stoke feelings of distrust between human-robotic teams (Admoni and Scassellati, 2017).

While anthropomorphism can lead to greater trust, some authors note caution when incorporating these features. Culley and Madhavan warn that individual differences can lead some individuals to place too great a trust in anthropomorphic robots. Individuals are even more likely to trust anthropomorphic robots if they perceive similarities to the robot and themselves regarding age, gender, and even similarity of movement (Verberne et al., 2013; Pak et al., 2014). Ososky further warns that trust in a robot's reliability alone is insufficient for better teamwork, and that over-trust combined with an inaccurate or incomplete mental model can lead to worse overall performance. This is especially important when considering changing levels of automation in the field of human-automation interaction, for example, an operator may not fully understand what they are currently responsible for if the robotic system is in control. This has led to numerous incidents leading to serious injury and death. Trust in human-robot teaming is slightly different than trust in automation, however, as robots are often seen as collaborating teammates rather than just an automated tool (Ososky et al., 2013).

2.4.2 Interviews with Subject Matter Experts

In order to provide a current assessment of the critical challenges associated with effective HARI systems across industries and domains, we also interviewed subject matter experts (SMEs). SMEs were chosen to represent a cross section of HARI related disciplines such as aerospace, industrial robotics, military applications, medical, and autonomous vehicles. SMEs were intentionally selected from different backgrounds, including military research, academia, and industry (robotics, medical, aerospace), in order to provide broad perspectives on the risks and challenges facing HARI technology, development, and research. We conducted ten phone interviews with SMEs who integrate humans, automation, and robotics

in their work. We asked each expert the following questions:

1. What technologies do you think are on the horizon in your field in the integration of humans, automation, and robotics?

2. How would you prioritize what technologies are in development involving the integration of humans, automation, and robotics?

   (a) What technologies would be the most responsive to increased research support?

   (b) For these technologies, what are the current TRLs? How much effort do you think it will take to raise the TRL over time?

3. What are you most concerned about for the integration of humans, automation, and robotics? What technologies do you think could mitigate these risks?

   (a) What risks do you see arising from inclusion of these technologies (what new risks do you anticipate)?

   (b) What risks currently have no technology solutions?

4. How will these technologies fill the gaps in our current abilities?

   (a) Where do you see additional automation as a plausible way to fill those gaps?

   (b) What other technology gaps should we be concerned about?

5. Based on this discussion, is there anything else we should know?

Based on the information gathered from literature and discussion with subject matter experts, the following specific HARI-related technologies and research topics emerged as areas for future research and development for the advancement of human automation and robotic interaction relevant to human spaceflight.

2.4.3 Specific Technologies

Non-invasive behavioral and physiological sensing

Non-invasive behavioral and physiological sensing includes a range of techniques. Some physiological sensing techniques include common place, if controversial, methods such as a

polygraph, electromyography (EMG), electroencephalogram (EEG), and electrocardiogram (EKG) sensing. Behavioral analysis techniques may rely extensively on video analysis, such as gait analysis, and more integrated technology covering additional modalities. This technology can be used to infer team member states. The use of artificial intelligence to combine these perceptions is relatively developed.

Implantable Biometrics

Compared to many of the other specific technologies we identified, implantable biometrics is a relatively young field which focuses on implantable biosensors for precision and personalized medicine. These sensors can provide continuous data on specific, targeted metrics which can allow for the immediate detection of problems or need for intervention. Implantable biometrics is especially important in the "diagnosis, monitoring, management and treatment of a variety of disease conditions" and can be used to detect changes in a person's health (Fitts, 1951). Further advances in miniaturization and nanotechnology are likely needed for this technology to become viable.

Autonomous obstacle detection/imaging

Autonomous obstacle detection/imaging is a combination of technologies designed to identify obstacles around a robot or other autonomous agent. Detection and imaging can make use of visual spectrum or other light sources, acoustic or magnetic sensors, or laser-based technologies such as LIDAR. Multiple techniques also take advantage of combining these technologies into multispectral sensors. These technologies are important for autonomous docking and landing of spacecraft but have also seen an enormous increase in interest from the self-driving car industry. One important side effect of increased demand of this technology in self-driving cars in the past few years is that the hardware has both rapidly miniaturized and dropped in price. Note that this technology is only concerned with detection, while resulting actions and path planning is captured elsewhere (autonomous path planning).

Autonomous path planning

In contrast to autonomous obstacle detection, autonomous path planning describes the resultant planning and action that is taken after an obstacle is sensed or an objective is determined. This technology benefits greatly from a good understanding of the robot or

autonomous agent's dynamics, the environment it acts in, other agents in the environment, and the objective's location. With regards to spaceflight, autonomous path planning is relevant when considering orbital proximity operations (including rendezvous and docking), surface landings, and rover movements. This technology has also seen great benefits from the self-driving car industry, especially regarding planning around other moving agents whose intent is often poorly understood.

Speech recognition

Speech recognition is a set of technologies that enable the translation of spoken words to text by computer software. Speech recognition has been actively developed since the 1970s and has a generally high rate of success. Despite this relatively long period of development, recent advancements in speech recognition have been made by integrating machine learning techniques. Transforming spoken word to text allows autonomous systems and robots to accept commands or infer human intent and is a common alternative to physical computer interfaces (Tsarouchi et al., 2016). It also allows for the detection of speech patterns and inflection classification to capture intent, trust, fatigue, or emotional states.

Intuitive control interfaces

Intuitive control interfaces consider ways of intuitively mapping human gestures to a resultant robotic action, and often takes human physiology, kinematics, and other elements of physical movement into consideration. This technology includes interface types such as joysticks, keyboards, touchscreens, and gesture recognition, among others.

Robotic/human information interfaces

Information displays must determine what information to transmit for any given task, which may be customized based on user preference, task or environment concerns, past experiences, or the presence of anomalies. These may include multimodal (visual, audio, and/or haptic) displays which display task relevant information to an operator. They may display 2D or higher-dimensional information and may be body-worn or mounted in the environment. These displays have elements designed by both human-computer interaction experts and machine learning algorithms. Ideally, such displays would be ubiquitous, capable of quickly and easily transferring information between stations, and able to appear on traditional monitors, tablets, smartphones, or augmented reality interfaces. While some elements

are well-defined and arguably in use today, others remain in early stages of development.

Augmented Reality and Virtual Reality

Augmented and virtual reality are a pair of technologies which provide a partial or fully virtual environment to a user, often in the form of a head mounted display. Augmented reality has also been developed to work with modern phones and tablets, and can provide additional, digital context to an otherwise physical object or environment. Virtual reality is increasingly used as a training tool, while augmented reality has begun to be used as a tool for both training and operations.

Robotic agents

This technology encompasses a large variety of robots, which include rovers, satellite or UAV swarms, robotic arms, and vehicles, among others. The relative TRL varies between relatively low, in the case of robotic swarms, to very high, in the case of rovers and robotic arms. These sets of technologies enable humans to complete tasks that they could not otherwise accomplish, either because they take place in an extreme, dangerous or difficult to reach environment (as is the case with Martian rovers), they require abilities humans do not (moving payloads required by robotic arms such as Canadarm2) or because they would take too long (such as the mapping or scouting of a region by a swarm of UAVs or satellites).

Assistive Robotics

In contrast to robotic agents, which largely replace the human or do not require a human to be present, assistive robotics describe robots that directly interface with humans to assist them in accomplishing a task. These robots include small assistive satellites such as Astrobee, a modern version of the Apollo Lunar Roving Vehicle with more advanced guidance capabilities, exoskeletons, or personal assistants. These robots enhance the already existing abilities of humans by enabling them to complete tasks that they otherwise could not, or by increasing performance in challenging tasks.

Artificial Intelligence

Artificial intelligence is intelligence demonstrated by machines, in contrast to human intelligence. Some of the major goals of AI include knowledge reasoning, planning, natural language processing, computer vision, robotics, and machine learning. In space HARI, AI

could primarily be leveraged in managing complex systems (i.e. diagnostics, prognostics, and maintenance of spacecraft) and in acting as assistants for crew completing science and activity tasks. By correctly interpreting human intent, AI can also control robots for payload and physical crew assistance.

Machine Learning

Machine learning describes a collection of algorithms which perform a specific task without using explicit instructions, instead relying on learned models. Common types of machine learning include supervised learning, in which a human trains the model, unsupervised learning, where the system learns on its own, and reinforcement learning, where the software takes actions in an environment to optimize a cost function. Machine learning has improved the performance of many varied technologies and is the foundation upon which artificial intelligence is being developed.

Flexible, Adaptive, or Adaptable Automation

As noted earlier in the report, flexible, adaptive, and adaptable automation are widely praised in the literature for their ability to provide dynamic levels of automation. The flexibility to provide different sets of automated features during different mission phases, for instance, is an effective requirement for many modern tasks. Chen and Barnes define flexible automation as "systems that invoke various levels of automation depending on the operator's state, critical events in the environment, or algorithms related to specialized problem sets."

2.4.4 Research Topics

Understanding human intent

The topic of understanding human intent is wide, and includes subtopics such as the robotic interpretation of human intent, understanding human intent unobtrusively, and improving human to robot communication. The interpretation of human intent by a computer or robot can be done in a variety of ways, including speech, gestures, and other forms of nonverbal communication. These techniques are at varied levels of development, from basic proof of concept to use in operations. Each technique can be broken down into several levels—gestures, for example, have four levels: sensor technologies, identification, tracking and classification (Liu and Wang, 2018). This area is also closely tied to interpreting behavioral

and human monitoring data and encompasses human/behavioral model research such as the prediction of intent from eye movements (Ruhland et al., 2015; Singh et al., 2018).

Autonomous/robotic system communication to humans

In contrast to the previous topic (understanding human intent) the research topic of autonomous/robotic system communication to humans addresses how these complex systems can best relay information back to a human operator. This topic includes both research of communication techniques and mitigation of miscommunications from the system to the human. This topic deals with discovering effective methods of providing information to a human user in an intuitive way, such that communication feels natural to a human operator. Human-robot communication is largely focused on developing multisensory methods to successfully communicate a robot's intent to humans. Human-autonomous system communications additionally deals with methods to successfully enable explainable and transparent autonomous system operation.

Ensuring human safety (physical)

This topic captures research which investigates how to enable safe human and robot operation in a shared environment in order to reduce risk. It specifically investigates methods to successfully prevent harm to humans in close physical proximity with robots and develops guidelines and recommendations as to how physical interaction between robots and astronauts can safely occur. This research topic benefits from the lessons learned in manufacturing settings, where humans and robots must often work nearby or directly with each other, as well as work done by autonomous car companies in avoiding pedestrians.

Continuous human performance monitoring

The topic of continuous human performance monitoring seeks to understand human-system performance and measure human performance unobtrusively. This research topic investigates which human-system performance measures and limits are required for spaceflight and seeks to validate novel methods and technologies for measuring a variety of aspects of human performance such as task performance, workload, and situational awareness. Research in this area also focuses on understanding the human performance effects resulting from adaptive automation and attempts to identify what are the performance differences between adaptable (human sets level of automation) versus adaptive (automation sets level

of automation).

HAR team performance optimization and function allocation

Human autonomous/robotics team performance optimization and function allocation investigates different ways of understanding human-robot teamwork and human-autonomous system robustness, decides whether a particular function will be accomplished by a person, technology (hardware or software) or some mix of person and technology (Fitts, 1951), and how to optimize that balance (Yanco et al., 2015). This research focuses on what social and teamwork elements enable successful human-robot collaboration, especially when it requires direct interaction between robots and astronauts. This area also captures research on how to measure robustness when humans are using a system in off-nominal conditions and identifies when these systems are off nominal.

Enabling command/control of complex robotic systems

This topic focuses on enabling command/control of complex robotic systems and enabling critical decision making. This includes research on methods to successfully allow humans to command and control multiple, mixed robotic agents with varying levels of autonomy and flexible function allocation. It also looks at new methods to enable humans to make time-critical decisions using autonomous systems across a variety of system dynamics and is required to evaluate methods for different autonomous systems with different functions (e.g., ECLSS vs. Power vs. Navigation).

Improving situation awareness in HAR systems

"Situation awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 2017a). This research topic focuses on techniques to both maintain and improve operator situation awareness when interacting with automation/robotics systems. Recent work by Endsley has further expanded early models of situation awareness, discussing the emerging problem of loss of operator situational awareness and out-of-the-loop performance problems associated with increasing system autonomy, reliability, and robustness. This new model for human-autonomy system oversight (HASO), incorporates situation awareness, trust, workload and automation interfaces among the key system design features influencing human cognitive processes involved in successful interaction with automated sys-

tems.

Improving training for HAR systems and tasks

The topic of improving training for HAR systems and tasks investigates what new methods are required or most effective to train humans to use complex, advanced autonomous and robotic systems. Research in this area explores different techniques and technologies to improve human performance and reduce workload, and often makes extensive use of mock-ups, simulations, hands-on walkthroughs, and human-in-the-loop studies. Many techniques have been explored to improve training, including various kinds of feedback, manual control adaptation, and the use of virtual and augmented reality.

Establishing appropriate trust in automation/robotics systems

This research topic focuses on techniques to establish appropriate trust in automation/robotics systems and to mitigate changes in trust between humans and these systems. It explores how trust changes with factors such as communication, reliability, workload, social acceptability, privacy, and transparency. As noted earlier, it can be challenging for operators to establish appropriate trust as these automation/robotics systems become more sophisticated (Chen and Barnes, 2014). This research has also focused on shared control between human-robot teams and how tasks are arbitrated to the robot when trust is established (Losey et al., 2018).

2.5 Trade Analysis

In addition to specific technologies and research topics, the information gathered from the literature review and the discussions with subject matter experts was used to identify factors relevant to the assessment of technology or research for future investment. With all of this information gathered, the factors were refined and used in a multi-dimensional trade study to assess the technologies and research topics as priorities for HARI investment.

2.5.1 Factor Assessment with NASA Stakeholders

In addition to interviewing the human, automation, and robotics integration SMEs, we also surveyed six NASA HARI stakeholders for their input on the trade study. As NASA stakeholders involved in human, automation, and robotic interaction, we asked them to

Factor	Weight
Task applicability	6
Task enabling	6
Potential for reducing risk	5
Potential for introducing risk	(-)4
External Investment (outside of NASA)	3
Technology Readiness Level (TRL)	2
Research Interest (within NASA)	1

Table 2.4: The ranking of seven factors resulting from feedback from our NASA stakeholders.

review eight factors and rank them from most important to least important in consideration of HARI technology for investment. We also had them rank additional "secondary criteria" for the factors related to risk.

Factors are characteristics of a technology that our team, in collaboration with the NASA HARI DS, has identified and selected because they are relevant to assessing HARI. These factors were generated from our review of the background literature and conversations with the SMEs. NASA stakeholders were informed that these factors would be weighted and used to conduct a trade study designed to help NASA in prioritizing which technologies and, consequently, which HARI research areas, should be further invested in to help with future long duration exploration missions. After our NASA stakeholders provided their input, we averaged and ranked their assessment of the factors. The ranked factors appear in Table 2.4.

Task applicability

Which tasks does the technology have an impact on? This factor characterizes how much impact the technology may have on the various HARI tasks identified for future exploration missions (Marquez et al., 2017). We determined the application of the various technologies to each task in order to measure their respective applicability to space HARI. These tasks, common to long duration orbital missions, deep space surface exploration missions, or both, are shown in Table 2.5.

Applicable to Orbit Operations	Maneuver/reboost/rendezvous
	Docking/undocking
	Spacecraft support, system maintenance
	Complex assembly, capture and berth
	Science and assigned activity support, payload assistance
	Science and assigned activity support, crew assistance–physical
	Science and assigned activity support, crew assistance–cognitive
Applicable to Surface Operations	Spacecraft support, system maintenance
	Spacecraft support, system preparation
	Site preparation assembly, excavation
	Complex assembly, heavy lift
	Drive/navigate
	Exploration, scouting
	Exploration, mapping
	Exploration, sampling/analyzing
	Science and assigned activity support, science/sample collection
	Science and assigned activity support, payload assistance
	Science and assigned activity support, crew assistance–physical
	Science and assigned activity support, crew assistance–cognitive

Table 2.5: HAR tasks for spaceflight.

Task enabling

Does the technology enable a new capability? This factor describes how much the technology enables one or more of the various HARI tasks identified for future exploration missions. HARI tasks are assumed to be critical and must be completed. We subjectively rated this by classifying the technology as: No effect relative to current technology (score of 0), Improves performance of current capability (score of 1), or Adds new capability (score of 2).

Potential for reducing risk

What is the benefit from risk reduction? This factor describes how risk might be reduced by the inclusion of the technology. Each type of risk was subjectively rated. Types of risks

(secondary criteria) are listed below:

- Improved safety: increase astronauts' safety.
- Reduced likelihood of system failure: increase overall robustness of system by predicting or preventing failures.
- Improved performance: astronauts can work more effectively and efficiently, including reducing physical and cognitive workload.

Potential for introducing risk

What is the cost from introduced risk? This factor describes how risk might be introduced by the inclusion of the technology. Each type of risk was subjectively rated. Types of risks are paired with the types of risk reduction. Note that, unlike all the other factors, a higher potential for introducing risk has a negative impact on the technologies overall score.

External investment (outside NASA)

What is the current research activity going on outside of NASA? This factor characterizes how much research and investment has recently and is currently going into the development of the technology by entities outside of NASA. This is overall research on the technology excluding HARI research investments. We measured this using the publication rate associated with each technology. The name of each technology was searched for on 6/24/2019 on Web of Science using the following search, where technology is substituted for each: ALL FIELDS: (technology) Timespan: 2013-2018. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. Similarly, each technology was also searched for in Google Scholar on the same date and with the same time span. The sums from both searches were used to determine scores for this factor (see Trade Study Approach 2.5.2).

Technology Readiness Level (TRL)

What is the current TRL? This factor characterizes the maturity level of the technology. We estimated the technology's current TRL using information gathered from the literature and provided by our SMEs. TRL was split into three categories: Below TRL 3, TRL 3-5, and TRL 6 or greater.

Research interest (within NASA)

What is the current research interest within NASA? This factor describes if the technology has any potential for infusion into NASA missions as determined by the NASA Technology roadmaps/NASA Strategic Technology Investment Plan. We qualify this by checking if the technology is present on the NASA technology roadmap.

2.5.2 Trade Study Approach

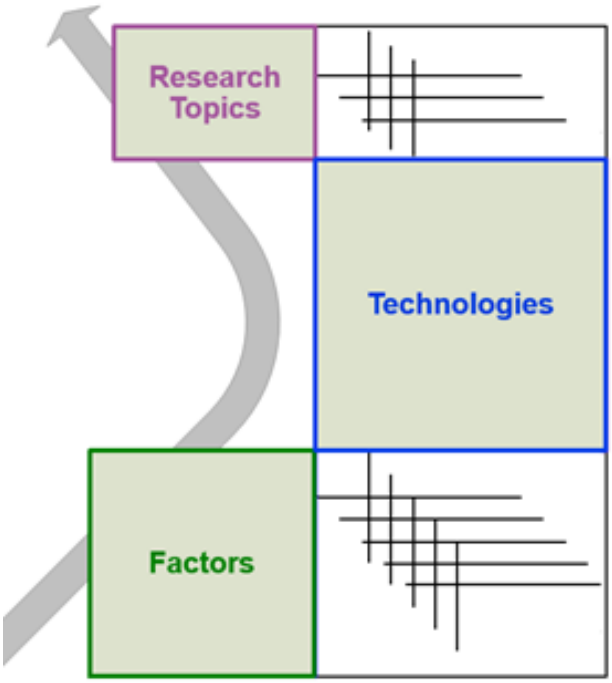
A multi-dimensional trade analysis was performed to objectively score HARI research topics and specific technologies in a recommended order of priority for NASA investment. The approach used was similar to a Relationship Matrix Decomposition Scheme (RMDS) (Boppe, 2010), see Figure 2.5. The factors for assessment described above pertain directly to HARI technologies, while research topics are assessed through direct relationships with those technologies, see Figure 2.5a. This parallels the RMDS approach of tracing assessment of system configurations and technology options based on objectives/goals through functional options, see Figure 2.5b. For the complete trade table used in this study, see Appendix B.

Research topics and technologies were defined as related if a given technology supports the research topic such that its development would fundamentally drive investigation of that topic. Each technology was given a score resulting from the technology-factors dimension of the trade. The scores for each related technology for a given research topic were summed to achieve the total score for that topic.

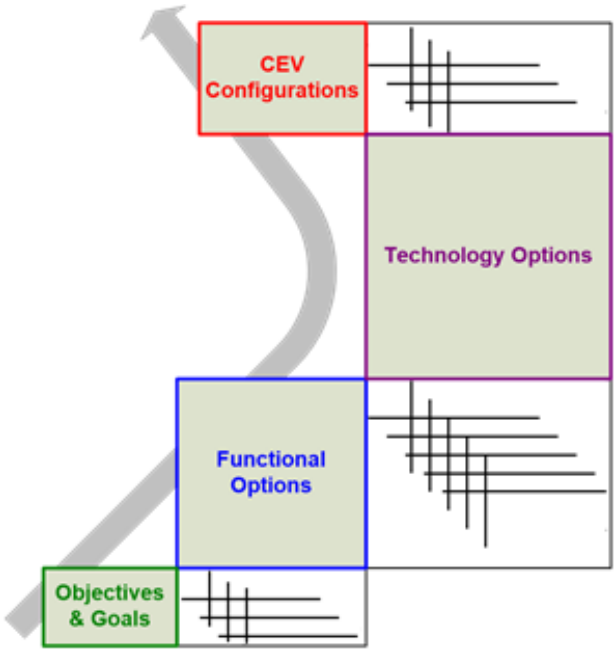
The technology total scores represent a roll-up of individual weighted factor scores for each specific technology. At a factor level, normalized scores for each technology were determined in a series of one-dimensional factor-technology trade studies. These individual factor-level trades used to compile the factor-technology dimension are found in Appendix B. The factor scores were multiplied by the factor weights as defined in the NASA Stakeholders section of this report and summed for each technology.

Factor-Level Trades

The factor-level trades for Risk Reduced, Risk Introduced, and TRL each assessed the specific technologies against three weighted options. Risk Reduced, for example assigned an individual score of 0 or 1 to each technology if it potentially reduced risk to crew, risk to



(a) Top-level trade study approach used in the HARI analysis.



(b) Structure of the RMDS trade study approach.

Figure 2.5: Trade study approaches.

mission/vehicle, or risk of loss of performance. Each potentially reduced risk was weighted (1 to 3) based on relative ranking as found by the NASA stakeholders. The total weighted scores for each technology (sum of weights x scores) were normalized by the highest possible score for a single technology to find the factor scores on a scale between 0 and 1 (see Equation 2.1, below).

$$\text{Normalized Score} = \frac{\Sigma(\text{Weighted Scores for a Technology})}{\Sigma(\text{Weights}) \times \max(\text{Individual Score})} \tag{2.1}$$

For Risk Introduced, normalized scores were multiplied by -1, as introduced risk were tallied as a negative contribution to overall technology assessment. The TRL trade was performed identically to Reduced Risk, with the exception that technologies could only be assigned to a single average TRL range. Assessment for Research Interest (within NASA) was simplified compared to other factor-level trades as scoring had a single binary level (trade table included in Appendix B).

In the Task Applicability and Task Enabling factor-level trades, scores were assigned

between technology and task, with equal weighting across tasks (all assigned a weight of 1). Technologies were assessed for Task Applicability with a score of 0 or 1, while for Task Enabling technologies were assigned a score for each task of 0, 1, or 2 as described in the NASA Stakeholders section.

In assessment of External Investment in NASA, as described previously, two search engines were used to find estimates on the number of recent publications pertaining to each technology. Searches with each service are known to poll databases of drastically different size. The relative database size used by both search engines was accounted for by applying a weight determined by the largest publication total found for each service. In this way, the relative publication totals for each service could be normalized independently prior to summing the results of the two searches.

2.6 Results

2.6.1 Research Topics

The final scores for research topics have been ranked, such that a high score represents a recommended higher priority for research investment by NASA (see Table 2.6). The top-ranking research topics are:

1. Improving training for HAR systems and tasks
2. Establishing appropriate trust in automation/robotics systems
3. Understanding human intent

Note that all these topics were identified as areas applicable to HARI concerns, regardless of the score. A high score here is highly dependent on the impact of the surveyed technologies on the research topic and suggests which research topics should be able to make relatively quick progress given the technology that is being developed now and in the next 5-10 years. A low score reflects research topics that either have relatively few technology solutions on the horizon, had relatively low factor scores for the technologies that are related to the topic, or both.

These top-ranking research topics were driven by their associated highly scoring technologies. Improving training for HAR systems and tasks touches on a variety of upcoming

Research Topic	Score	Rank
Improving training for HAR systems and tasks	84.66	1
Establishing appropriate trust in automation/robotics systems	79.27	2
Understanding human intent	77.13	3
Enabling command/control of complex robotic systems	72.80	4
HAR team performance optimization and function allocation	68.76	5
Autonomous/robotic system communication to humans	58.82	6
Ensuring human safety (physical)	50.66	7
Improving situation awareness in HAR systems	50.07	8
Continuous human performance monitoring	44.68	9

Table 2.6: The resulting prioritization of research topics from linking research topics to the upcoming technologies.

technologies ranging from machine learning to robotic/human information interfaces. These technologies ranked highly in their task applicability and potential for reducing risk and had relatively little potential for introducing new risks when compared to other technologies. By leveraging these upcoming technologies, researchers have ample opportunities to investigate novel techniques for improving training for these systems. These upcoming techniques will prove invaluable as HAR systems and tasks continue to increase in complexity, especially if, for example, they can provide crew with just-in-time training for critical tasks when they are far from the support provided by mission control. The topic of training came up many times in our conversations with subject matter experts, who often noted case examples of major failures in their explanations for why this training was needed. Similarly, many subject matter experts also mentioned the need for establishing appropriate trust in automation/robotics systems during our interviews, a topic which came up repeatedly in our review of the literature. Over-trust and under-trust in these complex systems were both noted as being dangerous and can result from inadequate training. Research in trust has often focused on its role in flexible, adaptive, and adaptable automation, where operators can be unclear which mode the system is in. By taking advantage of upcoming technologies in intuitive physical control and robotic/human information interfaces, researchers can help

to bring humans into the loop on what is happening within these complex systems.

In contrast to the top-ranking research topics, continuous human performance monitoring and improving situation awareness in HAR systems were among the lowest ranked topics. While both are important when considering future long duration exploration missions, neither were directly associated with many upcoming technologies. Despite being associated with the top-ranking technology, machine learning, continuous human performance monitoring ranked lowest. This is primarily because it was also associated with the two lowest ranking technologies, non-invasive behavioral and physiological sensing and implantable biometrics. Despite these technologies' clear relation to performance monitoring, they were among the lowest ranking when considering the task applicability and task enabling factors, which were considered the most important by our NASA stakeholders. Improving situation awareness in HAR systems lower score came as a surprise as situation awareness presents a challenge across all HAR applications.

2.6.2 Technologies

The resulting ranks of the technologies from the weighted factors are shown in Table 2.7. The top-ranking technologies identified from the trade study are:

1. Machine Learning
2. Autonomous obstacle detection/imaging
3. Robotic/human information interfaces
4. Artificial Intelligence

These top-ranking technologies have seen enormous advancements in research interest and development over the past few years, and all offer a large benefit to the tasks required by NASA on future LDEMs. In contrast to the top-ranking technologies, low ranking technologies show a trend of reflecting a combination of low research interest, task relevance, or TRL.

The top-ranking technologies all benefited from high marks across all our factors. Machine learning, our top-ranking technology, particularly stands out due to scoring highest in the External Investment (outside of NASA) factor, where it significantly outperformed

Technology	Score	Rank
Machine Learning	19.33	1
Autonomous obstacle detection/imaging	16.37	2
Robotic/human information interfaces	15.86	3
Artificial Intelligence	15.25	4
Intuitive physical control interfaces	13.98	5
Autonomous path planning	12.36	6
Augmented Reality and Virtual Reality	12.29	7
Robotic agents	12.25	8
Flexible, Adaptive, or Adaptable Automation	12.24	9
Assistive Robotics	9.69	10
Speech recognition	8.78	11
Non-invasive behavioral and physiological sensing	8.12	12
Implantable Biometrics	4.98	13

Table 2.7: The resulting prioritization of technologies using the trade study.

the other technologies. As we noted in the literature review, machine learning has been repeatedly forecast as being essential to the future of human-robotic interaction (Wang et al., 2018). It also came up extensively in our conversations with our subject matter experts, though several of them also stressed caution in assuming machine learning could solve any problem without issue. Artificial Intelligence was also mentioned by most of the SMEs but ranked lower due to its dramatically higher potential for introducing risk.

Autonomous obstacle detection/imaging was our second highest scoring factor but had little impact on our research topic recommendations. Despite being a well-established technology, the only topic it was ultimately related to was ensuring human safety (physical). Like our other high scoring technologies, however, it scored well due to its high task applicability, task enabling potential for reducing risk factors. This suggests that, while the technology should continue to be developed and refined, there is minimal applicability toward ongoing research that addresses outstanding HARI risks and challenges. Improvements resulting from

refinement in the commercial sector, especially regarding autonomous cars, should enable faster and safer algorithms in the future.

Several technologies were highly clustered in the middle of our rankings: autonomous path planning, augmented reality/virtual reality, robotic agents, and flexible/adaptive/adaptable automation also scored within a few tenths of a point from each other. These technologies all had relatively high potential for introducing risk but were otherwise highly applicable to the tasks related to space HARI. As noted previously, the two lowest ranking technologies, non-invasive behavioral and physiological sensing and implantable biometrics were among the lowest ranking when considering the task applicability and task enabling factors, which were considered the most important by our NASA stakeholders. These technologies were also those which did not score in the Research Interest (within NASA) factor, as they were not present in the NASA Technology roadmaps or NASA Strategic Technology Investment Plan. The third lowest ranking technology, speech recognition, despite being a widespread and high TRL technology, scored poorly because it was the lowest scoring in both the task applicability and task enabling factors.

2.7 Contribution (Relation to NASA HARI Gaps)

NASA has identified four gaps in HARI knowledge, as part of the larger HFBP characterization of human factors risks and associated knowledge gaps (Human Research Program, 2011). These gaps need to be closed in order to mitigate HARI related risk as it pertains to spaceflight. The NASA HARI Gaps are:

HARI-01 We need to evaluate, develop, and validate methods and guidelines for identifying human-automation/robot task information needs, function allocation, and team composition for future long duration, long distance space missions.

HARI-02 We need to develop design guidelines for effective human-automation-robotic systems in operational environments that may include distributed, non-collocated adaptive mixed-agent teams with variable transmission latencies.

HARI-03 We do not know how to quantify overall human-automation-robotic system performance to inform and evaluate system designs to ensure safe and efficient space mission

operations.

HARI-04 We need to identify and scope the critical human-automation/robotic mission activities and tasks that are required for future long duration, long distance space missions.

This study extends prior investigation of HARI tasks, specifically to address gap HARI-04 directly. This investigation and trade study identify prioritized lists of specific technologies whose advancement support the activities and tasks required for future space exploration missions, as well as research topics where investment will support both HARI task capabilities and closing of the other three HARI knowledge gaps. All the research topics identified in this report can assist with closing HARI-02, and most address HARI-03 as well. Table 2.8 provides a complete mapping of the relationships between research topics and HARI gaps. Although few of the research topics address HARI-01, the outstanding concerns identified by NASA for closing HARI-01 pertain directly to the topics of safety and function allocation, which are reflected here.

2.8 Recommendations

Based on the trade analysis performed, we recommend that NASA's HFBP Element prioritizes research investment in the topics of improving training for HAR systems and tasks, establishing appropriate trust in autonomous/robotic systems, and understanding human intent. It is important to note that all the identified HARI research topics have application toward mitigating HARI risk in spaceflight tasks for future missions. These topics, however, represent the highest-priority areas for investment.

Investigation and identification of methods to improve training for HAR systems has the potential for far-reaching impact on reducing risk in mission operations, with limited chance of introducing new risk. Training is also a ubiquitous concern across all HAR systems and tasks. Similarly, establishing trust between the human and robotic/autonomous system showed trends of tracing to high risk reduction potential, though risk introduction potential was more varied. Trust between the human and the autonomous system (or robotic agent) came up again and again in discussions with experts across different HARI related disciplines as critical to the success of HAR operations. Without appropriate trust, elements fundamental to other research topics, such as teamwork or performance, break down.

Research Topic	Gap HARI-01	Gap HARI-02	Gap HARI-03
Understanding human intent	✓	✓	✓
Autonomous/robotic system communication to humans		✓	✓
Ensuring human safety (physical)	✓	✓	✓
Continuous human performance monitoring		✓	✓
HAR team performance optimization and function allocation	✓	✓	✓
Enabling command/control of complex robotic systems		✓	
Improving situation awareness in HAR systems		✓	✓
Improving training for HAR systems and tasks		✓	
Establishing appropriate trust in automation/robotics systems		✓	✓

Table 2.8: Mapping of HARI related Research Topics to HARI Gaps identified by NASA.

While understanding human intent ranked highly as a research topic largely because of the number of technologies to which it was related, we believe this topic deserves its place in the prioritized rankings because, like training and establishment of trust, it stands out in overall potential to address HAR concerns for spaceflight. The ability to interpret human communication, input, need, and general intent is critical to the successful operation of any HAR system which interacts directly with a human user. Consequently, it is strongly tied to several of the other research topics defined (e.g. continuous human performance monitoring, enabling command/control of complex robotic systems) and investment in this area could bolster study in those lower-priority topics as well.

One of the primary outcomes from this research was to determine directions for HARI re-

search that will close HARI risk and support capabilities for HAR tasks in space exploration. Investigation of these research topics will provide a fundamental foundation for addressing challenges that face implementation of HARI technology solutions. Improvement of training, trust, and human intent interpretation in HAR systems enables capability for a wide range of HAR space exploration tasks, both for long duration orbital missions and future planetary surface exploration.

The first prioritized research topic, identification of methods to improve training in HAR systems, is a basis for which the research in the remainder of this dissertation work is based. Augmented feedback, as we discussed in the literature review in the previous Chapter, has been shown to greatly improve performance without increasing workload when implemented correctly. The efficacy of concurrent bandwidth feedback as a method to improve training in complex human automation/robotic systems will be further explored in the remaining Chapters of this dissertation.

Chapter 3

Augmented Reality Tracking Task

To further our analysis of concurrent bandwidth feedback, we first analyzed a relatively simple compensatory tracking task. For the study presented in this Chapter, we designed an experiment to evaluate if concurrent bandwidth feedback could improve human's ability to recognize stereoscopic cues presented by an augmented reality display. This task is conceptually and functionally simpler than the SAFER study presented in Section 1.2.2, allowing for more insight into the effects of concurrent bandwidth feedback. Portions of this chapter were originally published in the conference proceedings for AIAA Modeling and Simulation Technologies 2019 (Karasinski and Robinson, 2019b). The work presented in this Chapter was supported by the Link Foundation's Modeling, Simulation, and Training Fellowship under the title "Evaluating Augmented Reality for Space Telerobotics Training".

3.1 Introduction

3.1.1 Overview

Recent advances in computing hardware have enabled a new generation of mobile augmented reality devices which have the potential to improve human performance and reduce workload in a variety of tasks. The aim of this study was to investigate the effect of several factors on human performance and workload in a three-axis manual tracking task. The Microsoft HoloLens is a head-mounted, mobile augmented reality device which provides a stereoscopic 3D view to the user which is not available with traditional 2D displays. This lightweight headset presents an augmented reality overlay in the form of a transparent 2.3

megapixel widescreen display. The stereoscopic display presented by the HoloLens, combined with the advantage augmented reality allows over virtual reality in not occluding the real world, makes it a compelling option to include in future human-machine interfaces. To evaluate whether this new display technology could improve user performance in a tracking task, we designed an experiment that investigated if the depth cueing delivered by stereoscopic display could provide improved performance and decreased workload. We were also interested to see if a 2D display could gain the benefits of a depth cue by rotating the axis of the task such that the depth cue was more readily available. Research has shown that presenting three-dimensional information on a two-dimensional screen is not a simple task, and that the projection of the 3D information onto the 2D screen can cause large changes in the performance of the user (Kim et al., 1987b). In addition to these cues, we also investigated the effects of concurrent bandwidth feedback on task performance and workload as an alternate technique to improving performance. Concurrent bandwidth feedback alerts the operator when their real-time performance has drifted outside an acceptable, predefined window of performance. The use of this type of feedback has been shown to improve performance in a wide variety of motor control tasks (Salmoni et al., 1984; Sigrist et al., 2013; Karasinski et al., 2017).

3.1.2 Stereoscopic Displays

Stereoscopic displays are systems "in which two slightly different views of a scene are provided to a viewer, one image for each eye... allow[ing] the viewer's binocular visual system to extract depth information in a scene using this disparate information" (McIntire et al., 2014). Without the aid of the binocular depth cue presented by stereoscopic displays, viewers are instead entirely reliant on monocular clues such relative sizing, occlusion, and motion. One of the primary motivation for stereoscopic displays is that "[t]he visual scene of a 3D world is a more 'natural,' 'ecological,' or 'compatible' representation than that provided by 2D displays" (Wickens, 1990). As a result of this motivation, the effects of stereoscopic displays on human performance have been extensively studied in the literature. Several authors have attempted to classify which types of tasks may stand to benefit (Wickens et al., 1989; Wickens, 1990; Naikar, 1998; Dixon et al., 2009; McIntire et al., 2014). A recent review of 184 papers, for example, suggests that 60% of studies showed some benefit of 3D stereo

displays, 15% of tasks showed unclear or mixed benefits, and 25% of studies showed no clear benefits (McIntire et al., 2014). In their review, tasks involving finding/identifying/classifying objects and tasks involving real/virtual spatial manipulations of objects benefited the most, while learning/training/planning tasks were the least likely to show a benefit.

Kim et al. also performed a quantitative evaluation of perspective and stereoscopic displays in a three-axis manual tracking task. They investigated the differences between perspective and stereoscopic displays, the elevation angle, azimuth angle, and the effects of two visual enhancements: a grid and a reference line. They found very strong relationships between elevation and azimuth angles and tracking performance, with the best performance occurring at an elevation angle of 45 degrees and an azimuth angle of 0 degrees. Tracking performance decreased rapidly as the azimuth angle varied, and decreased less rapidly as the elevation angle varied. In general, they found that the stereoscopic display allowed for better tracking performance, though the inclusion of the reference line visual enhancement greatly decreased the benefit over the perspective display. Using only two subjects, they provided some insight into intrasubject and intersubject variability. In several instances, intrasubject variability showed 50% changes within the same experimental condition, while intersubject variability also appeared quite large in some conditions. Kim et al. repeated the evaluation of these parameters on a telerobotics pick-and-place study. They found similar results in this second study, suggesting that their results could be generalized and that three-axis tracking performance can be correlated with pick-and-place completion time.

Smallman et al. similarly investigated the effect of visual enhancements and 2D vs 3D displays for the development of a naval air warfare console. Participants viewed naval and aircraft tracks in either a conventional 2D top-down display or a 3D display, and then attempted to reconstruct track positions. They investigated the effectiveness of drop-lines and drop-shadows, and found that they significantly improved subjects ability to localize aircraft compared to when the enhancements were not present. Furthermore, in the absence of either visual enhancement, subjects performed better with the 2D display than the 3D display. Similar to Kim et al., they ultimately recommended that 3D stereoscopic displays include the use of a reference or drop-line for optimal performance.

3.1.3 Summary

Concurrent bandwidth feedback has been used in a large variety of motor control tasks, and has generally been found to improve performance. Until recently, however, only simple tasks such as physical movements or basic pursuit tasks have been investigated. More recent works, including the lane-keeping task by de Groot et al., and our previous work with the SAFER task, have indicated that concurrent bandwidth feedback can also be quite effective for complex tasks. The decrease in required learning time, improved performance, and decreased workload seen in our previously discussed SAFER task show that concurrent bandwidth feedback can prove to be most useful very early in training when subjects are first exposed to complex, highly dynamic tasks. While visual concurrent bandwidth feedback has been used in a variety of tasks, no researchers have investigated its effects on a three-axis tracking task.

Despite extensive previous research, to the authors' knowledge there exists no study in the literature addressing human performance or workload changes in manual tracking tasks between traditional computer monitors and mobile, augmented reality headsets. If operator performance while using augmented reality displays is improved–or at the very least, not degraded–then these devices could prove especially valuable in scenarios where it is impractical or otherwise difficult to provide a traditional computer interface. There are a variety of robotics tasks, such as pick-and-place tasks, for which performance may be improved by allowing an operator the mobility to move and view the scene from whatever position is convenient at a given time. Traditional robotics stations require the operator to remain in a single position, and typically only allow for several camera angles. Mobile augmented reality displays allow the operator to take advantage of their ability to move through the environment, without needing to manage external cameras.

3.2 Materials and Method

In this experiment, subjects were responsible for simultaneously completing three primary tracking tasks and a secondary task for assessing real-time workload. Each axis of the tracking task was disturbed by a sum-of-sines, resulting in a random appearing signal that was difficult for the subjects to predict. The two-choice secondary task appeared on a screen

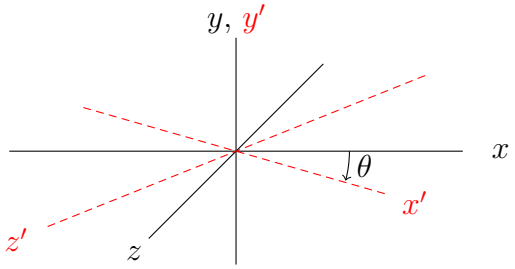


Figure 3.1: Perspective display of the coordinate frame for the tracking tasks, with the x, y, and z axes labeled. After rotating by θ around the y axis, the resulting reference frame of $x'y'z'$ is also labeled.

next to the tracking task, and asked subjects to respond to either a "LEFT" or "RIGHT" command by pressing a button. Subjects controlled the tracking task and responded to the two-choice task by using a Microsoft Xbox controller. The subjects used the left joystick on the controller to control the x and y axes tracking tasks. Subjects moved this joystick left and right to control the x axis, and up and down to control the y axis. The subjects used the right joystick on the controller to control the z axis. Subjects moved this joystick up and down to control the z axis. See Figure 3.1 for a visualization of the coordinate frame of the primary tracking task. Subjects used the left and right triggers on the controller to respond to the two-choice task, using the left trigger to indicate "LEFT" on the two-choice task, and the right trigger to indicate "RIGHT" on the two-choice task.

3.2.1 Hypotheses

This study assessed the influence of display type (2D with perspective vs. 3D stereoscopic), relative display attitude (zero degrees vs. thirty degrees), and concurrent bandwidth feedback (with vs. without) on performance and workload. Objective performance was measured using the root-mean-square error (RMSE) of the depth (z) axis. Objective workload was measured using the response time to the secondary task, and subjective workload was measured using the NASA-TLX. It was hypothesized that:

Hypothesis 1 Concurrent bandwidth feedback will improve performance in the depth (z) axis for both display types, and will decrease workload.

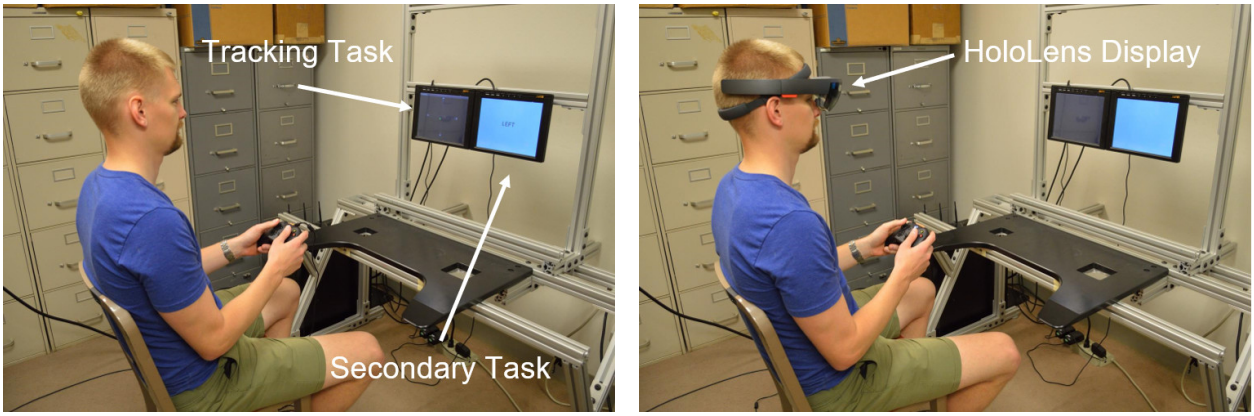
Hypothesis 2 Stereoscopic augmented reality displays improve performance in the depth (z) axis, but do not affect workload.

Hypothesis 3 Rotating the display improves performance in the depth (z) axis for both display types, and will decrease workload.

3.2.2 Procedure

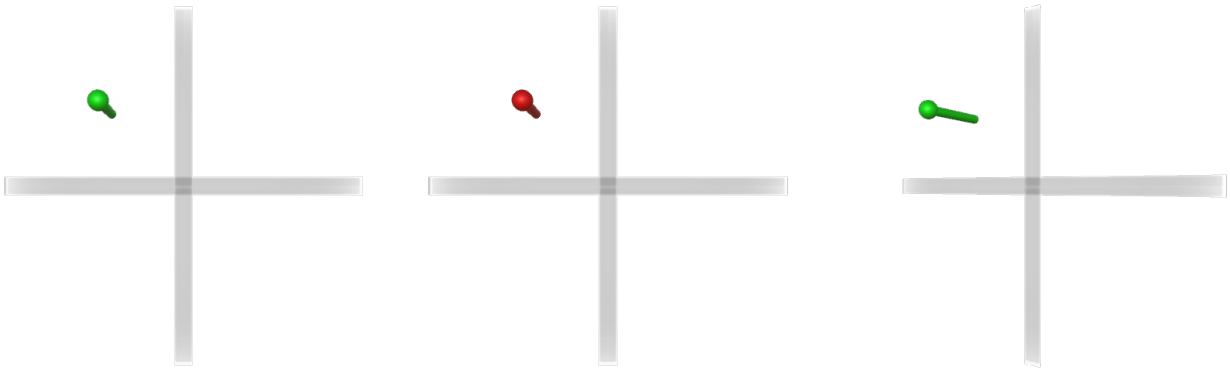
A total of 24 volunteer subjects (19 males, 5 females) were recruited in accordance with the University of California, Davis Internal Review Board (Project #1219474-1), and subjects were not compensated. There were 12 subjects in the 2D group, and 12 subjects in the 3D group. Subjects were students in the University of California, Davis, College of Engineering. All participating subjects had normal vision (no colorblindness, eyesight correctable to 20/20 vision) and full motor control of their hands.

A human-in-the-loop simulation was conducted using a fixed-base simulator (see Figure 3.2). The simulator consisted of two 10.4 inch LCD displays. The primary tracking task was shown on the left display, while the right display showed the two-choice secondary task. Subjects were seated for the duration of the experiment and were placed one meter perpendicular from the center of the left display. For subjects in the 3D group, the left LCD monitor was turned off, and the tracking task was instead displayed on the HoloLens. The secondary task on the right-hand monitor was visible to the subjects through the clear display of the HoloLens. For these subjects, the cross was placed at the same height as the



(a) 2D Group. **(b)** 3D Group.

Figure 3.2: The fixed-based simulator used by both groups. The left screen was used to display the primary task (in the 3D group, the primary task was placed here in the augmented reality display), and the right screen was used to display the secondary task (regardless of display type).



(a) Baseline. **(b)** Color Feedback. **(c)** Rotated (about the y axis).

Figure 3.3: The three different designs in the same error state. Subjects were instructed to control the indicator and return it to the origin of the axes. (b) The color feedback has been activated here, changing the guidance target from green to red.

subject's head, such that it was viewed with no relative attitude in the baseline condition (see Figure 3.3). Both groups viewed the guidance cross with a width and height of 5 inches by 5 inches. For subjects in the HoloLen's group, the z motion of the cross also could move up to 5 inches in either direction away from the center of the guidance cross. Subjects in both groups used the same Microsoft Xbox controller and control scheme to complete the task.

The disturbance function that drove the guidance indicator on the primary display was a sum of 13 sinusoids approximating a rectangular spectrum with a 2.0 rad/s cutoff frequency. The disturbing force, as a function of time, $d(t)$, was

$$d(t) = \sum_{i=1}^{13} A_i \sin\left(w_i t + \phi_i\right) \tag{3.1}$$

Each sine wave amplitude, frequency, and phase offset was borrowed from a similar experiment (Hess, 1984b). Table 3.1 lists the sine wave amplitude, frequency, number of cycles in a 60 second run, and phase offset for each sine wave in the disturbance force. This disturbance was the same for all subjects and trials, though the subjects were naive to this. The x, y, and z axes all experienced the same disturbance force generating function, but the y axis was temporally offset by 60 seconds and the z axis was offset by 120 seconds. This allowed for a very similar generation of disturbance forces for each axis. The RMSE of the disturbance force was normalized along each axis such that all three were the same.

i	A_i/A_1	w_i rad/s	No. of cycles in 60-s run	ϕ_i
1	1.0	0.18850	1	$\pi/6.5$
2	1.0	0.31416	3	$2 \times \phi_1$
3	1.0	0.50265	4	$3 \times \phi_1$
4	1.0	0.87965	8	$4 \times \phi_1$
5	1.0	1.44513	13	$5 \times \phi_1$
6	1.0	2.13628	20	$6 \times \phi_1$
7	0.1	3.07876	29	$7 \times \phi_1$
8	0.1	4.20973	40	$8 \times \phi_1$
9	0.1	5.78053	55	$9 \times \phi_1$
10	0.1	8.23097	78	$10 \times \phi_1$
11	0.1	11.24690	107	$11 \times \phi_1$
12	0.1	15.77079	150	$12 \times \phi_1$
13	0.1	23.93894	228	$13 \times \phi_1$

Table 3.1: The relative amplitude, frequency, number of cycles in each 60 second run, and phase offset each i^{th} sine, see Equation 3.1.

Design	θ (degrees)	Feedback
Baseline	0	No
Feedback	0	Yes
Rotated	30	No

Table 3.2: The factors that were modified between the different designs.

Three designs were presented to the subjects to evaluate: a baseline design, a color-based concurrent bandwidth feedback (CBF) design, and a rotated design. Figure 3.3 shows all three designs in the same error state. The three designs were very similar, having only minor differences between each other.

The baseline design consists of a flat cross with a center target point and a green sphere error indicator. This indicator also casts a green, variable-length rod perpendicular to the

plane of the cross, which allows for a visual estimation of the error in the z axis. The x axis is parallel with the horizontal cross, while the y axis is parallel with the vertical cross. The color feedback design was identical to the baseline design in every way, with the addition of visual concurrent bandwidth feedback (green or red) on the z axis. When the absolute value of the error on the z axis exceeded a threshold (upper or lower limit of a fixed bandwidth), the color of both the spherical indicator and the cylindrical rod changed from green to red (see Figure 3.3b). When the absolute value of the error on the z axis was lowered back within this fixed bandwidth, the indicator changed back to a green color (see Figure 3.3a). The rotated design was identical to the baseline design, but the relative attitude of the display was rotated about the y axis by 30 degrees (see Figure 3.1). This design was created to provide subjects with an enhanced ability to visually detect errors in the z axis, and 30 degrees was chosen after a brief pilot study.

Before entering the study, subjects were randomly placed in a display group (either the LCD monitor or HoloLens), and were then randomly placed into an order group (which consisted of Baseline-Feedback-Rotated, Feedback-Rotated-Baseline, and Rotated-Baseline-Feedback). This order group was created to remove any order effects that might arise due to training on a given display, and follows a standard Latin squares design. (The order of the designs was expected to be insignificant, but we will later discuss that this was found not to be the case.) After entering the experiment room, each subject was familiarized with the task, the three designs, NASA-TLX, and the controller through a twenty minute training session during which they were instructed to:

- Minimize the displacement of their guidance target from the center of the display axes
- Respond to the two choice task as accurately and quickly as possible

Subjects in the 3D group completed a short calibration program that adjusted the display to their interpupillary distance. All subjects were then allowed to complete two familiarization trials, during which time they could ask questions about how the controls worked, or any other aspects of the task. The proctor also used this time to ensure that subjects showed basic competency with the task by responding to both the tracking and two-choice tasks appropriately. All familiarizations were done with the baseline design, regardless of

which design the subjects evaluated first.

After this familiarization process, subjects completed ten trials with their first design. After completing these trials, they answered a brief survey which asked them if the design was adequate to complete the task. Subjects were also asked to subjectively rate their performance in a questionnaire after evaluating each design. Subjects were asked to rate "I found the tracking display adequate to complete the task." on a five point scale where 1 indicated "Strongly Disagree" and 5 indicated "Strongly Agree." After this survey, subjects then completed a NASA-TLX workload survey. Subjects then repeated this process with their second and third designs. At the conclusion of the three designs, subjects were also asked to complete a preference survey which inquired into what design the subjects preferred.

3.3 Results

We conducted three-way mixed ANOVAs between display (2D or 3D), design (Baseline, Feedback, or Rotated), and starting design (Baseline, Feedback, or Rotated) with repeated measures on the design factor. When significant effects were observed, post hoc comparisons using the Tukey Honest Significance Difference (HSD) test with a Bonferroni adjustment were completed to investigate which pairs of the factor were significant. In order to remove learning and fatigue effects, each subject's best performing five trials in each design were averaged together to produce one average score for each subject and design. Additionally, the first ten seconds of each sixty second trial were not included in the analysis to remove initial transient effects.

The root-mean-square error (RMSE) of the depth (z) axis was used to understand the differences between the three designs and the two devices. The RMS of the disturbance signal was calculated and used to normalize the RMSE. Under this definition, an RMSE of 1 indicates performance no better than no input, and an RMSE greater than 1 indicates worse than hands-off performance. It was expected that the baseline design would result in worse performance in the z axis than for the color feedback and rotated designs. It was also expected that, due to the stereoscopic nature of the display, the HoloLens would allow for better z axis performance than the 2D LCD monitor.

Results of the ANOVA on the z axis RMSE showed significant effects for design ($F(2, 36) =$

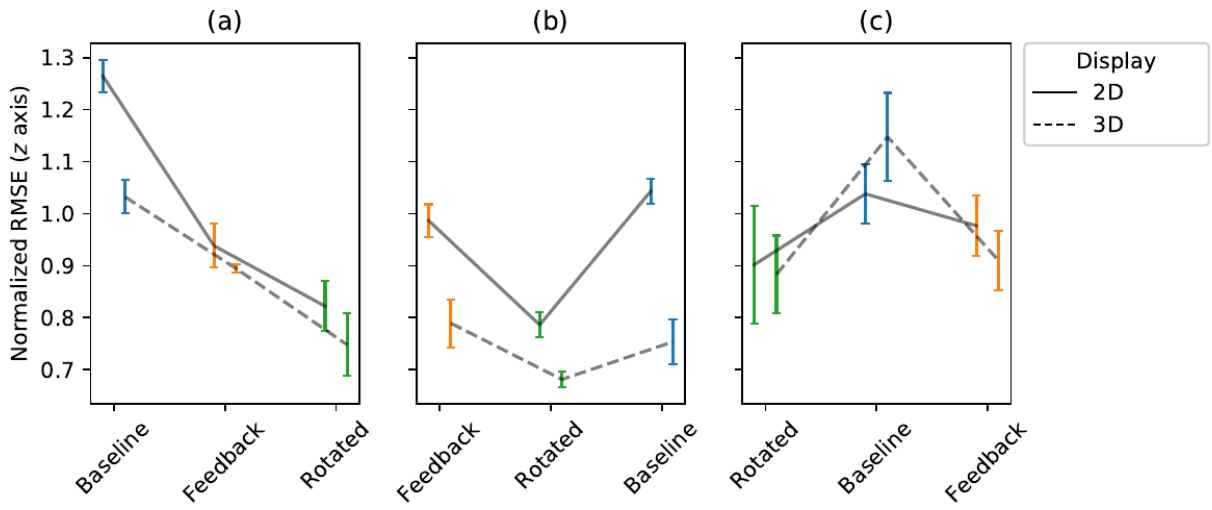


Figure 3.4: The resulting normalized RMSE along the z axis. Subjects started in either the (a) Baseline, (b) Feedback, or (c) Rotated design.

$84.92, p < .001$), device ($F(1, 18) = 7.22, p < 0.015$), and start design ($F(2, 18) = 4.81, p < 0.021$). The ANOVA also showed a significant interaction effect between design and starting design ($F(4, 36) = 8.55, p < 0.0001$), and a three way interaction between design, device, and starting design ($F(4, 36) = 5.57, p < 0.002$). Further investigation into the effect of starting design using pairwise comparisons showed significance differences between subjects that started in the concurrent bandwidth feedback group and those in the baseline ($p < 0.001$) or rotated ($p < 0.001$) designs, but no difference between subjects that started in the baseline and rotated designs ($p > .38$). The difference in performance between design, device, and starting design can be seen in Figure 3.4.

Due to the unanticipated but significant interaction effects in the starting design factor, the remainder of the analysis is split between subjects who started with the concurrent bandwidth feedback design and those that did not (e.g., those that started in the baseline or rotated design). For subjects that started in the CBF design, there was a significant effect of design ($F(2, 12) = 21.65, p < .0001$), a significant effect of display ($F(1, 6) = 36.73, p < 0.001$), and a significant interaction effect between design and device ($F(2, 12) = 5.42, p < 0.021$). For subjects that did not start in the CBF design, there was a significant effect of design ($F(2, 28) = 38.70, p < .0001$), but no significant effect or display ($F(1, 14) = 1.03, p >$

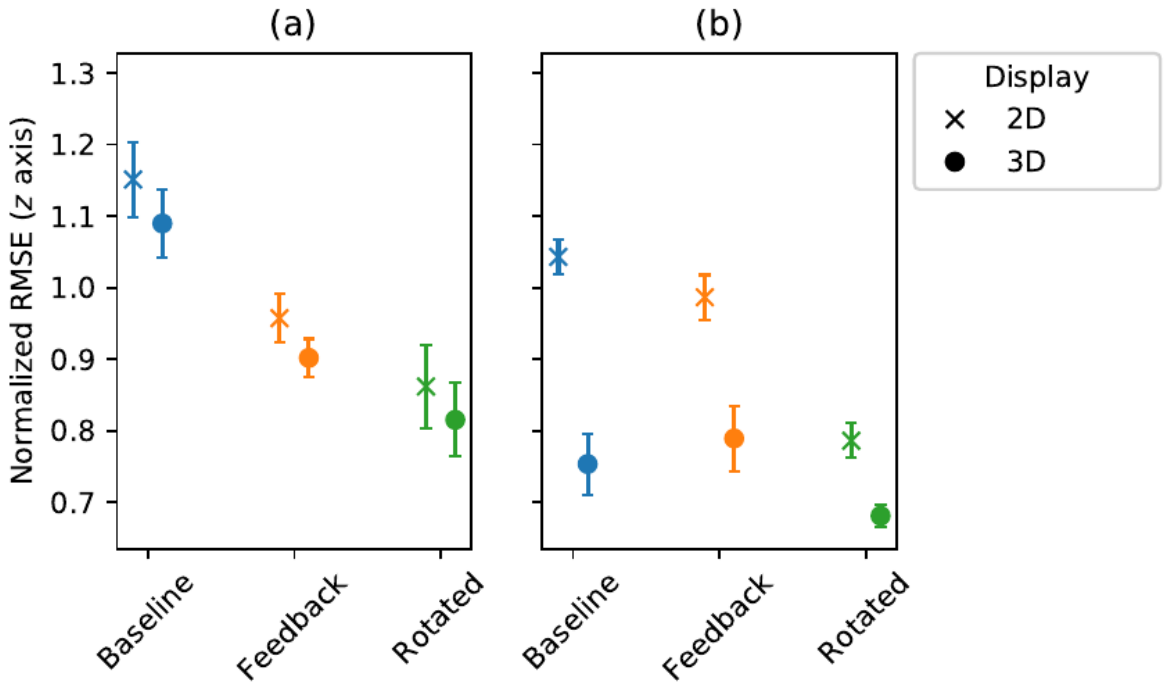


Figure 3.5: The resulting normalized RMSE along the z axis, grouping by subjects that started (a) without or (b) with feedback.

0.32), and no interaction effect ($F(2, 12) = 0.03, p > 0.97$). The resulting difference between subjects who started in the CBF design and those who did not is presented in Figure 3.5. For subjects who did not start in the CBF design, display had no significant effect, and subjects performed best in the rotated design, followed by the feedback design and finally the baseline design. Subjects who started in the CBF design performed significantly better in the 3D display than the 2D displays, and performed best in the rotated design, but comparably between the feedback and baseline designs.

The NASA-TLX was used to measure differences in subjective workload, and the reaction time to the secondary task was used to measure differences in objective workload between design, device, and starting design. There were no significant effects, nor interaction effects, found in the ANOVA for design, device, or starting design for the NASA-TLX measurements. There was a significant effect of design ($F(2, 36) = 7.93, p < 0.0014$) for the reaction time to the secondary task, though the magnitude of this effect was very small between designs (less than 100 ms difference) and was not significant during Tukey HSD tests. In general, there

were no significant effects found for workload measurements.

3.4 Discussions and Conclusion

To summarize these results, there were significant effects found in the z axis RMSE for design, with subjects generally performing the best using the rotated design, followed by the CBF design, and performing worst with the baseline design. There were significant effects found for the factors of device and starting design, though these must be interpreted carefully. Subjects who started with concurrent bandwidth feedback performed better than subjects who did not. We believe that this result reinforces the finding of our prior SAFER experiment, where subjects who were exposed to the CBF early on learned the task better and more quickly than those who were not exposed (Karasinski, 2016; Karasinski et al., 2016, 2017). An interesting effect of this exposure is that, after learning the task with CBF, subjects continued on to perform significantly better in the baseline condition than those subjects that did not start in the CBF design.

Subjects who started in the CBF design and who were wearing the HoloLens appear to have used the CBF to better learn the depth cue presented in the stereoscopic display. These subjects continued to perform significantly better than subjects who started with the CBF design but without the stereoscopic display when they continued to the baseline design. This indicates that even a brief exposure to the concurrent bandwidth feedback was sufficient to induce improved performance in the baseline design. Additionally, this also suggests that well-trained subjects could perform better using the stereoscopic display compared with the traditional display, but that subjects who were still learning the task could not take advantage of the additional depth cues provided by the display. Finally, there were no significant effects found for the NASA-TLX measurements, and there were significant but small effects found between the designs for reaction time.

In summary, we find partial agreement with all of our hypotheses in respect to the performance aspects of our experiment, while the workload was essentially unaffected by all of our experimental factors. For Hypothesis 1, subjects who completed the baseline design before the CBF design performed better in the CBF design, while subjects who completed the CBF design before the baseline performed approximately the same in both designs.

This indicates that CBF can both improve performance compared to a baseline design, and better train subjects such that, even after brief exposure, the feedback is no longer required. Workload was unaffected by the CBF.

For Hypothesis 2, subjects who started in the CBF design appear to have better learned the task and used the CBF to learn to interpret the depth cue provided by the stereoscopic display. Subjects who were not initially exposed to this feedback were unable to exploit the display, and did not perform significantly better than subjects without the display. For Hypothesis 3, subjects in the rotated design did perform better than those in the baseline design and appeared to be able to use this view to better interpret the depth of the target.

These results suggest that 3D displays such as the HoloLens have the potential to improve performance, but that simply donning a 3D display is not in itself wholly sufficient for improvement. Subjects performed the tracking task better when we exposed them to CBF early in the experiment, and were able to use the depth cues provided by the 3D display to sustain this improvement when the feedback was removed. These subjects achieved the best performance that we observed in this experiment, suggesting that the combination of CBF and 3D displays is an effective training technique for optimal performance. The rotated design generally resulted in excellent performance by all subjects, which suggests that a direct ($\theta = 0$) view should be avoided for three-axis tracking tasks. Care must be taken to provide subjects with the feedback required to adequately complete the task, and subjects should not expect to perform better simply because they have 3D displays.

Chapter 4

Surface Electromyography Task

The previous Chapter shows the benefits of applying concurrent bandwidth feedback to continuous motor control, but did not explore if this feedback can be successfully applied to more discrete tasks. To further investigate the utility of concurrent bandwidth feedback, we applied it to an electromyography-based control task. This experiment also investigated the difference between using concurrent and terminal feedback as potential training strategies to reduce the required time to use electromyography control. Portions of this chapter were originally published in the conference proceedings for AIAA SciTech 2020 (O'Meara et al., 2020).

4.1 Introduction

Training novice users to effectively use electromyography (EMG) control can be a long and difficult task (Skavhaug et al., 2016). EMG control has been applied to robotics (Artemiadis and Kyriakopoulos, 2010; Hussain et al., 2016; Hussain et al., 2016; Saraiji et al., 2018) and tele-operations (Chen et al., 2016), and has been suggested for UAV operation (Singh et al., 2019), but long learning times have generally prevented EMG from being used operationally. The use of augmented feedback strategies, however, have been shown to help reduce training times. Augmented feedback provides information that "cannot be elaborated without an external source; thus, it is provided by a trainer or a display" and has been shown to effectively improve performance in a wide variety of motor tasks (Sigrist et al., 2013). Biofeedback, which applies augmented feedback strategies to physiological

signals such as EMG, has proven to be a useful tool for improving performance and assisting in rehabilitation (Huang et al., 2006). Researchers have investigated various augmented biofeedback techniques and have found them to help subjects to "become more cognizant of their own EMG signal", allowing for better control (Basmajian, 1963). A recent review of the biofeedback literature suggests that "[b]iofeedback is more effective than usual therapy," though they also note that "[f]urther research is required to determine the long-term effect [biofeedback has] on learning" (Stanton et al., 2017).

Augmented feedback can be split into two large categories describing when the feedback is presented to a subject. Concurrent feedback is presented in real-time, as subjects complete a task, while terminal feedback is presented after the task is complete. In general, concurrent feedback has been shown to be more useful with increasing functional task complexity, while terminal feedback is often less useful when complexity is high (Sigrist et al., 2013). Concurrent feedback has recently shown great promise in EMG control, though less progress has been made comparing the effects of terminal feedback strategies or investigating long-term learning effects (Schweisfurth et al., 2016; Dosen et al., 2017; Markovic et al., 2017; Shehata et al., 2018).

Additionally, research into the bayesian theory of motor adaptation, which suggests that increased errors and decreased visual uncertainty lead to faster adaptation, may be a useful training methodology for reducing training times. Research into motor adaptation and surface EMG (sEMG) control has showed great promise compared to traditional techniques (Johnson et al., 2014). Recent research at UC Davis has further investigated the use of motor adaptation for two-dimensional myoelectric control (Lyons and Joshi, 2018). Lyons and Joshi found that subjects were more successful at hitting targets when exposed to control perturbations compared to a control group, and concluded by suggesting that "exposure to a variable mapping encourages exploratory behavior and underlies a change in adaption rate, which could potentially be used to train myoelectric control users." Motor adaptation has not previously been used as a training methodology, however, and it is unclear how it would compare to an augmented feedback based methodology.

4.1.1 Trust in Automation

Trust is an important factor when considering human-automation interaction, and inappropriate trust can lead to the disuse or misuse of automated systems. Trust is "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004). The reliability of a system, in particular, has been shown to be an important aspect of an operator's trust in a system (Wickens et al., 2000). While there have been many proposed models for trust, Hoff and Bashir's three layer model deserves particular attention. After performing a systematic review of the literature, they developed a three-layer model of trust which is split between dispositional trust, situational trust, and learned trust. Though researchers have little control over dispositional trust, they can affect situational trust by varying an experimental interface or environment and learned trust can be evaluated using repeated measures.

4.1.2 Summary

In this study, we address the effects of training methodology on performance, workload, and trust during a computer-based Fitts's Law cursor-to-target task (Fitts, 1954). The training methodologies evaluated were concurrent and terminal augmented feedback and motor adaptation based on variable mapping of control inputs. To test if the benefits provided by these training methodologies were acting to improve learning or simply being used as performance enhancers when present, we removed the feedback and variable mapping at the end of the study.

4.2 Materials and Methods

4.2.1 Subjects and Experimental Setup

55 subjects were recruited from the University of California, Davis's Psychology department. Subjects were excluded if they had a history of neuromuscular disorders, physical limitations of dominant arm, or prior sEMG control experience. Of the 55 subjects recruited, 48 completed the complete protocol and had an average age of 20.07 years (SD = 1.39). This study was approved by the University of California, Davis Institutional Review Board (Project #1461183-1), and subjects provided written consent and were compensated

in the form of university research credits.

The experimental setup for attaching the electrodes for the EMG control followed Lyons and Joshi, and two electrodes (ConMed 1620 Ag/AgCl center snap) approximately 2.5 cm apart were placed on the dominant hand side near the extensor digitorum proximal attachment. The signal from the electrodes was acquired as described in Lyons and Joshi (2016), and signal processing followed O'Meara et al. (2019). The raw signal was windowed, and the root-mean-square (RMS) value for each window was calculated, normalized by a manually set calibration constant, and incorporated into a moving average window to yield $\bar{x}$, see Equation 4.1.

4.2.2 Experimental Design and Subject Groups

Subjects in the experiment used EMG to control the cursor in a 2D center-out Fitts's Law task. The task took place in a normalized area of -1 to 1 units in each direction. To control the cursor, subjects input a series of three pulsive EMG commands which indicated a direction and velocity of the cursor. For the pulse to be registered by the system, it needed to cross threshold value, l_1, which was nominally set to 0.20. The first two pulses were each classified as "short" or "long", where long was indicated by holding the pulse for greater than 0.5 seconds, and different combinations of these pulses formed a 2-bit code. The different codes represented up, down, left, and right, and two "short" pulses, for example, selected the up command. After the first two pulses were entered and recognized as a movement code, the third pulse could be used to command the velocity of the cursor. The cursor velocity was defined by Equation 4.1,

$$v = v_c + (v_m - v_c)\left[\frac{(\bar{x} - l_2)}{(1 - l_2)}\right] \tag{4.1}$$

where v_c was the minimum velocity (0.05 units/second), v_m was the maximum velocity (0.50 units/second), l_2 was 0.30, and $\bar{x}$ was the filtered, averaged sEMG input. The interface and this control strategy was first presented in O'Meara et al. (2019), and further pilot studies were conducted to determine a maximum time of 60 seconds for subjects to complete a Trial.

Subjects were randomly assigned to one of four groups (12 subjects per group), and each was exposed to separate training methodology:

1. The Control group trained solely through repetition.

2. The Concurrent Feedback group received visual feedback during the task which indicated when the sEMG signal exceeded the threshold value, l_1, for selecting commands.

3. The Terminal Feedback group received visual feedback during the task after they successfully entered a command.

4. The Adaptive Threshold group was the same as the Control group, but the threshold value, l_1, varied Trial-to-Trial and was randomly selected for each Trial ($l_1 = 0.10, 0.15, 0.20, 0.25, 0.30$).

Despite the presence of visual feedback, the Concurrent and Terminal feedback groups viewed very similar interfaces to that of the Control Group, see Figure 4.1. The Concurrent Feedback group's feedback appeared on the cursor itself (see Figure 4.1b), while the Terminal Feedback group saw their visual feedback appear on the edge of the screen in the direction they commanded (see Figure 4.1c). An illustration of a potential signal input over time, along with an example of the feedback that would be presented for the two augmented feedback groups, is available in Figure 4.2.

Subjects completed 3 Code Accuracy Tests and 16 Blocks of testing during the experiment, see Figure 4.3. The Code Accuracy Test was designed to evaluate subject's ability to enter commands, and each Test consisted of 20 randomly ordered commands (5 of each of the 4 commands). The 16 Blocks were split into 12 Training Blocks, during which time the

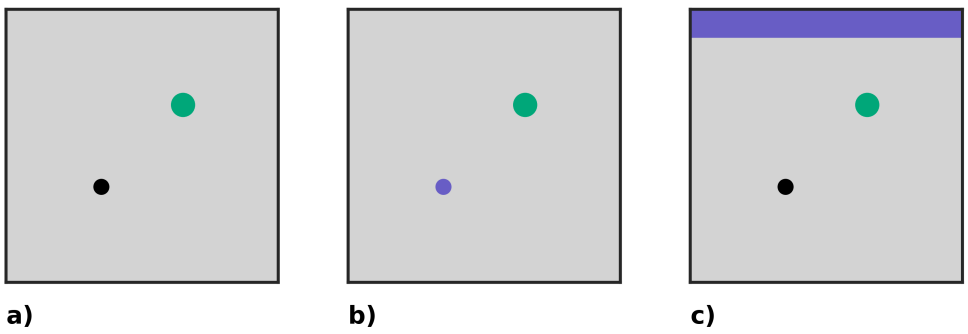


Figure 4.1: Cursor interfaces for a) the Control and Adaptive Threshold, b) Concurrent Feedback and c) Terminal Feedback groups (not to scale). The feedback displayed in c) indicates the "up" command. Concurrent Feedback and Terminal Feedback groups see their respective displays when the feedback is activated, otherwise like a). The target is shown in green and the cursor is the other, smaller circle.

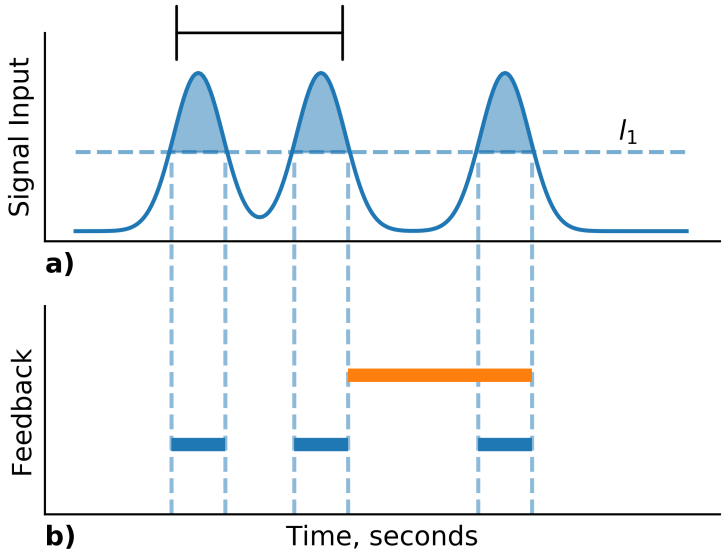


Figure 4.2: Illustrative signal input over time. a) The first two inputs select the command; the black bar represents the command time. The third input causes motion. The shaded areas indicate when the signal crosses the minimum threshold, l_1. b) Visual feedback is presented according to the bars (blue = Concurrent Feedback, orange = Terminal Feedback).

separate training methodologies were applied, and 4 Evaluation Blocks, during which time all subjects were exposed to same interface as the Control group. Each Block consisted of 10 cursor-to-target trials, after which subjects completed the workload and trust surveys and had a 30 second break.

4.3 Analysis and Hypotheses

This experiment investigated the effect of training methodology on performance, workload, and trust. The primary performance metric was the percentage of successful trials in a Block, and secondary metrics were the average trial time of successful trials and the throughput. While a majority of the our metrics are analyzed by Block, trial time was analyzed by Session as the randomization of the 40 target positions occurred over 4 Blocks. Trust and workload were each measured using surveys. Trust was evaluated using Jian et al.'s twelve statement questionnaire which measures trust between people and automated systems (Jian et al., 2000). Perceived workload was measured using Modified Bedford scale (Roscoe and Ellis, 1990). These surveys were administered after each Block. We also analyzed the results of the Command Accuracy Test, which was completed prior to training, after training, and

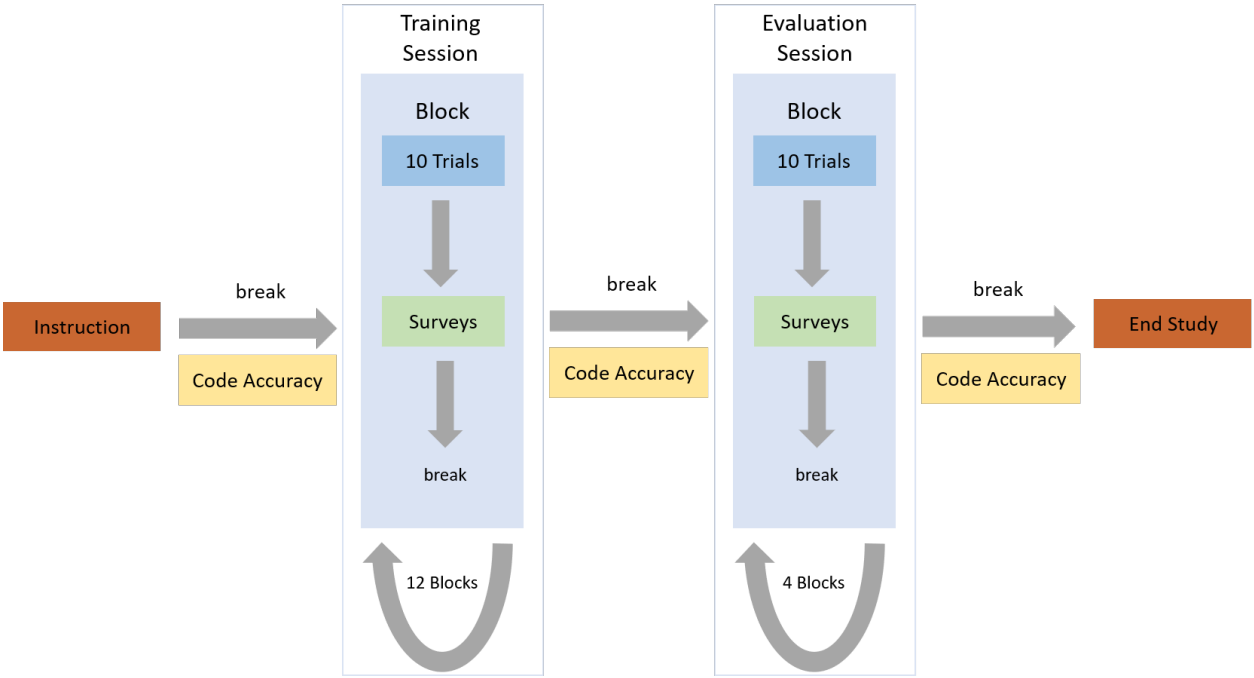


Figure 4.3: Experimental design flowchart.

after the retention phase at the end of the experiment. During the Command Accuracy Test, subjects were asked to input each of the four commands 5 times, and the percentage of successful inputs was used as a metric of performance.

4.3.1 Hypotheses

Based on our prior experience with augmented feedback and sEMG cursor control, we formed the following hypotheses.

1. During the training phase, the Concurrent Feedback group will have the highest performance followed by Terminal Feedback, then Control, and finally the Adaptive Threshold groups.

2. All groups will perform similarly in the retention phase.

3. The Concurrent Feedback and Terminal Feedback groups will have a high level of trust during training with some decrease during retention. Although, the trust level will still remain high during retention.

4. The Control group's trust will continually increase.

5. The Adaptive Threshold group will have lower trust during training, which will increase in retention.

6. The perceived workload will continually decrease during the training phase for all groups with the largest decreases for the Concurrent Feedback and Terminal Feedback groups.

7. There will be no significant difference in workload in the retention phase for all groups.

4.4 Results

We ran two-factor mixed models to investigate changes in performance, workload, and trust with one between-subjects factor, Group, and one within-subjects repeated measure, Block. When significant effects were observed, post hoc comparisons using the Tukey Honest Significance Difference (HSD) test were performed and considered significant at the $p < 0.05$ level, and the Satterthwaite method was used to calculate the degrees of freedom.

4.4.1 Performance Metrics

The percent success metric measured the percentage of successfully completed Trials within a Block; a Block contained 10 Trials. There were significant main factors of Group ($F(3, 44) = 8.18, p < 0.001$) and Block ($F(15, 660) = 31.80, p < 0.001$). There was also a significant interaction effect between Group and Block ($F(45, 660) = 3.90, p < 0.001$). Despite the presence of an interaction effect that resulted from subjects learning the task (as indicated by the Block factor), the main effect of Group could still be interpreted. A Tukey test showed that the subjects in the groups differed significantly, with subjects in the Concurrent Feedback group performing significantly better than those in the Control group ($p = 0.020$). The Tukey test also showed that subjects in the Adaptive Threshold group performed significantly worse than those in the Terminal Feedback and Concurrent Feedback groups ($p < 0.001, 0.01$, respectively).

The interaction effect resulted from different learning rates between the groups (see Figure 4.4), where subjects learned in the following order (fastest to slowest): Concurrent Feedback, Terminal, Control, and Adaptive Threshold. Compared to the Control group, the Concurrent Feedback group significantly outperformed them for the first 6 Blocks. Unlike the

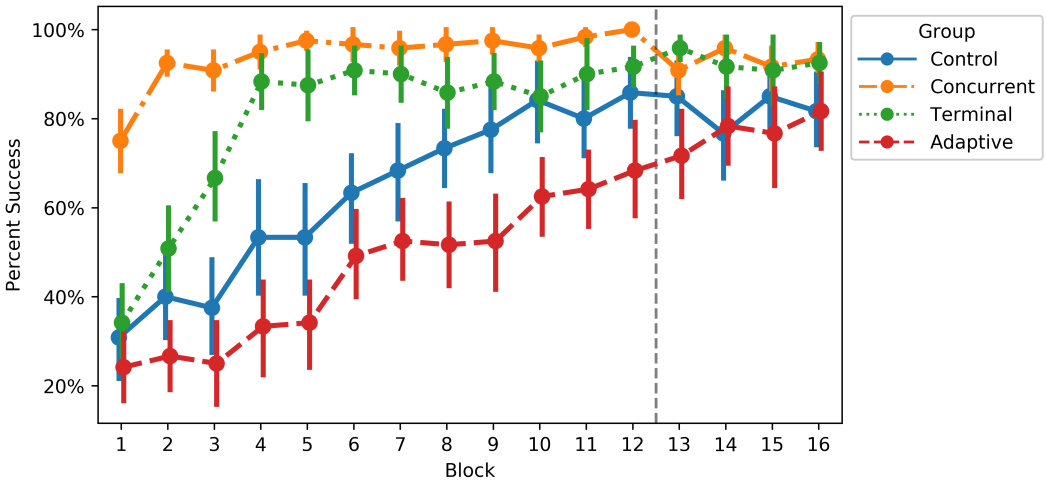


Figure 4.4: Percent success by Block across groups. The vertical dashed line represents the transition from the training phase to the retention one. Error bars shown are the standard error of the mean.

Concurrent Feedback group, the Terminal Feedback and Adaptive Threshold groups started with the same initial performance as the Control group. The Terminal Feedback group learned more quickly than the Control group, however, and significantly outperformed the Control group for Blocks 4 and 5. Compared to the Control group, all groups performed at statistically similar level after Block 6. Investigating the immediate retention effects when the group-specific treatments are removed in Block 13, the mixed model showed no change in performance for any of the groups ($p > 0.99$ for all groups). As such, the percentage of successfully completed Trials did not show any effect from the guidance hypothesis (i.e. the subjects did not rely on the feedback to complete the task and removing the feedback did not result in decreased performance).

The throughput was calculated for the retention phase and averaged across Blocks 13 through 16 (i.e. Session 4). Throughput is generally used to assess an input device, which should be measured when the subjects can complete the task. Since there were no significant differences in the retention phase for percent complete, it was logical to only calculate throughput at this time. There was no significant difference in throughput between the Groups ($F(3, 44) = 1.62, p < 0.20$). The mean throughput for all subjects was found to be 0.56 ± 0.02 bits/s ($\mu \pm \sigma$).

The randomization of the 40 target positions occurred over 4 Blocks, thus it seemed

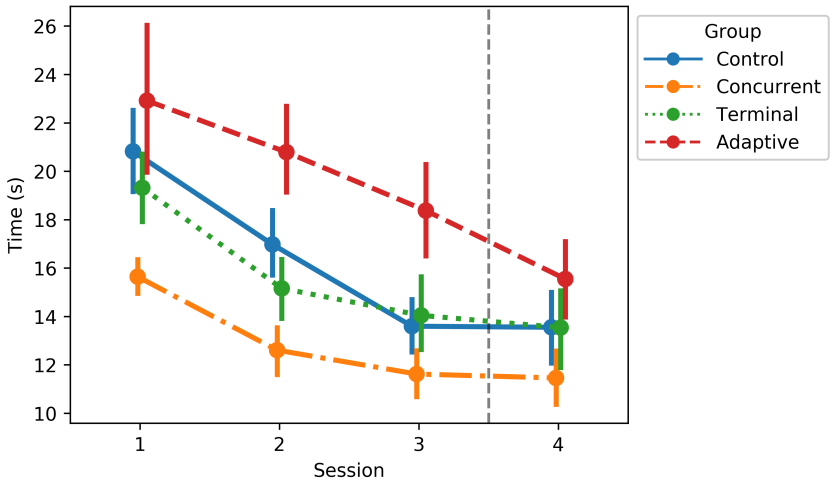


Figure 4.5: Average Trial time by Session (sets of 4 Blocks) across groups. The vertical dashed line represents the transition between the training and evaluation phases.

appropriate to average Trial time over a Session (i.e. set of 4 Blocks). Trial time was only defined for successfully completed Trials, and the Satterthwaite method was used to calculate the adjusted degrees of freedom using the lmerTest package in R (Bates et al., 2015). The results are displayed in Figure 4.5. There were significant main factors of Group ($F(3, 43.97) = 4.39, p < 0.01$) and Session ($F(3, 131.07) = 24.91, p < 0.001$). The interaction effect between Group and Session was not significant ($F(9, 131.07) = 0.78, p = 0.63$). A Tukey test showed that the Concurrent Feedback group performed significantly better than those in the Adaptive Threshold group ($p = 0.004$), which was the only significant difference between groups. No groups significantly outperformed the Control group. Analysis of the Session factor showed increased performance ($p < 0.05$) until the last two Sessions, which were not statistically different ($p = 0.65$). These results further supported that the guidance hypothesis did not occur.

4.4.2 Trust and Perceived Workload

The Modified Bedford Workload metric is a subjective measurement of perceived workload that ranges from 1-10, where 1 indicates low workload and 10 indicates high workload. There was a significant main factor of Block ($F(15, 660) = 18.29, p < 0.001$), but Group was not found to be significant ($F(3, 44) = 2.16, p = 0.106$). There was also a significant interaction effect between Group and Block ($F(45, 660) = 1.82, p = 0.001$) (see Figure 4.6).

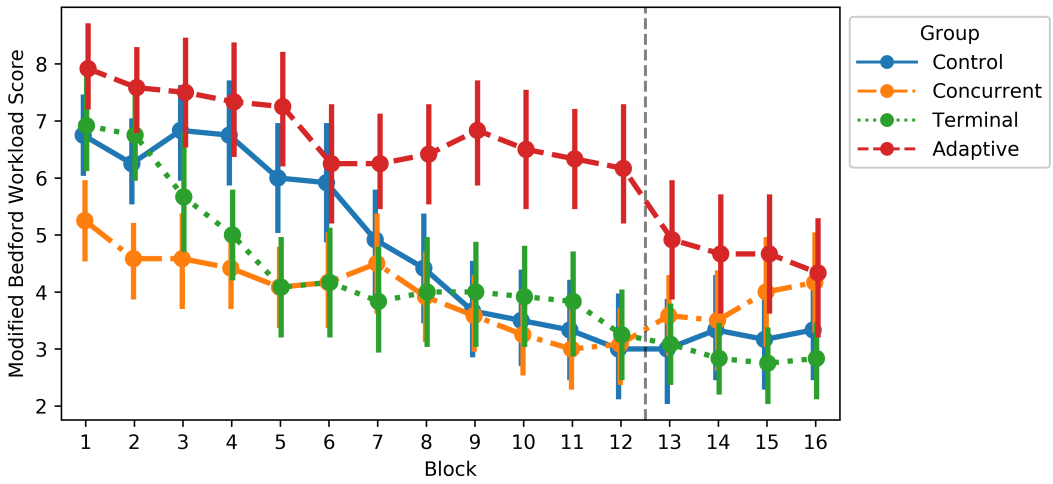


Figure 4.6: Modified Bedford Workload Score by Block across groups. The vertical dashed line represents the transition from the training phase to the retention one. Error bars shown are the standard error of the mean.

The interaction effect resulted from subjects reporting lower workload as they learn the task at different rates, as indicated by the Block factor. In further investigation of the interaction, we observed that the Adaptive Threshold group reported a significantly higher workload than the Concurrent Feedback group for Blocks 9, 10, and 11. This perception of high workload may possibly have resulted from the significantly worse performance of the Adaptive Threshold group. None of the groups showed a significant difference in workload compared to the Control group and all four groups reported statistically similar workloads in the retention phase.

Intragroup changes in workload were also of interest. The Concurrent Feedback group showed no statistically significant changes in performance between Blocks, though they did demonstrate a nonsignificant, increasing workload trend in the retention phase of the experiment. The Terminal Feedback group had statistically higher initial workload for Blocks 1, 2, and 3, but the remainder of the Blocks had a statistically similar level of workload. The Control group's workload was significantly higher for Blocks 1-6, possibly due to their slower learning rate, but leveled off for the remainder of the Trials. Finally, the Adaptive Threshold group reported the highest workload in Blocks 1-5, but also saw the largest improvement transitioning into the retention phase where their l_1 threshold stabilized to the same fixed value as the other groups.

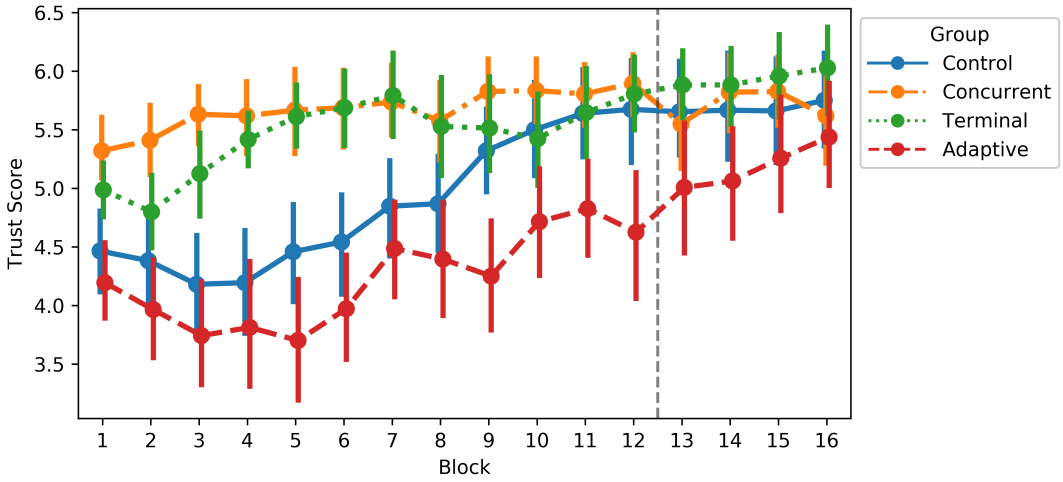


Figure 4.7: Trust Score by Block across groups. The vertical dashed line represents the transition from the training phase to the retention one. Error bars shown are the standard error of the mean.

Trust was measured using Jian's twelve question trust survey (Jian et al., 2000). Each question was rated on a 7-point Likert scale, the five reverse coded questions were reversed, and the results were averaged to create a single trust score (see Figure 4.7). There was a significant main factor of Block ($F(15, 660) = 13.05, p < 0.001$), but Group was not found to be significant ($F(3, 44) = 2.59, p = 0.065$). There was also significant interaction effect between Group and Block ($F(45, 660) = 2.00, p < 0.001$). The significant main effect of Block showed a gradual increase in trust throughout the duration of the study. After investigating the interaction effect, we saw that no group reported a significantly different trust level than the control group on any Block, but that the Adaptive Threshold group recorded a significantly lower trust than the Concurrent Feedback group on Blocks 3-6 and 9. Similar to workload, the primary interaction effects appeared driven by intragroup differences. The Concurrent Feedback group showed no statistically significant changes through the study, the Terminal Feedback and Control groups displayed significant increases in trust in Blocks 1-6, and the Adaptive Threshold group reported significantly higher trust in the retention Session than during early Blocks.

4.4.3 Command Accuracy

In contrast to previous results, the results in this section are not reported by Block or Session. The Command Accuracy Test occurred three times (before Block 1, after Block 12,

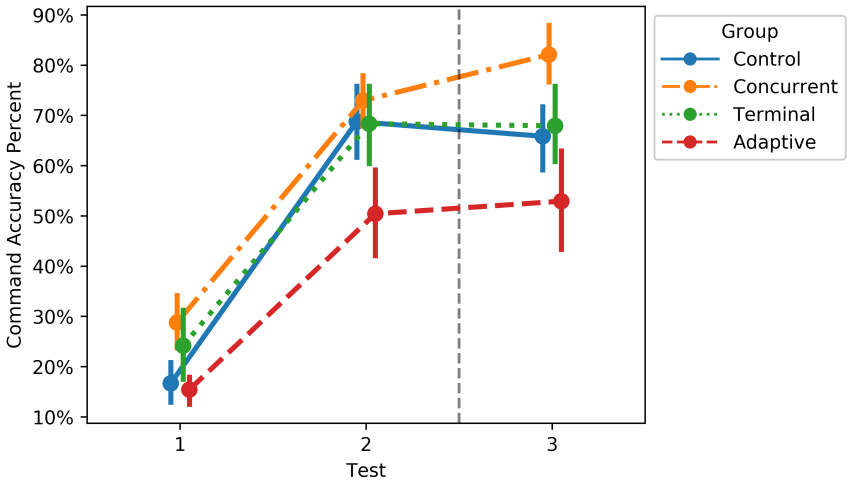


Figure 4.8: Command Accuracy results by Test across groups. Test 1, 2, and 3 occurred prior to the training phase, after the training phase, and after the retention phase, respectively.

and after Block 16), and the average was calculated for each Test.

In each Command Accuracy Test, subjects responded to prompts to perform commands. The command accuracy percent indicates the percentage of the 20 prompts in each Test that subjects correctly performed. Results were averaged by group (see Figure 4.8). There was a significant main factor of Test ($F(2, 88) = 108.48, p < 0.001$), but Group was not significant ($F(3, 44) = 2.63, p = 0.06$). The interaction effect between Group and Test was not significant ($F(6, 88) = 0.82, p = 0.55$). Investigation into the Test variable showed that subjects performed significantly better between Test 1 and 2, and between Test 1 and 3, but not between Test 2 and 3. These results demonstrated that there was a significant improvement in the percent of commands accurately entered after the training portion of the experiment was finished.

4.5 Discussion

The results elucidated a relationship between performance, workload, and trust that was influenced by the training methodology. The Concurrent Feedback group started with the highest performance, lowest workload, and highest level of trust. By the 5th Block, the Terminal Feedback group overlapped with the Concurrent Feedback group in terms of percent success, workload, and trust. These two groups then tracked each other for the remainder of the study. The Concurrent Feedback and Terminal Feedback groups did not have significantly

different average trial times throughout the study. The initial learning of these two groups differed with the Concurrent Feedback group demonstrating high performance with smaller, incremental gains and the Terminal Feedback group with larger Block-to-Block improvements. The Control and Adaptive Threshold groups did not appear to reach a performance plateau for the percent success and steadily improved in subsequent Blocks. Interestingly, the trust score also continually increased for both the Control and Adaptive Threshold groups. Although the changing cursor dynamics in the Adaptive Threshold group does not seem to adversely affected trust compared to the Control group, the Adaptive Threshold group did report a significantly higher perceived workload for Blocks 9-12—the last four Blocks in the training phase. Once the cursor dynamics stabilized, the perceived workload in the Adaptive Threshold group was not statistically different from the other groups. These results suggested that visual feedback led to earlier performance gains with improvements in trust and workload. The adaptive training methodology surprisingly did not adversely affect trust, but the cost was reflected in the performance and perceived workload. Overall, all training methodologies achieved statistically similar results during the retention phase.

Measuring performance by percent success assessed the ability to complete the cursor-to-target task in the allotted 60 s. The average trial time provided more insight into the group's performance. The trends in the average trial time may be explained by the Command Accuracy Test results. Higher percentages of command accuracy tracked with lower average trial times. This relationship was expected as less time was spent attempting to input the command when the subject was able to accurately convey the inputs. Although there were no statistically significant differences between groups in the retention phase for these metrics, it was notable that the Adaptive Threshold group consistently performed worse than the Concurrent Feedback group.

The Adaptive Threshold group did not outperform any group in any metric during the retention phase. The results from this group were interesting for two reasons. Firstly, an adaptive training methodology did not appear to cause adverse effects compared to the Control group. Unreliable automation behavior can lead to poor human-automation interaction (Dzindolet et al., 2003), but was not the case in this study. Secondly, the benefits of increased adaption induced by uncertainty did not manifest (e.g. generalization for novel

tasks). The adaption training methodology may be better tested with a different task with the same underlying structure instead of returning to a stable condition. For example, Braun et al. showed that subjects trained with an adaptive training strategy were able to quickly generalize to novel tasks with similar underlying structure. It is also possible that the cursor dynamics unpredictability was not sufficiently large compared to the inherent sEMG control noise.

Our results aligned with previously published research. The Concurrent Feedback and Terminal Feedback groups follow the findings of Basmajian that augmented feedback can improve performance. We also observed effects that may be explained by Hoff and Bashir's three layer model. At times, there were significantly different levels of trust between the groups, which indicated that the training methodologies altered situational trust. Subjects' trust levels also increased, which supported the notion of learned trust. While developing a brain computer interface (BCI) was not the study objective, the throughput values during the retention phase for all groups fell within published results for sEMG cursor control systems. Our previous single-site sEMG cursor control system with 2 DOF (counterclockwise rotation and forward) reported 2.24 bits/s and 0.23 bits/s for control methodologies that used different levels of automation (O'Meara et al., 2019). Multi-site systems have achieved 0.4 bits/s (Cler and Stepp, 2015), 0.84 bits/s (Williams and Kirsch, 2008), and 1.3 bits/s (Williams and Kirsch, 2016). The sEMG cursor control system used in this study had a throughput of 0.56 bits/s and may be of additional interest to the BCI community. However, the purpose of the sEMG cursor control system in this study was to provide a testbed that lent itself to motor learning adaptation and was sufficiently challenging to probe the relationship between performance, workload, and trust.

The study results largely supported our hypotheses. The percent success performance during the training phase followed the order of Concurrent Feedback, Terminal Feedback, Control, and Adaptive Threshold, but not for all times during that phase. All groups performed similarly in percent success during the retention phase. The Trial completion time only supported significant differences between the Concurrent Feedback and Adaptive Threshold groups. The subjects' trust followed our expectations with Concurrent Feedback and Terminal Feedback having the highest levels, and the Control group continually

increased. The Adaptive Threshold group had lower trust during training, and the trust increased to the level of the other groups during retention. The workload results also supported our hypotheses that Concurrent Feedback and Terminal Feedback groups would have the largest decrease in workload, and that all groups would have similar workload during retention. Interestingly, the Concurrent Feedback and Terminal Feedback groups converged across performance, workload, and trust by Block 5. This study provided insights on the relationship between performance, workload, and trust for various training methodologies, and highlighted the advantage of certain methodologies during the training phase.

In the greater context of this dissertation, this experiment illustrated that concurrent bandwidth feedback could effectively improve performance and decrease the required learning time of a discontinuous task. The extremely large and immediate gains in performance seen in Figure 4.4 mirror those that we observed in the SAFER task, see Figure 1.5a. Subjects in the Concurrent Feedback group were again able to immediately outperform those in the Control and sustained this performance when the feedback was removed. While slightly lower workload and higher trust was observed initially, this effect was not significant and faded with continued use of the system. The Terminal Feedback group's also showed rapid performance improvement over the Control group, though this took slightly longer than with the Concurrent Feedback group and showed a larger variance in ability. This suggests that subjects in the Concurrent Feedback group were able to use the feedback to more accurately and reliably control their EMG signal. Optimal performance was more quickly achieved by presenting the feedback concurrently with the subject's action instead of delaying it.

Chapter 5

Feedback for Training Flight Tasks

The literature has shown that a variety of tasks can benefit from concurrent feedback, and that more functionally complex tasks tend to see a larger performance improvement. This result stems from results of many different experiments and researchers, however, and it is uncommon for experiments to explore the interaction effect between feedback and functional task complexity directly. We designed this experiment to directly address this gap in the literature by evaluating the effects of feedback on an aircraft flight task with one, two, or three degrees of freedom. Portions of this chapter were originally published in the conference proceedings for the Human Factors and Ergonomics Society 2019 (Karasinski and Robinson, 2019a).

5.1 Introduction

Augmented feedback, information that relates an individual's performance to a desired performance, has been found to generally enhance motor learning in a wide variety of manual motor control tasks (Salmoni et al., 1984). Many feedback modalities and implementations have been investigated in the literature, some of which have been found to be more effective than others. One of the key aspects to successfully implementing feedback is knowing when to provide feedback to the participant. Feedback can be provided concurrently, in real-time as the task is executed; or terminally, after the task is completed. Visual concurrent feedback, for example, has been shown to greatly enhance motor learning as task complexity increases, while terminal feedback is better suited for tasks with low functional complexity (Sigrist

et al., 2013). As Sigrist et al. note in their review of augmented visual, auditory, haptic, and multimodal feedback, however, "[u]p to now, mostly low-dimensional, simple, and rather artificial labor tasks have been investigated even though, in real life, most motor tasks are multidimensional and complex."

Aircraft and spacecraft flight-control tasks are complex, multidimensional challenges for human manual control, and present both demanding learning requirements and high cognitive demands. In the pursuit of improving pilot performance during training, several researchers have investigated the effects of adding visual and/or audio feedback to flight displays. Bronkhorst et al. explored the use of auditory displays in a 3D flight task. By adding auditory displays to the existing simulation, they were able to reduce search time in an aircraft location and tracking task. Similarly, Tannen et al. showed that U.S. Air Force pilots could better maintain flight parameters and report reduced workload with the use of multisensory cueing. Of the many studied feedback strategies, concurrent bandwidth feedback is among the most promising for complex tasks. Concurrent bandwidth feedback is presented in real-time, during task execution, but only when some variable deviates outside of a defined bandwidth of acceptable values. Researchers have used concurrent bandwidth feedback to study participants ability to learn to drive a vehicle, having found it to be effective at improving lane keeping (de Groot et al., 2011).

At UC Davis, our recent experiments with concurrent bandwidth feedback in complex manual tasks have resulted in large improvements in human performance with an added benefit of reduced workload. Our experiment with simulated spacecraft-piloting investigated the effects of concurrent bandwidth feedback on a complex, four-degree-of-freedom manually controlled on-orbit inspection task (Karasinski et al., 2017). We found that simple visual feedback on the controlled degrees of freedom improved initial and fully trained performance while reducing inferred and self-reported workload. In fact, participants in the feedback group performed as well in their first trial as participants in the control group did after two hours of training. Our recent work also includes an investigation into the effectiveness of concurrent bandwidth feedback for learning a 3D joystick-controlled tracking task in augmented reality (Karasinski and Robinson, 2019b). This experiment investigated whether concurrent bandwidth feedback could teach participants to interpret 3D depth cues. Our re-

sults suggested that participants who were exposed to visual concurrent bandwidth feedback early on sustained improved performance through the duration of the experiment compared to participants that began in a baseline condition without feedback. Through these previous experiments, we have shown that concurrent bandwidth feedback can be effective at improving human performance for complex manual tasks. In the research reported here, we build upon our previous work and that in the literature by investigating an operationally relevant, joystick-commanded flight-control task with the objective of determining the effect of task complexity on the influence of concurrent bandwidth feedback.

In our current experiment, subjects controlled a simulated aircraft with realistic flight dynamics through a series of tasks of increasing functional complexity. By allowing for multiple levels of functional complexity, we can investigate what level of complexity is required to observe changes in human performance and cognitive workload. This experiment also investigated the effects of removing concurrent feedback after training to evaluate changes in performance and workload during participants' immediate retention. The retention portion of the experiment was performed to investigate the guidance hypothesis, which states that consistent feedback during the acquisition phase of learning leads to a dependency on the feedback (Salmoni et al., 1984).

5.2 Method

5.2.1 Task

Control Modes

Participants were tasked with flying a simulated Boeing 747 aircraft in three control modes. In order of increasing degrees of freedom and functional complexity, these three control modes were:

P Pitch only (low complexity)

PR Pitch and Roll (moderate complexity)

PRA Pitch, Roll and Altitude (significant complexity)

Depending on the control mode, participants were required to use a joystick to null disturbances in pitch, roll, and/or to maintain a constant altitude. Participants were informed

that all three tasks were equally important, and to try not to neglect or prioritize individual tasks.

Each participant completed a total of 36 trials; 12 in each of the three control modes. Each trial had a duration of 82 seconds, and participants self-initiated the trial by activating a trigger on the joystick. The trial order was designed such that each participant flew the simulator in increasing order of task complexity, with the sequence of P, PR, PRA, P, PR, ..., PRA. This design was chosen to provide exposure to each control mode as quickly as possible, such that we could capture the early learning phases of each mode.

Forcing Functions

Both the pitch and roll axes were affected by disturbance signals, resulting in a disturbance-rejection task. The disturbance signal took the form of a quasi-random sum of sines from Zaal et al. (2009). The same forcing function was used for disturbing both pitch and roll, though the roll disturbance function was temporally shifted by 85 seconds to minimize the correlation between resulting pitch and roll disturbances. Aircraft altitude varied as a result of pitch variation. The same disturbing function was used for every trial, though participants were naïve to this.

Secondary Task

To estimate participant workload more objectively than with a questionnaire, we added a secondary task that assesses the subject's cognitive margin available for attending to non-primary task execution. The secondary task was displayed to the right of the flight-guidance display (see Figure 5.1) and consisted of a teal colored indicator which changed color to blue or green at pseudorandom times. Ten 8-second secondary-task windows were displayed during each 82-second long trial. The indicator would randomly change during this interval, providing participants up to 5 seconds to respond. The pseudorandom times and colors for the secondary task were identical for each participant. This secondary task has been validated in previous studies, which have shown it to correlate well with participants' subjective workload estimates (Hainley et al., 2013).

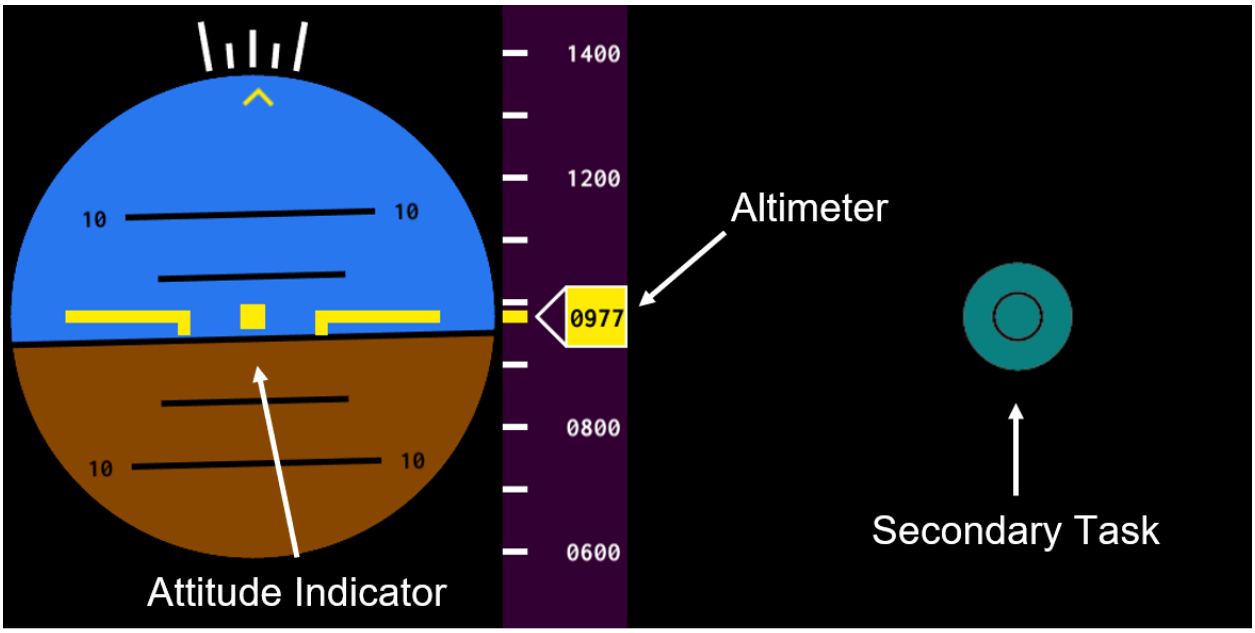


Figure 5.1: The user interface consisting of the attitude indicator, altimeter, and secondary task.

5.2.2 Simulator

Participants were seated at a fixed-base simulator for the duration of the experiment, see Figure 5.2. The simulator consisted of a 30-inch monitor and a single joystick. The user interface was centered in the display and presented to the user at 1000x500 pixels. The user interface consists of a traditional aircraft attitude indicator on the left, an altimeter in the middle, and the secondary task on the right. The Boeing 747 was modeled using data available in NASA CR-2144 (Heffley and Jewell, 1972). Flight Condition 2 was chosen for this experiment, and represents slow, sea-level flight, with the landing gear retracted and with flaps extended to 20 degrees. Longitudinal and lateral body axis derivatives were converted to the stability axis and implemented in a state-space model (Stevens et al., 2015). The values from NASA CR-2144 for this flight mode were plugged into the longitudinal and lateral dynamics matrices presented in Appendix C. Participants used the elevators and ailerons to control pitch and roll, respectively. Altitude was controlled by the subject with time-integrated pitch commands, as in a real aircraft.

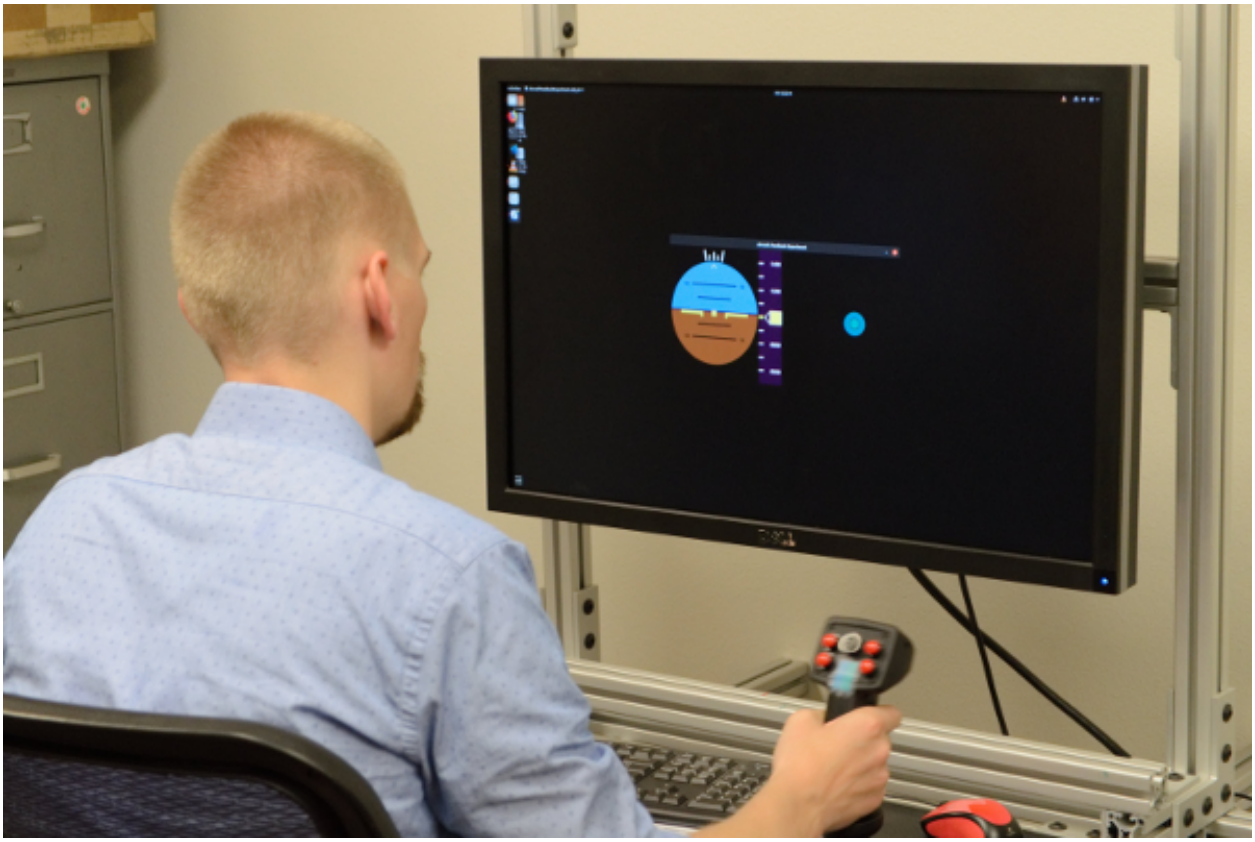


Figure 5.2: A participant seated in front of the simulator display, controlling the flight task with the joystick.

5.2.3 Experimental Design

Participants were evenly split into two groups: a control group and a feedback group. Participants in the feedback group received concurrent bandwidth feedback on the first 27 trials (9 in each control mode) and then flew the last 9 trials (3 in each control mode) without feedback to test their immediate retention. Participants in the control group never received concurrent bandwidth feedback, nor was it mentioned to them during the study.

For participants experiencing feedback, feedback was presented based on the flight control mode of their current trial. This feedback was implemented by changing the color of an indicator's elements from yellow to red when it deviated from outside of its allowed bandwidth and returning the indicator to yellow when it returned within the bandwidth. Acceptable bandwidths of 3 degrees for pitch and roll, and 30 feet for altitude were chosen based on preliminary testing. The choice of these bandwidths was based on preliminary pilot studies,

and is further explored in Chapter 6. Pitch feedback occurred on the center dot. Roll feedback was shown on the wings and the roll indicator at the top of the attitude indicator. Altitude feedback was enabled on the background color of the altimeter. In Figure 5.1, all three parameters are shown inside the acceptable bandwidth, and are therefore displayed in yellow.

Participants began the experiment by signing a consent form, then filled out a survey with demographic questions, then had a brief, pre-recorded training session. After this training, participants immediately began the experiment and progressed through the 36 trials at their own pace. Participants noted their workload on a piece of paper after each trial. Subjects in the feedback group were paused after the 27th trial, at which time the proctor explained that the feedback would no longer appear, but that they "should continue to perform the task to the best of [their] ability." Subjects in the feedback group also filled out a questionnaire after the end of the experiment trials which asked them about their experience with the feedback.

Independent variables

The three independent variables in this experiment were Group, Mode, and Trial. Group, a between subjects factor, described if subjects received feedback — Control or Feedback. Mode, a within subjects factor, was the three different control modes — P, PR, or PRA. Trial, also a within subjects factor, was the trial that subjects repeated 12 times in each mode.

Dependent measures

The root-mean-square error (RMSE) of pitch was calculated for every trial. This allowed for a consistent measurement across every trial as the same pitch disturbance was present regardless of the control mode. The roll RMSE was calculated for each PR and PRA trial, and the altitude RMSE was calculated for each PRA trial. The RMSE values provide an objective measurement of performance. The secondary task was activated ten times per trial. Participants had five seconds to correctly respond to the secondary task once it changed color. We recorded the rates for correct and incorrect responses, as well as lack of response. We used the average correct response time as objective indication of workload. The Modified Bedford Workload Scale is a ten-point subjective workload measurement tool (Roscoe and Ellis, 1990). Participants were asked to follow the scale and record their workload after each

trial, allowing us to observe changes in workload during training.

5.2.4 Hypotheses

We had three major hypotheses for this experiment:

H1. Participants in the feedback group will immediately outperform those in the control group. We expect this effect to be most pronounced for the most complex mode and to see little to no improvement for simple trials.

H2. Participants in the feedback group will have lower workload than participants in the control at the end of the experiment. We expect this effect to be most pronounced for the most complex mode and to see little to no improvement for simple trials.

H3. Participants in the feedback group will not suffer from the guidance hypothesis and will retain their performance and workload levels when the feedback is removed in the immediate retention trials.

These hypotheses were established from the literature, our previous experiments with feedback, and early pilot studies with this simulation framework.

5.3 Results

5.3.1 Participants

Participants in the experiment were 30 engineering students from the University of California, Davis (23 men, M = 23.0 years, SD = 4.4; 7 women, M = 22.6 years, SD = 3.0). All participants had normal or corrected-to-normal vision and full motor control of their upper bodies. Eighty percent of participants had previously used a joystick, 43% had spent time in flight simulators, and 30% had prior flight experience. Both gender and participants with flight experience were counterbalanced between the two groups. Written informed consent was obtained prior to testing in accordance with the University of California, Davis Institutional Review Board (Project #1399789-1).

5.3.2 Analysis

Mixed models were used to calculate the significance of factors in our analysis due to the presence of performance outliers which were removed from the analysis. The Satterthwaite

method was used to calculate the adjusted degrees of freedom using the lmerTest package in R (Bates et al., 2015). When significant effects were observed, post hoc comparisons using the Tukey Honest Significance Difference (HSD) test were performed and considered significant at the $p < .05$ level, and the Satterthwaite method was again used to calculate the degrees of freedom. Only 7 of the 1080 total trials (30 subjects with 36 trials per subject) were removed. These trials were extreme performance outliers, and including these trials does not change the primary results of the study. A three-factor (Group, Mode, and Trial) mixed model with two repeated measures (Mode and Trial) was run on the pitch root-mean-square error. There were significant main factors of group ($F(1, 27.97) = 6.3, p = 0.018$), mode ($F(2, 53.47) = 29.7, p < .001$), and trial ($F(11, 300.29) = 48.4, p < .001$). There were also significant interaction effects between group and trial ($F(11, 300.29) = 2.5, p < 0.01$) and between mode and trial ($F(22, 601.58) = 2.8, p < .001$). Despite the presence of interaction effects that result from participants learning the task (as indicated by the trial factor), the main effects can still be interpreted, see Figure 5.3. A Tukey test showed that the participants in the groups differed significantly, with the participants in the feedback group outperforming those in the control group ($M = 2.35, 3.05$, respectively, $SE = 0.20$). An additional Tukey test showed that the participants' performance in the modes differed significantly, and participants performed best in P mode, followed by the PR mode, and finally the PRA mode ($M = 2.30, 2.67, 3.14$ respectively, $SE = 0.15$).

This same analysis was completed on the roll root-mean-square error, with similar results. There were significant main factors of group ($F(1, 28.00) = 8.8, p < 0.01$), mode ($F(1, 27.93) = 6.8, p = 0.015$), and trial ($F(11, 308.22) = 19.6, p < .001$). There was also a significant interaction effect between mode and trial ($F(11, 308.06) = 4.6, p < .001$), see Figure 5.4. Tukey tests showed that the participants' performance between the groups and the modes each differed significantly, with the participants in the feedback group again outperforming those in the control group ($M = 1.96, 2.43$, respectively, $SE = 0.11$), and performance was best in the PR mode followed by the PRA mode ($M = 2.15, 2.24$, respectively, $SE = 0.08$). A two-factor (Group and Trial) mixed model with one repeated measure (Trial) was run on the altitude root-mean-square error. There were significant main factors of group ($F(1, 27.54) = 5.2, p = 0.030$) and trial ($F(11, 301.57) = 11.4, p < .001$). Tukey

tests showed that the participants' performance between the groups differed significantly, with the participants in the feedback group again outperforming those in the control group ($M = 28.3, 50.2$, respectively, $SE = 6.8$), and the trial effect showing learning throughout the experiment for both groups, see Figure 5.5. See Figure 5.6 for a plot of the root-mean-square error for each flight task in each mode. A three-factor (Group, Mode, and Trial) mixed model with two repeated measures (Mode and Trial) was run on the modified Bedford workload scores. There were significant main factors of mode ($F(2, 56) = 134.8, p < .001$), and trial ($F(11, 308) = 8.7, p < .001$), and a significant interaction effect between mode and trial ($F(22, 616) = 2.0, p < 0.01$). Tukey tests showed that the participants' workload between the modes differed significantly, with workload lowest in P, then PR, and finally PRA ($M = 3.80, 4.99, 6.16$, respectively, $SE = 0.28$), and the trial effect representing slightly reduced workload throughout the experiment.

5.4 Discussion

Our analysis showed that participants in the feedback group performed significantly better than the control group in every trial across every metric and flight control mode. For the first time that participants completed P, PR, and PRA trials, the feedback group performed 17.5%, 31.7%, and 37.4% better than the control group according to the pitch RMSE metric. Participants in the feedback not only immediately performed better, they also reached their peak performance much faster and had a final performance level which was significantly better than the control group, see Table 5.1. The largest performance improvement was seen in controlling altitude, where the feedback group had a final performance level which was 44.2% better than the control group, confirming H1. No group-related workload differences were

Mode	Pitch	Roll	Altitude
P	20.1%	-	-
PR	17.8%	21.1%	-
PRA	20.5%	26.1%	44.2%

Table 5.1: Performance improvement of the feedback group over the control group at the end of the experiment for each flight RMSE metric.

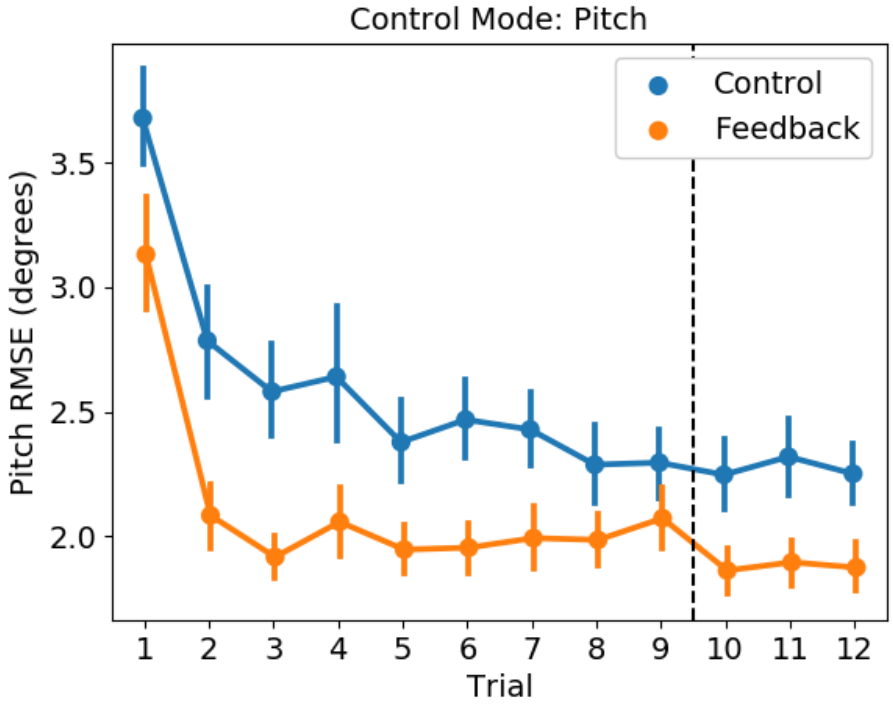


Figure 5.3: The mean Pitch RMSE for each trial for participants in the P control mode. Data points are the mean, and error bars are the standard error of the mean.

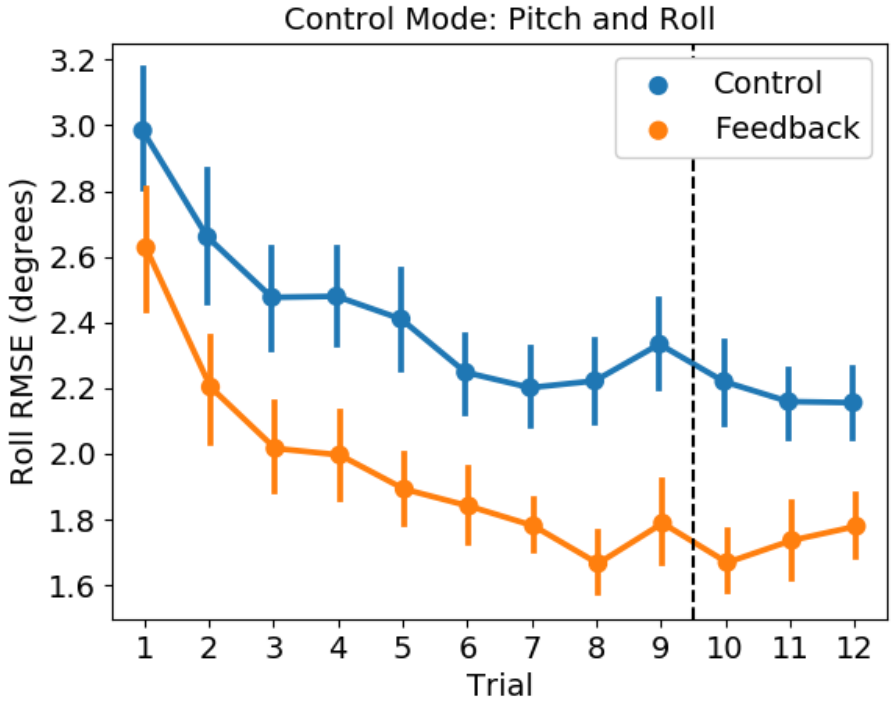


Figure 5.4: The mean Roll RMSE for each trial for participants in the PR control mode. Data points are the mean, and error bars are the standard error of the mean.

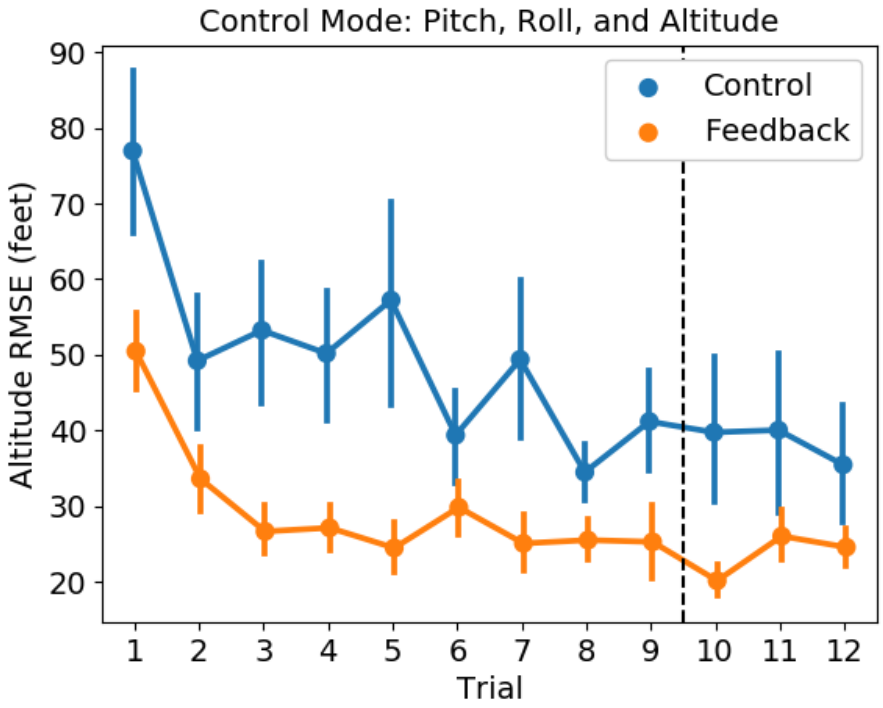


Figure 5.5: The mean Altitude RMSE for each trial for participants in the PRA control mode. Data points are the mean, and error bars are the standard error of the mean.

detected in either the Modified Bedford scores or in the secondary task reaction times. This suggests that, for tasks of this complexity, concurrent bandwidth feedback may not reduce workload (Karasinski and Robinson, 2019b). Thus, for our second hypothesis, H2—we find that concurrent bandwidth feedback does not reduce workload in our flight tasks, and that tasks with higher functional complexity may be required to observe these effects. Our third hypothesis was that subjects would primarily use the concurrent bandwidth feedback as a learning tool but that they would not become dependent on the feedback to the point that they required it to complete the task. Retention tests are commonly used in the augmented feedback literature to verify that participants are not dependent on the feedback techniques. Our analysis indicates no significant performance changes across any of our performance metrics when the feedback was removed, confirming H3. After the experimental trials, we asked participants in the feedback group to complete a survey designed to identify when the feedback was or was not useful and how participants thought their performance and workload changed in the retention phase of the experiment. One hundred percent of participants

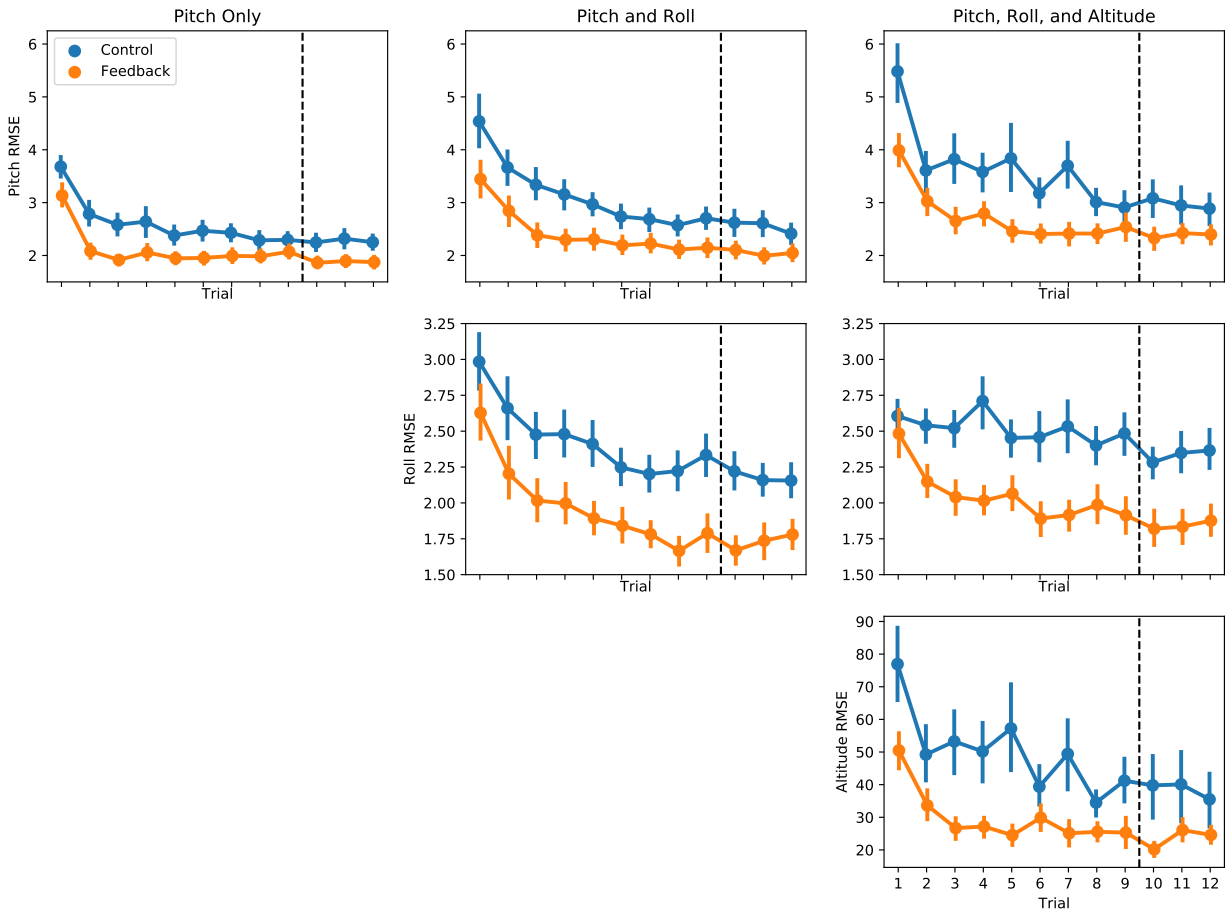


Figure 5.6: The root-mean-square error for each flight task in each mode. Data points are the mean, and error bars are the standard error of the mean.

thought that the feedback helped them perform the task, 80% reported that it helped them regain focus when their mind wandered, 73% reported that it helped them to learn a scan pattern, and 27% reported that the feedback was motivating.Survey results suggested that participants found the feedback especially useful in the roll and altitude tasks, which was reflected in the objective performance metrics.

5.5 Conclusions

Participants took part in a study investigating the effect of concurrent bandwidth feedback on flight performance and workload in three flight control modes of increasing complexity. The participants in the feedback group performed significantly better than those in the control group, generally performing better on their second trial than those in the control

group could by the end of the experiment. Subjective and objective workload metrics showed no change in participant workload between the groups. Survey questions identified that most participants found the feedback helpful in training them to establish a scan pattern, helping them to learn the task much quicker than those without feedback. Feedback was removed for immediate retention trials, and participants showed no changes in performance or workload, suggesting that participants did not suffer from the guidance hypothesis.

This experiment provided additional insight into many of the research questions posed in Section 1.3. Additionally, it confirms the result found generally in the literature and summarized in Figure 1.1, that concurrent visual feedback is more effective at improving performance as functional workload increases. By evaluating the effects of feedback on an aircraft flight task with multiple degrees of freedom, we were able to see 17.8% to 44.2% increases in performance, which increased as the degrees of freedom increased. The lack of performance degradation when the feedback is removed, and therefore the rejection of guidance hypothesis, further illustrates that the concurrent bandwidth feedback is truly acting as feedback, not additional guidance. This promotes the concept of the "instructor model" discussed in Section 1.2.3, where the feedback becomes increasingly rare as pilot performance increases with training.

Chapter 6

Feedback Bandwidth Study

As we mentioned in Section 1.2.3, several authors have noted that "[b]andwidth feedback has been shown to be effective; however, setting the error threshold is not trivial" (Timmermans et al., 2009; Ribeiro et al., 2011; Sigrist et al., 2013). While using an operational limit as the level of performance considered "acceptable" can make the choice of setting the error threshold simpler, this option is not always available. Additionally, mission designers often want to know what the optimum pilot performance is for mission planning, and instructors are interested in what levels of performance are acceptable during training. In order to provide insight into the effects of variable bandwidth and answer these questions, in this Chapter we treat the bandwidth as a variable in order to evaluate its effect on resultant subject performance and workload.

6.1 Introduction

Our previous work has shown that concurrent bandwidth feedback can be effective at improving performance, but our choice of bandwidth has thus far been ad-hoc in nature and primarily based on preliminary pilot studies. These pilot studies have shown that there may be a wide variety of acceptable bandwidths, but it is unclear if there is a single bandwidth that would lead to optimal performance or minimum learning time. It is clear, however, that extreme thresholds which require unrealistic performance will not improve performance and may instead lead to degraded performance when compared to a no feedback condition. Conversely, bandwidths that are too loose or not demanding enough will result in feedback

that is sparse and may result in performance that resembles a no feedback condition.

6.2 Method

We designed a follow-up experiment to the experiment presented in Chapter 5 which investigates the changes in performance that result from exposure to different bandwidth sizes. This experiment used the same simulator and flying task as the previous experiment with a few modifications. In this experiment the display, forcing functions, and secondary task were all identical to the previous experiment, however, subjects only completed trials in the pitch and roll mode.

6.2.1 Experimental Design

In this experiment, subjects were randomly assigned to one of four groups. A control group, which received no feedback and three feedback groups which received feedback during training. Subjects in the feedback groups received concurrent bandwidth feedback along both the pitch and roll axes when these states exceeded a bandwidth of 2, 3, or 4 degrees. As with the previous experiment, the feedback on each axis was activated independently when the individual state crossed over the threshold bandwidth. Unlike the previous experiment, which only had a training and immediate retention phases, this experiment has four phases spread across two experiment sessions. During the first session, subjects completed 16 training trials followed by 8 immediate retention trials during which time subjects in feedback conditions had their feedback removed. Subjects were dismissed after completing the immediate retention trials and asked to return approximately 24 hours later for the second experiment session. In the second session, subjects completed 8 retention trials without feedback, then completed 8 transfer task trials without feedback. The only change between the transfer task and the other trials were the system dynamics of the controlled vehicle and the magnitude of the disturbance force. In the transfer task, subjects were responsible for flying a Navion aircraft instead of a Boeing 747. The Navion was modeled using data available in NASA CR-96008 (Teper, 1969). Compared to the Boeing 747, the Navion is a much smaller and lighter aircraft and requires a significantly different control strategy to effectively pilot. The disturbance force consisted of the same functions as in the rest of the experiment but was scaled down such that maximum pitch deflection under an uncontrolled

scenario was approximately the same as with the Boeing 747 trials.

Independent variables

The two independent variables in this experiment were Group and Trial. Group, a between-subjects factor, described if and when subjects received feedback. Trial, a a within-subjects factor, was the trial that subjects repeated over the course of the experiment.

Dependent measures

The root-mean-square error (RMSE) of pitch and roll was calculated for every trial to provide an objective measurement of performance. As in the previous experiment, the secondary task was activated ten times per trial and participants had five seconds to correctly respond to the secondary task once it changed color. We used the average correct response time as the objective indicator of workload. Participants were asked to follow the Modified Bedford scale and record their workload after each trial, allowing us to observe changes in workload during training.

6.2.2 Hypotheses

We developed four hypotheses for this experiment:

H1. Participants in the feedback groups will immediately outperform those in the control group.

H2. Participants will report the same workload between groups, which will gradually decrease with training.

H3. Participants in the feedback group will not suffer from the guidance hypothesis and will retain their performance and workload levels when the feedback is removed in the immediate retention and 24 hour retention trials.

H4. Participants will continue to perform at relatively similar levels when they complete the transfer task.

6.3 Results

6.3.1 Participants

In December 2019, the novel coronavirus SARS-CoV-2 which results in a disease called COVID-19 began to spread and quickly resulted in a global pandemic. This infectious disease caused by severe acute respiratory syndrome coronavirus is highly contagious and is mainly spread during close contact. As a result of this pandemic, the University of California, Davis was closed and the Institutional Review Board recommended that we act to "limit transmission of the virus by delaying or otherwise modifying non-essential interactions," postponing subject testing (Mohapatra, 2020). We had anticipated testing approximately 10-15 subjects per group, for a total of 40-60 subjects. The remainder of the results discussed here are preliminary and based on the 19 subjects whose data was collected before a shelter-in-place was ordered.

Participants in the experiment were 19 engineering students from the University of California, Davis (17 men, 1 woman, 1 decline to state) with an average age of 21.3 years (SD = 2.2). All participants had normal or corrected-to-normal vision and full motor control of their upper bodies. Participants with flight experience were counterbalanced between the two groups. Of the 19 subjects, 18 returned the next day after an average of 24.93 hours (SD = 0.6). This study was exempted by the University of California, Davis Institutional Review Board (Project #1537932-1), and subjects were not compensated for their time.

6.3.2 Analysis

As in the previous study, linear mixed models were used to calculate the significance of factors in our analysis due to the presence of performance outliers. These outliers were removed from the analysis and the Satterthwaite method was used to calculate the adjusted degrees of freedom using the lmerTest package in R. When significant effects were observed, post hoc comparisons were performed using the Tukey Honest Significance Difference (HSD) test and considered significant at the $p < .05$ level, and the Satterthwaite method was again used to calculate the degrees of freedom. A two-factor (Group and Trial) mixed model with one repeated measure (Trial) was run on the pitch root-mean-square error. There was a significant main factor of Trial ($F(31, 434) = 3.6, p < 0.001$) and Group was not significant

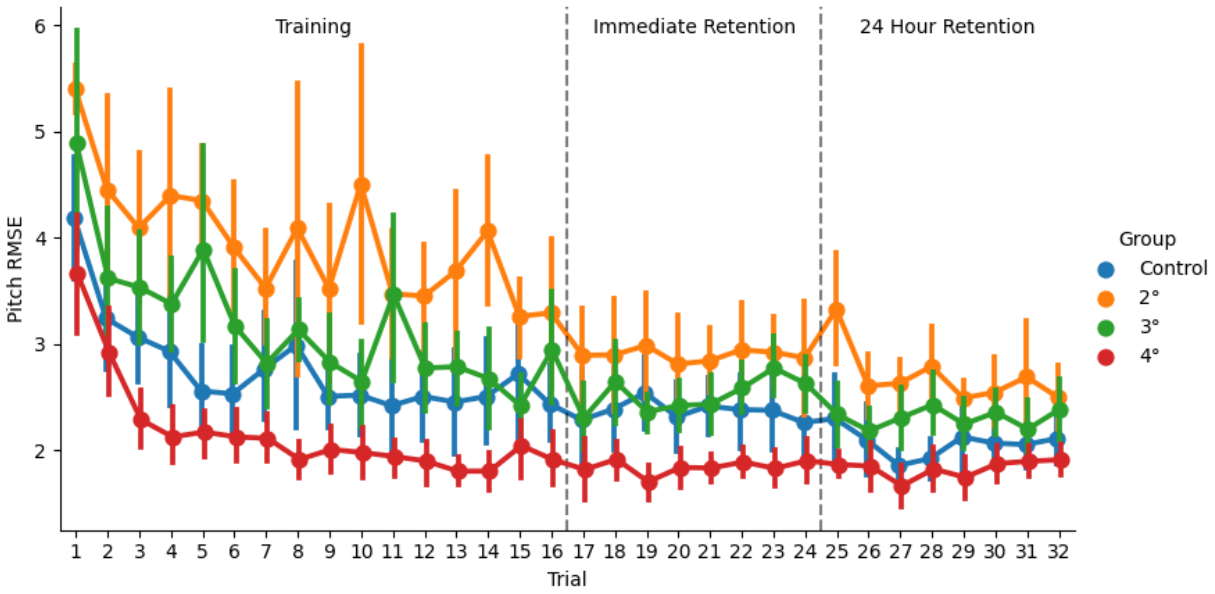


Figure 6.1: The mean Pitch RMSE for each trial for participants. Data points are the mean, and error bars are the standard error of the mean.

($F(3, 14) = 2.6, p = 0.09$). Additionally, the interaction effect between Group and Trial was not significant ($F(93, 434) = 0.9, p = 0.78$), which indicates that subjects learned to perform the task over time but that the effect of Group was not significant given the effect size and current number of subjects. Even with one half to one third of the anticipated number of subjects, however, the main effect of Group is already trending towards significance, as we would expect based on the results of the previous study. Differences in groups are beginning to emerge when examining the plot, see Figure 6.1, though the error bars are too large to draw definite conclusions.

This analysis can also be performed on the other dependent measures, with similar results. Most metrics closely track those seen in the previous study, though the low sample size makes it difficult to draw conclusions at this stage. Due to their extremely preliminary nature, these results are not included here.

6.4 Discussion

Based on the preliminary findings presented in Figure 6.1, choosing an appropriate bandwidth can potentially have an effect on the efficacy of concurrent bandwidth feedback. While

we have yet to observe statistically significant effects between groups, current data suggests that having a looser bandwidth improves performance. While it was expected that there may be an optimal choice of bandwidth to maximally reduce training times and improve performance, we hypothesized that this value would be lower than or approximately equal to three degrees based off our pilot studies.

While we cannot make any definite claims based on this preliminary data and analysis, we can revisit data from the previous experiment to make further predictions. Increasing the acceptable bandwidth of feedback results in the feedback being presented less often. The repetition in training results in subjects improving their performance and reducing their deviations during compensatory tracking. We can use the disturbance force to estimate the percentage of time that subjects are exposed to the feedback because we presented subjects with the same disturbance force during each trial. Figure 6.2 presents the percentage of time that pitch feedback would be active if subjects did not provide any control input whatsoever. This suggests that, at the beginning of training, subjects with a bandwidth of two, three, and four degrees would experience feedback 76%, 62%, and 47% of the time, respectively, if they provided no input.

The previous experiment's feedback group subjects had their bandwidth set at three degrees and were instructed to try and improve their performance whenever they saw that the feedback was activated. By looking at the percentage of pitch errors that were above a given value, we can infer what percentage of time feedback would have been on when subjects had finished their training phase, regardless of chosen bandwidth. Figure 6.3 presents the percentage of time that pitch feedback was active at the end of training, showing 12% for the three-degree bandwidth that subjects experienced. It is unclear if subjects continue to improve until they reach a point where feedback appears approximately 12% of the time, or if this is simply a limit of human motor control for this task. In either case, the 4% of time feedback would have been presented if a four-degree bandwidth was chosen is clearly within human control. If the 12% target is a true motivational limit, 4% would likely inspire greater confidence in subjects, who may go on to report lower workload values. The two-degree bandwidth, however, would only be attainable 29% of the time by fully trained subjects in the previous experiment and would likely have a demotivating or otherwise negative effect.

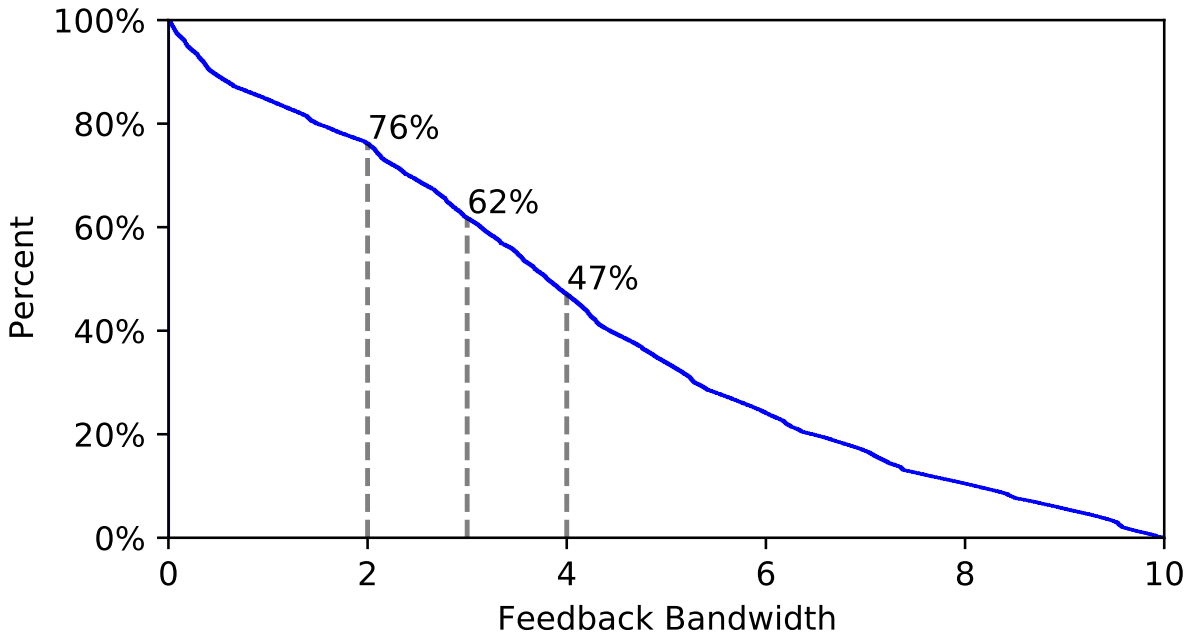


Figure 6.2: The percentage of active pitch feedback with no input present. This represents the resulting aircraft motion due to the disturbance force.

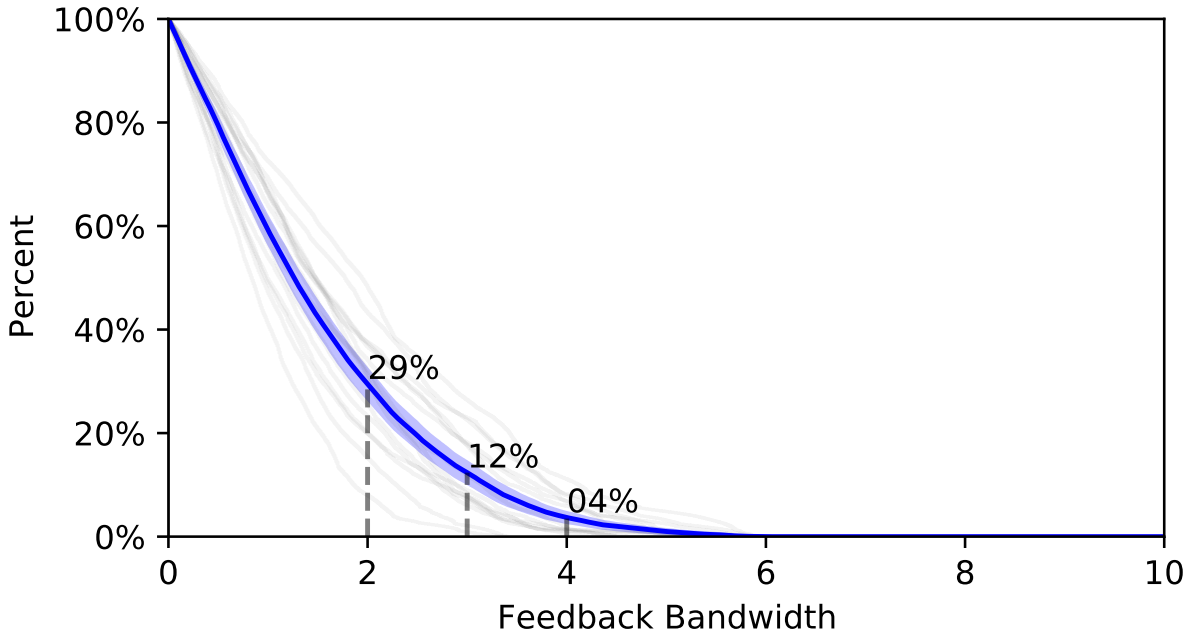


Figure 6.3: The percentage of active pitch feedback time at the end of the previous study, when the bandwidth was set to three degrees. The black line is the mean, and the shaded region is the standard error of the mean.

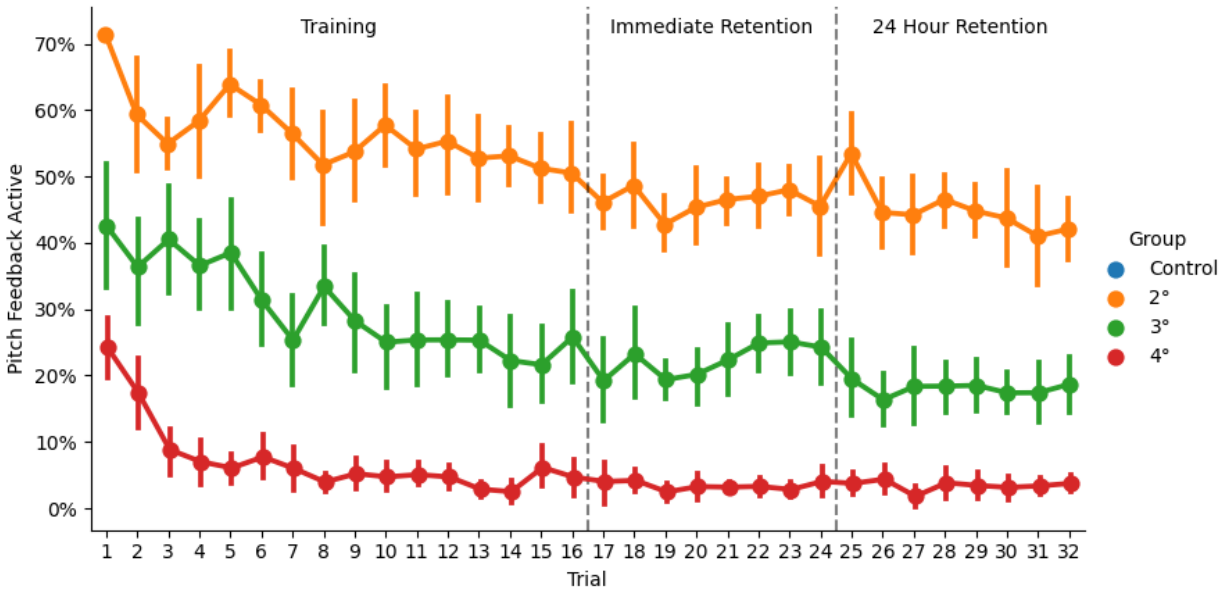


Figure 6.4: The percentage of active pitch feedback time for each trial for participants. Note that the Control group never received feedback. Data points are the mean, and error bars are the standard error of the mean.

Finally, we can take a preliminary look at how often subjects in this experiment experienced their feedback between the different bandwidth groups. The percent of time that feedback was active at the beginning and end of training should provide some insight into how subjects respond to different bandwidth levels. Figure 6.4 shows the percentage of active feedback time for each trial for subjects in the three feedback groups. We can see how effectively subjects used the feedback on the first trial by looking at the difference between the value on this plot and that predicted by the no feedback case. On the first trial, subjects in the three feedback groups tended to reduce their anticipated feedback exposure when the bandwidth was larger. Subjects in the four-degree bandwidth group, for instance, nearly halved their anticipated feedback exposure, as could be expected from their pitch RMSE presented in Figure 6.1. Additionally, subjects in this group asymptote very closely to the 4% value predicted by subjects in the previous study. The two- and three-degree bandwidth groups, however, continue to show much larger feedback exposure percentages. The results for subjects in the three-degree bandwidth group are especially unexpected, as they do not asymptote to the 12% number shown by subjects in the previous experiment. While

differences in experimental procedures could be the cause of this, we believe that the low sample size (5 subjects thus far in this experiment compared to 15 subjects in the previous experiment) is more likely the cause.

6.5 Conclusions

While the preliminary data collected for this experiment can be used to begin to explain the effects of employing different bandwidths in concurrent bandwidth feedback, more subjects are needed to make any definitive claims. Briefly revisiting our hypotheses, we can start to predict the results of this study.

H1. While subjects in the three- and four-degree bandwidth groups will likely immediately outperform those in the control group, the two-degree bandwidth presents too much of a challenge for subjects to effectively use the concurrent bandwidth feedback. This indicates researchers should be careful to choose attainable bandwidths, or they may incorrectly conclude that concurrent bandwidth feedback, in general, does not work for their task.

H2. Based on our early results, participants will report the same workload between groups, which will gradually decrease with training.

H3. Participants in the feedback group will not suffer from the guidance hypothesis and will retain their performance and workload levels when the feedback is removed in the immediate retention and 24 hour retention trials.

H4. It is still too early to make claims on the effects of concurrent bandwidth feedback on transfer of training.

Finally, we note that this study will resume testing when the University of California, Davis Institutional Review Board deems that testing can continue.

Chapter 7

Modeling the Effects of Feedback

The second Aim of this dissertation is to develop a model of the human pilot that includes the effects of concurrent bandwidth feedback. After evaluating many of the human pilot models presented in the literature, the Structural Model was selecting for adaptation because it has been very effective in predicting pilot performance for a variety of system dynamics and requires a relatively small number of parameters. The Structural Model employs classical feedback control theory and includes both the vehicle dynamics and the human operator's control input response to task demands and disturbance forces. The previous two Chapters detailed human-in-the-loop experiments where subjects trained on simplified aircraft simulators, providing a rich data set to use as modeling verification. This Chapter outlines the extension of Hess's Structural Model of the human pilot to include the effects of concurrent bandwidth feedback.

7.1 Introduction

7.1.1 Motivation

The model presented here extends Hess's 1997 Structural Model of the human pilot. The Structural Model has successfully predicted human performance through a variety of system dynamics and can predict how performance changes during a pilot's adaptation to changing dynamics. Hess developed adaptive logic for the human pilot in a pursuit task, which triggers when the pilot notices that vehicle dynamics have changed, see Figure 7.1 (Hess, 2009). This logic is based on several criteria, which "must be predicated upon information available to

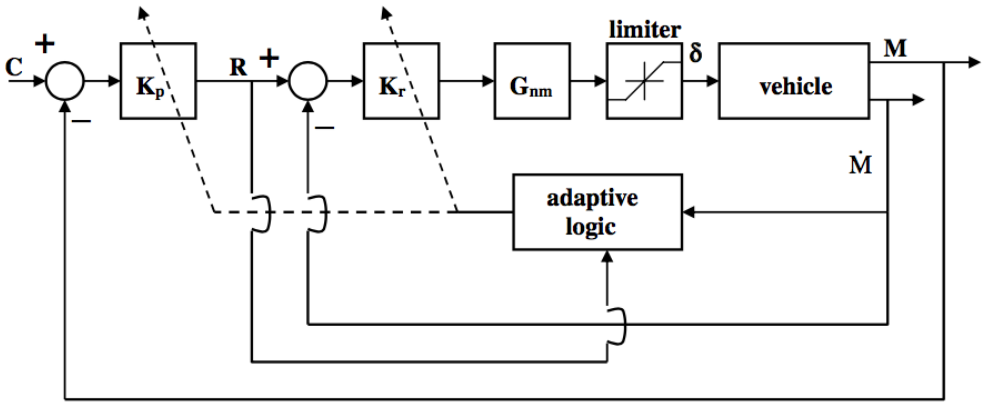


Figure 7.1: Hess's model of the adaptive human pilot, from Hess (2009).

the human [and] the postadapated pilot models must follow the dictates of the crossover model of the human pilot" (Hess, 2009). The crossover model is defined as

$$Y_c(s)Y_p(s) \approx \frac{w_c e^{-s\tau_e}}{s} \tag{7.1}$$

relating the operator and controlled element transfer characteristics. Here Y_c is the controlled element transfer function, Y_p is the human operator transfer function, w_c is the crossover frequency, and τ_e is the effective time delay of the pilot. The primary result of the adaptive logic is to increase the resulting crossover frequency of the pilot, effectively making them more responsive, which could be interpreted as more focused on the task.

While the Structural Model of the adaptive pilot has been successful in predicting changes in performance for a well-trained subject, it does not consider how a pilot would behave when they are still in the early stages of training. The modified model presented here includes two major changes to Hess's current model:

- The adaptation logic is changed to focus on concurrent bandwidth feedback
- The timescale of the adaptation is significantly longer

The logic and reasoning behind these modifications are developed in the following sections. In this work, we propose that the adaptive logic should trigger when the pilot is receiving concurrent bandwidth feedback, instead of when a change in system dynamics occurs as in the original model. This requires an additional feedback loop in the the Structural

Model that triggers when the bandwidth feedback is activated. This loop is based around the K_e gain, which is the primary way of setting the crossover frequency in the Structural Model, see Figure 7.4. This change implies that the subjects in our experiments do their primary learning while they are receiving the qualitative feedback that their current level of responsiveness is not sufficient to complete the task.

Based on the experimental results described in previous Chapters, the change in performance we see when subjects use concurrent bandwidth feedback happens relatively rapidly–within a few minutes. This is reflected in the performance delta between subjects in the different groups of our SAFER experiment (see Figure 1.5a), EMG experiment (see Figure 4.4), and aircraft experiment (see Figure 5.3). Hess's model results in pilots adapting within a very short time period, on the order of 5 seconds (Weir and Phatak, 1966). Conversely, the results of our experiments in the past three chapters suggest relatively long adaptation times, on the order of a few minutes.

In the remainder of this chapter, we briefly describe the experimental methods and results from Chapter 5 to refresh the reader. We then apply a novel identification technique for finding the values of the parameters of the Structural Model. Using an Autoregressive Model with Exogenous Variables (ARX) approach, we validate the accuracy of the resultant crossover frequencies from this identification technique. Finally, using the difference between the control and feedback groups in identified parameters, we develop a modified Structural Model that includes the effects of concurrent bandwidth feedback.

7.2 Method

7.2.1 The Piloting Task

Training operators to perform complex manual control tasks is expensive and time consuming. Moreover, poorly trained operators can cause catastrophic failures. The use of concurrent feedback techniques has been shown to improve performance and reduce training times but has not previously been evaluated for complex, real-world tasks such as flying aircraft. The application of CBF to flight tasks has the potential to reduce costs and risks by reducing training time and producing better pilots. In our experiment (fully described in Chapter 5), thirty participants were evenly split into two groups. Participants were tasked

with flying a simulated Boeing 747 aircraft in three control modes, each of increasing degrees of freedom and functional complexity:

P Pitch only (low complexity)

PR Pitch and Roll (moderate complexity)

PRA Pitch, Roll and Altitude (significant complexity)

Each participant completed a total of 36 trials, 12 in each of the three control modes. Each trial had a duration of 82 seconds, and participants self-initiated the trial by activating a trigger on the joystick. The complete results of this experiment can be seen in Chapter 5.

The two groups of participants were control and feedback, and the interface that participants saw during the experiment is available in Figure 7.2. When the vehicle was well-controlled (i.e., small errors in the controlled degrees of freedom), the interface was identical for participants in both groups. As errors increased above the threshold, participants in the feedback group experienced concurrent bandwidth feedback on various elements of the attitude indicator and altimeter, depending on the control mode. This feedback was indicated by changing the display element from the default yellow color to a red color when the error deviated outside of a predetermined bandwidth. For the Pitch-only control mode focused on in this Chapter, the display changed from yellow to red when the pitch deviated

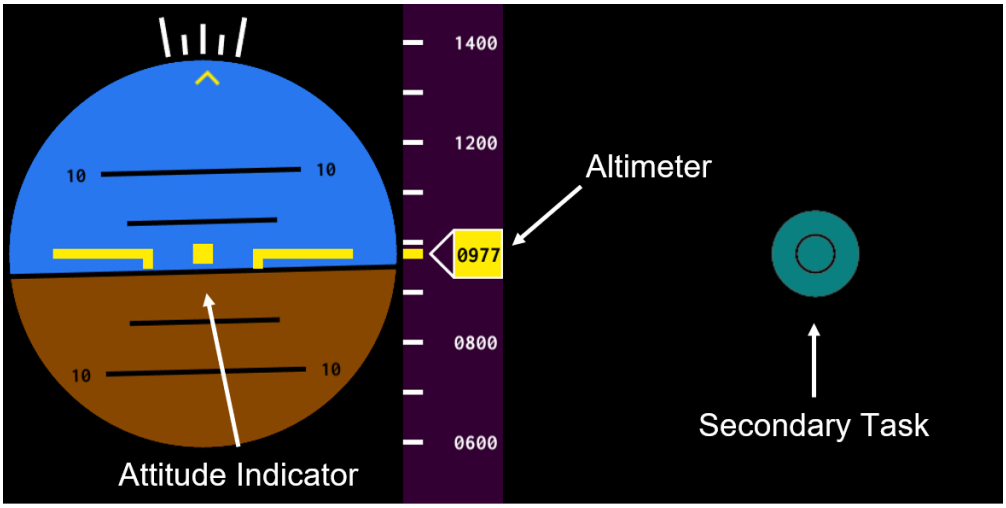


Figure 7.2: The interface that participants were presented with during the experiment. On the left, the flight display shows deviations in pitch, roll, and altitude. On the right, and the secondary, two-choice task.

by greater than three degrees. For both groups, performance measurements were evaluated to determine the effects of the feedback on subject learning rate and maximum skill level.

To assess short-term retention of learned skill for the feedback group, the concurrent bandwidth feedback was removed after the 9th trial, and performance was again evaluated. Following the 9th trial, subjects in the feedback group no longer experienced feedback, and the interfaces for the control and feedback groups were identical. This was done in order to investigate how well the feedback subjects could retain their performance benefit.

Experiment Results

The resulting root-mean-square error (RMSE) for each group of participants by trial is presented in Figure 7.3. Both groups of participants had a relatively high initial RMSE which was reduced through repeated training on the task. Participants in the feedback group, however, performed significantly better than those in the control group and learned the task much faster. Subjects in the feedback group were able to maintain and further improve their performance after the feedback was removed.

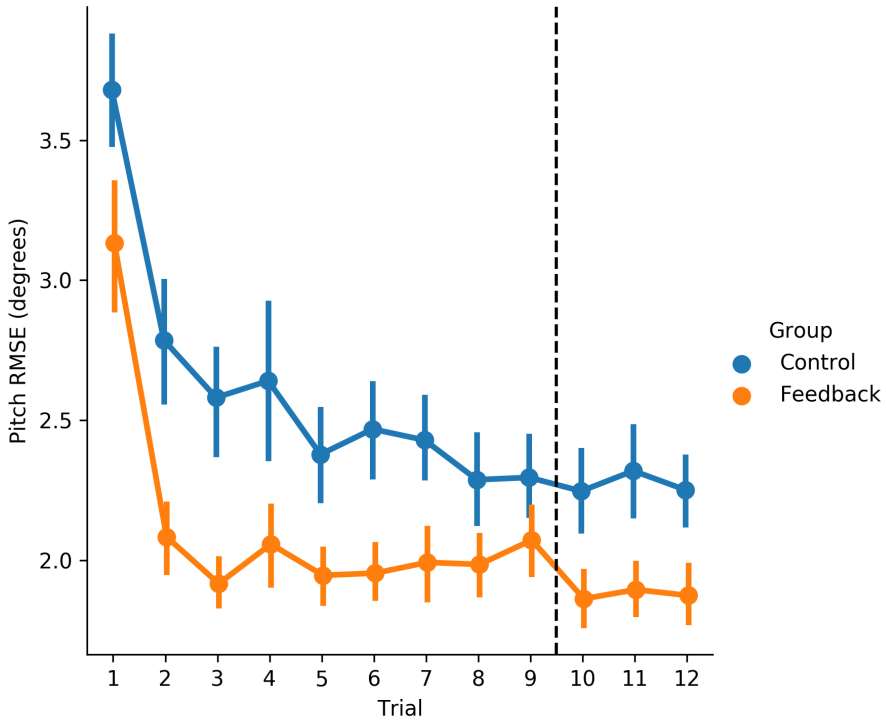


Figure 7.3: Pitch root-mean-square error. Feedback was removed from subjects in the feedback group after the 9th trial in order to investigate immediate retention. Error bars are the standard error of the mean.

7.2.2 Modeling Techniques

We investigated two techniques for creating models of a human pilot in order to better understand the results of the aircraft training study. Both techniques produce an estimated transfer function for the human operator, Y_p, which can be combined with the controlled system dynamics, Y_c, in order to explore the crossover model characteristics discussed in Section 1.2.5. The crossover frequency can be identified using

$$Y_c Y_p = \frac{w_c e^{-s\tau_e}}{s} \tag{7.2}$$

which provides an indication of how hard the operator is working.

These two techniques are an Autoregressive Model with Exogenous Variables (ARX) technique and the Structural Model.

Autoregressive Model with Exogenous Variables (ARX)

Autogregressive (AR) models are commonly used to represent time-varying processes in which an output variable linearly depends on its previous value and a stochastic error term (Yule, 1927). When a process also depends on an external input variable (often referred to as an exogenous variable), the resulting model is known as an autoregressive model with exogenous variables (ARX). The structure of an ARX model is given by

$$y(t) + a_1 y(t-1) + \ldots + a_{n_a} y(t-n_a) = b_1 u(t-n_k) + \ldots + b_{n_b} u(t-n_b-n_k+1) + e(t) \tag{7.3}$$

where $y(t)$ is the output variable at time t, n_a is the number of poles, n_b is the number of zeros, n_k is the number of input samples which occur between input and output (also known as the dead time or time delay in the system), $y(t-1) \ldots y(t-n_a)$ are the previous outputs on which the current output depends, $u(t-nk) \ldots u(t-n_b-n_k+1)$ are the delayed inputs on which the current output depends, and $e(t)$ is a white-noise error term.

An autoregressive technique to identify a transfer function from an input/output time sequence pair was previously developed in Hess et al. (2002). Using this technique, provided in a script called *gettf1*, we identified transfer functions representing pilot models for the data trials from our experiment. The inputs to this technique are the input and output time series and three integers describing the resulting model (n_a, n_b, and n_k). For the results presented in this Chapter, $n_a = n_b = 10$ and $n_k = 0$. This technique uses an ARX model which is fit

using least-squares and results in an estimation of a discrete transfer function model of the input/output system. The script also produces a continuous transfer function using a Tustin approximation Tustin (1947). The result of this process is a transfer function of the pilot with ten poles and zeros describing the human operator's response to the displayed flight dynamics.

Structural Model

The complete Structural Model in Figure 7.4a can be reduced to Figure 7.4b by setting the following standard parameter values.

$$S_1 = S_2 = S_3 = \downarrow$$

$$K_{\dot{e}} = \epsilon = K_{\dot{m}} = K_{\ddot{m}} = 0$$

These switches are set to their standard locations for normal error sensing (Hess, 1997). $K_{\dot{e}}$ becomes irrelevant as it no longer appears in the loop due to the orientation of the switches, and ϵ is set to zero for simplicity. $K_{\dot{m}}$ and $K_{\ddot{m}}$ are set to zero because the participants did not experience any motion, so the vestibular feedback loop is not needed.

The resulting model can be parameterized by defining Y_{NM}, Y_{FS}, and Y_{PF}. Y_{NM} describes the open-loop dynamics of the neuromuscular system driving the joystick, and can be defined

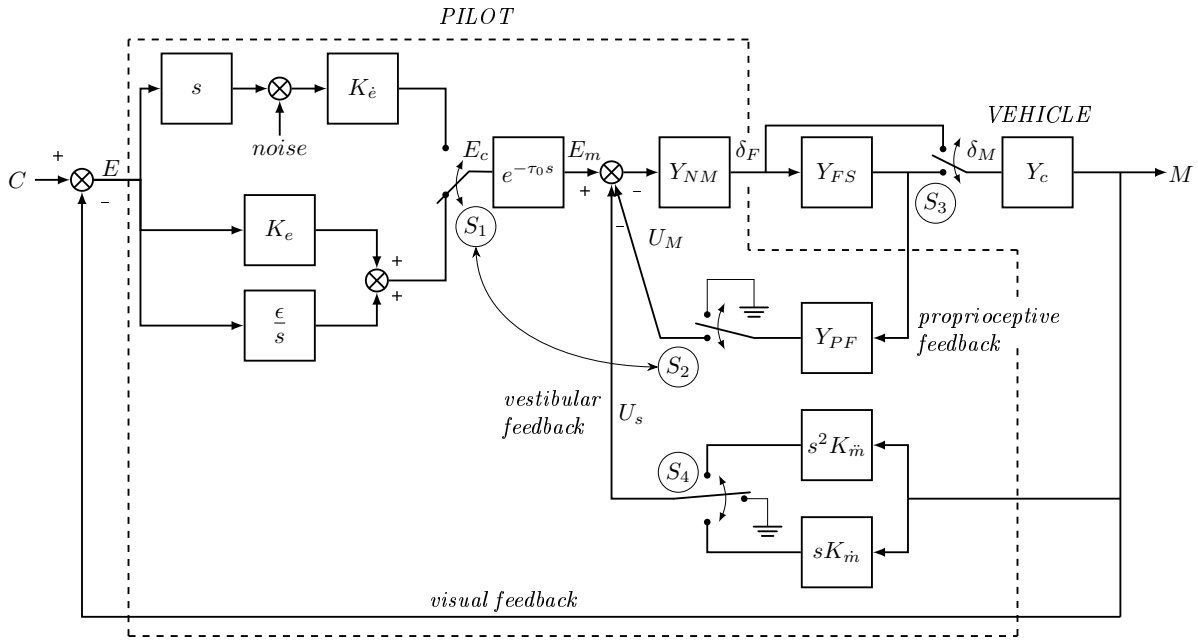
$$Y_{NM} = \frac{\omega_{NM}^2}{s^2 + 2\zeta_{NM}\omega_{NM}s + \omega_{NM}^2} \tag{7.4}$$

where ω_{NM} and ζ_{NM} are the undamped natural frequency and damping ratio of the neuromuscular system. Y_{FS} is set to 1 here, as for a simple joystick it takes the form of a gain which is absorbed by the other elements in the model. The proprioceptive feedback block is defined as

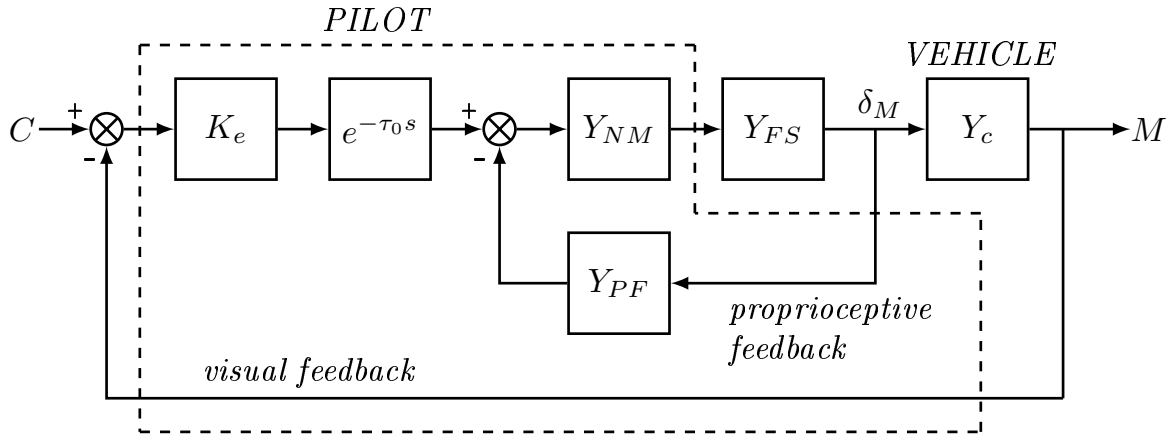
$$Y_{PF} = \frac{K}{s + a} \tag{7.5}$$

which is chosen to satisfy $Y_{PF} \propto sY_c(s)$ and represents the operator's internal model of the system dynamics (Hess, 1997). Finally, Y_c describes the flight dynamics for the task. The dynamics of this task describe a Boeing 747 in Flight Condition 2 (slow, sea-level flight), and resultant elevator-to-pitch dynamics take the form (Heffley and Jewell, 1972):

$$Y_c = \frac{0.5716(s + 0.5535)(s + 0.03952)}{(s^2 + 0.006158s + 0.01512)(s^2 + 1.12s + 0.8006)} \tag{7.6}$$



(a) The Structural Model, adapted from Hess (1997). This model includes proprioceptive, vestibular, and visual feedback within the pilot's model, and allows researchers to model the effects of pilot induced oscillations.



(b) The reduced structural model used in this analysis, which only includes the proprioceptive feedback loop.

Figure 7.4: The Structural Model of the Human Pilot.

This result is obtained by substituting the values for this flight mode from NASA CR-2144 into the longitudinal dynamic matrix presented in Appendix C.

7.3 Results

7.3.1 ARX

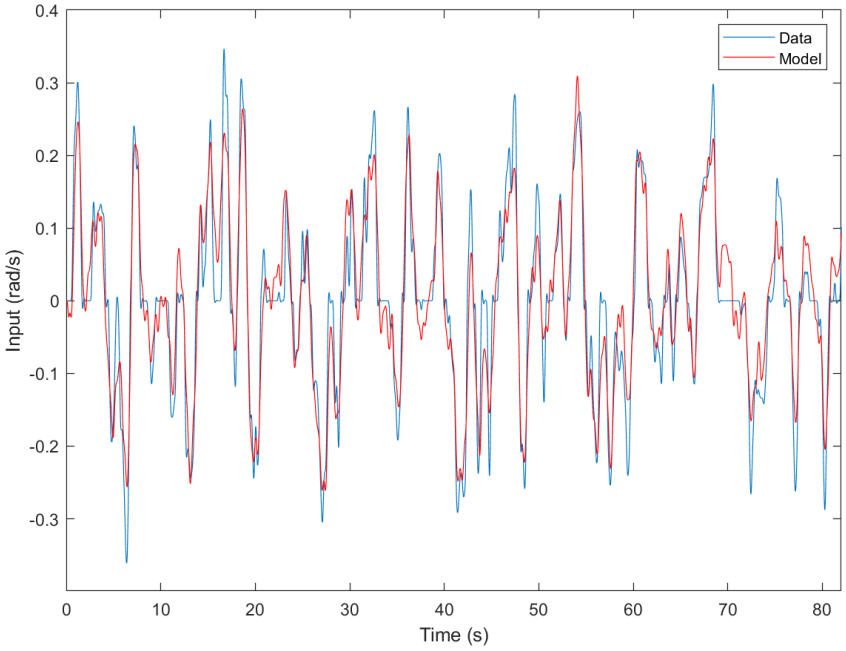
The original input to the system from the experiment participants (u) is fed into the resulting transfer function to produce a simulated output (u_{sim}). The output from the pilot models are compared with the experimental data from which they are generated in order to determine the quality of fit. The variance accounted for (VAF) is a commonly used metric to quantify the quality of the fit between the outputs of the model and the experimental data. The VAF is calculated by

$$\text{VAF} = \left(1 - \frac{\sum |u - u_{sim}|^2}{\sum u^2}\right) \times 100\% \tag{7.7}$$

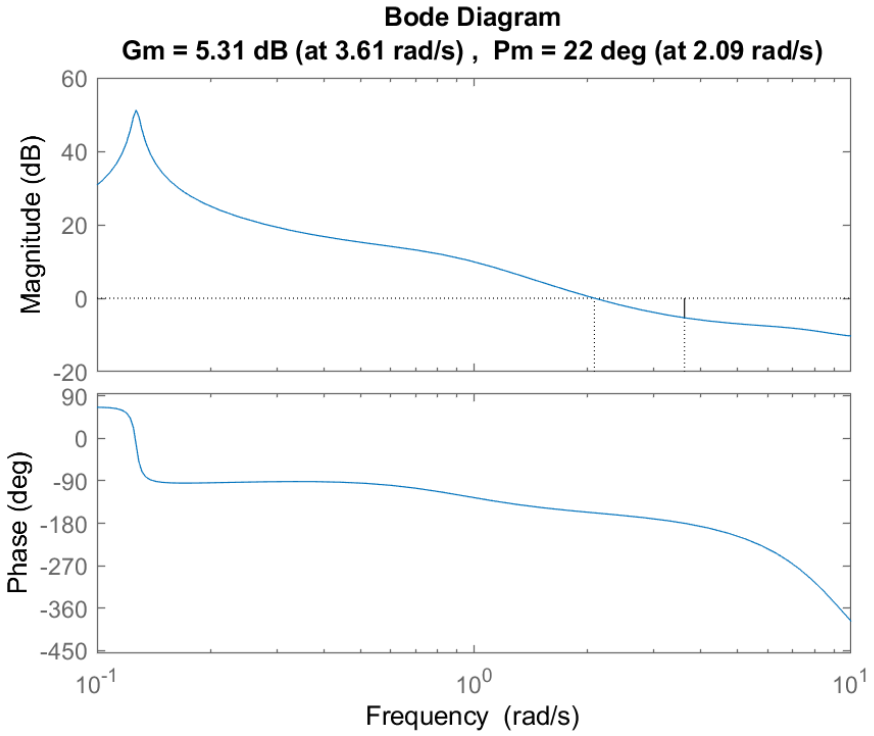
where u is the participant's command input to the vehicle's elevators, and u_{sim} is the model's simulated input. The VAF ranges between 0 (no correlation) and 1 (perfect match).

Applied to the results of our experiment, the VAF of the feedback group's models immediately jumps up after the first trial to an average of $\approx .75$, indicating a good fit. The control group's VAF increases more slowly to a value of approximately $\approx .68$. Here, an increased VAF indicates that the pilot is behaving in a more linear fashion, and these results suggest that the concurrent bandwidth feedback can immediately assist participants. The RMSE between the simulated and actual data is similar between groups and trends slightly up during training. Visual inspection of the experimental and simulated data also shows very good agreement, see Figure 7.5a.

By combining the above pilot models with the aircraft system dynamics, the crossover frequency (Figure 7.5b) and gain and phase margins can be found. The crossover model has long been used as the standard model for human control tasks, where the crossover frequency represents "how hard" the pilot is working (McRuer and Krendel, 1957). A larger crossover frequency indicates that the pilot is working harder. The results of these parameters reflect what was found in the pitch RMSE performance analysis: subjects in the feedback group immediately show increased crossover frequency and decreased phase and gain margins.



(a) Comparison between estimated transfer function and experimental data, showing a good variation accounted for.



(b) Combined pilot/vehicle open-loop bode diagram showing the standard crossover model characteristics, with a crossover frequency of 2.09 rad/s.

Figure 7.5: Example results from using the estimated pilot models from *gettf1*.

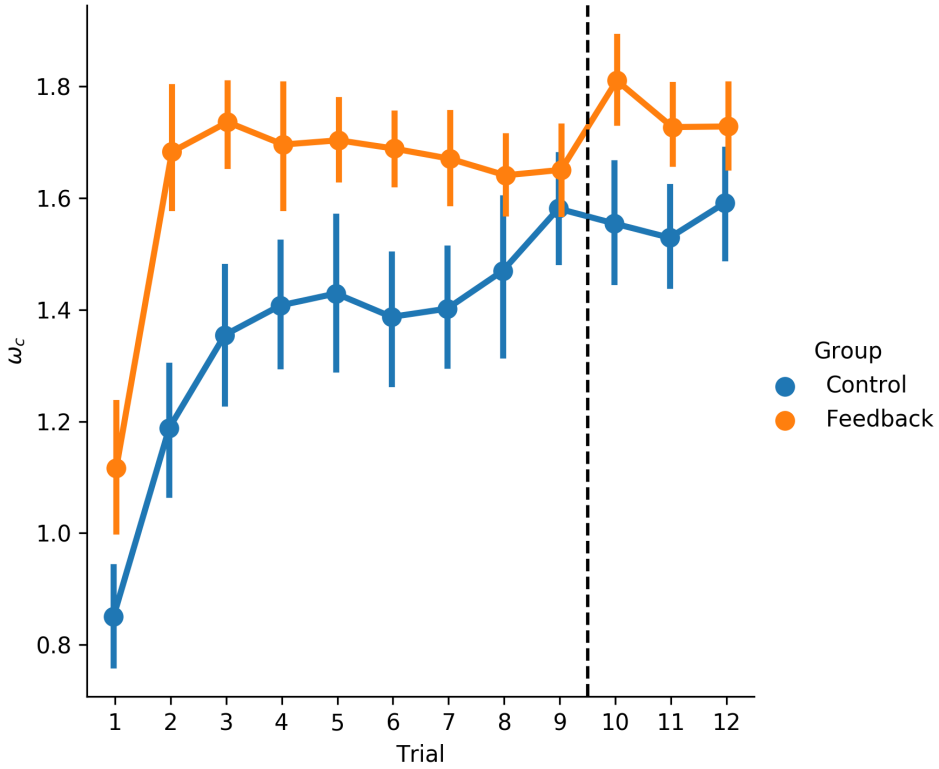


Figure 7.6: Crossover frequency (rad/s) predicted by the ARX technique for the pitch only trials. The dashed line indicates where the feedback was removed. Error bars are the standard error of the mean.

These results are sustained during retention and again show an immediate improvement when the feedback is removed. The crossover frequency identified by the ARX technique for the pitch only trials is available in Figure 7.6.

Linear mixed models were used to calculate the significance of factors in our analysis due to the presence of outliers which were removed from the analysis, and the Satterthwaite method was used to calculate the adjusted degrees of freedom. When significant effects were observed, post hoc comparisons using the Tukey Honest Significance Difference (HSD) test were performed and considered significant at the $p < .05$ level, and the Satterthwaite method was again used to calculate the degrees of freedom.

A three-factor (Group, Mode, and Trial) mixed model with two repeated measures (Mode and Trial) was run on the pitch crossover frequency. There were significant main factors of group ($F(1, 28.00) = 4.9, p = 0.036$), mode ($F(2, 55.62) = 17.1, p < 0.001$), and trial ($F(11, 306.23) = 39.4, p < 0.001$). There were also significant interaction effects between

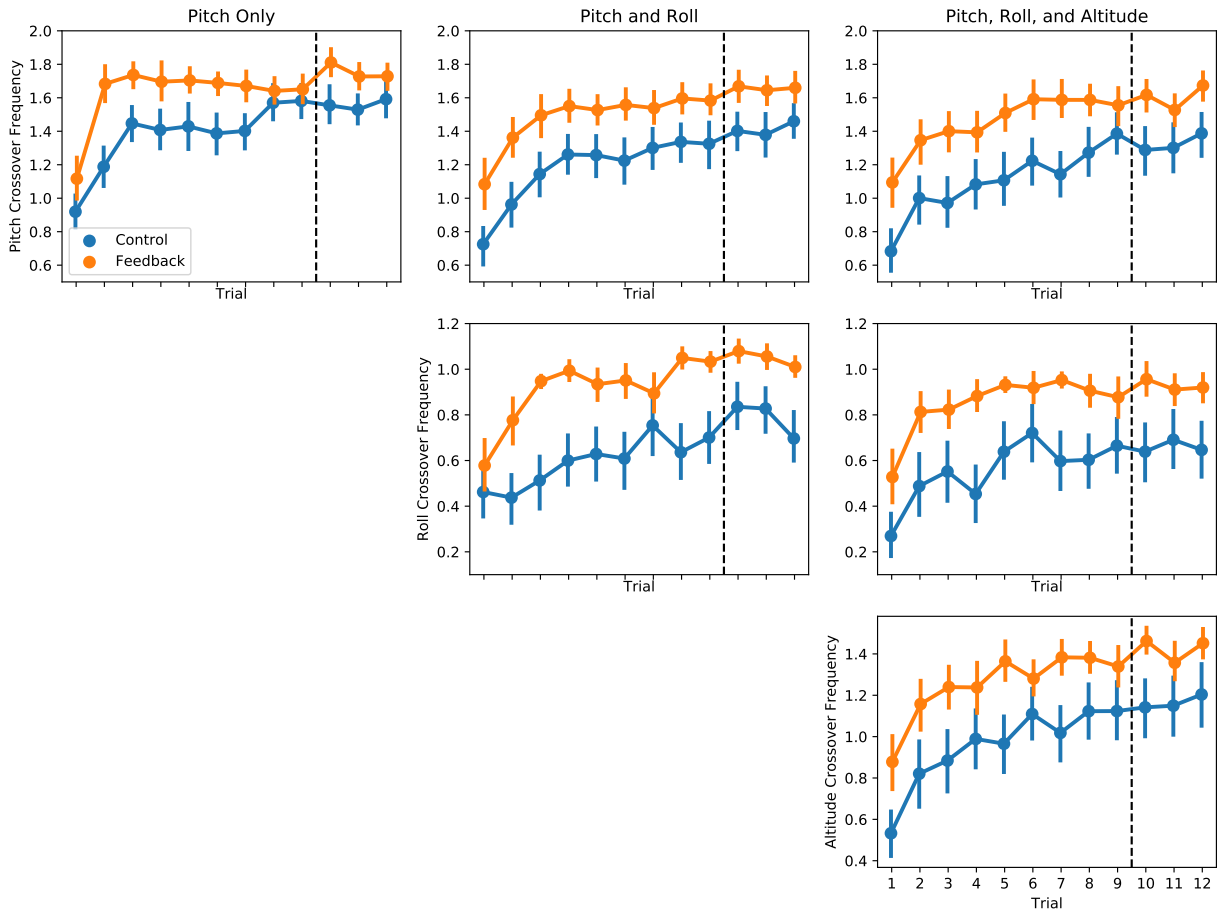


Figure 7.7: The crossover frequency (rad/s) of the estimated pilot/vehicle open-loop transfer functions for each group, trial, and control task. The dashed line indicates where the feedback was removed. Error bars are the standard error of the mean.

group and trial ($F(11, 306.23) = 2.0, p = 0.025$) and between mode and trial ($F(22, 610.08) = 2.8, p < 0.001$). Despite the presence of interaction effects that result from participants learning the task (as indicated by the trial factor), the main effects can still be interpreted. A Tukey test showed that the groups differed significantly, with the participants in the feedback group outperforming those in the control group ($M = 1.25, 1.56$, respectively, $SE = 0.10$). An additional Tukey test showed that the participants' crossover frequencies between the modes differed significantly, with the largest crossover frequency in the P mode, followed by the PR mode, and finally the lowest in the PRA mode ($M = 1.53, 1.38, 1.31$ respectively, $SE = 0.07$). This same analysis was completed on the roll crossover frequency with similar results. There were significant main factors of group ($F(1, 27.82) = 8.2, p < 0.01$), mode

($F(1, 25.49) = 23.2, p < 0.001$), and trial ($F(11, 292.40) = 16.3, p < 0.001$). Tukey tests showed that the participants' crossover frequencies between the groups and the modes each differed significantly, with the participants in the feedback group again outperforming those in the control group ($M = 0.55, 0.89$, respectively, $SE = 0.08$), and crossover frequency was best in the PR mode followed by the PRA mode ($M = 0.77, 0.67$, respectively, $SE = 0.06$).

A two-factor (Group and Trial) mixed model with one repeated measure (Trial) was run on the altitude crossover frequency. There were significant main factors of group ($F(1, 27.92) = 4.4, p = 0.046$) and trial ($F(11, 301.95) = 18.3, p < 0.001$). Tukey tests showed that the participants' crossover frequencies between the groups differed significantly, with the participants in the feedback group again outperforming those in the control group ($M = 0.98, 1.29$, respectively, $SE = 0.11$), and the trial effect showing learning throughout the experiment for both groups. See Figure 7.7 for the plots of the crossover frequency of the estimated pilot/vehicle open-loop transfer functions for each group, trial, and control task.

7.3.2 Structural Model

Value Identification Technique

Once the Structural Model is parameterized to the version presented in Figure 7.4b, the remaining effort in achieving Aim 2 then becomes fitting the remaining parameters to best match the data collected in the experiment. This process is more complex than the ARX technique as there are many parameters to identify that interact nonlinearly, and the fit function has many local maximums.

A four-step process was developed to address the challenge of finding the optimal model parameters. This approach was validated by comparison to the results of the ARX technique. The general outline of the technique was:

1. Conduct a brute-force parameter search over the six remaining model parameters.
2. Select the set of parameters that results in a maximum VAF as an initial best guess.
3. Iterate on the initial guess to identify a final optimal set of parameters.
4. Identify the crossover frequency using the system dynamics and identified optimal parameters.

Each of these steps is described in detail below.

The first step of the value identification technique was a brute force parameter search used to identify the value of the remaining parameters. Each combination of the parameters K_e, τ_0, K, a, ω_{NM}, and ζ_{NM} were simulated in Simulink using the same flight model and disturbance profiles as the participants experienced in the experiment, see Table 7.1. The values from this table were chosen based on a coarse preliminary analysis such that they covered the expected range of values with sufficient granularity. The resulting 102,600 combinations were simulated and compared to each trial in order to identify the set of parameter values resulting in the largest VAF. Each combination of experimental subject data and simulations was then compared to find a set of simulation parameters that maximized the VAF. This solution represented a globally maximum set of a parameters, mitigating the issue resulting from the parameter's nonlinear interactions with each other. Once this initial global best fit was identified, the parameters were further tuned in MATLAB using *fmincon* to find the maximum of the constrained nonlinear multivariable function, which was the VAF. The solver was constrained by the bounds of the global brute-force search. In general, the optimizer did not result in large parameter changes from the initial values.

The optimizer initially experienced convergence issues due to an interaction effect between trying to simultaneously fit ω_{NM} and ζ_{NM}. This issue was resolved by leaving ζ_{NM} out of the initial optimization step and tuning it independently at the end of the optimization process. Finally, the optimal set of six parameters was inserted into the model and

Parameter	Values
K_e	1, 2, 3, ..., 30
τ_0	.20, .25, .30 s
K	1.0, 1.5, 2.0, ..., 10.0
a	.0, .25, .50, ..., 3.0
ω_{NM}	6, 7, 8, ..., 12 rad/s
ζ_{NM}	0.7 rad/s

Table 7.1: Structural Model parameters used for the initial global optimal fit.

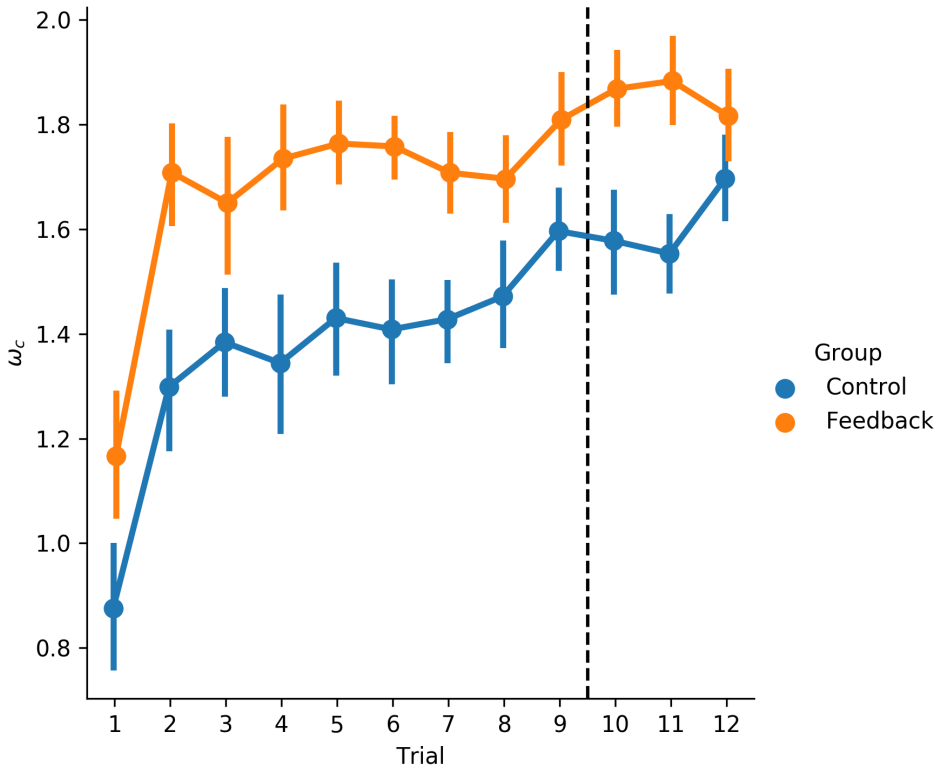


Figure 7.8: Crossover frequency (rad/s) predicted by the Structural Model parameter identification for the pitch only trials. The dashed line indicates where the feedback was removed.

combined with the controlled system dynamics to identify the resultant crossover frequency (Figure 7.8), which compares very favorably with that found using the ARX technique (Figure 7.6). Both techniques show an initially small crossover frequency which approaches the generally accepted upper limit of $\approx 1.7 - 2$ rad/s (Hess, 1984a).

By investigating the control mode with the smallest effect of the feedback, we show that this technique is sensitive enough to detect small changes between the two groups. For the purposes of this work, only the Pitch only control mode was analyzed. The technique developed here could be extended to all three control loops however, and the results should be similar and more pronounced for modes of higher complexity. A two-factor (Group and Trial) mixed model with repeated measures (Trial) was run on each of the six parameters to identify which parameters significantly changed between groups and over the course of training. As with the ARX technique, the Satterthwaite method was used to calculate the adjusted degrees of freedom. The results of this analysis are presented in Table 7.2.

Evaluating the results of the linear mixed effects models presented in Table 7.2, we immediately see that τ_0, K, and ω_{NM} have no significant effects dividing either Group or Trial. From this we can determine that these variables have no significant effect resulting from either training or the difference in performance between the feedback and control groups. Of the remaining variables, a and ζ_{NM} show significant effects in the trial factor, suggesting that they change over time during training. Further investigation of these variables showed that a was only significantly different for the very first trial, a result which was ignored for the remainder of this analysis to simplify the resulting model. ζ_{NM} also varied significantly over the course of training, but closer inspection revealed that the effects of varying this parameter did not significantly affect the resulting outputs of the model. As a result of this, ζ_{NM} was considered a secondary term and not included in the feedback model. Finally, the K_e variable showed significant effects between both the Group and Trial variables, suggesting that it was the primary parameter responsible for changes in performance between the groups

Parameter	Effect	df_{num}	df_{den}	F	Pr(>F)
K_e	Group	1	27	7.27	0.01
	Trial	11	308	8.67	< 0.001
τ_0	Group	1	27	0.02	0.90
	Trial	11	308	0.61	0.82
K	Group	1	27	1.65	0.21
	Trial	11	308	1.03	0.42
a	Group	1	27	0.01	0.94
	Trial	11	308	2.68	< 0.01
ω_{NM}	Group	1	27	0.05	0.83
	Trial	11	308	1.21	0.28
ζ_{NM}	Group	1	27	0.19	0.66
	Trial	11	308	1.87	0.04

Table 7.2: Results of the linear mixed models of the identified Structural Model parameters.

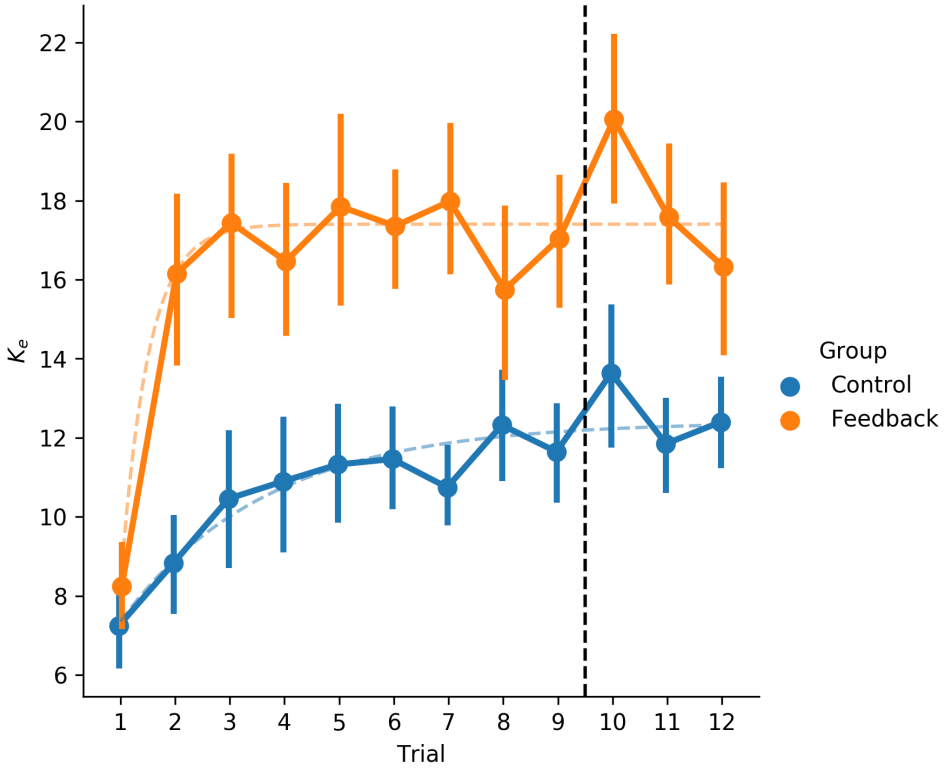


Figure 7.9: K_e is greater for subjects exposed to feedback. The vertical dashed line indicates where the feedback was removed from participants in the feedback group, and the colored dashed lines indicate exponential fits to the data. Error bars are the standard error of the mean.

and repeated training in the simulator. As the most dominating factor, understanding the effect of K_e is of primary interest, and the remainder of this Chapter focuses on understanding its connection to increases in performance and the modification this parameter to explain the results of increased performance in the feedback group.

7.4 Extending the Structural Model

From the analysis of the linear mixed effects models, we determined that all the parameters of the Structural Model can be treated as constants with the exception of K_e, which changes with both Trial and Group. Values for these secondary parameters were obtained by averaging and are identified in Table 7.3, see below. Note that the standard error of the mean for the identified secondary values was quite low, further reinforcing how little they changed between subjects, trials, and group. Inspecting the trend in Figure 7.9, the primary

Parameter	Mean	SEM
τ_0 (s)	0.238	0.006
K	5.927	0.163
a	1.301	0.041
ω_{NM} (rad/s)	7.768	0.185
ζ_{NM} (rad/s)	0.707	0.001

Table 7.3: Identified optimal parameters of the Structural Model for the Aircraft Flight Task.

Group	A	B	C
Control (K_{eC})	-7.344	0.373	12.404
Feedback (K_{eF})	-67.323	1.994	17.402

Table 7.4: Identified K_e parameters from Equation 7.8 for the two groups.

parameter, K_e, can be modeled as an exponential function that increases with training. K_e was modeled for each group as a function of trial with three parameters, such that

$$K_e = Ae^{(-B*\text{Trial})} + C \tag{7.8}$$

where A is the scale factor, B describes how rapidly K_e changes with trial, and C is a baseline value. Best-fit values for these constants were obtained using least squares and are presented in Table 7.4. The resulting fit accurately represents the values identified by the value identification technique, and reflects what we have observed throughout this analysis—the feedback group learns much more quickly and performs better than the control group.

We propose that the K_f, the difference between the identified K_e of the control and feedback groups, determines the change in performance of subjects exposed to the feedback. We further propose that this K_f is a result of the accumulation of exposure to feedback, such that

$$K_f = K_{eF} - K_{eC} \tag{7.9}$$

$$K_f \propto \int \text{Feedback}(t)dt \tag{7.10}$$

$$\dot{K_f} \propto \text{Feedback}(t) \tag{7.11}$$

where K_{eC} and K_{eF} are the K_e variables identified from exponential fits to the K_e variable for the control and feedback groups, respectively. Here we have defined K_f to be proportional to the total accumulated time that operators were exposed to the feedback, and $\dot{K}_f$ is defined as the time rate of change in K_f trial-to-trial which results from the time exposed to the feedback in a given trial.

We can test this proposal by plotting the average time that subjects in the feedback group were exposed to the feedback in each trial compared to the value of $\dot{K}_f$ calculated from the derivative of the exponential fits. This comparison is available in Figure 7.10. Here we see that the data from the experiment is directly proportional to the modeled value $\dot{K}_f$, suggesting that our proposal is a viable explanation of the observed effects. $\dot{K}_f$, which can also be interpreted as the change in the amount of advantage that subjects in the feedback group have over the control group in a given trial, sharply decreases over the first few trials to a steady value. This confirms our experimental observation that subjects receive the majority of the benefit from feedback very rapidly, in just a few trials. Further analysis of

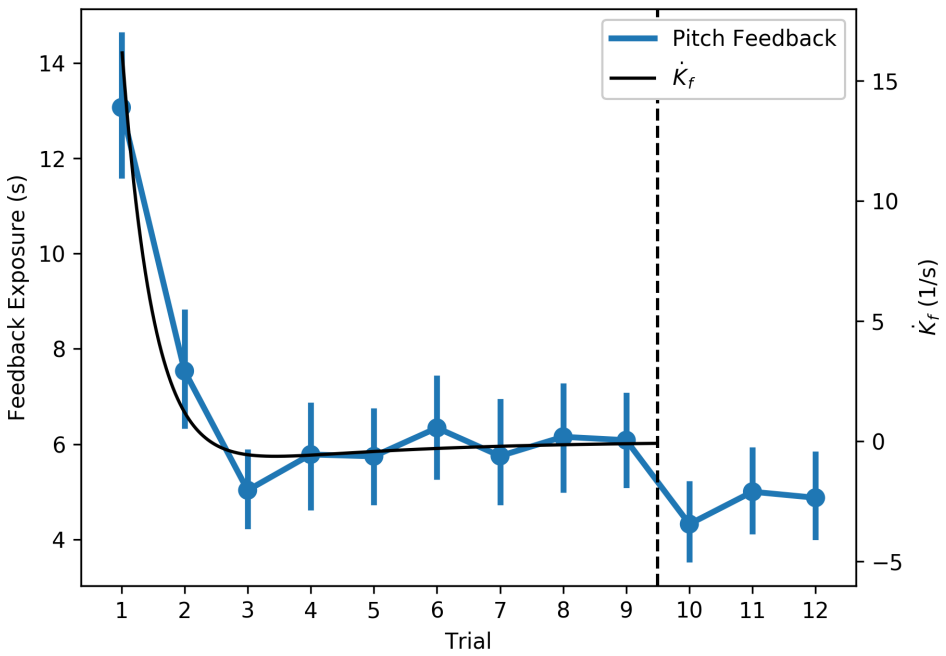
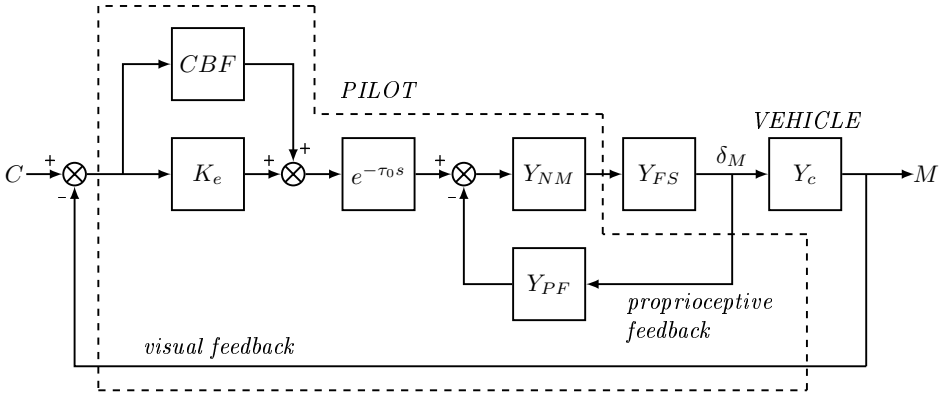
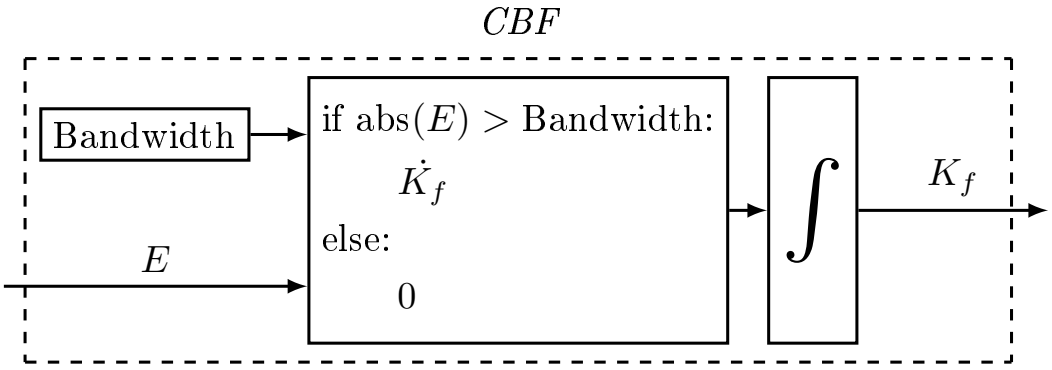


Figure 7.10: The feedback exposure time directly correlates with $\dot{K}_f$, which confirms that the accumulation of exposure to feedback drives K_f. The dashed line indicates where the feedback was removed. Error bars are the standard error of the mean.



(a) The Structural Model with concurrent bandwidth feedback.



(b) The adaptive logic of the addition to the Structural Model.

Figure 7.11: The proposed addition to the Structural Model to account for concurrent bandwidth feedback.

Figure 7.10 shows that, for this task, an exposure period of approximately 6 seconds (or $6/82 \approx 0.07\%$) in a trial leads to no change in performance. It is currently unclear if this percentage of time is task dependent or if, for example, it simply represents the average boundary of human control for this task. As a result of this analysis, the modified version of the Structural Model which incorporates the improvements derived in this analysis is available in Figure 7.11a.

7.5 Discussion

We were interested in using a pilot modeling approach to explain the differences in performance that we observed from the aircraft flight task study explored in Chapter 5. In this study, subjects repeatedly trained on flight tasks of increasing functional complexity, and one

group of subjects received visual concurrent bandwidth feedback. We explored two different modeling strategies, a time-domain autoregressive exogenous (ARX) based technique called *gettf1* and the Structural Model.

A ARX identification technique was used to estimate pilot transfer functions. The estimated combined pilot/vehicle open-loop transfer functions were evaluated to determine the evolution of the crossover frequency for each control loop throughout training. Results indicate that participants in the feedback group had a significantly lower root-mean-square error and higher crossover frequency than those in the control group, indicating better performance. The ARX identification technique provided a consistently high variance accounted for (VAF), indicating that it can identify transfer functions representing a variety of operators at various levels of training. These results are more pronounced for modes with increased functional complexity and persisted in retention testing when the feedback was removed, indicating that participants were not reliant on the feedback.

The Structural Model of the human pilot was also investigated, which "describe[s] the underlying structure which contributes to human pilot dynamics." The Structural Model is of interest for interpreting the effects of concurrent bandwidth feedback as it incorporates multiple sensory channels and models of visual acuity and the time-varying human pilot. We developed a novel approach for identifying the parameters in the Structural Model from experimental subject data. This approach was validated by comparing the crossover frequency identified from the ARX technique, which showed that both techniques identified a similar evolution of the crossover frequency with training and between groups. The result of a linear mixed effects model showed that, of the six parameters which were fit, only K_e, the leading visual gain, varied between both group and trial.

An extension to the Structural Model was proposed which explains the effects observed when operators are exposed to concurrent bandwidth feedback. This proposed model includes the addition of a single additional parameter, K_f, which accumulates when the operator is exposed to feedback. This parameter is added in the primary control loop next to the K_e term, and results in operators performing the task with a statistically lower root-mean-square error. Operators exposed to the feedback also immediately approach the limit of human control, with identified crossover frequency values of $\approx 1.7 - 2$ rad/s. We showed

that the Structural Model can accurately model adaptation during the training phase of a simulated flight task and incorporate exposure to concurrent bandwidth feedback.

Chapter 8

Conclusion

We laid out two aims for this research in Section 1.3 which were accomplished with this dissertation work:

Aim One Investigate the effects of concurrent bandwidth feedback on human performance and workload in complex manual control tasks.

Aim Two Extend the Structural Model of the human pilot to include the effects of concurrent bandwidth feedback.

We designed, recruited subjects for, and analyzed data from four human-in-the-loop experiments to address the first aim. Starting with a relatively simple compensatory tracking task, we then investigated an EMG controlled Fitts's task, then explored the effects of functional task complexity and the choice of feedback bandwidth in an aircraft flight task. Subjects exposed to the concurrent bandwidth feedback outperformed those in a control group in each of these experiments without reporting an increase in workload. Additionally, subjects also did not suffer from the guidance hypothesis when feedback was removed in retention trials, suggesting that they were able to use concurrent bandwidth feedback as a training tool and that they did not rely on it as a form of guidance.

To address the second aim, we extended the Structural Model of the human pilot to include the effects of concurrent bandwidth feedback and validated this model with the data from the experiment presented in Chapter 5: Feedback for Training Flight Tasks. This extended Structural Model is successfully able to predict pilot performance for a variety of bandwidths, and is achieved by the relatively straightforward addition of a single parameter.

8.1 Summary

Appropriate integration between automation and robotics systems and their human operators is essential for future space exploration. The Human Factors and Behavioral Performance Element of NASA's Human Research Program requires a systematic understanding of the critical human-automation/robotic integration, or HARI, design challenges for future space exploration. In Chapter 2: Trade Study, we present a trade study which reports the results of a systematic assessment of the spaceflight-relevant HARI technologies and research topics addressing critical gaps in knowledge, and prioritizes research required for successful human performance and HAR integration. We reviewed relevant literature across the past ten years and interviewed ten subject matter experts to investigate the current state of HARI technology, challenges facing development, the state of HARI research across a wide range of fields, and opportunities for advancing the state of the art through directed research. This information was used to identify relevant HARI technologies and research topics, as well as factors to assess relative priority of HARI technologies. We worked with NASA stakeholders to weight the factors relevant to assessing HARI specific technologies. A multi-dimensional trade analysis was performed to objectively score HARI research topics and specific technologies to recommend investment priorities for NASA.

Recent advances in computing hardware have enabled a new generation of mobile augmented reality devices which have the potential to improve human performance and reduce workload in a variety of tasks. The aim of Chapter 3: Augmented Reality Tracking Task was to investigate the effect of several factors on human performance and workload in a three-axis manual tracking task. Twenty-four engineering students at the University of California, Davis were randomly placed into a one of two device groups ($n = 12$ per group): a 2D or 3D display. Subjects in both groups evaluated three different displays in a random order: a baseline display, a concurrent bandwidth feedback display, and a rotated display. The objective performance of individual subjects was evaluated using the root-mean-square error (RMSE) of the depth (z) axis. Objective workload was measured by the response time to a two-choice task, and subjective workload was evaluated using the NASA Task Load Index (NASA-TLX). Results of ANOVA analysis on the z axis RMSE showed significant effects for design ($F(2, 36) = 84.92, p < 0.001$), device ($F(1, 18) = 7.22, p < 0.015$), and start

design ($F(2, 18) = 4.81, p < 0.021$). The ANOVA also showed a significant interaction effect between design and starting design ($F(4, 36) = 8.55, p < 0.0001$) and a three way interaction between design, device, and starting design ($F(4, 36) = 5.57, p < 0.002$). In general, there were no significant effects found for workload measurements. Both concurrent bandwidth feedback and a rotated display resulted in superior performance when compared to a baseline display. Providing a 3D display did not, in general, improve performance. Subjects in the 3D display group that had early exposure to the concurrent bandwidth feedback, however, were able to use the feedback to achieve superior performance.

It is critical to evaluate the effects of training methodologies on performance and the evolution of workload and trust in order to achieve a seamless human-automation system. In Chapter 4: Surface Electromyography Task, we developed a surface EMG control task to observe the effects of training methodology on the development of performance, workload, and trust. We were specifically interested in comparing the efficacy of concurrent and terminal feedback and a motor adaptation group. After evaluating 48 subjects in a Fitts's Law cursor-to-target task, we found that the concurrent feedback group significantly outperformed the control group. The feedback group immediately outperformed the control group, reaching nearly perfect performance in only two trials, and the control group could not match this performance until over an hour later. We found significant interaction effects when evaluating workload and trust, where the concurrent feedback group reported lower workload and higher trust when compared to the motor adaptation group. Subjects did not suffer from the guidance hypothesis when the feedback was removed, suggesting that they were not reliant on the feedback but instead used it merely as a tool for accelerating learning. By evaluating concurrent bandwidth feedback in the discrete task of using myoelectric control to control a cursor, we showed that this feedback can be extremely effective for a variety of tasks.

In Chapter 5: Feedback for Training Flight Tasks, the effects of concurrent bandwidth feedback on operator performance and workload were analyzed during training for an aircraft flight control task. In the experiment, participants completed a simulated flight task consisting of three complexity levels using traditional flight instruments. Thirty participants were divided into equal-sized control and feedback groups. The control group controlled simulated

aircraft motion with visual guidance for pitch, roll, and altitude provided by traditional flight instruments. The feedback group received additional visual concurrent bandwidth feedback for each controlled degree of freedom. For both groups, performance and workload measurements were evaluated to determine the effects of the feedback on subject learning rate and maximum skill level. To assess short-term retention of learned skill for the feedback group, the concurrent feedback was removed, and performance was again evaluated. Statistical analyses showed that participants in the feedback group immediately performed better than those in the control group, that the performance difference between the two groups was more pronounced for more complex tasks, and that final performance levels for the feedback group significantly exceeded that of the control group. We found that concurrent bandwidth feedback does not reduce workload in our flight tasks. For the short periods tested, participants continued to perform at the same performance and workload levels when the feedback was removed.

In Chapter 6: Feedback Bandwidth Study, we presented the preliminary results of a study designed to evaluate the effect of choosing the bandwidth for our concurrent bandwidth feedback on pilot flight performance and workload. We theorized that there is an optimal bandwidth which leads to peak performance and minimum subject workload, and that there are bandwidths which do not improve — and can actually degrade performance and workload compared to a no feedback condition. Based on the preliminary findings presented, it appears that choosing an appropriate bandwidth determines the efficacy of concurrent bandwidth feedback. The current data suggests that having a looser bandwidth further improves performance; a surprising finding considering wider thresholds result in subjects experiencing less feedback, and eventually lead to them effectively not experiencing any feedback at all. These findings are still preliminary, however, and we look forward to publishing full results when subject testing is completed.

Finally, in Chapter 7: Modeling the Effects of Feedback, the Structural Model of the human pilot was investigated to model the results of the aircraft feedback study to better understand how the subjects' performance was changed by exposure to the feedback. We estimated the values of the parameters in the model using a novel parameter identification technique and used the model to estimate the crossover frequency of the combined pilot/ve-

hicle system. The results of this technique were validated using the crossover frequencies identified by an ARX technique, the comparison with which showed good agreement. Linear mixed effect models were used to evaluate how each structural model parameter changed between groups over the course of the experiment. The result of this statistical test led us to propose the extension of the model to include a new term, K_f, which accumulates with exposure to feedback, and which is linearly added to the normal error sensing and gain compensation, K_e. We found evidence that this term accumulates when subjects are exposed to a sufficient amount of feedback over the course of a given trial and that including this term allows for the difference in performance between the control and feedback groups.

8.2 Research Questions

In Chapter 1.3, we listed six research questions that we sought to answer in this dissertation As we initiated this research, our primary questions were

1. Can concurrent bandwidth feedback (CBF) improve human performance in complex manual control tasks?
   (a) Can CBF reduce the required training time to peak performance?
   (b) Can CBF be removed after reaching peak performance without reducing subject performance (i.e., does the guidance hypothesis not hold)?
   (c) Can performance be increased without increasing workload?
2. Can we develop a model of human performance that includes the effects of concurrent bandwidth feedback?
   (a) Can we use this model to estimate operational limits?

Here we summarize our answers to these questions.

Can concurrent bandwidth feedback improve human performance in complex manual control tasks?

Yes. The results are presented in the four research studies in this dissertation: the augmented reality tracking task, surface electromyography task, and aircraft flight tasks tested over one hundred subjects. In each of these tasks, subjects exposed to concurrent bandwidth

feedback had improved performance in their control tasks compared to the control groups. The magnitude of performance improvement varied between the tasks but appears to increase with functional task complexity. Performance improvement varied between 17.8% and 44.2% for the aircraft flight task, with the largest benefit appearing in the most complicated task mode (see Table 5.1). In the augmented reality tracking task, subjects that trained with the feedback while wearing the augmented reality headset were able to better perform the task when the feedback was removed, which was not the case for subjects that did not wear the headset.

Can CBF reduce the required training time to peak performance?

Yes. Subjects exposed to concurrent bandwidth feedback generally performed better by their second trial than subjects in the control group did by the end of the experiments. This is especially clear in the percent of successful trials completed in the surface electromyography study. In this experiment, subjects exposed to CBF achieved a nearly perfect score by the second block of trials, a feat that subjects in the control group never achieved (see Figure 4.4). This was also the case in the aircraft study, where subjects saw immediate performance improvements after the first trial (see Figures 5.3, 5.4, and 5.5).

Can CBF be removed after reaching peak performance without reducing subject performance (i.e., does the guidance hypothesis not hold)?

Yes. Many forms of augmented feedback have been plagued by the "guidance hypothesis". This term describes when subjects become reliant on the augmented feedback provided during training such that they rely primarily upon the augmented feedback rather than other indicators available in the task, reducing their performance when the augmented feedback is removed. The concurrent bandwidth feedback used in these experiments does not result in the guidance hypothesis. This was hinted at early on in the augmented reality tracking task, where subjects that were initially exposed to the feedback continued to perform at the same level during their other conditions. While there was a small, nonsignificant reduction in performance during the EMG task when the feedback was removed (see Figure 4.4), we also saw a small, nonsignificant increase in performance when the feedback was removed in the aircraft task (see Figures 5.3). These results indicate that visual concurrent bandwidth feedback does not suffer from the guidance hypothesis, and that this form of augmented

feedback can be used to help train tasks of this type and difficulty.

The reason that this feedback is effective may be that it naturally fades away as subjects perform the task better (see Figure 7.10, for example, which shows the average feedback exposure time as a function of trial). Fading feedback has usually been considered effective, but knowing how to set the fading rate has continued to be an elusive issue. While bandwidth feedback of this type only replaces that issue with another (how does one best set the bandwidth for a given task?), it does provide an easy way to directly tie the fading rate to subject performance. Additionally, many tasks have established operational limits which can be used to set a bandwidth, providing a simple way to justify bandwidth selection.

Can performance be increased without increasing workload?

Yes. Subjective workload was measured using the NASA-TLX and/or Modified Bedford Workload Scores in all four of our experimental studies. No significant changes in workload were detected between the control and feedback conditions in any of our studies, though we were able to detect changes in workload as a function of training and functional complexity. This suggests that the augmented feedback developed in this work is exceptionally effective as **it can simultaneously increase performance without also increasing workload**.

Concurrent bandwidth feedback likely does not increase workload because it is only present on elements of the display that are already available to the control group. By pointing out elements of the display that subjects should be paying attention to at any given time (much in the way that an expert instructor might), we are simply redirecting attention to the most vital elements of the display. As we are not adding additional tasks for subjects to attend to; they can use the feedback to do more (perform better) without adding demands to their cognitive load.

Can we develop a model of human performance that includes the effects of concurrent bandwidth feedback?

Yes. In Chapter 7 we were able to develop a model of human performance based on Structural Model of the Human Pilot, which includes the effects of concurrent bandwidth feedback (Hess, 1997). This model modifies the Structural Model to include one additional gain, K_f, which accumulates as subjects are exposed to the concurrent bandwidth feedback. Evidence from the aircraft study was used to develop this model, which shows that the

amount of time exposed to the feedback in each trial leads to the change in the K_f term from trial to trial. Using our modified Structural Model, differences between groups of subjects with and without feedback can be successfully modeled as their performance changes through the training process.

Can we use this model to estimate operational limits?

Yes. The primary effect of increasing K_f is to increase the crossover frequency of the combined pilot/vehicle system. This means that pilots exposed to the concurrent bandwidth feedback effectively work harder without reporting increased workload. One can use this model to explore different feedback bandwidths and see the resulting effect it has on the crossover frequency. As the upper limit of human performance is around 2 rad/s, any bandwidth which results in a higher value would result in unsustainable levels of workload. Further research is required, however, to determine if $\dot{K}_f$ is task dependent.

8.3 Future Work

This research focused on the effect of one type of augmented feedback, visual concurrent bandwidth feedback, on human performance and workload in complex manual control tasks. We primarily investigated how feedback works with manual tracking tasks, where we showed that performance increases were positively correlated with functional task complexity. The largest performance gains occurred in the most complex tasks, where subjects used the feedback to improve their performance dramatically over very short periods of time when compared to subjects in the control group. An extension of this work could further investigate the utility of concurrent bandwidth feedback to improve performance in emergency or other off-nominal scenarios. If operators could use concurrent bandwidth feedback in actual operation, this research may have the potential to reduce aviation accidents by increasing operator situational awareness or reducing inflight workload. One limitation of this work, however, was that it focused only on inexperienced and untrained undergraduate students. Future work should investigate if concurrent bandwidth feedback can improve performance in well-trained operators. If feedback only works for naive subjects, then it may only be useful in the very early stages of training.

In this research, we showed that concurrent bandwidth feedback does indeed resemble the

presence of an experienced instructor rather than imposing additional guidance demands. As a result, concurrent bandwidth feedback did not cause subjects to suffer from the guidance hypothesis. While we investigated immediate retention, it is unclear how much performance benefit would be retained in longer term experiments. We also provided a uniform number of trials with feedback for all subjects and did not attempt to schedule or remove the feedback preemptively for subjects that were already performing well. Future work could investigate the minimum number of trials that a subject needs to be exposed to concurrent bandwidth feedback to retain a performance benefit.

We also did not investigate the effect of changing the acceptable performance bandwidth as subjects improved through training. It is possible that future work could investigate the effect of progressively narrowing the bandwidth as subjects improve their performance. This "adaptive bandwidth" could allow for further improvements in performance but may ultimately lead to impossible performance restraints if not scheduled carefully.

This research focused exclusively on feedback that was visually displayed to an operator while they were completing tasks. Additional strategies to provide multi-sensory feedback include audio, haptic, or multimodal feedback. Future researchers could investigate what the effect of the modality of concurrent bandwidth feedback has on subject performance and could identify if multimodal feedback has greater advantages over unimodal feedback. Multimodal feedback is generally thought to be more effective, but an experiment investigating the interaction between feedback modality and task complexity could show the optimal way to provide feedback. Our previous experiment with the SAFER vehicle (4 degree of freedom task) showed workload improvements, while the pitch, roll, and altitude task presented in this dissertation (3 degree of freedom task) did not. In addition to having one additional degree of freedom, the SAFER tasks also involved navigation. The aircraft flight task only involved stabilizing the vehicle and holding an altitude, however, and making direct comparisons between the two is difficult. Investigating tasks of increasing difficulty would also allow researchers to identify what level of task complexity is needed for subjects to report reduced levels of cognitive workload when exposed to feedback.

This work has spawned a number of new questions worth investigating. The questions brought up in this section include:

1. Can the concurrent bandwidth feedback be utilized in emergency scenarios to assist operators to quickly recover from off-nominal conditions?

2. Does the concurrent bandwidth feedback improve performance for well-trained operators, or does this technique only work for naive or inexperience operators?

3. What is the minimum number of trials that subjects need to be exposed to concurrent bandwidth feedback before it can be removed without degrading performance? (Can the path to "expertise" be shorted by concurrent bandwidth feedback?)

4. Is there an optimal way that the bandwidth can be scheduled as subjects improve through trials?

5. What is the effect of providing the concurrent bandwidth feedback through other modalities (audio, haptic, or a multimodal approach integrating these and/or visual)?

6. What level of functional complexity is needed to see reductions in workload for subjects exposed to the concurrent bandwidth feedback?

Bibliography

H. Admoni and B. Scassellati. Social Eye Gaze in Human-robot Interaction: A Review. *J. Hum.-Robot Interact.*, 6(1):25–63, May 2017. ISSN 2163-0364. doi: 10.5898/JHRI.6.1.Admoni. URL https://doi.org/10.5898/JHRI.6.1.Admoni.

M. Ahmad, O. Mubin, and J. Orlando. A Systematic Review of Adaptivity in Human-Robot Interaction. *Multimodal Technologies and Interaction*, 1(3):14, Sept. 2017. doi: 10.3390/mti1030014. URL https://www.mdpi.com/2414-4088/1/3/14.

A. Andreea-Irina and I. Achim. Prediction of the handling qualities and pilot-induced oscillation rating levels. *INCAS BULLETIN*, 6(Special 1):3–13, Apr. 2014. ISSN 20668201, 22474528. doi: 10.13111/2066-8201.2014.6.S1.1. URL http://bulletin.incas.ro/files/afloare_a__ionita_a__vol_6_spec_iss_1.pdf.

P. K. Artemiadis and K. J. Kyriakopoulos. An emg-based robot control scheme robust to time-varying emg signal features. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):582–588, May 2010. ISSN 1558-0032. doi: 10.1109/TITB.2010.2040832.

E. Bachelder, R. Hess, M. Godfroy-Cooper, and B. Aponso. Modeling Pilot Pulse Control. In *43rd European Rotorcraft Forum*, Milan, Italy, Sept. 2017. URL https://ntrs.nasa.gov/search.jsp?R=20170010679.

E. N. Bachelder, R. A. Hess, M. Godfroy-Cooper, and B. L. Aponso. Linking the Pilot Structural Model and Pilot Workload. In *2018 AIAA Atmospheric Flight Mechanics Conference*, Kissimmee, Florida, Jan. 2018. American Institute of Aeronautics and Astronautics. ISBN 978-1-62410-525-8. doi: 10.2514/6.2018-0533. URL https://arc.aiaa.org/doi/10.2514/6.2018-0533.

S. Baron, D. L. Kleinman, and W. H. Levison. An optimal control model of human response part II: Prediction of human performance in a complex task. *Automatica*, 6(3):371–383, May 1970. ISSN 0005-1098. doi: 10.1016/0005-1098(70)90052-X. URL http://www.sciencedirect.com/science/article/pii/000510987090052X.

J. V. Basmajian. Control and training of individual motor units. *Science*, 141(3579):440–441, 1963. ISSN 0036-8075. doi: 10.1126/science.141.3579.440.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48, 2015. ISSN 1548-7660. doi: 10.18637/jss.v067.i01.

J. M. Beer, A. D. Fisk, and W. A. Rogers. Toward a Framework for Levels of Robot Autonomy in Human-robot Interaction. *J. Hum.-Robot Interact.*, 3(2):74–99, July 2014. ISSN 2163-0364. doi: 10.5898/JHRI.3.2.Beer. URL https://doi.org/10.5898/JHRI.3.2.Beer.

C. Boppe. Program/project decision-aiding methodology training. In *Charles Stark Draper Laboratory*, May 2010.

D. A. Braun, A. Aertsen, D. M. Wolpert, and C. Mehring. Motor task variation induces structural learning. *Current Biology*, 19(4):352 – 357, 2009. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2009.01.036.

A. W. Bronkhorst, J. A. H. Veltman, and L. V. Breda. Application of a three-dimensional auditory display in a flight task. *Human Factors*, 38(1):23–33, 1996. doi: 10.1518/001872096778940859. PMID: 8682519.

J. Chen, M. Glover, C. Li, and C. Yang. Development of a user experience enhanced teleoperation approach. In *2016 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 171–177, Aug 2016. doi: 10.1109/ICARM.2016.7606914.

J. Y. Chen and M. J. Barnes. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, 44(1):13–29, 2014.

M. J. Cler and C. E. Stepp. Discrete versus continuous mapping of facial electromyography for human–machine interface control: Performance and training effects. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(4):572–580, July 2015. ISSN 1558-0210. doi: 10.1109/TNSRE.2015.2391054.

D. O. Cote, B. H. Williges, and R. C. Williges. Augmented feedback in adaptive motor skill training. *Proceedings of the Human Factors Society Annual Meeting*, 22(1): 105–109, 1978. doi: 10.1177/107118137802200127. URL https://doi.org/10.1177/107118137802200127.

H. P. Crowell and I. S. Davis. Gait retraining to reduce lower extremity loading in runners. *Clinical Biomechanics*, 26(1):78 – 83, 2011. ISSN 0268-0033. doi: https://doi.org/10.1016/j.clinbiomech.2010.09.003. URL http://www.sciencedirect.com/science/article/pii/S0268003310002512.

K. E. Culley and P. Madhavan. A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior*, 29(3):577–579, May 2013. ISSN 0747-5632. doi: 10.1016/j.chb.2012.11.023. URL http://www.sciencedirect.com/science/article/pii/S0747563212003287.

S. de Groot, J. C. de Winter, J. M. L. García, M. Mulder, and P. A. Wieringa. The effect of concurrent bandwidth feedback on learning the lane-keeping task in a driving simulator. *Human factors*, 53(1):50–62, 2011.

E. J. de Visser, S. S. Monfort, R. McKendrick, M. A. B. Smith, P. E. McKnight, F. Krueger, and R. Parasuraman. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331–349, 2016. ISSN 1939-2192(Electronic),1076-898X(Print). doi: 10.1037/xap0000092.

S. Dixon, E. Fitzhugh, and D. Aleva. Human factors guidelines for applications of 3d perspectives: a literature review. In *Display Technologies and Applications for Defense, Security, and Avionics III*, volume 7327, page 73270K. International Society for Optics and Photonics, May 2009. doi: 10.1117/12.820853. URL

https://www.spiedigitallibrary.org/conference-proceedings-of-spie/7327/73270K/Human-factors-guidelines-for-applications-of-3D-perspectives--a/10.1117/12.820853.short.

S. Dosen, M. Markovic, M. Strbac, M. Belić, V. Kojić, G. Bijelić, T. Keller, and D. Farina. Multichannel electrotactile feedback with spatial and mixed coding for closed-loop control of grasping force in hand prostheses. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(3):183–195, March 2017. ISSN 1558-0210. doi: 10.1109/TNSRE.2016.2550864.

M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6): 697 – 718, 2003. ISSN 1071-5819. doi: https://doi.org/10.1016/S1071-5819(03)00038-7. Trust and Technology.

G. Ellis and A. Roscoe. The airline pilot's view of flight deck workload: A preliminary study using a questionnaire. Technical report, ROYAL AIRCRAFT ESTABLISHMENT FARNBOROUGH (ENGLAND), 1982.

M. R. Endsley. Toward a theory of situation awareness in dynamic systems. In *Situational Awareness*, pages 9–42. Routledge, 2017a.

M. R. Endsley. From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors*, 59(1):5–27, Feb. 2017b. ISSN 0018-7208. doi: 10.1177/0018720816681350. URL https://doi.org/10.1177/0018720816681350.

Federal Aviation Administration. Press Release – FAA Boosts Aviation Safety with New Pilot Qualification Standards. https://www.faa.gov/news/press_releases/news_story.cfm?newsId=14838, 2013. [Online; accessed 21-Jan-2020].

R. Fischer and J. Miller. Differential redundancy gain in onset detection versus offset detection. *Perception & Psychophysics*, 70(3):431–436, Apr. 2008. ISSN 1532-5962. doi: 10.3758/PP.70.3.431. URL https://doi.org/10.3758/PP.70.3.431.

P. M. Fitts, editor. *Human engineering for an effective air-navigation and traffic-control system*. Human engineering for an effective air-navigation and traffic-control system. National Research Council, Div. of, Oxford, England, 1951.

P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954. ISSN 0022-1015.

V. J. Gawron. *Human Performance, Workload, and Situational Awareness Measures Handbook*. CRC Press, Mar. 2008. ISBN 978-0-429-14984-9. doi: 10.1201/9781420064506. URL https://www.taylorfrancis.com/books/9780429149849.

M. Gestwa and J.-M. Bauschat. On the Modelling of a Human Pilot Using Fuzzy Logic Control. *Computational Intelligence in Control*, pages 148–167, 2003. doi: 10.4018/978-1-59140-037-0.ch009. URL https://www.igi-global.com/chapter/modelling-human-pilot-using-fuzzy/6836.

N. B. Gordon and M. J. Gottlieb. Effect of supplemental visual cues on rotary pursuit. *Journal of Experimental Psychology*, 75(4):566–568, 1967. ISSN 0022-1015(Print). doi: 10.1037/h0025122.

O. L. Grant, K. A. Stol, A. Swain, and R. A. Hess. Handling Qualities of a Twin Ducted-Fan Aircraft: An Analytical Evaluation. *Journal of Guidance, Control, and Dynamics*, 38(6): 1126–1131, 2015. ISSN 0731-5090. doi: 10.2514/1.G000826. URL https://doi.org/10.2514/1.G000826.

M. J. Griffin. The validation of biodynamic models. *Clinical Biomechanics*, 16:S81–S92, Jan. 2001. ISSN 0268-0033. doi: 10.1016/S0268-0033(00)00101-7. URL http://www.sciencedirect.com/science/article/pii/S0268003300001017.

M. A. Guadagnoli and T. D. Lee. Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior*, 36 (2):212–224, 2004. doi: 10.3200/JMBR.36.2.212-224. URL https://doi.org/10.3200/JMBR.36.2.212-224. PMID: 15130871.

J. Guiochet, M. Machin, and H. Waeselynck. Safety-critical advanced robots: A survey. *Robotics and Autonomous Systems*, 94:43–52, Aug. 2017. ISSN 0921-8890. doi: 10.1016/j.robot.2017.04.004. URL http://www.sciencedirect.com/science/article/pii/S0921889016300768.

C. J. Hainley, K. R. Duda, C. M. Oman, and A. Natapoff. Pilot Performance, Workload, and Situation Awareness During Lunar Landing Mode Transitions. *Journal of Spacecraft and Rockets*, 50(4):793–801, 2013. ISSN 0022-4650. doi: 10.2514/1.A32267. URL https://doi.org/10.2514/1.A32267.

S. G. Hart. NASA-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, Oct. 2006. ISSN 1541-9312. doi: 10.1177/154193120605000909. URL https://doi.org/10.1177/154193120605000909.

S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock and N. Meshkati, editors, *Advances in Psychology*, volume 52 of *Human Mental Workload*, pages 139–183. North-Holland, Jan. 1988. doi: 10.1016/S0166-4115(08)62386-9. URL http://www.sciencedirect.com/science/article/pii/S0166411508623869.

Z. He. Gesture recognition based on tri-axis accelerometer using 1d gabor filters. In *2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*, pages 146–151, Dec 2018. doi: 10.1109/PAAP.2018.00033.

R. K. Heffley and W. F. Jewell. Aircraft handling qualities data. Technical Report AD-A277031, NASA-CR-2144, NASA, December 1972.

R. A. Hess. Structural Model of the Adaptive Human Pilot. *Journal of Guidance, Control, and Dynamics*, 3(5):416–423, Sept. 1980. ISSN 0731-5090. doi: 10.2514/3.56015. URL https://arc.aiaa.org/doi/10.2514/3.56015.

R. A. Hess. Analysis of aircraft attitude control systems prone to pilot-induced oscillations. *Journal of Guidance, Control, and Dynamics*, 7(1):106–112, 1984a.

R. A. Hess. Effects of time delays on systems subject to manual control. *Journal of Guidance, Control, and Dynamics*, 7(4):416–421, July 1984b. ISSN 0731-5090. doi: 10.2514/3.19872. URL https://arc.aiaa.org/doi/10.2514/3.19872.

R. A. Hess. Model for human use of motion cues in vehicular control. *Journal of Guidance, Control, and Dynamics*, 13(3):476–482, May 1990. ISSN 0731-5090. doi: 10.2514/3.25360. URL https://arc.aiaa.org/doi/10.2514/3.25360.

R. A. Hess. Unified Theory for Aircraft Handling Qualities and Adverse Aircraft-Pilot Coupling. *Journal of Guidance, Control, and Dynamics*, 20(6):1141–1148, 1997. ISSN 0731-5090. doi: 10.2514/2.4169. URL https://doi.org/10.2514/2.4169.

R. A. Hess. Modeling Pilot Control Behavior with Sudden Changes in Vehicle Dynamics. *Journal of Aircraft*, 46(5):1584–1592, Sept. 2009. ISSN 0021-8669. doi: 10.2514/1.41215. URL https://arc.aiaa.org/doi/10.2514/1.41215.

R. A. Hess. Modeling Human Pilot Adaptation to Flight Control Anomalies and Changing Task Demands. *Journal of Guidance, Control, and Dynamics*, 39(3):655–666, 2016. ISSN 0731-5090. doi: 10.2514/1.G001303. URL https://doi.org/10.2514/1.G001303.

R. A. Hess and R. Joyce. Analytical Investigation of Transport Aircraft Handling Qualities. In *AIAA Atmospheric Flight Mechanics (AFM) Conference*, Boston, MA, Aug. 2013. American Institute of Aeronautics and Astronautics. doi: 10.2514/6.2013-4505. URL http://arc.aiaa.org/doi/10.2514/6.2013-4505.

R. A. Hess, Y. Zeyada, and R. K. Heffley. Modeling and simulation for helicopter task analysis. *Journal of the American Helicopter Society*, 47(4):243–252, 2002.

N. J. Hodges and A. M. Williams. *Skill Acquisition in Sport: Research, Theory and Practice*. Routledge, 2020.

K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570. PMID: 25875432.

R. Hosman and H. Stassen. Pilot's perception in the control of aircraft motions. *Control Engineering Practice*, 7(11):1421–1428, Nov. 1999. ISSN 0967-0661. doi: 10.1016/S0967-0661(99)00111-2. URL http://www.sciencedirect.com/science/article/pii/S0967066199001112.

R. J. Hosman. *Pilot's Perception and Control of Aircraft Motions*. PhD Thesis, Ph. D. Thesis, Delft University of Technology, Delft, The Netherlands, 1996.

H. Huang, S. L. Wolf, and J. He. Recent developments in biofeedback for neuromotor rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 3(1):11, Jun 2006. ISSN 1743-0003. doi: 10.1186/1743-0003-3-11.

M. Huet, D. M. Jacobs, C. Camachon, C. Goulon, and G. Montagne. Self-controlled concurrent feedback facilitates the learning of the final approach phase in a fixed-base flight simulator. *Human Factors*, 51(6):858–871, 2009. doi: 10.1177/0018720809357343. URL https://doi.org/10.1177/0018720809357343. PMID: 20415160.

Human Research Program. HRR - Risk - Risk of Inadequate Design of Human and Automation/Robotic Integration. https://humanresearchroadmap.nasa.gov/Risks/risk.aspx?i=163, 2011. [Online; accessed 05-Aug-2019].

I. Hussain, G. Salvietti, G. Spagnoletti, and D. Prattichizzo. The soft-sixthfinger: a wearable emg controlled robotic extra-finger for grasp compensation in chronic stroke patients. *IEEE Robotics and Automation Letters*, 1(2):1000–1006, July 2016. ISSN 2377-3774. doi: 10.1109/LRA.2016.2530793.

I. Hussain, G. Spagnoletti, G. Salvietti, and D. Prattichizzo. An emg interface for the control of motion and compliance of a supernumerary robotic finger. *Frontiers in Neurorobotics*, 10:18, 2016. ISSN 1662-5218. doi: 10.3389/fnbot.2016.00018.

J.-Y. Jian, A. M. Bisantz, and C. G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 2000. ISSN 1088-6362.

R. E. Johnson, K. P. Kording, L. J. Hargrove, and J. W. Sensinger. Does emg control lead to distinct motor adaptation? *Frontiers in Neuroscience*, 8:302, 2014. ISSN 1662-453X. doi: 10.3389/fnins.2014.00302.

D. B. Kaber, C. M. Perry, N. Segall, C. K. McClernon, and L. J. Prinzel. Situation awareness implications of adaptive automation for information processing in an air traffic control-related task. *International Journal of Industrial Ergonomics*, 36(5):447–462, May 2006. ISSN 0169-8141. doi: 10.1016/j.ergon.2006.01.008. URL http://www.sciencedirect.com/science/article/pii/S0169814106000229.

J. Karasinski, S. Robinson, P. Handley, and K. Duda. Real-Time Performance Feedback in a Manually-Controlled Spacecraft Inspection Task. In *AIAA Modeling and Simulation Technologies Conference*, Grapevine, Texas, Jan. 2017. American Institute of Aeronautics and Astronautics. ISBN 978-1-62410-451-0. doi: 10.2514/6.2017-1314. URL http://arc.aiaa.org/doi/10.2514/6.2017-1314.

J. A. Karasinski. Real-Time Performance Feedback for the Manual Control of Spacecraft. Master's thesis, University of California, Davis, United States – California, 2016. URL https://search.proquest.com/docview/1872339216/abstract/1D8D65D9AB6E4725PQ/1.

J. A. Karasinski and S. K. Robinson. Utility of concurrent bandwidth feedback in training aircraft flight tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1):1729–1733, 2019a. doi: 10.1177/1071181319631097.

J. A. Karasinski and S. K. Robinson. Evaluating Augmented Reality in a Three-Axis Manual Tracking Task. In *AIAA Scitech 2019 Forum*, San Diego, California, Jan. 2019b. American Institute of Aeronautics and Astronautics. ISBN 978-1-62410-578-4. doi: 10.2514/6.2019-1227. URL https://arc.aiaa.org/doi/10.2514/6.2019-1227.

J. A. Karasinski, S. K. Robinson, K. R. Duda, and Z. Prasov. Development of real-time performance metrics for manually-guided spacecraft operations. In *2016 IEEE Aerospace Conference*, pages 1–9, Mar. 2016. doi: 10.1109/AERO.2016.7500734.

B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg. A Survey of Research on Cloud Robotics and Automation. *IEEE Transactions on Automation Science and Engineering*, 12(2):398–409, Apr. 2015. ISSN 1545-5955. doi: 10.1109/TASE.2014.2376492.

W. Kim, F. Tendick, and L. Stark. Visual enhancements in pick-and-place tasks: Human operators controlling a simulated cylindrical manipulator. *IEEE Journal on Robotics and Automation*, 5(3):418–425, 1987a. ISSN 0882-4967. doi: 10.1109/JRA.1987.1087127. URL https://www.infona.pl//resource/bwmeta1.element.ieee-art-000001087127.

W. S. Kim, S. R. Ellis, M. E. Tyler, B. Hannaford, and L. W. Stark. Quantitative Evaluation of Perspective and Stereoscopic Displays in Three-Axis Manual Tracking Tasks. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(1):61–72, Jan. 1987b. ISSN 0018-9472. doi: 10.1109/TSMC.1987.289333.

R. G. Kinkade. A differential influence of augmented feedback on learning and on performance. Technical report, OHIO STATE UNIV RESEARCH FOUNDATION COLUMBUS LAB OF AVIATION PSYCHOLOGY, 1963.

D. L. Kleinman, S. Baron, and W. H. Levison. An optimal control model of human response part I: Theory and validation. *Automatica*, 6(3):357–369, May 1970. ISSN 0005-1098. doi: 10.1016/0005-1098(70)90051-8. URL http://www.sciencedirect.com/science/article/pii/0005109870900518.

A. Kolling, P. Walker, N. Chakraborty, K. Sycara, and M. Lewis. Human Interaction With Robot Swarms: A Survey. *IEEE Transactions on Human-Machine Systems*, 46(1):9–26, Feb. 2016. ISSN 2168-2291. doi: 10.1109/THMS.2015.2480801.

A. J. Kovacs and C. H. Shea. The learning of 90° continuous relative phase with and without lissajous feedback: External and internally generated bimanual coordination. *Acta Psychologica*, 136(3):311 – 320, 2011. ISSN 0001-6918. doi: https://doi.org/10.1016/j.actpsy.2010.12.004. URL http://www.sciencedirect.com/science/article/pii/S000169181000226X.

D. Laurillard. *Rethinking University Teaching : A Conversational Framework for the Effective Use of Learning Technologies*. Routledge, Sept. 2002. ISBN 978-0-203-16032-9. doi: 10.4324/9780203160329. URL https://www.taylorfrancis.com/books/9780203160329.

J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392. PMID: 15151155.

G. Lintern. Transfer of landing skill after training with supplementary visual cues. *Human Factors*, 22(1):81–88, 1980. doi: 10.1177/001872088002200109. URL https://doi.org/10.1177/001872088002200109. PMID: 7364448.

G. Lintern, S. N. Roscoe, J. M. Koonce, and L. D. Segal. Transfer of landing skills in beginning flight training. *Human Factors*, 32(3):319–327, 1990. doi: 10.1177/001872089003200305. URL https://doi.org/10.1177/001872089003200305.

G. Lintern, H. L. Taylor, J. M. Koonce, R. H. Kaiser, and G. A. Morrison. Transfer and quasi-transfer effects of scene detail and visual augmentation in landing training. *The International Journal of Aviation Psychology*, 7(2):149–169, 1997. doi: 10.1207/s15327108ijap0702_4.

H. Liu and L. Wang. Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, 68:355–367, Nov. 2018. ISSN 0169-8141. doi: 10.1016/j.ergon.2017.02.004. URL http://www.sciencedirect.com/science/article/pii/S0169814117300690.

D. P. Losey, C. G. McDonald, E. Battaglia, and M. K. O'Malley. A Review of Intent Detection, Arbitration, and Communication Aspects of Shared Control for Physical Human–Robot Interaction. *Applied Mechanics Reviews*, 70(1):010804–010804–19, Feb. 2018. ISSN 0003-6900. doi: 10.1115/1.4039145. URL http://dx.doi.org/10.1115/1.4039145.

Z. Lu, R. Happee, C. D. Cabrall, M. Kyriakidis, and J. C. de Winter. Human factors of transitions in automated driving: A general framework and literature survey. *Transportation research part F: traffic psychology and behaviour*, 43:183–198, 2016.

K. R. Lyons and S. S. Joshi. Real-time evaluation of a myoelectric control method for high-level upper limb amputees based on homologous leg movements. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6365–6368, Aug 2016. doi: 10.1109/EMBC.2016.7592184.

R. K. Lyons and S. S. Joshi. Effects of mapping uncertainty on visuomotor adaptation to trial-by-trial perturbations with proportional myoelectric control. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5178–5181, July 2018. doi: 10.1109/EMBC.2018.8513412.

R. J. Lysaght, S. G. Hill, A. O. Dick, B. D. Plamondon, and P. M. Linton. Operator Workload: Comprehensive Review and Evaluation of Operator Workload Methodologies. Technical Report TR-2075-3, ANALYTICS INC WILLOW GROVE PA, June 1989. URL https://apps.dtic.mil/docs/citations/ADA212879.

M. Markovic, H. Karnal, B. Graimann, D. Farina, and S. Dosen. GLIMPSE: Google glass interface for sensory feedback in myoelectric hand prostheses. *Journal of Neural Engineering*, 14(3):036007, mar 2017. doi: 10.1088/1741-2552/aa620a.

J. J. Marquez, B. D. Adelstein, M. L. Chang, S. R. Ellis, K. A. Hambuchen, and R. L. Howard. Future exploration missions' tasks associated with the risk of inadequate design

of human and automation/robotic integration. Technical Report NASA/TM-2017-219516, ARC-E-DAA-TN40802, NASA, 2017.

J. P. McIntire, P. R. Havig, and E. E. Geiselman. Stereoscopic 3d displays and human performance: A comprehensive review. *Displays*, 35(1):18–26, Jan. 2014. ISSN 0141-9382. doi: 10.1016/j.displa.2013.10.004. URL http://www.sciencedirect.com/science/article/pii/S0141938213000929.

D. T. McRuer and E. S. Krendel. DYNAMIC RESPONSE OF HUMAN OPERATORS. Technical report, KELSEY-HAYES CO INGLEWOOD CA CONTROL SPECIALISTS DIV, Oct. 1957. URL https://apps.dtic.mil/docs/citations/AD0110693.

D. T. McRuer and E. S. Krendel. Human pilot dynamics in compensatory systems. Technical report, SYSTEMS TECHNOLOGY INC HAWTHORNE CA, July 1965. URL https://apps.dtic.mil/docs/citations/AD0470337.

D. T. McRuer and E. S. Krendel. Mathematical Models of Human Pilot Behavior. Technical Report AGARD-AG-188, ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT NEUILLY-SUR-SEINE (FRANCE), Jan. 1974. URL https://apps.dtic.mil/docs/citations/AD0775905.

P. Mohapatra. Campus Guidance on Reducing On-Campus Research Activities Due to COVID-19 (March 17, 2020), Mar 2020. URL https://research.ucdavis.edu/campus-guidance-on-reducing-on-campus-research-activities-due-to-covid-19-2/.

B. M. Muir and N. Moray. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, 1996.

N. Naikar. Perspective Displays: A Review of Human Factors Issues. Technical Report DSTO-TR-0630, DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION CANBERRA (AUSTRALIA), Feb. 1998. URL https://apps.dtic.mil/docs/citations/ADA360645.

D. A. Norman and S. W. Draper. *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1986. ISBN 978-0-89859-781-3.

S. M. O'Meara, M. C. Shyr, K. R. Lyons, and S. S. Joshi. Comparing two different cursor control methods which use single-site surface electromyography*. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1163–1166, March 2019. doi: 10.1109/NER.2019.8716903.

S. M. O'Meara, J. A. Karasinski, C. L. Miller, S. Joshi, and S. K. Robinson. The effects of training methodology on performance, workload, and trust during human learning of a computer-based task. In *AIAA Scitech 2020 Forum*, Orlando, Florida, Jan. 2020. American Institute of Aeronautics and Astronautics. doi: 10.2514/6.2020-1110.

S. Ososky, D. Schuster, E. Phillips, and F. G. Jentsch. Building Appropriate Trust in Human-Robot Teams. In *2013 AAAI Spring Symposium Series*, Mar. 2013. URL https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5784.

R. Pak, A. C. McLaughlin, and B. Bass. A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults. *Ergonomics*, 57(9):1277–1289, Sept. 2014. ISSN 0014-0139. doi: 10.1080/00140139.2014.928750. URL https://doi.org/10.1080/00140139.2014.928750.

R. Parasuraman and C. D. Wickens. Humans: Still vital after all these years of automation. *Human factors*, 50(3):511–520, 2008.

R. B. Payne and G. T. Hauty. Effect of psychological feedback upon work decrement. *Journal of Experimental Psychology*, 50(6):343–351, 1955. ISSN 0022-1015(Print). doi: 10.1037/h0045068.

E. Phillips, K. E. Schaefer, D. R. Billings, F. Jentsch, and P. A. Hancock. Human-animal Teams As an Analog for Future Human-robot Teams: Influencing Design and Fostering Trust. *J. Hum.-Robot Interact.*, 5(1):100–125, Mar. 2016. ISSN 2163-0364. doi: 10.5898/JHRI.5.1.Phillips. URL https://doi.org/10.5898/JHRI.5.1.Phillips.

A. Ramaprasad. On the definition of feedback. *Behavioral Science*, 28(1):4–13, 1983. ISSN 1099-1743. doi: 10.1002/bs.3830280103. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bs.3830280103.

S. S. Rautaray and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, Jan. 2015. ISSN 1573-7462. doi: 10.1007/s10462-012-9356-9. URL https://doi.org/10.1007/s10462-012-9356-9.

D. C. Ribeiro, G. Sole, J. H. Abbott, and S. Milosavljevic. Extrinsic feedback and management of low back pain: A critical review of the literature. *Manual Therapy*, 16(3):231 – 239, 2011. ISSN 1356-689X. doi: https://doi.org/10.1016/j.math.2010.12.001. URL http://www.sciencedirect.com/science/article/pii/S1356689X10002201.

A. H. Roscoe and G. A. Ellis. A Subjective Rating Scale for Assessing Pilot Workload in Flight: A decade of Practical Use. Technical Report RAE-TR-90019, Royal Aerospace Establishment Farnborough (United Kingdom), Mar. 1990. URL https://apps.dtic.mil/docs/citations/ADA227864.

K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum*, 34(6):299–326, 2015. ISSN 1467-8659. doi: 10.1111/cgf.12603. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12603.

A. W. Salmoni, R. A. Schmidt, and C. B. Walter. Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin*, 95(3):355–386, 1984. ISSN 1939-1455(Electronic),0033-2909(Print). doi: 10.1037/0033-2909.95.3.355.

M. Y. Saraiji, T. Sasaki, K. Kunze, K. Minamizawa, and M. Inami. Metaarms: Body remapping using feet-controlled artificial arms. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, pages 65–74, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5948-1. doi: 10.1145/3242587.3242665.

K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, 58(3):377–400, May 2016. ISSN 0018-7208. doi: 10.1177/0018720816634228. URL https://doi.org/10.1177/0018720816634228.

M. A. Schweisfurth, M. Markovic, S. Dosen, F. Teich, B. Graimann, and D. Farina. Electrotactile EMG feedback improves the control of prosthesis grasping force. *Journal of Neural Engineering*, 13(5):056010, aug 2016. doi: 10.1088/1741-2560/13/5/056010.

B. D. Seppelt and J. D. Lee. Keeping the driver in the loop: Dynamic feedback to support appropriate use of imperfect vehicle control automation. *International Journal of Human-Computer Studies*, 125:66 – 80, 2019. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2018.12.009.

A. W. Shehata, L. F. Engels, M. Controzzi, C. Cipriani, E. J. Scheme, and J. W. Sensinger. Improving internal model strength and performance of prosthetic hands using augmented feedback. *Journal of NeuroEngineering and Rehabilitation*, 15(1):70, 2018. ISSN 1743-0003. doi: 10.1186/s12984-018-0417-4.

T. B. Sheridan. Human–Robot Interaction: Status and Challenges. *Human Factors*, 58 (4):525–532, June 2016. ISSN 0018-7208. doi: 10.1177/0018720816644364. URL https://doi.org/10.1177/0018720816644364.

R. Sigrist, G. Rauter, R. Riener, and P. Wolf. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic Bulletin & Review*, 20 (1):21–53, Feb. 2013. ISSN 1531-5320. doi: 10.3758/s13423-012-0333-8. URL https://doi.org/10.3758/s13423-012-0333-8.

P. Simoens, M. Dragone, and A. Saffiotti. The Internet of Robotic Things: A review of concept, added value and applications. *International Journal of Advanced Robotic Systems*, 15(1):1729881418759424, Jan. 2018. ISSN 1729-8814. doi: 10.1177/1729881418759424. URL https://doi.org/10.1177/1729881418759424.

G. Singh, R. N. Roy, and C. Ponzoni Carvalho Chanel. Towards Multi-UAV and Human Interaction Driving System Exploiting Human Mental State Estimation. In *10th International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS 2019)*, Prague, Czech Republic, Feb. 2019.

R. Singh, T. Miller, J. Newn, L. Sonenberg, E. Velloso, and F. Vetere. Combining planning with gaze for online human intention recognition. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 488–496, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems. URL http://dl.acm.org/citation.cfm?id=3237383.3237457.

I.-M. Skavhaug, K. R. Lyons, A. Nemchuk, S. D. Muroff, and S. S. Joshi. Learning to modulate the partial powers of a single semg power spectrum through a novel human–computer interface. *Human Movement Science*, 47:60 – 69, 2016. ISSN 0167-9457. doi: https://doi.org/10.1016/j.humov.2015.12.003.

G. K. Slocum, B. H. Williges, S. N. Roscoe, and H. Stanley. Meaningful shape coding for aircraft switch knobs. *Aviation Research Monographs*, 1(3):27–40, 1971.

H. S. Smallman, E. Schiller, and M. B. Cowen. Track Location Enhancements for Perspective View Displays. Technical Report SSC-TR-1847, SPACE AND NAVAL WARFARE SYSTEMS CENTER SAN DIEGO CA, Dec. 2000. URL https://apps.dtic.mil/docs/citations/ADA386408.

R. Stanton, L. Ada, C. M. Dean, and E. Preston. Biofeedback improves performance in lower limb activities more than usual therapy in people following stroke: a systematic review. *Journal of Physiotherapy*, 63(1):11 – 16, 2017. ISSN 1836-9553. doi: https://doi.org/10.1016/j.jphys.2016.11.006.

B. L. Stevens, F. L. Lewis, and E. N. Johnson. *Aircraft control and simulation: dynamics, controls design, and autonomous systems*. John Wiley & Sons, 2015.

R. S. Tannen, W. T. Nelson, R. S. Bolia, J. S. Warm, and W. N. Dember. Evaluating adaptive multisensory displays for target localization in a flight task. *The International Journal of Aviation Psychology*, 14(3):297–312, 2004. doi: 10.1207/s15327108ijap1403_5.

G. L. Teper. Aircraft stability and control data. Technical Report N69-31783, NASA CR-96008, NASA, April 1969.

E. L. Thorndike. The Law of Effect. *The American Journal of Psychology*, 39(1/4):212–222, 1927. ISSN 0002-9556. doi: 10.2307/1415413. URL https://www.jstor.org/stable/1415413.

A. A. Timmermans, H. A. Seelen, R. D. Willmann, and H. Kingma. Technology-assisted training of arm-hand skills in stroke: concepts on reacquisition of motor control and therapist guidelines for rehabilitation technology design. *Journal of NeuroEngineering and Rehabilitation*, 6(1):1, Jan. 2009. ISSN 1743-0003. doi: 10.1186/1743-0003-6-1. URL https://doi.org/10.1186/1743-0003-6-1.

E. Todorov, R. Shadmehr, and E. Bizzi. Augmented feedback presented in a virtual environment accelerates learning of a difficult motor task. *Journal of Motor Behavior*, 29 (2):147–158, 1997. doi: 10.1080/00222899709600829. URL https://doi.org/10.1080/00222899709600829. PMID: 12453791.

P. Tsarouchi, S. Makris, and G. Chryssolouris. Human–robot interaction review and challenges on task planning and programming. *International Journal of Computer Integrated Manufacturing*, 29(8):916–931, Aug. 2016. ISSN 0951-192X. doi: 10.1080/0951192X.2015.1130251. URL https://doi.org/10.1080/0951192X.2015.1130251.

A. Tustin. An investigation of the operator's response in manual control of a power driven gun. *C.S. Memorandum*, 169, 1944.

A. Tustin. A method of analysing the behaviour of linear systems in terms of time series. *Journal of the Institution of Electrical Engineers-Part IIA: Automatic Regulators and Servo Mechanisms*, 94(1):130–142, 1947.

G. Tzetzis, E. Votsis, and T. Kourtessis. The effect of different corrective feedback methods on the outcome and self confidence of young athletes. *Journal of sports science & medicine*, 7(3):371–378, Sep 2008. ISSN 1303-2968. URL https://pubmed.ncbi.nlm.nih.gov/24149905.

M. Vagia, A. A. Transeth, and S. A. Fjerdingen. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied ergonomics*, 53:190–202, 2016.

K. Vassigh, D. Weigel, and T. Mack. USA simplified aid for EVA rescue (SAFER) operations manual. *NASA Johnson Space Center, Systems Division, EVA Systems Group, JSC-26283 Rev. A*, 1998.

F. M. F. Verberne, J. Ham, A. Ponnada, and C. J. H. Midden. Trusting Digital Chameleons: The Effect of Mimicry by a Virtual Social Agent on User Trust. In S. Berkovsky and J. Freyne, editors, *Persuasive Technology*, Lecture Notes in Computer Science, pages 234–245. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-37157-8.

T.-M. Wang, Y. Tao, and H. Liu. Current Researches and Future Development Trend of Intelligent Robot: A Review. *International Journal of Automation and Computing*, 15(5):525–546, 2018.

D. H. Weir and A. V. Phatak. Model of human-operator response to step transitions in controlled element dynamics. In *Proceedings of the second annual NASA-University conference on manual control*, pages 65–83. Citeseer, 1966.

C. D. Wickens. Three-dimensional stereoscopic display implementation: guidelines derived from human visual capabilities. In *Stereoscopic Displays and Applications*, volume 1256, pages 2–12. International Society for Optics and Photonics, Sept. 1990. doi: 10.1117/12.19883.

C. D. Wickens, S. Todd, and K. Seidler. Three-Dimensional Displays: Perception, Implementation, and Applications. Technical Report CSERIAC-SOAR-89-001, CREW SYSTEM ERGONOMICS INFORMATION ANALYSIS CENTER WRIGHT-PATTERSON AFB OH, Oct. 1989. URL https://apps.dtic.mil/docs/citations/ADA259937.

C. D. Wickens, K. Gempler, and M. E. Morphew. Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors*, 2(2):99–126, 2000. doi: 10.1207/STHF0202_01.

N. Wierner. *Cybernetics (or the control and communication in the animal and the machine)*, volume 1. MIT Press, 1948. ISBN 9780262730099. doi: 10.1037/h0051026.

A. C. Williams and G. E. Briggs. On-target versus off-target information and the acquisition of tracking skill. *Journal of Experimental Psychology*, 64(5):519–525, 1962. ISSN 0022-1015(Print). doi: 10.1037/h0044468.

M. R. Williams and R. F. Kirsch. Evaluation of head orientation and neck muscle emg signals as command inputs to a human–computer interface for individuals with high tetraplegia. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(5):485–496, Oct 2008. ISSN 1558-0210. doi: 10.1109/TNSRE.2008.2006216.

M. R. Williams and R. F. Kirsch. Case study: Head orientation and neck electromyography for cursor control in persons with high cervical tetraplegia. *Journal of Rehabilitation Research & Development*, 53(4), 2016. ISSN 0748-7711.

R. C. Williges and W. W. Wierwille. Behavioral Measures of Aircrew Mental Workload. *Human Factors*, 21(5):549–574, Oct. 1979. ISSN 0018-7208. doi: 10.1177/001872087902100503. URL https://doi.org/10.1177/001872087902100503.

G. Wulf and C. H. Shea. Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, 9(2):185–211, 2002. ISSN 1531-5320. doi: 10.3758/BF03196276. URL https://doi.org/10.3758/BF03196276.

G. Wulf, C. H. Shea, and S. Matschiner. Frequent feedback enhances complex motor skill learning. *Journal of Motor Behavior*, 30(2):180–192, 1998. doi: 10.1080/00222899809601335. URL https://doi.org/10.1080/00222899809601335. PMID: 20037033.

S. Xu, W. Tan, A. V. Efremov, L. Sun, and X. Qu. Review of control models for human pilot behavior. *Annual Reviews in Control*, 44:274–291, Jan. 2017. ISSN 1367-5788. doi: 10.1016/j.arcontrol.2017.09.009. URL http://www.sciencedirect.com/science/article/pii/S136757881730024X.

H. A. Yanco, A. Norton, W. Ober, D. Shane, A. Skinner, and J. Vice. Analysis of Human-robot Interaction at the DARPA Robotics Challenge Trials. *Journal of Field Robotics*, 32(3):420–444, 2015. ISSN 1556-4967. doi: 10.1002/rob.21568. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21568.

G. U. Yule. On a method of investigating periodicities in disturbed series with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London*, 1927.

P. M. T. Zaal, D. M. Pool, J. de Bruin, M. Mulder, and M. M. van Paassen. Use of pitch and heave motion cues in a pitch control task. *Journal of Guidance, Control, and Dynamics*, 32(2):366–377, 2009. doi: 10.2514/1.39953.

M. Zamora, E. Caldwell, J. Garcia-Rodriguez, J. Azorin-Lopez, and M. Cazorla. Machine Learning Improves Human-Robot Interaction in Productive Environments: A Review. In I. Rojas, G. Joya, and A. Catala, editors, *Advances in Computational Intelligence*, Lecture Notes in Computer Science, pages 283–293. Springer International Publishing, 2017. ISBN 978-3-319-59147-6.

K. Zaychik, F. Cardullo, and G. George. A Conspectus on Operator Modeling: Past, Present and Future. In *AIAA Modeling and Simulation Technologies Conference and Exhibit*, Keystone, Colorado, Aug. 2006. American Institute of Aeronautics and Astronautics. ISBN 978-1-62410-047-5. doi: 10.2514/6.2006-6625. URL http://arc.aiaa.org/doi/10.2514/6.2006-6625.

Appendices

Appendix A

Workload Surveys

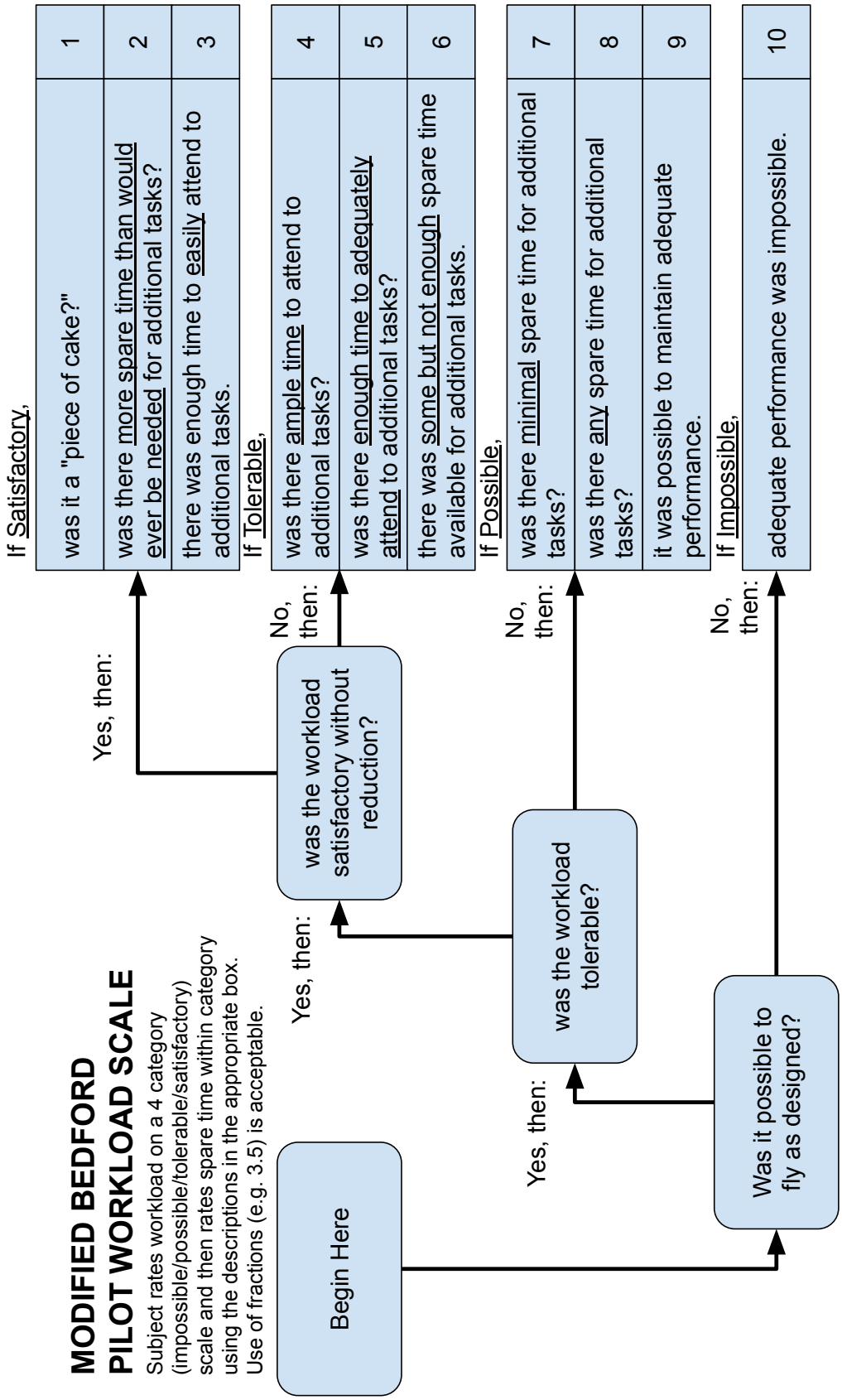


Figure A.1: The Modified Bedford Workload Scale, adapted from Roscoe and Ellis (1990).

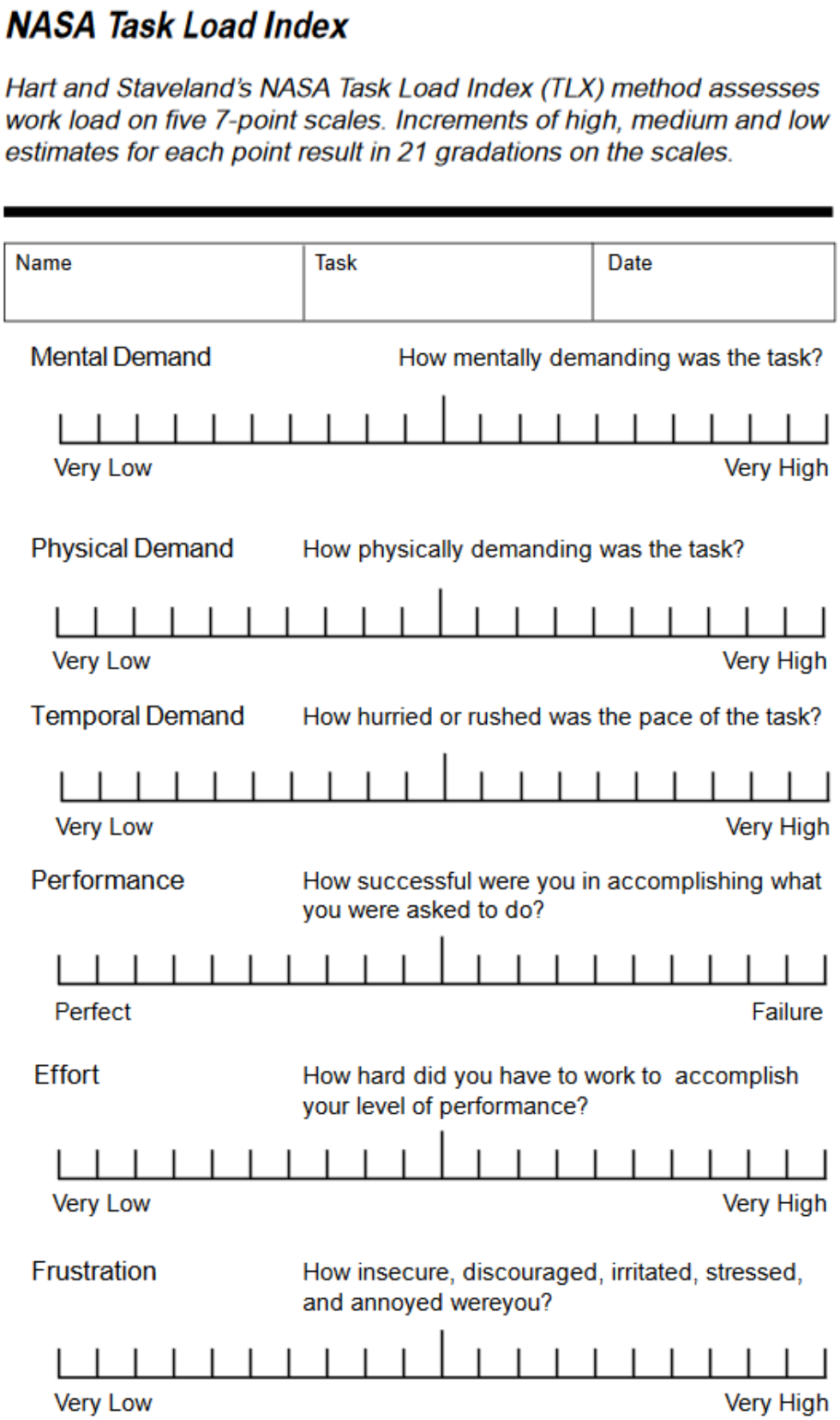


Figure A.2: The NASA Task Load Index (NASA-TLX), from Hart and Staveland (1988).

Appendix B

Trade Analysis Tables

Research Topics														Score	Rank
Understanding human intent		1	1			1	1				1		1	1	77.13	3
Autonomous/robotic system communication to humans						1		1		1	1			1	58.82	6
Ensuring human safety (physical)				1	1					1	1				50.66	7
Continuous human performance monitoring		1	1										1	1	44.68	9
HAR team performance optimization and function allocation										1	1	1	1	1	68.76	5
Enabling command/control of complex robotic systems						1	1	1		1	1			1	72.80	4
Improving situation awareness in HAR systems								1	1		1			1	50.07	8
Improving training for HAR systems and tasks								1	1	1	1	1	1		84.66	1
Establishing appropriate trust in automation/robotics systems							1	1		1	1	1		1	79.27	2

Factors	Weight	Technology: Non-invasive behavioral and physiological sensing	Implantable Biometrics	Autonomous obstacle detection/imaging	Autonomous path planning	Speech recognition	Intuitive physical control interfaces	Robotic/human information interfaces	Augmented Reality/Virtual Reality	Robotic agents (rovers, swarms, arms...)	Assistive Robotics	Artificial Intelligence	Machine Learning	Flexible/Adaptive/Adaptable Automation
Task applicability	6	2.84	2.53	5.37	5.37	1.26	3.79	6.00	4.74	5.05	3.79	6.00	6.00	4.42
Task enabling	6	1.26	1.11	3.63	2.84	0.95	2.84	3.16	2.37	3.79	2.84	3.95	3.00	3.32
Potential for reducing risk	5	3.33	3.33	5.00	5.00	3.33	5.00	5.00	5.00	5.00	3.33	5.00	5.00	5.00
Potential for introducing risk	4	-0.67	-2.67	-0.67	-4.00	-0.67	-0.67	-0.67	-2.67	-4.00	-2.67	-4.00	-0.67	-4.00
External Investment (outside of NASA)	3	0.02	0.01	0.04	0.15	0.91	0.02	0.03	0.51	0.07	0.05	1.97	3.00	0.51
Technology Readiness Level (TRL)	2	1.33	0.67	2.00	2.00	2.00	2.00	1.33	1.33	1.33	1.33	1.33	2.00	2.00
Research Interest (within NASA)	1	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Score:		8.12	4.98	16.37	12.36	8.78	13.98	15.86	12.29	12.25	9.69	15.25	19.33	12.24
Rank:		12	13	2	6	11	5	3	7	8	10	4	1	9

Figure B.1: Top-level trade table with final research topic scores (top right), final technology scores based on factors (bottom) and weighted factor-level scores for each technology.

Task Applicability		Technology	Non-invasive behavioral and physiological sensing	Implantable Biometrics	Autonomous obstacle detection/imaging	Autonomous path planning	Speech recognition	Intuitive physical control interfaces	Robotic/human information interfaces	Augmented Reality/Virtual Reality	Robotic agents (rovers, swarms, arms…)	Assistive Robotics	Artificial Intelligence	Machine Learning	Flexible/Adaptive/Adaptable Automation
		Weight													
Robotic Operations, Orbit	Maneuver/reboost/rendezvous	1	0	0	1	1	0	1	1	1	0	0	1	1	0
	Docking/undocking	1	0	0	1	1	0	1	1	1	1	0	1	1	0
	Spacecraft support, system maintenance	1	1	1	1	1	0	0	1	1	1	1	1	1	1
	Complex assembly, capture and berth	1	0	0	1	1	0	1	1	1	1	0	1	1	0
	Science and assigned activity support, payload assistance	1	1	1	1	1	0	1	1	1	1	1	1	1	1
	Science and assigned activity support, crew assistance–physical	1	1	1	1	1	0	1	1	1	1	1	1	1	1
	Science and assigned activity support, crew assistance–cognitive	1	1	1	0	0	1	0	1	1	0	1	1	1	1
Robotic Operations, Surface	Spacecraft support, system maintenance	1	0	0	1	1	0	0	1	0	1	1	1	1	1
	Spacecraft support, system preparation	1	0	0	1	1	0	0	1	0	1	0	1	1	1
	Site preparation assembly, excavation	1	0	0	1	1	0	1	1	0	1	1	1	1	0
	Complex assembly, heavy lift	1	0	0	1	1	0	1	1	0	1	1	1	1	0
	Drive/navigate	1	1	0	1	1	1	1	1	1	1	1	1	1	1
	Exploration, scouting	1	0	0	1	1	0	1	1	1	1	0	1	1	1
	Exploration, mapping	1	0	0	1	1	0	0	1	1	1	0	1	1	1
	Exploration, sampling/analyzing	1	0	0	1	0	0	0	1	1	1	0	1	1	1
	Science and assigned activity support, science/sample collection	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Science and assigned activity support, payload assistance	1	1	1	1	1	0	1	1	1	1	1	1	1	1
	Science and assigned activity support, crew assistance–physical	1	1	1	1	1	0	1	1	1	1	1	1	1	1
	Science and assigned activity support, crew assistance–cognitive	1	1	1	0	1	1	0	1	1	0	1	1	1	1
	Total Weighted Score:		9	8	17	17	4	12	19	15	16	12	19	19	14
	Normalized Score:		0.47	0.42	0.89	0.89	0.21	0.63	1	0.79	0.84	0.63	1	1	0.74

Figure B.2: Technology to Task Applicability factor-level trade table.

Task Enabling		Technology	Non-invasive behavioral and physiological sensing	Implantable Biometrics	Autonomous obstacle detection/imaging	Autonomous path planning	Speech recognition	Intuitive physical control interfaces	Robotic/human information interfaces	Augmented Reality/Virtual Reality	Robotic agents (rovers, swarms, arms…)	Assistive Robotics	Artificial Intelligence	Machine Learning	Flexible/Adaptive/Adaptable Automation
		Weight													
Robotic Operations, Orbit	Maneuver/reboost/rendezvous	1	0	0	1	1	0	1	1	1	0	0	1	1	0
	Docking/undocking	1	0	0	1	1	0	1	1	1	1	0	1	1	0
	Spacecraft support, system maintenance	1	1	1	1	1	0	0	1	1	2	1	1	1	1
	Complex assembly, capture and berth	1	0	0	1	1	0	1	1	1	1	0	1	1	0
	Science and assigned activity support, payload assistance	1	1	1	2	1	0	2	1	1	2	2	2	1	2
	Science and assigned activity support, crew assistance–physical	1	1	1	2	1	0	2	1	1	1	2	2	1	2
	Science and assigned activity support, crew assistance–cognitive	1	1	1	0	0	2	0	2	1	0	2	2	1	2
Robotic Operations, Surface	Spacecraft support, system maintenance	1	0	0	1	1	0	0	1	0	2	1	1	1	1
	Spacecraft support, system preparation	1	0	0	1	1	0	0	1	0	2	0	1	1	1
	Site preparation assembly, excavation	1	0	0	2	2	0	2	1	0	2	1	1	1	0
	Complex assembly, heavy lift	1	0	0	1	1	0	1	1	0	1	1	1	1	0
	Drive/navigate	1	1	0	2	1	1	2	1	1	1	1	1	1	1
	Exploration, scouting	1	0	0	1	1	0	1	1	1	1	0	1	1	1
	Exploration, mapping	1	0	0	1	1	0	0	1	1	2	0	1	1	1
	Exploration, sampling/analyzing	1	0	0	1	0	0	0	1	1	1	0	1	1	1
	Science and assigned activity support, science/sample collection	1	0	0	1	1	1	1	1	1	1	1	1	1	2
	Science and assigned activity support, payload assistance	1	1	1	2	1	0	2	1	1	2	2	2	1	2
	Science and assigned activity support, crew assistance–physical	1	1	1	2	1	0	2	1	1	2	2	2	1	2
	Science and assigned activity support, crew assistance–cognitive	1	1	1	0	1	2	0	1	1	0	2	2	1	2
	Total Weighted Score:		8	7	23	18	6	18	20	15	24	18	25	19	21
	Normalized Score:		0.21	0.18	0.61	0.47	0.16	0.47	0.53	0.39	0.63	0.47	0.66	0.5	0.55

Figure B.3: Technology to Task Enabling factor-level trade table.

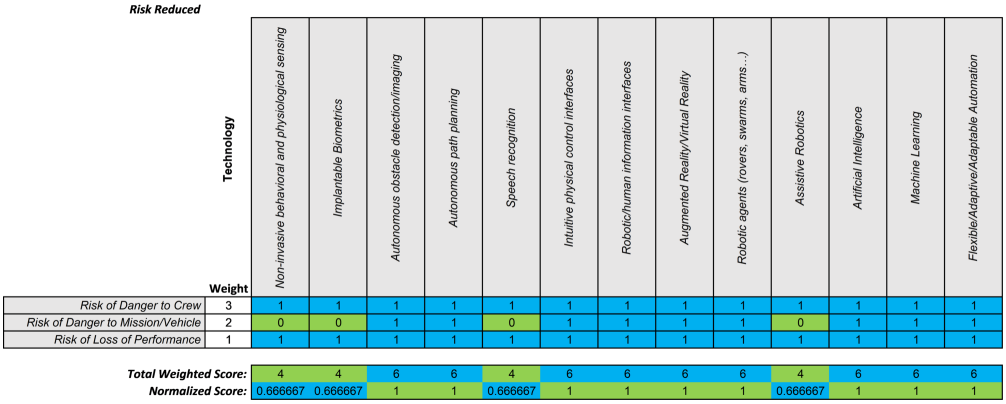


Figure B.4: Technology to Risk Reduced factor-level trade table.

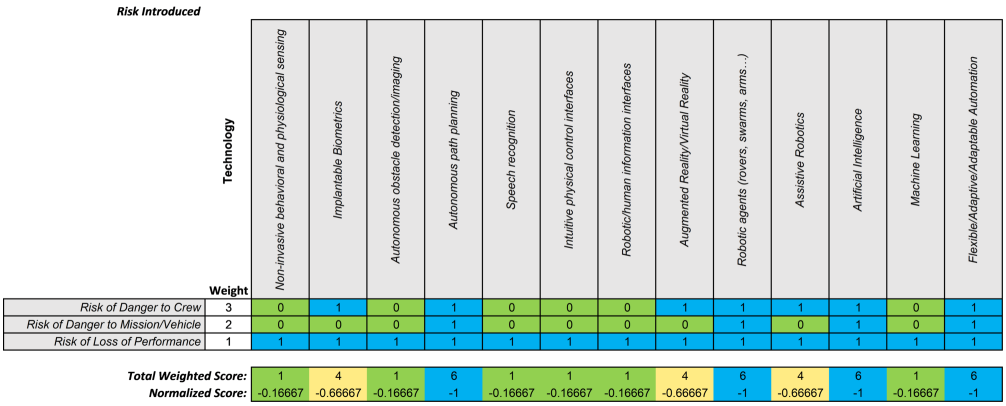


Figure B.5: Technology to Risk Introduced factor-level trade table.

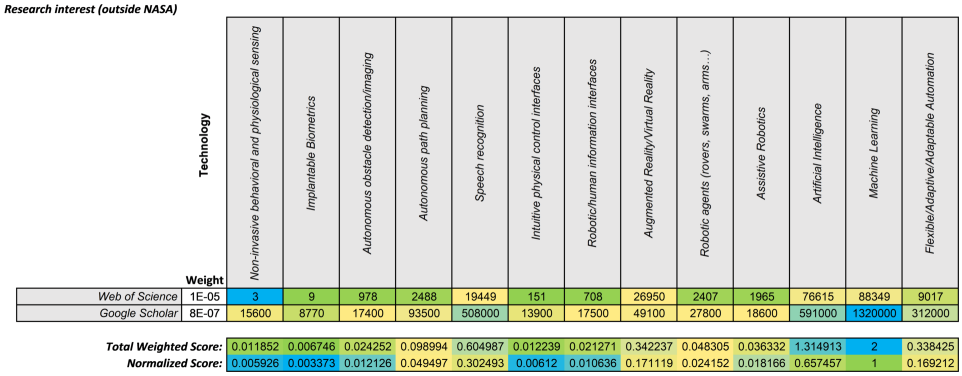


Figure B.6: Technology to Research Interest (outside NASA) factor-level trade table.

TRL	Technology	Non-invasive behavioral and physiological sensing	Implantable Biometrics	Autonomous obstacle detection/imaging	Autonomous path planning	Speech recognition	Intuitive physical control interfaces	Robotic/human information interfaces	Augmented Reality/Virtual Reality	Robotic agents (rovers, swarms, arms...)	Assistive Robotics	Artificial Intelligence	Machine Learning	Flexible/Adaptive/Adaptable Automation
Weight														
TRL less than 3	1	0	1	0	0	0	0	0	0	0	0	0	0	0
TRL 3-5	2	1	0	0	0	0	0	1	1	1	1	1	0	0
TRL 6 or greater	3	0	0	1	1	1	1	0	0	0	0	0	1	1
Total Weighted Score:		2	1	3	3	3	3	2	2	2	2	2	3	3
Normalized Score:		0.666667	0.333333	1	1	1	1	0.666667	0.666667	0.666667	0.666667	0.666667	1	1

Figure B.7: Technology to TRL factor-level trade table.

Research interest (within NASA)	Technology	Non-invasive behavioral and physiological sensing	Implantable Biometrics	Autonomous obstacle detection/imaging	Autonomous path planning	Speech recognition	Intuitive physical control interfaces	Robotic/human information interfaces	Augmented Reality/Virtual Reality	Robotic agents (rovers, swarms, arms...)	Assistive Robotics	Artificial Intelligence	Machine Learning	Flexible/Adaptive/Adaptable Automation
	Weight													
Research interest (within NASA)	1	0	0	1	1	1	1	1	1	1	1	1	1	1
Total Weighted Score:		0	0	1	1	1	1	1	1	1	1	1	1	1
Normalized Score:		0	0	1	1	1	1	1	1	1	1	1	1	1

Figure B.8: Technology to Research Interest (within NASA) factor-level trade table.

Appendix C

Aircraft Dynamics

C.1 Longitudinal Dynamics

$$\dot{\vec{x}}_{long} = \begin{bmatrix} X_u & X_w & X_q & -g\cos\theta_0 & 0 \\ Z_u & Z_w & Z_q + U_0 & -g\sin\theta_0 & 0 \\ M_u + M_{\dot{w}}Z_u & M_w + M_{\dot{w}}Z_w & M_q + M_{\dot{w}}(Z_q + U_0) & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & U_0 & 0 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta w \\ \Delta q \\ \Delta\theta \\ \Delta z \end{bmatrix} + \begin{bmatrix} X_{\delta_e} & X_{\delta_{th}} \\ Z_{\delta_e} & Z_{\delta_{th}} \\ M_{\delta_e} + M_{\dot{w}}Z_{\delta_e} & M_{\delta_{th}} + M_{\dot{w}}Z_{\delta_{th}} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta\delta_e \\ \Delta\delta_{th} \end{bmatrix}$$

C.2 Lateral Dynamics

$$\dot{\vec{x}}_{lat} = \begin{bmatrix} Y_v & Y_p & Y_r - U_0 & g\cos\theta_0 & 0 \\ L'_v & L'_p & L'_r & -g\sin\theta_0 & 0 \\ N'_v & N'_p & N'_r & 0 & 0 \\ 0 & 1 & \tan\theta_0 & 0 & 0 \\ 0 & 0 & \sec\theta_0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta v \\ \Delta p \\ \Delta r \\ \Delta\phi \\ \Delta\psi \end{bmatrix} + \begin{bmatrix} Y_{\delta_a} & Y_{\delta_r} \\ L'_{\delta_a} & L'_{\delta_r} \\ N'_{\delta_a} & N'_{\delta_r} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta\delta_a \\ \Delta\delta_r \end{bmatrix}$$