DISSERTATION

ENGAGEMENT AND NOT WORKLOAD IS IMPLICATED IN AUTOMATION-INDUCED LEARNING DEFICIENCIES FOR UNMANNED AERIAL SYSTEM TRAINEES

Submitted by

John G. Blitch

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2014

Doctoral Committee:

Advisor: Benjamin Clegg

Edward Delosh
Kurt Kraiger
Daniel Robinson

ABSTRACT

ENGAGEMENT AND NOT WORKLOAD IS IMPLICATED IN AUTOMATION-INDUCED LEARNING DEFICIENCIES FOR UNMANNED AERIAL SYSTEM TRAINEES

Automation has been known to provide both costs and benefits to experienced humans engaged in a wide variety of operational endeavors. Its influence on skill acquisition for novice trainees, however, is poorly understood. Some previous research has identified impoverished learning as a potential cost of employing automation in training. One prospective mechanism for any such deficits can be identified from related literature that highlights automation's role in reducing cognitive workload in the form of perceived task difficulty and mental effort. However three experiments using a combination of subjective self-report and EEG based neurophysiological instruments to measure mental workload failed to find any evidence that link the presence of automation to workload or to performance deficits resulting from its previous use. Rather the results in this study implicate engagement as an underlying basis for the inadequate mental models associated with automation-induced training deficits. The conclusion from examining these various states of cognition is that automation-induced training deficits observed in novice unmanned systems operators are primarily associated with distraction and disengagement effects, not an undesirable reduction in difficulty as previous research might suggest. These findings are consistent with automation's potential to push humans too far "out of the loop" in training. The implications of these findings are discussed.

DEDICATION

For MGB, RMD, JC[4], and everything they represent.

TABLE OF CONTENTS

Chapter I

Introduction

Automation has been defined as "the execution by a machine agent (usually a computer) of a function that was previously carried out by a human" (Parasuraman & Riley, 1997). It seems evident that performing tasks on behalf of humans can be quite beneficial to those engaged in complex and demanding activities where fatigue and skill limitations can result in degraded and perhaps even dangerous performance of professional duties (Wiener, 1988). If automation is constantly performing tasks for humans, however, it seems equally evident that those same benefactors may be lulled into a false sense of security and become over reliant on it (Parasuraman & Manzey, 2010).

So what is the right balance or level of automation to be implemented in pursuit of maximum performance by systems that require both mechanical precision and human skill in a particular operational setting? Is that same balance appropriate for the acquisition of skill itself, or are adjustments to level of automation required as humans become more proficient at a task? If so, what are the boundary conditions and metrics to be used in pursuit of that balance?

The dissertation that follows investigated these overarching questions with emphasis on how the well-established costs and benefits associated with automation's influence on operational environments might transfer into the education and training arena. This investigation was initially focused exclusively on the relationship between automation and mental workload in the context of learning deficits observed in novice unmanned system pilots under compressed (four hours in this case) training time. As the study progressed across three experiments however, its scope was expanded to explore the apparent influence that one of the instruments used to measure workload had on performance itself.

The description of this investigation starts with an overview of previous research investigating the role of automation in training. This is followed by an analysis of two experiments that endeavored to replicate automation-induced training deficits observed in previous research while

monitoring workload with a well-established subjective self-report instrument, the NASA Task Load Index - commonly referred to as simply the "TLX". A discussion of its apparent influence on performance is presented next, along with a brief review of alternative workload measurement instruments that might be appropriate for further investigation of learning deficits observed in unmanned aerial systems training. A description of the final experiment in the study is then presented with an emphasis on trial-by-trial EEG data collected in a successful replication of these deficits. These data provide evidence for disengagement and distraction as primary factors in the development of learning deficits in this particular training paradigm while workload was not. The document concludes with a discussion of implications derived from this discovery, and their potential influence on the development of future automation.

Automation in Training

The ubiquitous popularity of unmanned systems throughout the military and remote sensing community has presented a daunting challenge for training managers. The U.S. Army, for example, has fielded thousands of tactical robots with tremendous variability in capability and Operator Control Unit (OCU) design that calls for extensive training transfer across platforms (Antal, 2009). The Air Force on the other hand is struggling to keep up with training demands for its most prominent Unmanned Aerial System (UAS), the Predator (Gramm & Papp, 2009). It seems from these examples that automation is an inherent component of the emerging world, and future generations will need to learn how to interact with it effectively and efficiently. Commensurate an increasing demand for rapid skill acquisition of this nature, one cannot help but wonder how the costs and benefits of automation observed in the operational regime might manifest in the training environment.

While the most obvious goal for training with automation involves the pursuit of improved performance, questions regarding the duration of this improvement and its applicability to related tasks that are slightly different than those experienced in training substantially complicate the issue.

When the primary objective is to achieve maximum performance regardless of accumulated cost in terms of training time and associated educational resources, it can be said that maximum effectiveness is the goal. If, however, high cost is not acceptable and resources are constrained, it is more reasonable to express the goal in terms of training efficiency, or the pursuit of maximum performance given a fixed amount of training time and access to other resources and assets.

Healy et al. (2005a) make it clear, however, that while these two concepts play off against each other in many regimes, training value may be diminished to negligible levels unless the skills practiced therein can generalize to other tasks of direct operational relevance and last long enough to provide an appreciable return on one's investment. The relationships between these four training objectives (effectiveness, efficiency, transfer, and durability) are complex, and provide ample opportunity to leverage the benefits of automation established in the operational environment against a broad spectrum of challenges in the education and training realm. It is important to note, however, that while automation may be used to achieve improvement along one or more of these axes of value; it may not necessarily contribute to advancement along all of them. In fact, it is not uncommon for benefits emerging from pursuit of one objective to result in costs associated with others and vice versa (Healy et al., 2010). In keeping with the rapid skill acquisition objectives established earlier, our focus for the remainder of this particular investigation remains on the pursuit of maximum training efficiency for novice operators seeking to acquire a new skill in the face of constrained resources. Given this focus on efficiency, diminished value and costs are recognized and accepted within the realm of effectiveness, transfer, and durability.

There are cases, however, in which automation presents unique learning opportunities regardless of cost, because they simply could not occur in its absence. Accurately implemented physics based flight simulators, for example, have been designed to allow aircrews to practice emergency procedures in highly dynamic and complicated scenarios without necessarily forfeiting the lives of or well-being of those who fail to reach an established performance criterion. They also

may enable the development of *new* procedures and emergency protocols through iterative evaluation of alternative crash recovery techniques that are difficult if not impossible to practice in real world settings without the loss of life. In each of these situations, automation has enabled the acquisition and/or improvement of new emergency response skills that could not be obtained any other way (Singh, Molloy, Mouloua, Deaton, & Parasuraman, 1998; Wiener, 1988).

In situations where unique learning opportunities like these are enabled, automation may be worthy of training integration at virtually any cost. There are unquestionably a number of similar benefits from operational automation that can be realized in training and educational environments as well. In keeping with the established focus of this study on automation-induced training deficits, however, we will forgo discussion of those applications and continue to concentrate on cost instead.

Training Costs and Deficits Induced by Automation

Despite the abundance of empirical evidence raising concern for automation's negative influence in operational settings, few have dealt with this issue from a learning perspective other than via explorations of trust and confidence (Lee, 1992; Madhavan & Wiegmann, 2007; Sheridan & Parasuraman, 2000; Wickens, Dixon, & Ambinder, 2006). A number of factors to consider in a training context have emerged from an operational body of empirical evidence; including levels and stages of automation (Sheridan & Parasuraman, 2000), functionality (Parasuraman & Wickens, 2008), and comprehensive perspectives that span the gamut of human-computer interaction in the form of situational awareness (Endsley & Kaber, 1999). The vast majority of these studies are based, however, on a critical assumption – that the people involved have reached an appreciable degree of competence in the tasks at hand before automation is introduced. Although there has been some consideration given to autonomy removal during refresher training, particularly in the field of aviation safety (Parasuraman, 1996; Wiener, 1988), little consideration has been given to the impact of automation on the acquisition of skill itself.

One study that did manage to explore the nature of automation assistance for skill acquisition was conducted by Clegg, Heggestad, and Blalock (2010) on participants who were trying to learn how to efficiently manage a complex control process involved in the pasteurization of orange juice. Performance was measured on how much "good" juice was pasteurized or "bad" juice spoiled in a given amount of time under different levels of automated assistance. This study found an initial advantage presented by assistive automation, particularly for low aptitude trainees. As training progressed, however, this advantage diminished to an insignificant level indicating that automation's value decreased as task familiarity increased. Furthermore, when automation was removed participants suffered a substantial drop in performance - suggesting that an overly dependent relationship had been developed between the trainees and automation - especially when autonomous modules had been activated by the system itself.

One explanation for the development of over-reliance on automation during training may have to do with confidence effects observed in previous studies. It has been well established for example, that humans are notoriously overconfident in their learning activities (Joseph, 2009; Koriat & Bjork, 2005; Koriat, Ma'ayan, Sheffer, & Bjork, 2006). Trainees who over-estimate the level of capability they have attained may be inclined to remove themselves from training and turn over tasks to automation prematurely, thus denying themselves potential access to higher performance and retention levels than might otherwise be achieved. In the context of the research that follows, this removal tendency suggests that trainees may feel comfortable offloading tasks to automation before they have acquired enough experience and skill to effectively revert back to manual control when needed to compensate for system failures or emergent situations that automated modules were not designed to handle.

An opposite situation - trainees who underestimate their skill level - may also have a similar effect resulting in overuse of automation. It has been well established that operators with low confidence in the task they are performing have a tendency to rely on automation when they perceive

task demands to exceed their level of capability – particularly when a computer is assumed to have access to a large information data base that the human is either denied access to, or cannot search rapidly enough to be effective (Cummings, 2004; Skitka, Mosier, & Burdick, 1999). This form of bias toward an automated solution is often reinforced by situations where a historical record of high reliability has been established leading the operator to believe that it is infallible (Manzey, Bahner, & Hueper, 2006; Mosier, Skitka, Heers, & Burdick, 1998; Parasuraman & Manzey, 2010). Regardless of whether over confidence or under confidence was involved, it's clear from the results reported in Clegg and Heggestad (2010) that automation had a detrimental impact on trainees' ability to respond to a sudden automation failure and return to manual control with appreciable effectiveness – a potentially disastrous situation if the process involved nuclear power production instead of orange juice pasteurization.

Although this work established a case for automation-induced decrements in a training regime similar to those previously observed in operational environments, it was unknown whether those effects reflected general tendencies that were specific to the task or the nature of the process control domain. In order to examine this issue in the context of unmanned systems control, Blitch and Clegg (2011) used a drone simulator made available by the Air Force Research Laboratory (AFRL) to explore the potential for automation to enable rapid skill acquisition under compressed (four hour) time constraints. This effort endeavored to replicate automation-induced training deficits through the use of a partial auto-pilot function administered in the middle of unmanned system flight simulator training. As such it presented two significant differences from Clegg et al. (2010).

First and foremost, this study would explore whether effects observed in a process control regime would manifest in a fine motor control task as well. It could be argued, for example, that the orange juice pasteurizer task primarily involved automation administered in stage three (decision selection), but only the earliest aspects of stage four (implementation) in the form of mouse clicks and textual inputs that were already quite familiar to participants. For participants in the flight

simulator study, the concurrent inputs from the HOTAS control assembly represented a novel task requiring the development of fine motor skills that lay toward the end of the final stage of automation in which high fidelity control inputs are required for successful execution of the task. Empirical evidence supporting the theory of direct manipulation indicated that this distinction would have different effects in how well a transition into and out of automated conditions would be handled (Ballas, Heitmeyer, & Pérez-Quiñones, 1992).

The second objective was to determine whether the performance drop observed by Clegg et al. (2010) immediately after automation was removed was simply a temporary phase, or could be mitigated by a return to full manual control that might compensate for the apparent over-reliance on automation that had developed during training. By offering auto-pilot assistance only in the middle of training, it was possible that automation-assisted participants might have had time to recover from their over-dependence before being tested on a novel landing task at the very end of training (Blitch, 2012).

In this experiment, automation was invoked during the second of three training blocks while attempting to train novice pilots how to land a Predator Unmanned Aerial Vehicle (UAV). This effort employed a physics based Predator simulator called the Synthetic Training Environment (STE) shown in Figure 1. This system imposed significant time delays on the interface between input device (stick & throttle) manipulations and simulated sensor displays (cameras & instruments) used by the trainee in a manner consistent with bandwidth and distance limited Radio Frequency (RF) communications between a distant Unmanned Aerial System (UAS) and its operator. As such it presented a more difficult control paradigm than typical manned aircraft flight training packages, or even sophisticated military simulators (Martin, Lyon, & Schreiber, 1998; USAF/AFRL, 2002).

Figure 1: Synthetic Training Environment (STE) Setup

Despite predictions of improved learning made by proponents of classic divide and conquer strategies (Russell & Norvig, 1995; Winston, 1984) and Cognitive Load Theory (van Merriënboer, Kester, & Paas, 2006), this study found no evidence that reducing intrinsic cognitive load for novices through the invocation of autopilot assistance provided any immediate performance benefit while automation was active, or during subsequent training after it had been removed. In fact the results at test suggested quite the opposite – automation-assisted participants performed worse during the final test of skill transfer to a novel yet related task – landing the aircraft. Not only did the automation-assisted group record substantially more glide slope error on average than the manual control group while landing the aircraft, but this effect occurred long after automation had been removed during the second half of training (Blitch & Clegg, 2011).

Although this research provided evidence for the continued of automation-induced training deficits reported by Clegg et al. (2010), it is clear that the effects occurred for different reasons. In the process control study, for example, complacency and over-reliance were the suspected source of deficits in skill acquisition based on excessive dependence on automation that was revealed as soon as its assistance was withdrawn. The lack of a significant performance drop observed when

automation was removed in the UAV simulator study suggested the presence of a different mechanism.

Automation as a Part-Task Training Agent

Given the definition of automation as the execution of a task previously performed by human(s) (Parasuraman & Riley, 1997) it seems clear that one outcome associated with its injection into training is to perform subtasks or portions of a complex mission in a manner that enables a human to focus their practice on specific aspects of a particular skill set that may be most beneficial at the time. If, for example, the operator of an unmanned drone involved in a search and rescue mission wants to develop an ability to slew a sensor pod back and forth between two potential targets, it might be useful for them to practice this task in straight and level flight over relatively flat surfaces in good weather before taking on the added (and potentially overwhelming) complexity of doing so in mountainous terrain with turbulent weather patterns that require complicated maneuvers just to keep the platform airborne and prevent a crash. In this type of situation an auto-pilot might be used to compensate for weather conditions and perform control inputs associated with holding the aircraft's altitude steady while the operator concentrates on manipulating its ground track in pursuit of an optimal search pattern.

The complexity reduction approach to skill acquisition was decomposed into three types of part-task training characterized by Wightman and Lintern (1985) in the context of aircraft carrier landings. *Segmentation* of parts is achieved in a serial fashion whereby practice is focused on one task component at a time and later chained back together in performance of the whole task. In the context of an aircraft landing, this might involve practicing turns that line up an approach path before shifting to drills that alter throttle and pitch settings associated with the descent and flare phases of the task. *Fractionalization* of a task allows a participant to focus on a smaller number of control components in training than what are simultaneously required in the whole task. As such, trainees are presented with a reduced time sharing demand on cognitive resources compared to the whole task

when fractions are practiced simultaneously (Wickens, Hutchins, Carolan, & Cumming, 2012b). *Simplification* also presents the trainee with a reduced load on cognitive assets in the form of lower task difficulty. Rather than breaking the task into serial or parallel components, a diluted or less complicated version of the task is presented to the trainee for practice before transitioning to the full difficulty of the whole task. In the search and rescue task described above, this might involve having novices fly in good weather at short ranges without significant signal latency before introducing them to turbulent weather conditions in distant environments that impose significant lag time on their control inputs.

While part-task training applications have shown value in terms of acquisition and transfer of some skill sets, the literature is mixed with regard to its return on investment. In their initial assessment of part task training for carrier landing skill acquisition, for example, Wightman and Lintern (1985) reported positive value for segmentation procedures, particularly when chained in a backward fashion from the final task. However, limited value was found for fractionalization and simplification methods. More recent reviews present evidence that whole task training is generally superior to part task training except for special cases involving low aptitude learners and/or when complexity is high and organization (inter-element dependence) is low (Fontana, Mazzardo, Furtado Jr, & Gallagher, 2009; Gutzwiller, Clegg, & Blitch, 2013a; Lim, Reiser, & Olina, 2009). Similar conclusions were reached in what is perhaps the most comprehensive meta-analysis conducted on the subject to date, with emphasis on the notion that time shared tasks typically addressed by fractionalization are not appropriate for part-task training whereas segmented tasks are (Wickens et al., 2012b). This analysis also extends into training strategies that manipulate difficulty. As with part-task training, increased difficulty strategies were found to be effective for novices, but not skilled trainees, and only when adaptive increases were implemented as opposed to those applied on a fixed schedule (Wickens et al., 2012b).

Based on a review of the part-task literature, a more detailed analysis was conducted on the UAV simulator data reported in Blitch and Clegg (2011). A trial-by-trial repeated-measures ANOVA was conducted on all data recorded during the three basic maneuver training blocks as well as the five landing test trials. This analysis revealed a consistent pattern of impoverished learning that only occurred after automation was introduced in training block two. It also revealed a slight trend toward equivalent performance by the end of the fifth and final landing trial, suggesting that the automation-assisted participants might be able to essentially catch up to the manual control participants if given enough additional training on the landing task itself (Blitch, 2012).

There are two important aspects of the Synthetic Training Environment flight simulator which deserve prominent attention at this point. The first is that its entire training program inherently follows a part-task training paradigm whereby participants are exposed to basic maneuver training blocks which isolate components of aerodynamic control in a serial fashion before they are presented with a fully integrated landing task. At the beginning of each basic maneuver trial, the simulator starts in straight and level flight with a ten second "lead in" period in which the participant shares control of the aircraft before taking over the task on their own for the remaining sixty seconds. The general idea behind this structure is to avoid overwhelming novice pilots with a multitude of skill acquisition challenges all at once.

The second is that the auto-pilot function in the Synthetic Training Environment can only be invoked and/or modified at the beginning of each trial with a combination of airspeed and altitude hold functions. As such this simulator presents an imperfect or uncoordinated aeronautic solution during turns. In other words, the altitude and airspeed hold functions enable the pilot to turn the aircraft without having to worry about causing a stall and having the aircraft fall out of the sky, but they do not compensate for induced drag in a manner that would enable criterion to be reached in the form of minimal control error. Given that a detailed error profile for each error axis (altitude, heading, and airspeed) is provided to the participant on the feedback display at the end of each trial,

it should become obvious to each automation-assisted participant that the two error components that they are told to ignore during training block two (altitude and airspeed) are not being managed with perfection.

Based on the absence of an autonomy removal effect during the last third of training, and the assumption that automation-assisted participants gradually became aware that the auto-pilot functions were performing imperfectly, it is doubtful that complacency or over-reliance on automation were the source of automation-induced training deficits in the initial UAV operator experiment. The conclusion from this study based on a detailed trial-by-trial analysis of both training and test data was that automation essentially functioned as a part-task training agent that achieved its complexity reduction goal at a cost of desirable difficulty that is necessary for efficient learning to occur (Bjork & Bjork, 2006; Blitch, 2012; Healy et al., 2005c; Schneider, Healy, & Bourne Jr, 2002).

In retrospect it is clear that while the segmentation and simplification aspects of the UAV simulator used in Blitch and Clegg (2011) were held constant across groups, the automation invoked in the second training block acted as a fractionalization-style part-task training agent that failed to produce any positive transfer value in a manner consistent with previous literature. Questions remain, however, regarding why such a manipulation would actually cause a deleterious influence on training to the extent that a negative transfer was observed. Given this curious state of affairs, the notion that automation may have somehow pushed or pulled participants out of a desirable difficulty sweet spot deserved further exploration of the relationship between cognitive workload and training (Blitch, 2012).

Cognitive Workload in Automation-Assisted Training

While the measurement of physical workload is a relatively straightforward process to conduct in terms of force vector summation, evaluations of mental effort and cognitive "work" are much more difficult to characterize. Indicators that a person has become physically overwhelmed are

usually quite obvious, while evidence of mental overload is much less apparent. An over-extended athlete who reaches one sort of physical limitation or another, for example, may stop performing the task at hand in a prominent display of distress. In extreme cases where a physical performance "red line" is reached, they may even breakdown and lapse into unconsciousness. Specific identification of a cognitive "red line" remains elusive, however, except perhaps in the case of excessive multi-tasking (Grier et al., 2008). Not only are there a number of abstract concepts implied by the term cognition that are difficult to measure directly, but there appears to be substantial variance in the literature regarding which of these activities constitute "work" vs. alternative activity of a recreational or perhaps even restful nature. These issues, combined with a lack of unifying theory makes the concept of mental workload hard to define (Jex, 1988).

Despite the challenges noted above, Hancock and Meshkati (1988) chose to characterize the concept of mental workload in terms of cognitive demand. Commensurate with development of his multiple resource theory, Yeh and Wickens (1988) took a highly structured approach and expressed the concept in terms of a specific demand for cognitive activity based on the nature of the task to be accomplished, the environment it was to be performed in, and the mental resources available from the person attempting it. Building on the notion that cognitive resources are limited and do not necessarily share the same reservoir to draw from, Mane and Wickens (1986) attempted to capture the relationship between task difficulty, workload and training effectiveness in the form of four general hypotheses. Based on an initial review of training literature, they concluded that empirical support for the popular notion of walk-before-run strategies that emphasized easy to hard training transfer was often weak and mixed at best. On the other hand, it was clear that increasing difficulty beyond a specific task/environment/participant dependent threshold was detrimental to learning as well.

After analyzing the characteristics of the task environments and types of mental resources involved in this literature, Mane and Wickens (1986) concluded in their first hypothesis that

increased difficulty during training could be beneficial, but only when applied to resource-limited (vs. data limited) components that were directly related to task accomplishment. Their second hypothesis is naturally derived from the first in that it predicts that increased difficulty applied to activities not directly related to skill acquisition would be counter-productive. In manipulations that slowed tasks down enough to allow peripheral activities such as data search and perhaps even the information processing requirement of learning itself (Newell, 1976), skill acquisition improved. Their third hypothesis asserts that the efficient human tendency to conserve resources can make workload shedding strategies adopted by novices placed under excessive demand inappropriate for learning during later experience at higher levels of complexity. Their fourth and final hypothesis calls for cognitive resource demands that are changed during training to be matched with changes in task demand in order to ensure effective learning (Mane & Wickens, 1986; Newell, 1976).

It seems clear from this review that all types of difficulty are not created equal with respect to their influence on training. The proponents of Cognitive Load Theory took a similar approach to resolving the apparent dichotomy between easy to hard and hard to easy training strategies. This theory builds on the notion of cognitive limitations with observations that constrained memory resources make instructional methods appropriate for simple tasks actually counterproductive in skill acquisition required for complex procedures in operational environments (Sweller, van Merrienboer, & Paas, 1998; van Merriënboer et al., 2006). In a theme consistent with the desirable difficulty literature (Schneider, 1985) and more recent theories of disuse in learning (Bjork & Bjork, 2006), empirical results support the notion that increasing mental demand or load while practicing simple tasks with more variability and less feedback can be beneficial, whereas reducing cognitive load during complex task training with partially worked examples and other methods is generally the superior approach. What sets Cognitive Load Theory apart, however, is how it distinguishes between the various *types* of difficulty or load to be increased or decreased.

The basis for notion of intrinsic load, for example, resides in the characteristics of the information to be learned or skill to be acquired. Information chunks or skill components which are typically dependent upon each other are considered to be task elements with a high degree of interactivity. This interactivity makes it difficult (if not impossible) to learn one element without the simultaneous involvement of one or more other dependent elements in focused practice or other aspects of the training process. Higher interactivity between elements of a task presents the performer with increased difficulty in the form of high intrinsic load.

One example of how this kind of load can manifest in the acquisition of a motor skill might be viewed in the context of learning how to drive a car with a manual transmission. The relationship between the engine, the clutch and the accelerator are highly dependent on one another. Letting the clutch out quickly without depressing the accelerator will cause the engine to stall. Due to the high interactivity between these elements, getting the car moving forward without stalling the engine presents a subtask with high intrinsic load compared to other components of the driving task such as changing lanes or activating a turn signal.

According to Cognitive Load Theory, various relationships between task elements are assembled into schemata that in turn are integrated into a mental model of various task requirements. In the driving scenario discussed above, a student might initially develop a schema for gradual acceleration by slowly letting out the clutch while barely depressing the accelerator. As the student gains experience and eventually earns the right to practice driving on their own, a separate schema might be developed for rapid acceleration or a "peel out" maneuver that involves depressing the accelerator to the floor and suddenly releasing the clutch entirely. The mental demand associated with this process of schema formation and mental model development is generally referred to as germane load since it contributes directly to the task at hand.

Another category of cognitive load called "extraneous load" has to do with peripheral knowledge and motor activity that may be necessary to implement various aspects of the training

process, but do not directly contribute to skill development per se. This type of load typically involves search tasks and other activities that simply provide access to information without doing anything to actually process or integrate it into the task at hand – a distinction that forms the basis for "germane load" – which is considered to be a desirable attribute of good learning practices. An example of extraneous load considered within the driving scenario might be the mental demand associated with looking through the operator's manual to locate various switches on the vehicle dash board so that electrically operated rear view mirrors can be adjusted. While the task needs to be done in order to enable effective training to be conducted safely, the mental demand associated with finding the switch and/or actually adjusting the mirrors does not directly contribute to skillful driving.

In a manner similar to Wickens' multiple resource theory (Mane & Wickens, 1986; Wickens, 2002; Wickens, 2008a), Cognitive Load Theory proposes that the key to effective and efficient learning lies not in adjusting difficulty or mental demand along a single axis or through consumption of a shared resource pool, but rather in developing the right balance between these types of load. It predicts, for example, that a reduction in extraneous load for virtually any combination of task complexity (high v. low) or student experience (novice v. expert) will lead to increased learning since it avoids waste of directly relevant cognitive resources that might be diverted elsewhere. Conversely, germane load is generally considered a good thing that should be increased substantially when complexity is low and student experience is high – a position that it shares with the desirable difficulty literature discussed earlier (van Merriënboer et al., 2006).

In situations where novices are attempting to learn complex tasks, however, the theory predicts that too much intrinsic load can have a detrimental impact on learning. When the density of interactivity between elements makes their relationship difficult to grasp, high intrinsic load can interfere with proper schema development and the mental model integration process and learning suffers. If these relationships are made more clear through the use of partially worked examples or

haptic assistance, however, performance increases and learning improves (Griffiths & Gillespie, 2005; Hutchins, Wickens, Carolan, & Cumming, 2013; van Merriënboer et al., 2006).

Returning now to the UAV simulator experiment conducted by Blitch & Clegg (2011), it is evident that while extraneous load reduction was not specifically addressed by the Synthetic Training Environment, the computer-based comprehensive tutorials administered before both the training phase and test phase ought to be considered as a step in the right direction by minimizing extraneous search requirements for enabling information. The video clips contained in each tutorial showed the interaction between control inputs conveyed to the aircraft via stick and throttle manipulations and the corresponding influence they have on both the control surfaces of the aircraft (e.g. aileron movement) and its behavior in flight (e.g bank left or right). In any case, extraneous load was held constant across both groups.

Since the only difference in treatment of participants throughout the entire study was the invocation of altitude and airspeed hold functions during the second training block, it appears that the only type of load influenced by automation in this particular experiment was the intrinsic load associated with interactivity between throttle settings and the pitch / roll axes on the stick. Rudder pedals were disconnected from the Synthetic Training Environment for this experiment and by default yaw was controlled by the simulator for both groups (Blitch, 2012). Although haptic feedback was not used to provide any assistance, one could view this auto pilot manipulation as a motor skill equivalent to the partially worked examples strategy championed by Sweller and colleagues (Sweller et al., 1998; van Merriënboer et al., 2006).

Given the complex nature of this task, its high interactivity between elements, and the low level of experience for the participants involved, the results obtained by Blitch and Clegg (2011) are counter to what might be predicted by Cognitive Load Theory in terms of intrinsic load reduction for novices (van Merriënboer et al., 2006). They do, however, support the notion of misapplied difficulty discussed by Mane and Wickens (1986) and an undesirable drop in germane load (Ayres & Paas,

2007; Paas, Tuovinen, van Merriënboer, & Darabi, 2005) as a potential source of the automation-induced training deficits previously observed (Blitch, 2012).

In any case, it is clear that the relationship between workload and performance represents a critical component of automation's role in skill acquisition. As such, a focus on workload monitoring during replication of Blitch & Clegg (2011) presented the logical next step in the exploration of automation's influence on training.

Chapter II

Experiment 1: UAS Training with NASA-TLX and Pre/Post Test

For automated systems involved in training, one design goal is to reduce the level of cognitive demand or workload distributed across a multitude of complex task components in order to focus learning on a particular topic or shortfall (Gutzwiller et al., 2013a). This complexity reduction could arguably mitigate cognitive overload situations which have been shown to hinder skill acquisition (Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Sweller et al., 1998; van Merriënboer, 2005). However, as Sarter, Woods, and Billings (1997) pointed out, the influence of automation on mental workload is far from straightforward.

The relationship between automation and mental workload has been studied extensively in the workplace with skilled performers. In a study of aircrews responding to emergency response scenarios, for example, Bowers, Thornton, Braun, Morgan Jr, and Salas (1998) presented results with lower levels of subjective workload reported when an auto-pilot was active, but overall performance improvement on only one of four performance metrics recorded under the same conditions. In another study of thirty six licensed pilots flying both dual and single UAV military missions, Dixon, Wickens, and Chang (2005) showed significant benefits from an auto-alert system that resulted in both lower workload and improved performance. In a study of air traffic controllers struggling to handle heavy densities of self-separating aircraft, Galster, Duley, Masalonis, and Parasuraman (2001) presented a case for automated assistance applied across the entire spectrum of information processing and decision implementation. Within the training realm, however, questions remain about the cost/benefit potential of automation injected into training (Gutzwiller et al., 2013a; Heggestad, Clegg, Goh, & Gutzwiller, 2012).

The following experiment examined the nature and magnitude of automation's influence on cognitive workload as a critical component of skill acquisition in the context of unmanned aerial

vehicle control. It is important to note that "control" in this case is considered to imply a direct relationship between human influence on specific aspects of flight characteristics (such as pitch, roll, and thrust inputs on a stick and throttle) as opposed to resource allocation activities that are commonly implemented via mouse clicks on a computer display.

Experiment 1 Goal

This experiment endeavored to monitor cognitive workload during the replication of the automation-induced training deficits initially identified by Clegg and colleagues in the context of process control (Clegg & Heggestad, 2010; Clegg et al., 2010), and further explored by Blitch and Clegg in the realm of unmanned aerial vehicle control (Blitch, 2012; Blitch & Clegg, 2011). One theoretical explanation for the occurrence of training deficits associated with automation involves the notion that it can act as a software agent that pushes or pulls trainees out "sweet spot" of desirable difficulty championed by Bjork and colleagues within the training literature (Bjork & Bjork, 2006; Healy et al., 2002; Kornell & Bjork, 2009). By making the task too easy, automation might allow trainees to backslide into a comfort zone where they are not challenged enough to make substantial progress in the task at hand. In order to test this hypothesis, Experiment 1 endeavored to examine workload alongside performance in the same experimental paradigm used by Blitch and Clegg (2011) in which a UAV flight simulator (the USAF STE) recorded control error across three training blocks and one test block.

Experiment 1 Research Questions and Hypotheses

In order to examine the desirable difficulty aspect of automation-assisted learning, experiment one assessed whether or not the performance deficits previously reported by (Clegg et al., 2010) might be associated with a concurrent drop in cognitive workload. One common approach to workload assessment involves the use of subjective self-reports, such as the Task Load Index (or TLX) developed by NASA (Hart & Staveland, 1988). A digital version of the original paper survey requires participants to retrospectively rate each workload component of the task they just

completed across six subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration) along with relative weights assigned to each as described by Cao, Chintamani, Pandya, and Ellis (2009). This instrument has been prominently featured in more than 550 studies spanning more than 20 years of research, and is considered to be easy to use and reliably sensitive to a wide range experimentally important manipulations (Hart, 2006). Since the TLX has proven to be popular in the human factors field and was available in a digital format that was relatively easy to administer (Cao et al., 2009), it was employed to assess whether automation's detrimental impact on training effectiveness noted earlier (Blitch & Clegg, 2011; Clegg & Heggestad, 2010) is associated with fluctuations in workload measured in terms of perceived effort.

In order to examine the magnitude of the predicted training deficits, a pre-test / post-test approach was used in a manner consistent with recent educational research conducted in the context of analogy based laboratory instruction (Yildirim, Ayas, & Küçük, 2013), case based learning in physiology (Gade & Chari, 2013), and creative learning in technology enriched game play (Randolph, Kangas, Ruokamo, & Hyvönen, 2013). Two test conditions were implemented for comparison of error acquired early in training before the automation "treatment" was applied in the second training block, and after it was removed in the third training block. The Landing Task Pre-Test (LTPT) condition recorded control error collected after participants attempted a single initial landing trial following a brief review of a landing tutorial but without any hands on practice. The Landing Task Last Trial (LTLT) condition reflects the amount of error recorded on the very last "test" trial each participant flew after approximately sixty minutes of cumulative hands on practice in the UAV flight simulator.

Six hypotheses were tested during this experiment under the general research question of how automation influences training efficiency. Given the findings from prior research (Blitch & Clegg, 2011; Clegg et al., 2010), it was expected that the automation-assisted group would again record significantly more control error during the Landing Task Test (H1), but not during the initial pre-test

(H2). It was also expected that the automation-assisted group would present higher error levels recorded between the two test conditions (LTPT, LTLT), especially on the glide slope metric (H3) since that was where the most error was recorded by Blitch and Clegg (2011).

Given that automation by definition is designed to take over tasks that humans previously performed (Parasuraman & Riley, 1997), it was also expected that the automation-assisted group would report significantly reduced workload scores when automation was invoked during training block 2 (H4). As automation-assisted participants returned to full manual control in training block 3, it was expected that they would report higher workload scores than the manual control (MC) group since they would be forced to regain control of inputs that were previously handled by the computer (H5). It was expected that this higher level of perceived workload would also be evident during the LTT test block for the same reason (H6). Based on these hypotheses, it was expected that experiment one would reveal that a reduction in perceived workload would be implicated as a primary factor in the development of automation-induced training deficits.

Experiment 1 Method

Participants. Seventy eight experimentally naïve undergraduate students were involved in this study for optional, partial course credit, forty two of whom were dropped from final analysis due to procedural errors with the newly acquired Synthetic Training Environment simulator, and four who were dropped due to experience as a private pilot or with flight simulators in general. Participants were randomly assigned to one of two groups: one requiring manual control during the entire training and test process, and another that was provided auto-pilot assistance during the middle of three training blocks.

The relatively large percentage of data dropped during this study was due to experimenter training challenges and equipment issues, and should not be misconstrued as an indicator of anomalous participant selection procedures. Attrition was evenly distributed across groups and thus exerted no undue influence on the results subsequently recorded.

Apparatus and Equipment. This experiment was conducted using the Predator STE (Synthetic Training Environment) simulation provided by the Air Force Research Laboratory (AFRL) installed on a Dell Pentium 4 desktop computer with a HOTAS joystick assembly and secondary monitor. Input stimuli of interest were provided by the Predator STE simulator as described by Martin et al. (1998). A pictorial representation of this experimental setup is provided in Figure 1. Workload was measured by a digital version of the NASA TLX described by Cao et al. (2009) installed on a SONY VAIO laptop running Windows XP as an operating system.

Procedure. This experiment used a between subjects design with a manual control (MC) group and automation-assisted (AA) group whose performance was compared in a pre-test / post-test fashion. Participation in this experiment was conducted in four sessions each lasting approximately one hour in duration: three Training Blocks (TB1, TB2, TB3) and one Landing Task Test (LTT) block.

Upon completion of informed consent documentation, session one began with each participant completing the 15 minute Landing Task tutorial that is included with the STE software package. Following this tutorial, each participant was required to perform one Landing Task trial as a pre-test with simple "just-do-the-best-you-can" instructions. After completing this pre-test trial, each participant then reviewed a 25 minute Basic Maneuver (BM) tutorial on basic aeronautic concepts and specific flight characteristics of the Predator UAS.

Following this tutorial, each participant then received a one page scenario description of basic maneuver scenario one. This task required them to reduce airspeed while holding heading and altitude constant. Participants then performed twenty hands-on training trials on this scenario, each lasting one minute in length for a total of twenty minutes of "stick time" during training block one. Performance for each trial was automatically recorded and displayed for the participant at the end of each trial in the form of a Root Mean Square Error (RMSE) compared to optimal flight control inputs indicated by the yellow text and graph plots in Figure 2.

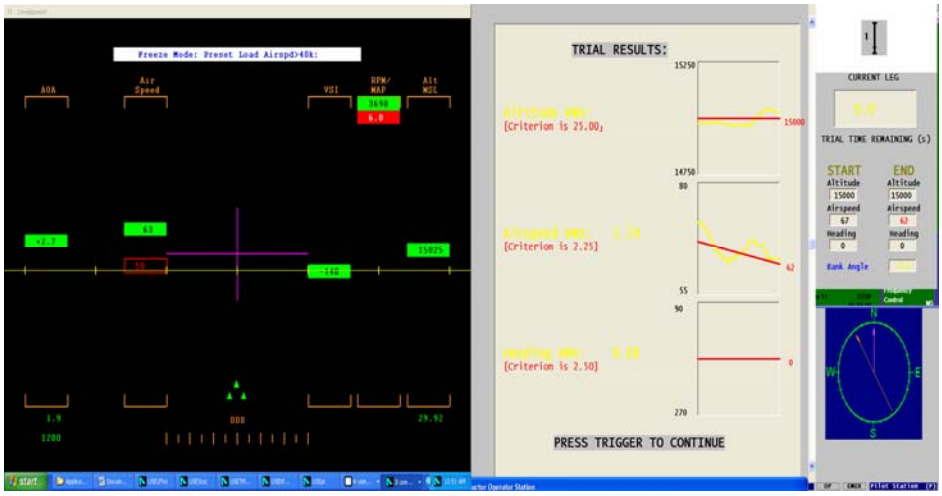

Figure 2: Typical training performance feedback provided by the U.S. Air Force Predator Unmanned Aerial Vehicle Synthetic Training Environment on the left and right monitors

Training block two began with a single page scenario description for basic maneuver scenario two which required a 180 degree turn while holding altitude and airspeed constant. Participants who had been randomly assigned to the manual group were provided with scenario description 2M (for manual control) that contained standard instructions. Those randomly assigned to the autonomy assisted group were provided with the scenario description 2A (for automation assistance) which explains that both the throttle and pitch inputs for this scenario were handled by the aircraft's autopilot. Thus the sole performance concern for the automation-assisted group was the roll axis controlled via side pressure on the joystick. These participants were instructed to ignore the performance feedback for altitude and airspeed, and focus on heading data instead. All participants completed twenty basic maneuver two trials for total "stick time" during training block two of twenty minutes.

During training block 3 participants were provided with a single page scenario description for basic maneuver task five which requires a straight line descent (reduction of airspeed and altitude while holding heading constant) under full manual control. Note that basic maneuver three and four involved ascent / climbing tasks that were skipped in deference to the compressed training constraint

of four hours imposed by university research pool policy. All participants completed twenty descent trials for a total cumulative "stick time" of twenty minutes during training block three.

Upon arrival for session four, participants were allowed five minutes to review the landing task (LT) tutorial in the same fashion as the pretest during training block one. It should be noted that unlike the basic maneuver tutorial, the landing task tutorial contained a detailed description of the task at hand (two 90 degree descending turns while on approach, and a slow descent / flare to land on the runway indicated by Figure 3, so no paper description of it was administered.

Once the landing task tutorial was completed, each participant was required to perform four additional landing task trials for a total of approximately twelve additional minutes of stick time acquired during the test session. This resulted in a total of five landing trials (including the pre-test) completed for the entire study. Performance feedback was provided on screen after each landing task trial in a manner similar to the basic maneuver trials. After all four landing task trials were completed the participant was asked to fill out a short (approximately five minute) spatial experience survey intended to validate their status as a novice trainee and identify any other sources of anomalous performance such as private pilot training, experience with other simulators, etc.

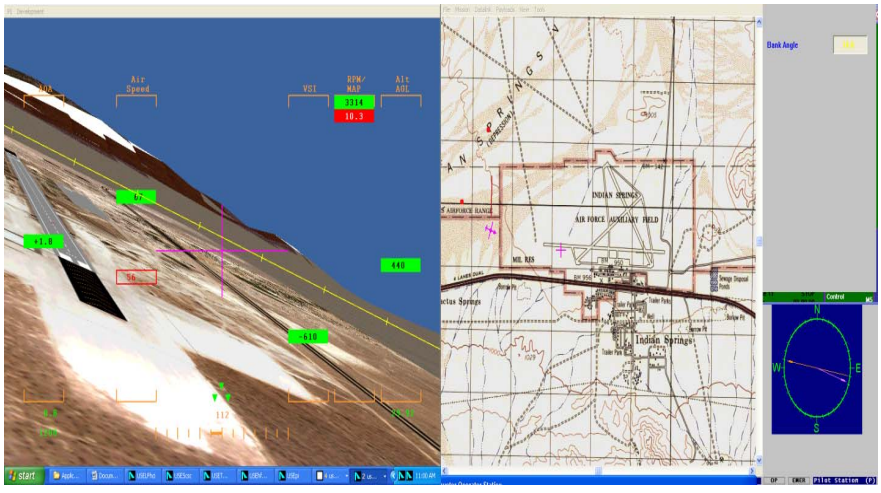

Figure 3: Typical Landing Task display provided by the U.S. Air Force Predator Unmanned Aerial Vehicle Synthetic Training Environment on the left and right monitors

Measures. Training performance in this experiment was automatically calculated by the Synthetic Training Environment using a Root Mean Square Error (RMSE) procedure and captured in three separate metrics as follows:

Altitude Error: was measured as vertical displacement along the z axis from the optimal altitude trace that a pilot candidate should have pursued in relation to various instrument readings displayed on the simulator monitor screens during each training scenario.

Heading Error: was measured as a deviation in degrees from the optimal heading trace that a pilot candidate should have pursued in relation to various instrument readings displayed on the simulator monitor screens during each training scenario.

Indicated Airspeed (IAS) Error: was measured as deviation in knots from the optimal airspeed trace that a pilot candidate should have pursued in relation to various instrument readings displayed on the simulator monitor screens during each training scenario.

Criterion Pass/Fail: training performance during each trial was monitored for excessive error in all three of the control input metrics described above. When participants were able to maintain error levels within a pre-established criterion dictated by the simulator protocols, they received a passing grade and were informed of this during via post trial feedback. Otherwise the trainee received an indication of the control input metric(s) exceeding criterion via a flashing numerical display of error magnitude as indicated in Figure 2.

Test performance recorded during landing task trials in this experiment was also calculated automatically by the Synthetic Training Environment using the same Root Mean Square Error (RMSE) procedure but was captured in three related, but slightly different metrics as described below:

Approach ground track error: was measured as an x/y horizontal displacement from the optimal approach path that a pilot candidate should have pursued in relation to various terrain

features in the vicinity of the airport while turning the aircraft through three waypoints enroute to their final landing activity.

Final ground track error: was measured as an x/y horizontal displacement from the optimal approach path that a trainee should have taken in relation to the runway once they had made their final turn toward the airport and were on the final leg of their landing.

Glide slope error: was measured as a +/- vertical displacement along the z axis of an optimal descent slope that the trainee should have pursued through each LTT trial from start to finish (touchdown).

Experiment 1 Results

Test Performance Data. Performance analyses for the final landing test data collected in this experiment were conducted using a 2x2 ANOVA based on group and test condition as indicated in Figure 4. The between subjects variable was established as control type (manual vs. automation-assisted) and the within subjects variable was test condition (pre vs. final). The main effect for test condition was marginally significant in both approach ground track $F(1,30)=2.85$, $p=.05$, $\eta^2_p=.04$, and glide slope $F(1,30)=3.95$, $p=.05$, $\eta^2_p=.06$ metrics, suggesting that a slight reduction in error may have occurred over the course of training in these two metrics. No main effect for test condition was observed in the final ground track metric $F(1,30)<1$, $\eta^2_p=.00$ between pre-test and final test conditions, however, suggesting that collective error rates for all participants did not fluctuate significantly over training in this particular metric.

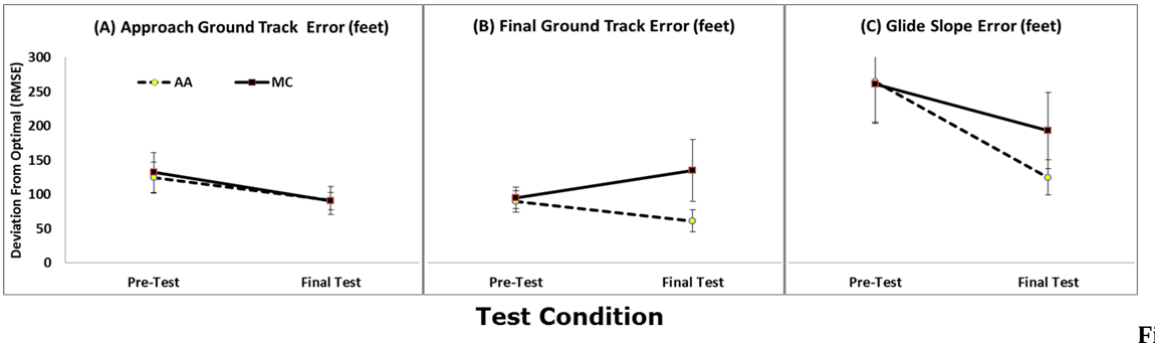

Figure 4: *Experiment 1 Test Performance on Landing Task by Automation and Test Condition.* Deviation in root mean square error from optimal control during the Landing Task Pre-Test and Landing Task Final Test for three metrics: Panel (A) Ground Track on Approach, Panel (B) Final Ground Track, and Panel (C) Glide Slope.

Dashed line with open circles shows participants who received automation assistance (AA) from the autopilot activated in training block 2. Solid line with filled squares shows participants who performed in manual control mode (MC) throughout. Test Condition shows error values recorded by participants before (Pre-Test) and after (Final Test) receiving hands-on practice across training blocks. Error bars show standard error.

Contrary to the predictions of hypothesis one, no main effect for group was observed in approach ground track $F(1,30)<1$, $\eta^2_p=.00$, final ground track $F(1,30)=2.05$, $p>.05$, $\eta^2_p=.03$, or glide slope $F(1,30)<1$, $\eta^2_p=.00$, nor was any interaction between group and test condition observed in approach ground track $F(1,30)<1$, $\eta^2_p=.00$, final ground track $F(1,30)=1.52$, $p>.05$, $\eta^2_p=.025$, or glide slope $F(1,30)<1$, $\eta^2_p=.01$. Taken in combination, these data indicate that participants in both groups were able to significantly reduce their approach ground track error and glide slope error during training, but without any differences observed between subjects based on level of automation invoked during training.

Training Performance Data. Analysis of training data was conducted using a repeated-measures ANOVA across all three training blocks as indicated in Figure 5. It should be noted that comparisons of altitude, airspeed, and criterion pass/fail were only valid in training blocks one and three, since the auto-pilot controlled these functions on behalf of the automation-assisted group during training block two.

No main effect for group was observed in altitude $F(1,30)<1$, $\eta^2_p=.00$, heading $F(1,30)=1.454$, $p>.05$, $\eta^2_p=.05$, indicated airspeed $F(1,30)<1$, $\eta^2_p=.01$, or criterion passed $F(1,30)=1.46$, $p>.05$, $\eta^2_p=.05$, nor was any interaction of group with block observed in altitude $F(1,30)<1$, $\eta^2_p=.01$, heading $F(1,30)=1.18$, $p>.05$, $\eta^2_p=.04$, indicated airspeed $F(1,30)<1$, $\eta^2_p=.01$, or criterion passed $F(1,30)<1$, $\eta^2_p=.05$.

No main effect for block was observed in altitude $F(1,30)<1$, $\eta^2_p=.02$, or indicated airspeed $F(1,30)=2.79$, $p>.05$, $\eta^2_p=.09$, but a large main effect for block was observed in heading $F(1,30)=74.78$, $p>.05$, $\eta^2_p=.71$ and criterion passed $F(1,30)=30.74$, $p<.01$, $\eta^2_p=.51$, suggesting the presence of a large fluctuations between training tasks in these last two metrics. Given that scenario in training block two is the only one to require a turning type of activity, there is not much analytical

value to be discussed in terms of the heading data other than to say that it reflects the nature of the task arrangement chosen in the design of the simulator program at hand. The criterion data, however, suggest a substantial drop in the number of participants in both groups who were able to maintain their control error within acceptable levels during training block three as compared to when they first started in training block one.

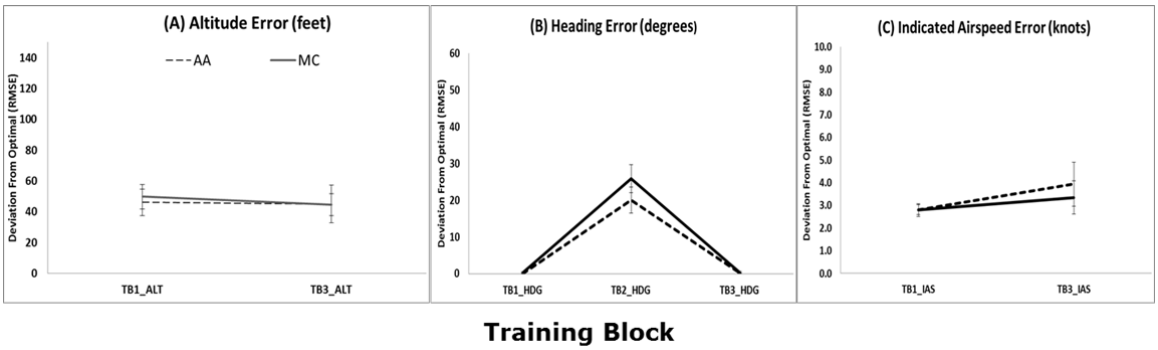

Figure 5: ***Experiment 1 Training Performance by Training Block (TB) and Automation Group.*** **Deviation in root mean square error from optimal control during training tasks for three metrics: Panel (A) Altitude Error, Panel (B) Heading Error, and Panel (C) Indicated Airspeed (IAS) Error. Dashed lines show participants who received automation assistance (AA) from the autopilot activated in Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

Workload Data by Block. In order to examine workload fluctuations throughout the training process, weighted NASA TLX scores were analyzed using a repeated-measures ANOVA conducted across all survey events as indicated in Figure 6, but focused primarily on scores obtained after each hands-on training block. Greenhouse-Geisser corrections were made for violations of sphericity assumptions where appropriate. These data revealed a main effect for block with a large effect size $F(5.40,30)=19.04$, $p<.01$, $\eta^2_p= .39$, indicating that workload fluctuated substantially between surveys administered throughout training. There was no main effect for workload observed between groups, however, $F(1,30)<1$, $\eta^2_p= .01$, nor was any workload interaction observed between group and block $F(5.16,43)<1$, $\eta^2_p= .02$.

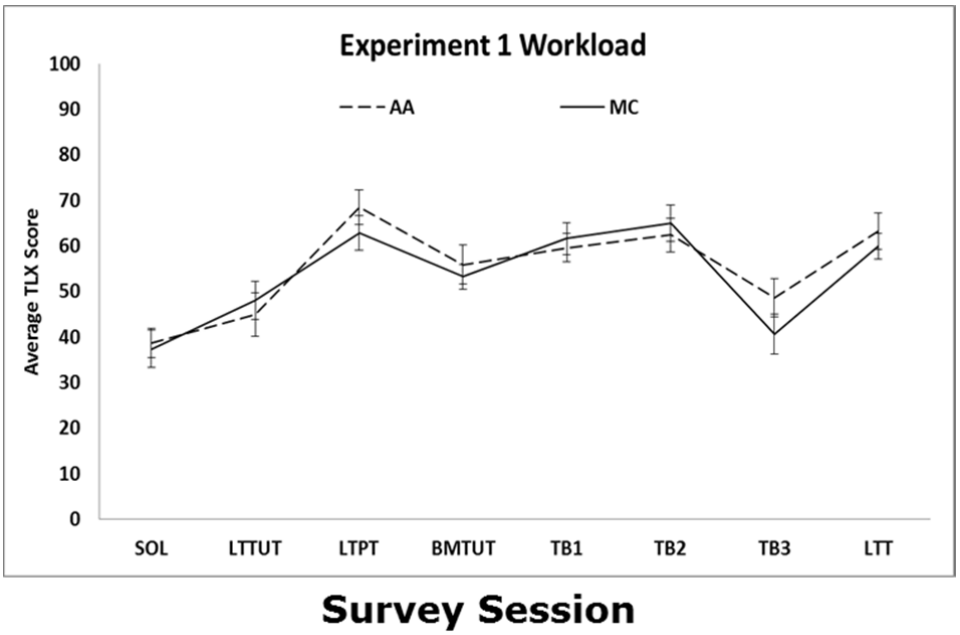

Figure 6: ***Experiment 1 Workload Measured by Average Task Load Index (TLX) Score Across Training Session.*** **Dashed lines show participants who received automation assistance (AA) from the autopilot activated in Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Survey Session shows the training period after which the TLX was administered starting with a Solitaire computer familiarization session (SOL), the Landing Tutorial (LTTUT), the Landing Task Pre-Test (LTPT), the Basic Maneuver Tutorial (BM TUT), each of the three hands-on Training Blocks (TB1, TB2, TB3), and the Landing Task Test (LTT). Error bars show standard error.**

Experiment 1 Discussion

The workload data plotted across training (see Figure 6) indicate that the NASA TLX was sensitive to differences in the task type as it varied throughout training. The non-significant main effect for group within the workload measure, however, along with negligible effect sizes and the lack of any interaction between group and training block indicates that the NASA TLX was not sensitive to the introduction of automation in training block two, its removal in training block three, or its continued absence during the landing task test trials.

The fact that none of the data collected in this experiment provided any support for any of the hypothesized connections between workload and performance was surprising on a number of levels. First and foremost the experiment failed to replicate automation-induced training deficits despite the large effect size (Cohen's d = 0.91) reported by Blitch and Clegg (2011), and more than double the

sample size of the previous experiment (n=45 in experiment 1 vs. n=20 in Blitch & Clegg). This lies counter to what one would expect from an experiment with increased statistical power.

Secondly, NASA TLX apparently lacked the sensitivity to pick up workload differences between manual control participants required to maintain a delicate balance between three control inputs (pitch, roll, and throttle / thrust) simultaneously and others (in the automation-assisted group) who enjoyed a substantial degree of autopilot assistance on two of those three axes. This insensitivity to automation's introduction is remarkable in and of itself, but the failure of the TLX to detect any workload differences after its subsequent removal is particularly surprising since this was when automation-induced training deficits were most evident during the process control training initially reported by Clegg et al. (2010).

In order for automation-induced training deficits to be diminished or washed out in comparison to previous research, there had to be a differential influence exerted across groups in which more error was induced in the manual group, and/or less error was accumulated by the auto assisted group. Given that the only two factors (the pre-test and the TLX) were added to experiment one's protocol compared to the previous research conducted by Blitch and Clegg (2011), it was logical to explore the nature of this replication failure by isolating these factors in the next experiment. Since error reduction differences can be gleaned from trial-by-trial analysis during training / landing task post-test, and any pre-treatment differences between groups would be revealed during training block one, it made sense to eliminate the pre-test from the next experiment and focus on the TLX as the primary variable of interest in further examination of the relationship between workload and automation-induced training deficits.

In summary, the goal of this experiment was to monitor workload for an undesirable drop in difficulty while replicating the automation-induced training deficits previously reported in the context of a simulated pasteurization process control task and UAV control task (Blitch & Clegg, 2011; Clegg & Heggestad, 2010). This goal was not met, most likely due to a combination of low

statistical power, and the introduction of two modifications to the experimental protocol; the pre-test and a subjective self-reported workload survey. Research pool limitations and further decomposition of the factors involved in this surprising result presented an obvious design modification as a logical next step in the experiment to follow – elimination of the pretest from experimental protocols, and examination of TLX influence through injection of a spatial ability test battery as a control for potential interference from a subjective instrument that was previously assumed to be non-invasive in nature.

Chapter III

Experiment 2: UAS Training with and without Subjective Self-Reported Workload

As a recap of issues from experiment one, it is suspected that one (or perhaps even both) of the modifications made to the initial protocols used by Blitch and Clegg (2011) to explore automation-induced training deficits in unmanned system control may have prevented a successful replication of the results therein. Commensurate with concerns established within instructional design literature for experimental confounds associated with the pre/post testing paradigm reported by Kromann, Jensen, and Ringsted (2009), it made sense to temporarily remove that factor from current procedures in order to focus on the potential influence of the second protocol modification – the use of a subject self-report instrument for workload measurement.

Despite evidence elsewhere that the NASA TLX is a reliable and sensitive workload measurement instrument (Hart, 2006; Hart & Staveland, 1988), the results from experiment one prompted a secondary review of the literature that investigated its potential for interference with task performance. That review could not identify a single study which compared a repeated-measures application of the NASA-TLX to a control condition that did not involve the requirement for subjective self-reporting. This stands as indirect support for Annett's skepticism regarding the validity of subjective self-report methods (Annett, 2002) and presents a direct challenge to the common assumption that retrospective self-report surveys administered after each block of performance are inherently non-invasive (Rubio, Díaz, Martín, & Puente, 2004).

Experiment 2 Goal

Given the surprising lack of a performance difference between groups in experiment one, the goal of this second experiment was to attempt another replication of the automation-induced training deficits previously observed by Blitch and Clegg (2011) and (Clegg et al., 2010) while monitoring workload.

Experiment 2 Research Questions and Hypotheses

While the primary research question regarding whether automation-induced training deficit are associated with an undesirable drop in difficulty remained the focus of activity in experiment two, a secondary research question was pursued in order to determine whether or not the mere presence of a subjective self-report instrument administered on a repeated-measures basis could influence the nature of automation's influence on learning performance.

In order to compensate for the time required to administer the TLX on multiple occasions throughout training, a control condition was established using a set of task relevant spatial ability tests that kept participants on the same approximate schedule, but avoided the retrospective nature of the TLX. This procedural adjustment left four of the previous six hypotheses from experiment one intact, while substituting a new proposition to test the influence of a subjective self-report instrument itself.

Commensurate with the notion of desirable difficulty from Bjork's theory of disuse (Bjork & Bjork, 2006), hypotheses one through four remained as follows: H1)The automation-assisted group would record substantially more control error in the final Landing Task Test than the manual control (MC) group, H2) the AA group would report lower workload levels when automation was introduced in TB2, H3) the automation-assisted group would report higher workload levels after automation was removed in TB3, and H4) the former automation-assisted group would report higher workload levels while automation was withheld during the Landing Test trials. The fifth hypothesis (H5) was addressed by the subjective self-report manipulation discussed above, namely that performance levels would fluctuate in the presence of the NASA TLX compared to a control group which was administered alternative (spatial ability) surveys in lieu thereof .

Experiment 2 Method

Participants. Fifty seven experimentally naïve undergraduate students participated in this experiment for optional, partial course credit, twenty three of whom were dropped from final

analysis due to procedural errors with the Synthetic Training Environment simulator, and two who were removed due to previous experience as a private pilot or with flight simulators in general.

Apparatus and Equipment. This experiment was conducted using the Predator STE simulator and NASA TLX as previously described in experiment one. Additional spatial ability survey instruments were obtained from online resources at University of California Santa Barbara (Hegarty, Richardson, Montello, Lovelace, & Subbiah, 2002; Hegarty & Waller, 2004; Kozhevnikov & Hegarty, 2001), and Psychometric Success (Newton & Bristoll, 2010) for use as alternatives to the NASA TLX in the control condition.

Design. This experiment used a 2x2 between subjects design with random assignments to the same autonomy assisted (AA) and manual control (MC) groups discussed in experiment one, with the addition of two survey conditions. The (W_TLX) condition involved the use of the NASA TLX to measure self-reported workload after each tutorial and training / test block. The (NO_TLX) condition substituted a subjective self-reported Sense of Direction instrument, a perspective taking Spatial Orientation examination, and a battery of other spatial ability tests in lieu of the NASA TLX.

Procedure

Training for all participants was organized into the same four training tasks and test sessions as experiment one: three Training Blocks (TB1, TB2, and TB3) and one Landing Task Test (LTT) block. Procedures used for experiment two were essentially identical to those used by Blitch and Clegg (2011) and experiment one except for the substitution of spatial ability surveys in the NO_TLX condition.

Experiment 2 Results

Test Performance Data. Analysis of test performance for this experiment was initially conducted using an omnibus 2x2 ANOVA across group and workload conditions. This analysis revealed no main effect for group in approach ground track $F(1,28)<1$, $p>.05$, $\eta^2_p= .004$, final ground track $F(1,28) =1.00$, $p>.05$, $\eta^2_p= .034$, or glide slope $F(1,28)<1$, $p>.05$, $\eta^2_p= .029$. Nor was a main

effect for workload condition observed in approach ground track $F(1,28)<1$, $p>.05$, $\eta^2_p= .015$, final ground track $F(1,28)<1$, $p>.05$, $\eta^2_p= .031$, or glide slope $F(1,28)<1$, $p>.05$, $\eta^2_p= .007$. Although evidence of an interaction between group and workload condition fell short of statistical significance in all three performance metrics, the large effect sizes observed in final ground track $F(1,28)=3.10$, $p>.05$, $\eta^2_p =.100$, and moderate effect sizes recorded for approach ground track $F(1,28)=1.92$, $p>.05$, $\eta^2_p=.064$ and glide slope $F(1,28)=2.10$, $p>.05$, $\eta^2_p =.070$ suggested the possibility of a crossover relationship that warranted further investigation.

Although the interaction between survey condition and automation group fell short of statistical significance in this omnibus analysis, the large effect size observed in the final ground track metric and moderate effect sizes observed in the other two performance metrics still suggest that an underlying interaction may be present that escaped numerical detection due to the relatively small sample size involved. Given this situation, a secondary analysis was conducted which considered group effects in isolation within each survey condition.

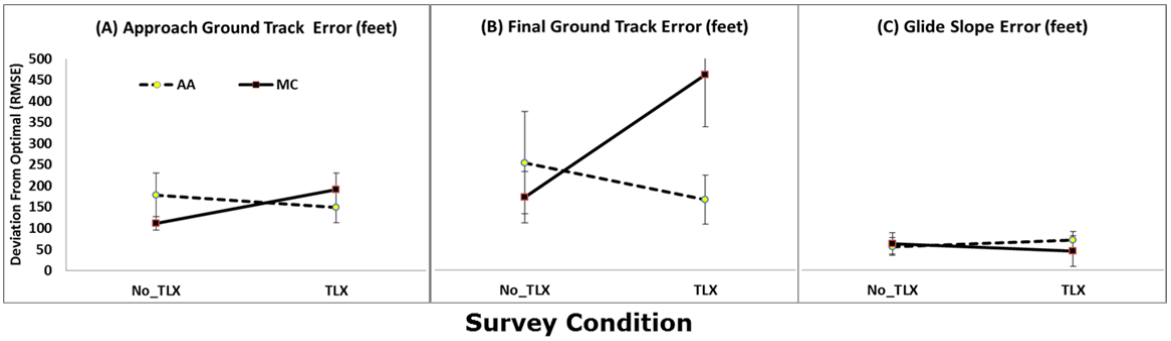

Figure 7: ***Experiment 2 Test Performance on Landing Task by Automation Group and Self-Report Workload Measure Presence or Absence.*** **Deviation in root mean square error from optimal control in the Landing Task is provided for three metrics: Panel (A) Ground Track on Approach, Panel (B) Final Ground Track, and Panel (C) Glide Slope. Dotted lines with open circles show participants who previously received automation assistance (AA) from the autopilot activated in training block 2. Solid lines with filled squares show participants who performed in manual control mode (MC) throughout. Survey Condition shows whether participants completed a self-report TLX survey between blocks ("TLX"), or a control survey ("No TLX"). Error bars show standard error.**

One way ANOVAs were conducted on data collected under each of the survey conditions described earlier: a NO_TLX condition in which participants took spatial abilities tests and evaluations in lieu of subjective self-report requirements, and a W_TLX condition which required

each participant to retrospectively rate their workload level after each major training event. Considering the NO_TLX condition first, there was more error and variance recorded in approach ground track by the AA group (M=179, SD=127) than the MC group (M=112,SD=42), as indicated in Figure 7 and supported by Levene's test $F(5.95,11)=12.39$, $p<.01$. This variance difference was accounted for by a t-test with unequal variance assumed that fell short of statistical significance $t(5.96)=1.24$, $p>.05$, despite the presence of a large effect size (Cohen's $d = 0.72$) presumably due to variance inequality and small sample size.

There were no significant differences noted between the AA group (M=254,SD=295), and MC group (M=172,SD=162) in final ground track $F(1,11)<1$, $p>.05$, $\eta^2_p=.034$, or glide slope AA(M=191,SD=51), MC(M=174,SD=68), $F(1,11)<1$, $p>.05$, $\eta^2_p=.021$ despite the appearance of the general (albeit statistically non-significant) trend toward more error in the AA group that is evident in Figure 7.

An identical analysis conducted on data collected under the W_TLX condition revealed a trend in the opposite direction with more error and variance reported by the MC group. A large accumulation of final ground track error recorded by the MC group (M=462, SD=406) dwarfed that recorded for the AA group (M=167, SD=163). This difference proved to be statistically significant with a very large effect size even after correcting for a homoscedasticity violation indicated by Levene's test $F(13.96,17)=9.94$, $p<.01$) with a t-test assuming unequal variance $t(13.96,17)=-2.18$, $p<.05$, Cohen's $d=0.84$.

While the test performance differences by group in approach ground track AA (M=150,SD=103), MC (M=192,SD=129, $F(1,17)<1$, $p>.05$, $\eta^2_p= .03$) and glide slope AA (M=132,SD=55), MC (M=206,SD=120, $F(1,17)=2.58$, $p>.05$, $\eta^2_p=.13$, fell short of statistical significance, the effect size for the latter suggests that the MC group performed worse in managing glide slope as well but escaped numerical detection due to its small sample size..

Training Performance Data. In order to further explore the group differences identified in the test data described above, analysis of training data in this experiment was conducted with a repeated-measures ANOVA applied across all three training blocks in both survey conditions. Greenhouse Geisser corrections were once again applied for violations of sphericity where appropriate.

A significant interaction between group and training block was recorded alongside large effect sizes for all four performance metrics: altitude $F(2,28)=4.056$, $p=.05$, $\eta^2_p=.13$, heading $F(2,28)=4.100$, $p=.05$, $\eta^2_p=.13$, indicated airspeed $F(2,28)=3.956$, $p=.05$, $\eta^2_p=.124$, and criterion passed $F(2,28)=6.894$, $p<.01$, $\eta^2_p=.20$ suggesting the existence of a differential influence of survey condition on performance throughout training. This observation is further supported by the large main effect for group observed in heading $F(2,28)=11.346$, $p<.01$, $\eta^2_p=.29$, with widening differences between survey conditions after automation was introduced in training block two as indicated in Figure 8.

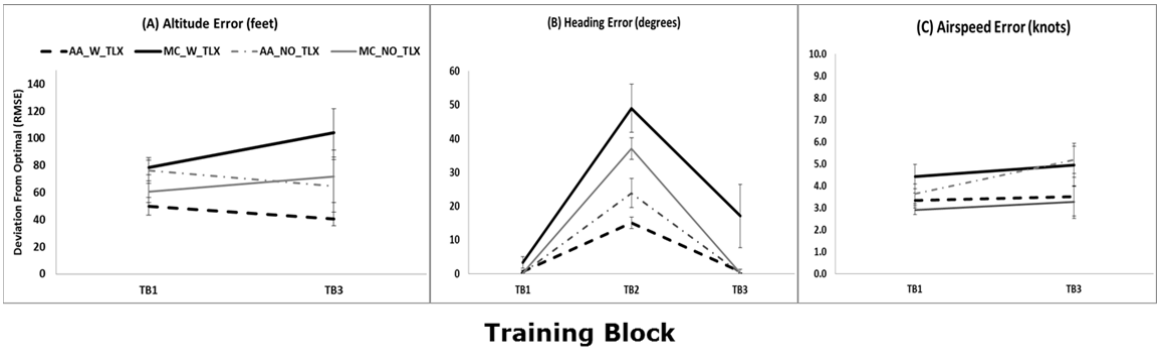

Figure 8: ***Experiment 2 Training Performance by Training Block (TB), Automation Group, and Self-Report Workload Measure Presence or Absence.*** **Deviation in root mean square error from optimal control during training tasks is provided for three metrics: Panel (A) Altitude Error, Panel (B) Heading Error, and Panel (C) Indicated Airspeed (IAS) Error. Dashed or dotted lines show participants who previously received automation assistance (AA) from the autopilot activated in Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Survey Condition shows whether participants completed a self-report TLX survey between blocks ("W_TLX" indicated by black lines), or a control survey ("No TLX" indicated by grey lines). Error bars show standard error.**

An interaction between block and group was evident in both heading $F(2,28)=5.18$, $p<.01$, $\eta^2_p=.16$, and criterion passed $F(2,28)=5.18$, $p<.01$, $\eta^2_p=.16$, providing additional evidence of a

differential influence exerted by automation in training block two as indicated in Figure 8. As in experiment one, a main effect for block was also observed in both heading $F(2,28)=11.35$, $p<.01$, $\eta^2_p=.61$, and criterion passed $F(2,28)=20.16$, $p<.01$, $\eta^2_p=.42$ although these data are more indicative of fluctuation across tasks rather than automation influence. There were no other significant within subject effects observed.

Workload Data by Block. NASA TLX scores were analyzed in a manner identical to the procedure used in experiment one with group means compared via repeated-measures AVOVA taken across tutorials and hands on training sessions for each of the three training blocks as well as the final test block. Once again these data reveal a main effect for trial $F(6,17) = 8.84$, $p<.01$, $\eta^2_p=0.34$ by virtue of a large effect size, but no main effect for group $F(1,17)<1$, $p>.05$, $\eta^2_p=0.04$, nor any evidence of an interaction of group with trial $F(6,17) <1$, $p>.05$, $\eta^2_p= 0.02$) as indicated in Figure 9.

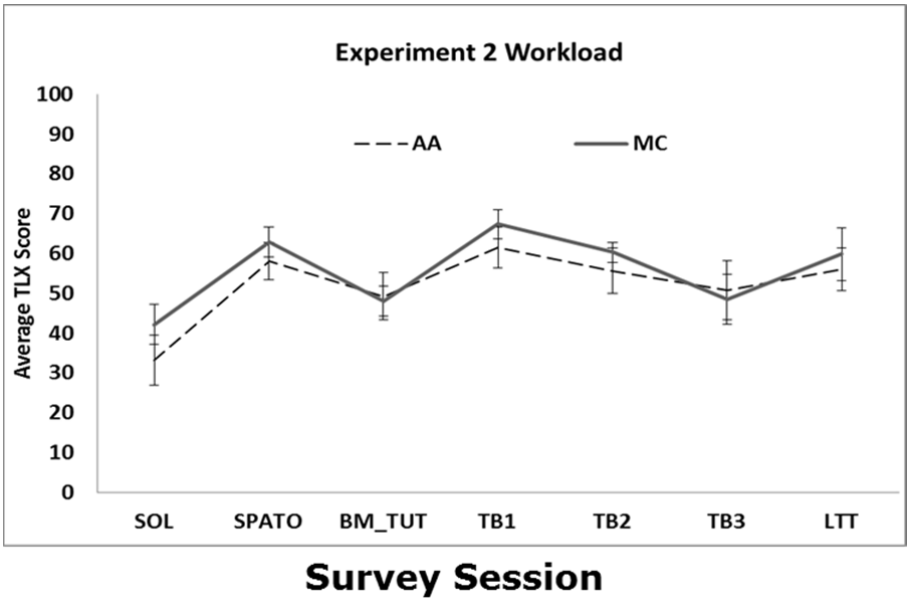

Figure 9: ***Experiment 2 Workload Measured by Average Task Load Index (TLX) Score Across Training Session.*** **Dashed lines show participants who received automation assistance (AA) from the autopilot activated in Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Survey Session shows the training period after which the TLX was administered starting with a Solitaire computer familiarization session (SOL), a Spatial Orientation (SpatO) test, the Basic Maneuver Tutorial (BM TUT) session, each of the three hands-on Training Blocks (TB1, TB2, TB3), and the Landing Task Test (LTT). Error bars show standard error.**

Experiment Two Discussion

Neither performance nor workload data provide support for any of the stated hypotheses except H5, which predicted that performance would fluctuate in the presence of the TLX. It is interesting to note, however, that a trend toward more error recorded by the automation-assisted group in the No_TLX survey condition revealed a large effect size and pattern consistent with previous research (Blitch, 2012; Blitch & Clegg, 2011; Clegg et al., 2010), suggesting that automation-induced training deficits may actually have been replicated if the TLX were not present and a larger sample size were available to increase statistical power.

The substantial increase in control error for the manual control group that occurred only in the presence of the TLX, shows a trend that is opposite of that observed in previous research and presents the surprising implication that administering a subjective self-report workload instrument on a repeated-measures basis might actually influence training performance. On the one hand it is considered unlikely that a retrospective survey like the TLX which was always administered *after* a particular training session could influence performance at all. It is possible for participants who are administered self-report surveys on a repeated basis, however, to build up an expectation of future reporting that could potentially divert cognitive resources otherwise devoted to the task at hand toward the anticipated retrospection requirement. Such an influence is not without historical precedent and theoretical backing.

In their examination of cell phone use while driving, for example Horrey and Wickens (2006) postulate that self-monitoring of workload could act as a secondary task, changing the nature of the primary task. Repetitive self-report activities have also been shown to disrupt the association of successive events (Heuer & Schmidtke, 1996) in a manner that might interfere with the transfer of skills acquired during part-task training into the integrated mental models required for successful performance during the final whole-task test administered at the end of training (Gutzwiller et al., 2013a; Wightman & Lintern, 1985). The potential influence of such a distraction on training

effectiveness would be consistent with introspective links to the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990) where verbal productions related to visual memories cause decreased performance. Potential shifts in the focus of attention from external events to internal conceptualization (Carpenter, Lohse, Healy, Bourne, & Clegg, 2013) might also impede learning for robot control tasks where an accurate and rapid response to external stimuli is the key to success.

The very large effect sizes associated with the MC group's increased error in the W_TLX condition provides strong evidence that a subjective self-report instrument of this nature may be more invasive than previously thought. What's perhaps most interesting is the differential nature of this invasiveness with the manual control group incurring far more error in the presence of the TLX, whereas error recorded by automation-assisted group appears to be equivalent across both conditions.

In consideration of the workload data itself, the pattern of average weighted TLX scores recorded across tutorials and hands on training blocks in Figure 9 presents an almost identical trend to that observed in experiment one. The statistical significance and large effect size associated with the main effect for trial provides strong evidence that the TLX is quite sensitive to differences in difficulty by task type as it varied throughout training. The non-significant main effect and very small effect size for workload by group, however, along with the lack of any significant interaction and effect size between group and training block provides supporting evidence that the NASA TLX remained insensitive to the level of automation as it was introduced in training block two, removed prior to training block three, and withheld once again during the landing task test trials.

In retrospect, the goal of this experiment was partially achieved in that subjective self-report influences were controlled for by the inclusion of a non-subjective testing battery which also helped verify the assumption of random assignment in spatial ability across individual differences between participants. Once this was done a clear trend toward replication of previous research was observed

in the NO_TLX survey condition although low statistical power arguably limits how conclusive that argument can be made.

Sample size and statistical significance notwithstanding, the prominent reversal of these effects observed in the presence of the TLX along with its demonstrated insensitivity to automation's presence raise question as to how appropriate it may be as a cognitive measurement instrument for this particular instructional paradigm. It may be that workload doesn't actually fluctuate in this specific training arrangement, or it may be that the observed changes were simply too subtle to pick up on a subjective basis when averaged across all twenty trials in each training or test block. In any case, the effects reported above provided ample motivation to incorporate the use of alternative workload monitoring techniques into future experimental design as discussed in the following chapter.

Chapter IV

Transition from Subjective to Neurophysiological Workload Measures

Despite the enormous potential for scientific insight to be gained through examination of mental effort and exploration of its relationship with task difficulty, definitive characterization and measurement of cognitive workload has proven to be a challenging endeavor in its own right (Hancock & Meshkati, 1988; Jex, 1988). For more than two decades researchers have developed, modified, and redesigned a variety of methods to address the need for accurate and timely measurement of cognitive workload across a broad spectrum of demanding tasks (Parasuraman & Wilson, 2008). These methods fall into four general categories: subjective self-reports, secondary task assessments, physiological measures, and neuro-physiological monitoring.

While subjective self-report instruments have been quite popular in behavioral research (Hart, 2006), claims of their reliability, sensitivity and scientific value may have been over stated (Annett, 2002; Rubio et al., 2004). In his comprehensive examination of this issue, Annett (2002) poses the question of whether the application of subjective measures to behavioral research presents more of an expression of art than a contribution to science. In support of the artistic side of this question, Annett points to a wealth of evidence indicating the presence and profound influence of human bias in any form of subjective assessment. The substantial variance imposed by this factor is amplified even further when such judgments are made in relation to one's assessment of their own cognitive workload (Guastello, Shircel, Malon, & Timm, 2013).

Although a targeted literature review supports the notion that self-reported ratings like those recorded with the NASA-TLX have been put to good use across a relatively wide variety of research endeavors, particularly within the aviation community (Hart, 2006), a number of studies have reported negative results regarding its sensitivity to workload fluctuations that are apparent in other measures. In their examination of cockpit activity, for example, (Hankins & Wilson, 1998) reported

that physiological measures provided extensive insight to mental demand during flight and correlated well with task activity, but subjective measurements with the TLX provided few statistically significant differences – none of which correlated with observed performance. In more recent research, Teo, Schmidt, Szalma, Hancock, and Hancock (2013) showed a similar insensitivity to workload when the TLX was used to explore workload effects in vigilance training. They report that although different feedback mechanisms via knowledge of results manipulations had a positive impact on performance, no differences in global workload were found when measured with the TLX.

Given the notion that the subjective self-report approach could only provide workload measurement in a post hoc manner that is not always sensitive to differences in task difficulty in real time (Moroney, Biers, & Eggemeier, 1995), an alternative secondary task monitoring method was developed to assess workload for multi-task scenarios that are typical of cockpit activity in the aviation and mass transportation communities (Griffiths & Gillespie, 2005). This approach proceeds with the assumption that performance on a secondary task starts to degrade substantially when the availability of surplus cognitive resources becomes limited due to increased consumption by the primary task. In other words, secondary task performance degradation implies increased primary task workload – but only when the *type* of resources involved are common to both task requirements (Wickens, 2008b).

This secondary task approach has been valuable in providing timely assessments regarding the relationship between workload distribution and situational awareness for air traffic controllers (Endsley & Kaber, 1999; Kaber & Endsley, 2004), shared control inputs for haptic interfaces (Griffiths & Gillespie, 2005), functional display differences (Smith, Fadden, & Boehm-Davis, 2005), and attentional processes associated with cell phone driving distractions (Horrey & Wickens, 2006) and imperfect cockpit alerts (Wickens & Colcombe, 2007). Unfortunately, however, the very fact that the secondary task diverts cognitive resources away from the primary task complicates the

examination of automation induced effects in that one cannot be sure whether its influence is confined to either task or a combination of the two.

In an examination of the potential dissociation between subjective measures and performance for example, Yeh and Wickens (1988) offered an explanation of why cognitive workload levels may become increasingly difficult to self-report accurately as demands increases throughout complex tracking and memory tasks required in challenging operational fields such as air traffic control. Commensurate with predictions made under Multiple Resource Theory, allocation of cognitive resources becomes more contentious as demand increases – especially when two or more tasks are competing for the same type of resource like working memory. At low levels of mental demand, sufficient cognitive resources may be available to perform the task at hand and simultaneously maintain awareness of one's workload level in the process. As demands increase, however, participants have to choose between satisfactory performance and accurate self-reporting, usually with deference to the former (Yeh & Wickens, 1988).

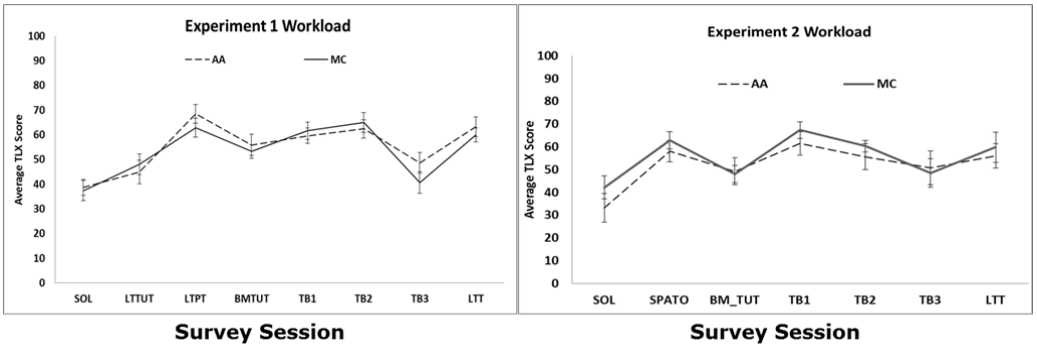

Figure 10: Experiment 2 Workload Measured by Average Task Load Index (TLX) Score Across Training Session. Dashed lines show participants who received automation assistance (AA) from the autopilot activated during Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Survey Session shows the training period after which the TLX was administered. Error bars show standard error.

The results from experiment one and two provide support for this notion. The data plots in Figure 10 show that TLX scores consistently showed remarkable sensitivity to variations in task difficulty between the solitaire baseline task, tutorials, and hands on training blocks across both experiments. Performance, however, was shown to suffer much more in the manual control group

than the automation group when the self-reported TLX was added to the experimental protocols. Given that the automation-assisted group had two thirds of its primary task demand handled by the auto-pilot in training block two, it is likely that those participants were able to report workload accurately without any degradation of performance on the primary task. Given their requirement to maintain simultaneous control of three finely tuned motor control inputs during a coordinated turning maneuver (vs. only one for the automation-assisted group) it makes sense that the manual control group would suffer more primary task degradation if they got the impression via repeated administration of the TLX that reporting workload might be more important to the experiment than actually flying the aircraft effectively.

It is important to note at this point that all of the participants who volunteered for this particular study did so with the expectation of receiving academic research credit. Since few endeavored to be professional robot operators or drone pilots in their future careers it is unlikely that any of them considered the skills they were trying to acquire to be of prominent professional value. Given that state of affairs it is logical to conclude that at least some of the participants might be inclined to consider the subjective workload reporting task of be of equal if not higher importance than the aircraft control task – especially since psychology students are ostensibly familiar with surveys of this sort. It might be interesting to replicate this study in a population sample of professional pilot trainees to see if this effect could be reversed.

In any case there is reason to believe that the NASA-TLX administered repeatedly throughout the first two experiments in this study functioned as a secondary task not directly related to the acquisition of skill necessary to land a Predator UAV during the final test. As such it represents the type of extraneous activity that Mane and Wickens (1986) discuss in the misapplication of difficulty, and proponents of Cognitive Load Theory would most likely classify as extraneous load (Paas et al., 2003; Sweller et al., 1998; van Merriënboer et al., 2006). Both accounts speak to the diversion of mental resources away from the primary task and provide ample motivation

to pursue alternative measures of cognitive load in the performance of complex tasks as suggested by (Antonenko, Paas, Grabner, & van Gog, 2010; Lee, 2013; Paas et al., 2003).

While the methodology of the secondary task approach seems relatively straightforward, the accuracy of its conclusion is also highly conditional upon the relationship between the primary and secondary task. In addition to the many dimensions with which resource contention and sharing can occur between these two tasks throughout the stages, modalities, and codes of processing (Wickens, 2008b), one has to consider how relevant the secondary task is to the overall training mission in order to avoid the deleterious influences described earlier by inappropriate difficulty distribution (Mane & Wickens, 1986), and excess levels of extraneous load (Paas et al., 2005; van Merriënboer et al., 2006). In complex tasks examined in pursuit of high ecological validity this may be particularly problematic since the nature of that relevance and even the goals of the secondary task may change dynamically in a rapid and complicated fashion.

Consider the flight training example discussed earlier in experiments one and two, and the situation where a participant is asked to intermittently report value changes from a particular dashboard gauge as a secondary task to measure workload on the primary task of maneuvering the aircraft through a mountainous environment enroute to an airport located at high altitude. If the gauge chosen for the secondary task happens to be the tachometer (which measures engine RPM), increases or decreases observed would provide direct feedback on one of the three control inputs (throttle setting) that form the basis of primary task skill acquisition. As such, this particular secondary task would involve high dependence and shared resource consumption with the primary learning task.

If the secondary task involves monitoring fuel consumption rates, however, a less direct relationship would be maintained between the primary task and secondary task since other factors besides throttle control can influence it. In good weather with a large fuel reserve on board, the relevance of this task (or level of germane load) to the overall mission might be small, despite the

moderate level of dependence (or intrinsic load) that exists between the primary and secondary tasks. As the weather remains calm but a powerful headwind develops that consumes more fuel than anticipated and threatens the pilot's ability to make it to the destination, however, the relevance of the secondary task may increase, potentially altering performance even though the difficulty of the primary task remains relatively stable. Similar claims can be made regarding the density of radio traffic monitoring as pilots approach busy airports, visibility constraints, etc.

Under dynamic conditions like those described above, the accuracy of secondary task performance as a workload reporting method comes into question. Since previous research suggests that the TLX might have functioned as a secondary task in experiments one and two (Horrey & Wickens, 2006), and exploration of the relationship between primary and secondary task performance in complex tasks can become as complicated as examination of learning itself (Hutchins et al., 2013; Parasuraman, Sheridan, & Wickens, 2008; Wickens, 2008a), this method was not considered to be a viable alternative for use in the third experiment of this study.

During the latter half of the 20th century, the need for crisp, non-invasive characterization of mental effort in real time gave rise to a number of physiological measures that have been used within the human factors community to monitor cognitive workload across a variety of challenging task sets. Fournier, Wilson, and Swain (1999), for example used a combination of eye blinks, respiration patterns, and heart rate variation alongside EEG, behavioral, and subjective methods to monitor workload fluctuation during a variety of tracking and vigilance tasks typically experienced in the aviation cockpit. While heart rate, behavioral, and subjective measures were found to be sensitive to changes in workload, only eye blink rate and behavioral measures were sensitive to training over time.

A similar study conducted by Hankins and Wilson (1998), found that heart rate variation was generally indicative of task difficulty but not very useful in determining the specific cause of workload variation during flight. Eye activity, by comparison, was found to be sensitive only to

visual tasks and therefore presented a more specific diagnostic value. Subjective measures, by comparison, showed general trends toward workload fluctuation, but fell short of statistically significant differences with conclusive results. Marshall (2002), provided similar support for physiological measures in the form of a pupillary reaction technique called the Index of Cognitive Activity which attempted to dissociate light sensitivity from workload based dilation effects. Although this work provided evidence for the notion that pupilometry may be robust in the presence of arousal effects, it also acknowledged the potential for stress and fatigue to present confounds in workload measurement, especially as they pertain to intensive vigilance and/or tracking tasks that might produce eye strain. It is with this potential for physiological contamination of cognitive processing data that Parasuraman & Wilson (2008) consider the emergence of direct neurological monitoring as perhaps the most promising workload assessment method of all.

In a recent comparison of neurophysiological with self-reported instruments de Guinea, Titah, Leger, and Micheneau (2012) conducted a multi trait / multi method analysis in order to examine accuracy in the measurement of cognitive load, engagement and arousal across a variety of task environments. Although they concluded that neurophysiological methods suffer less measurement error than subjective measures, they stopped short of claiming them to be superior because their inherently narrow focus on neurology makes it difficult for them to tap into the entire construct space that participants may experience subjectively. They advocate for a combination of both subjective and physiological measures to be used in order to avoid what they call mono-method bias. In earlier research, the combined workload measure developed by Ryu and Myung (2005) based on eye blink intervals, heart rate variability and electro encephalography presented strong evidence that physiological workload measurement augmented with neurological data (considered hereafter as neurophysiological measurement) provided exceptional sensitivity to task difficulty, but also enabled discrimination of workload associated with arithmetic processing from visual tracking tasks.

If the cost of all instruments was equal and small, then it would probably make sense to employ multitude of devices to assess workload and thereby control for as many variables as possible and arrive at the greatest accuracy. But costs (especially in terms of time availability) are not equal in workload measurement. While the setup time for many neurophysiological methods should not be ignored, they typically present very little cost to research during the data collection process itself. Not only does this maximize the efficiency of measurement, but it preserves cognitive momentum and train of thought between successive training activities rather than interrupting them in a potentially invasive manner (Antonenko et al., 2010; Parasuraman & Wilson, 2008; Popovic, Stikic, Berka, Klyde, & Rosenthal, 2013).

While the multi-instrument approach proposed by (de Guinea et al., 2012) may be appropriate for exploratory studies where the factors involved are unclear, this study has already made two unsuccessful attempts to replicate previous research with subjective measures attempting to confirm theoretical backing that implicates workload in the onset of automation-induced training deficits. Given the time and resource limitations described at the beginning of this study, therefore, the decision was made to focus on the neurophysiological approach for the next experiment in pursuit of maximum efficiency and minimum complication of the factors involved in examining the nature of workload's relationship with training automation.

In describing the recent emergence of the relatively new field of neuroergonomics within the broad realm of behavioral science, Parasuraman and Wilson (2008) point out that neurophysiological devices have recently come into favor as a workload monitoring process that can balance the need for crisp temporal accuracy with non-intrusive comfort and precision. For additional reviews, see Dorneich et al. (2007), Baldwin and Penaranda (2012), and Wilson, Estepp, and Davis (2009). One approach in particular championed by Berka et al. (2004), Wilson and Russell (2003), and others involves the use of EEG signals collected from the scalp surface to provide classifications of mental activity and cognitive state. Not only does this approach avoid attentional disruption from invasive

introspection and/or secondary task effects, but it allows for dynamic assessment of cognitive activity in real time.

One such instrument made available for this study was the wireless B-Alert system (Advanced Brain Monitoring, Incorporated, Carlsbad CA), which has been used to measure mental workload across a number of dynamic and challenging task domains ranging from submarine crew interaction (Gorman, Martin, Dunbar, Stevens, & Galloway, 2013) to intelligence analyst evaluations (Cowell et al., 2007). One important feature of this system is the wireless nature of the hardware which all but eliminates participants' awareness of its presence within just a few minutes of donning it in a lightweight hat/helmet fashion. This presents a significant contrast to high density EEG systems that are cable intensive and typically require electrodes to be placed on the face and other highly sensitive dermal regions which remain prominent in the perceptual realm of participants.

Whereas other researchers developed classification models based on Artificial Neural Network (ANNs) and other methods that were highly sensitive to individual differences but involved extensive setup time, Berka et al. (2004) endeavored to produce rapidly acquired yet generalizable classifiers by discriminate function analyses carried out on historical data sets and later adjusted to accommodate individual differences. The B-Alert system uses a proprietary epoch by epoch classification method that models cognitive workload as an internal process based primarily (but not exclusively) on working memory while engagement is modeled in terms of cognitive states more closely (but not orthogonally) devoted to perceptual processing and external perspective.

To summarize this classification approach, the B-Alert headset typically acquires 9 channels of EEG and ECG. The sensor locations for the EEG comprise: Fz, Cz, POz, F3, F4, C3, C4, P3, and P4. Data are sampled at 256 Hz with a band pass from 0.5 Hz to 65 Hz (at 3 dB attenuation) obtained digitally with Sigma-Delta A/D converters. The RF link is frequency-modulated to transmit at a rate of 57 kBaud in the 915 MHz ISM band. By utilizing a bidirectional mode, the firmware allows the host computer to initiate impedance monitoring of the electrodes, select the transmission channel (so

two or more headsets can be used in the same room), and monitor battery power of the headset. Data are acquired across the RF link on a host computer via an RS232 interface. Data acquisition software then stores the EEG data on the host computer. The proprietary acquisition software used in this process also includes artifact decontamination algorithms for eye blink, muscle movement, and environmental/electrical interference such as spikes and saturations.

In order to utilize the cognitive state algorithms within the B-Alert system for engagement and workload, a set of three neurocognitive assessments are administered by the Alertness and Memory Profiler (AMP) software package. This custom software was developed to time the presentation and capture of each stimulus response, while creating a file that stored the simultaneously acquired electroencephalographic (EEG) signals for subsequent classification and comparison during the operational or research task of interest.

The first of these "baseline" assessments involves a three choice vigilance task (3CVT) which requires subjects to discriminate one primary geometric shape (with 70% occurrence) from two secondary geometric shapes (with 30% occurrence each) displayed on a computer screen with a stimulus presentation interval of 200 ms over a 5-minute test period and an inter-stimulus interval range from 1–3 seconds. Participants are instructed to respond as quickly as possible to each stimulus presentation by selecting the left keyboard arrow to indicate target stimuli, and the right keyboard arrow to indicate non-target stimuli.

The second assessment, an eyes-open (EO) finger tapping task, requires participants to respond to a visual stimulus (a 10 cm red circular image positioned in the center of a computer monitor presented for 200 ms, and repeated every 2 seconds for 5 minutes) by hitting the keyboard spacebar in time with target appearance. This is followed by the third and final baseline assessment comprised of an eyes-closed (EC) finger tapping task that requires participants to respond to an auditory tone presented every 2 seconds for 5 minutes by hitting a keyboard spacebar in time with the tone onset.

For every epoch of EEG data five variables are computed within each 1-Hz bin between 3Hz and 40Hz: 1) the logged PSD 2) relative power compared to total power (between 3-40Hz), and 3-5) Z-scores computed from the means and standard deviations of each of the baseline tasks indicated above. A stepwise analysis of 381 variables (5 x 38 bins for each FzPOz and CzPOz pair plus eye blink speed) results in the selection of 19 predictors compared in matrices to arrive at the probability that each epoch is dominated by one of 4 engagement states: 1) Probability of Sleep Onset (pSLEEP) based on comparisons with previously collected sleep deprivation data and correlations with eye blink interval, duration, etc., 2) Probability of Distraction (pDSTRCT) based on comparisons with the EC task profile, 3) Probability of Low Engagement (pLoEng) based on comparisons with the EO (passive vigilance) task profile, and 4) Probability of High Engagement (pHiEng) based on comparisons with the 3CVT (active vigilance) task profile. The system also produces a single workload measure based on an adjustment of "global" working memory profile from each individual's 3CVT profile and historical data collected from Forward/Backward Digit Span (FBDS) assessments. This measure is labeled as p_WKLD or the probability of high working memory saturation based on Forward / Backward Digit Span profiling.

While many EEG collection systems are designed around a cable intensive clinical setup that is typically limited to laboratory environments, the B-Alert system was initially intended to be worn by drivers and pilots in their actual operating environment (e.g. aircraft cockpit, truck cab, etc.) in order to provide real time alerts of drowsiness or sleep onset that threatened vehicular operations and safety (Berka et al., 2004). Over time, the system came to be used in another fashion - to monitor workload levels and cognitive states in a number of dynamic and challenging environments from Aegis missile simulators (Berka et al., 2005) to elementary school classrooms (Stevens, Galloway, & Berka, 2007). Its classifiers have been validated across a broad range of cognitive measures (Berka et al., 2007; Johnson et al., 2011; Popovic et al., 2013) with demonstrated dissociations between workload and engagement metrics in teamed map reproduction (Stevens & Galloway, 2013) math

skills (Galán & Beal, 2012) and submarine crew coordination (Stevens, Galloway, Berka, & Sprang, 2009; Stevens, 2012).

It should be noted here that despite the validation and utility record of the B-Alert system, Wilson and colleagues have shown superior performance when using artificial neural networks to comprehensively model and account for individual differences in EEG signal patterns as they pertain to neurophysiological workload measurement (Russell, Wilson, Rizki, Webb, & Gustafson, 2005; Wilson et al., 2009) and time of day influences (Wilson, Russell, Monnin, Estepp, & Christensen, 2010). This approach to EEG classification has also been demonstrated to significantly enhance UAV operator capabilities when used as a cueing mechanism for adaptive automation (Wilson & Russell, 2007).

Despite these advantages, an artificial neural network EEG classifier was not available for this study, nor was acquisition or development of one feasible within financial constraints of the university laboratory involved. It may also be important to note that tuning an artificial neural network to accommodate the many combinations of EEG in each individual's cognitive profile can be a tedious process that takes quite a bit of time. In the case of full time pilots or aircrew who are participating in research that may directly improve performance in their professional endeavors, this investment of personal time is unquestionably worthwhile. When measuring EEG of novice college research participants with strict time limits imposed by institutional policy however, it was important to use a device with minimal setup cost in order to maximize time available for data collection with only a slight compromise in fidelity of individual differences profiling. For this reason and the fact that it was immediately available at no cost, a B-Alert X-10 system was chosen as the workload monitoring instrument for the remaining experiment in this study.

Chapter V

Experiment 3: UAS Simulator Training with EEG Based Workload and Engagement Measures

As a recap of research objectives, previous research identifying the potential for automation to have a deleterious influence on training efficiency (Blitch & Clegg, 2011; Clegg & Heggestad, 2010; Clegg et al., 2010; Gutzwiller et al., 2013a) served as a motivation for further exploration of automation-induced deficits via concurrent workload monitoring. The two subsequent attempts to investigate this influence in the context of automation's relationship with workload failed to replicate previous results, quite possibly due to interference in cognitive resource allocation invoked by introspection demands posed by subjective self-reporting that was required on a repeated-measures basis.

After exploring alternative methods for measuring workload that avoided the apparently invasive nature subjective self-report instruments observed in experiments one and two, a third attempt at replicating automation-induced training deficits was made using an EEG based instrument applied to the same general protocols within the Synthetic Training Environment. It was hoped that the non-invasive nature of the neurophysiological instrument chosen for this task would not only provide a workload profile that was more sensitive to the presence of automation in training, but also do so with high enough resolution to enable inspection of each participant's cognitive state on a trial-by-trial basis.

Experiment 3 Goal

The goal of the third experiment in this study was to further explore the general research question regarding automation's influence on training by developing understanding of the mechanism(s) that underlie automation-induced decrements in training effectiveness - particularly as they relate to mental workload of the novice operators involved. In order to do this, a replication of

those deficits had to be achieved while simultaneously monitoring workload in a non-invasive manner.

Experiment 3 Research Questions and Hypotheses

The general research question for this experiment remained the same as in the two prior experiments, namely whether an undesirable reduction in cognitive workload can be implicated as a potential source of automation-induced training deficits in novice UAV operators. Based on prior research linking skill acquisition deficits (Blitch & Clegg, 2011; Clegg et al., 2010; Gutzwiller et al., 2013a; Gutzwiller, Clegg, Smith, Lewis, & Patterson, 2013b; Healy, Wohldmann, Parker, & Bourne, 2005b), to learning theory (Bjork & Bjork, 2006; Kornell & Bjork, 2008; Schmidt & Bjork, 1992; Shiffrin & Schneider, 1984), it was logical to expect that when the auto-pilot was invoked in the second training block, neurophysiological workload measures would drop substantially, thereby providing evidence that a participant had fallen out of the desirable difficulty "sweet spot" previously discussed. Of course, it was also expected that performance would suffer at some point soon thereafter although the latency between automation onset and performance drop had not been specifically addressed. Given the epoch by epoch classification potential of the equipment made available for this experiment, it was expected that the specific timing of this influence might be revealed as well. Since prior research also identified spatial orientation score as the strongest correlation between individual differences and performance on these particular training tasks (Blitch, Bauder, Gutzwiller, & Clegg, 2012) the assumption of random assignment was checked using this particular metric.

Within the rubric of this general design a slightly modified set of hypotheses were established. Once again, it was expected that the automation-assisted group would record substantially more control error on a trial-by-trial basis during the final Landing Task Test than the manual control group (H1), and that this performance difference would be accompanied by reduced workload levels recorded on a neurophysiological basis (H2) when automation was introduced

during training block two. Commensurate with previous research using this specific simulation paradigm, it was not expected that performance differences would be observed once automation was removed in training block three (H3), but it was expected that the automation-assisted group would report higher workload levels since they were no longer provided with auto-pilot assistance and had to resume control of airspeed and altitude with inputs provided by pitch on the stick and adjustment of throttle levels (H4). It was also expected that this difference in neurophysiological workload would persist long enough to manifest in the landing task trials as well, although at a diminished level compared to the previous (third) training block (H5). In accordance with the assumption that random assignment would distribute participants of equivalent spatial ability between participants, it was also expected that no significant difference between spatial orientation scores would be observed between the automation-assisted and manual control groups (H6). Despite evidence from Endsley and colleagues that automation might induce out of the loop effects in skilled participants performing tasks in quasi-operational non-educational settings (Kaber & Endsley, 1997), it was expected that the influence of automation on engagement levels for novice operators would be negligible primarily because of task novelty, and the fact that only a moderate level of imperfect automation was being used (H7).

Experiment Three Method

Participants. Forty-four experimentally naïve undergraduate students participated in this study for optional, partial course credit, with seven dropped from analyses due to technical issues with the Predator Synthetic Training Environment simulator or because participants reported previous flight experience. Another four participants were dropped due to problems acquiring EEG baseline classification data that fell below acceptable signal quality thresholds established within the B-Alert EEG processing software package. It may be worth noting that the research assistants responsible for collecting data in this experiment were more experienced in setup and operation of the Synthetic Training Environment simulator and thus less inclined to make procedural mistakes

resulting in dropped performance data. These same research assistants were less experienced in collecting data with the B-Alert system, however, so the number of participants dropped due to excessive baseline noise should not be misconstrued as an indication of confounds from individual differences or other sources of variance.

Apparatus and Equipment. Performance data was collected during this experiment using the Predator Synthetic Training Environment (STE) simulator used in related research (Martin et al., 1998), and discussed earlier in experiments one and two. Neurophysiological workload and engagement data was collected using the B-Alert wireless EEG classification system developed by Advanced Brain Monitoring (ABM) of Carlsbad CA that was discussed in the previous chapter.

Procedure. Participants were first fitted with the B-Alert device and then performed the three baseline vigilance tasks developed by Berka et al. (2004) in order adjust the centroids of each individual's EEG profile for workload and engagement. Participants then reviewed a basic maneuver tutorial which provided an overview of control inputs and essential aerodynamic concepts associated with fixed wing flight training. Hands on basic maneuver training sessions were then conducted in a sequence of three component training blocks, each comprised of fifteen one minute trials. This reduction of five trials within each training block was the only significant procedural modification of the experimental paradigm followed by Blitch & Clegg (2011), and was necessary to accommodate the amount of time for EEG baseline tasks to be completed by each participant and stay within the institution's four hour research pool limitation.

As in experiments one and two above, the first training block required the trainee to reduce airspeed while holding altitude and heading constant. The second block involved a180-degree heading change, holding altitude and airspeed constant, and the third block involved a descent task during which both altitude and airspeed were reduced simultaneously while holding heading constant. Following a short (fifteen minute) tutorial, participants performed five landing task trials,

each lasting three to four minutes in duration. Final performance was measured in this novel "test" block that combined the trained maneuvers into an integrated landing task.

Participants were randomly assigned to two conditions, an automation-assisted group and a manual control group. Automation was invoked only during training block two, and involved two of the simulator's three auto pilot functions to hold altitude and airspeed steady. This allowed the automation-assisted group to focus exclusively on the roll axis during the turning task, whereas the manual control group was required to also manage airspeed and altitude via throttle and pitch inputs. Both groups returned to full manual control for training block three and the landing task test.

Measures. Performance feedback provided to the operator after each trial was automatically recorded by the simulator, and used the same metrics described in experiments one and two above. Since sleep onset and distinctions between active and passive vigilance lie beyond the scope of the current study, analysis of EEG data recorded from the B-Alert classifier discussed earlier focused on probabilities of distraction, engagement, and workload. As a recap of those measures, workload is classified for each participant based on historical data sets of forward / backward span tasks for the general population modulated by frequency differentials recorded during individual baseline tasks. High engagement, low engagement, and distraction classifications are determined by comparisons of each second of EEG collected during flight tasks with baseline frequency profiles established by each participant's individual performance on the three choice vigilance task, the eyes-open vigilance task, and the eyes-closed spacebar task respectively.

Experiment Three Results

In order to specifically repeat the analytical procedures used in Blitch and Clegg (2011) an initial analysis of results for this experiment was conducted on performance and EEG data averaged across all five final landing task test trials using a one way ANOVA procedure except when significant results obtained on Levene's test for unequal variance required a t-test to be performed with equal variance not assumed. Given the potential for averaging procedures to mask fluctuations

in performance and EEG at a higher level of granularity (Haider & Frensch, 2002), this was followed by a more detailed investigation of both test and training data on a trial-by-trial basis as indicated in the previous discussion of experimental design.

Average Performance and EEG Analysis for the Landing Task Test. Initial analysis of averaged performance data did not reveal any significant differences in approach ground track error between the automation-assisted (AA) group (M=126,SD=72) and the manual control (MC) group (M=130,SD=89, $F(1,31)<1$, $p>.05$, $\eta^2=.00$). Nor were there any group differences noted in final ground track error AA (M=149,SD=142), MC (M=96,SD=117, $F(1,31)=1.39$, $p>.05$, $\eta^2=.04$).

A much larger accumulation of glide slope error was recorded by the automation-assisted group (M=194,SD=86), however, compared to the manual control group (M=120,SD=54), $t(24.88,31)=-2.96$, $p<.01$), equal variance not assumed based on Levene's test $F(1,31)=9.58$, $p<.01$). A large effect size (Cohen's $d = 0.93$) was observed for this data in a nearly identical performance pattern to that observed by Blitch and Clegg (2011). Despite this strong support for the replication hypothesis (H1) recorded in the performance data, no differences in workload were observed during the landing task test block between the automation-assisted group (M=.567,SD=.134) and the manual control group (M=.586,SD=.090), $F(1,31)= 0.23$, $p>.05$, $\eta^2=.00$). A surprisingly low probability of engagement was recorded for the automation-assisted group (M=.869, SD=.084), however, compared with the manual control group (M=.962, SD=.032), $t(19.03,31)=4.20$, $p<.01$, equal variance not assumed based on Levene's test $F(31)=11.20$, $p<.01$. This between subjects difference was accompanied by an even larger effect size ($d=1.20$) than that observed in the performance data. The automation-assisted group also recorded a higher average probability of distraction (M=.072,SD=.092) than the manual control group (M=.021,SD=.021), $t(16.53,31)=-2.17$, $p<.05$), equal variance not assumed based on Levene's test $F(31)=8.47$, $p<.01$, with a large effect size (Cohen's $d =0.73$) observed on this metric as well.

In pursuit of greater insight into the specific relationship between learning rate fluctuations and cognitive state change, performance and workload data were analyzed on a trial-by-trial basis using a repeated-measures ANOVA procedure with Greenhouse-Geisser corrections made for violations of sphericity assumptions where appropriate.

Trial-by-trial Performance and EEG Analysis for the Landing Task Test. Performance on the five landing task test trials is depicted in Figure 11. It should be noted that no participant in this study ever reached criterion on the landing task which provides testament to just how difficult this skill acquisition challenge is for novice trainees. Main effects for trial in approach ground track $F(2.03,31)=4.96$, $p<.05$, $\eta^2_p=.12$, final ground track $F(2.08,31)=4.39$, $p<.05$, $\eta^2_p=.12$, and glide slope $F(2.96,31)=3.16$, $p<.05$, $\eta^2_p=.09$, provide evidence that significant error fluctuations occurred with an observed reduction trend, suggesting that learning was taking place throughout these trials despite the "test" label place upon them. The main effect for group in glide slope $F(1,31)=5.12$, $p<.05$, $\eta^2_p=.14$ suggests that the automation-assisted trainees continued to suffer from an automation-induced training deficit despite the fact that full manual control had been restored for the entire second half of the experiment – the final thirty of the total sixty minutes of "stick time" provided to all participants (3 x 15 x 1 minute training trials + 5 x 3 minute test trials).

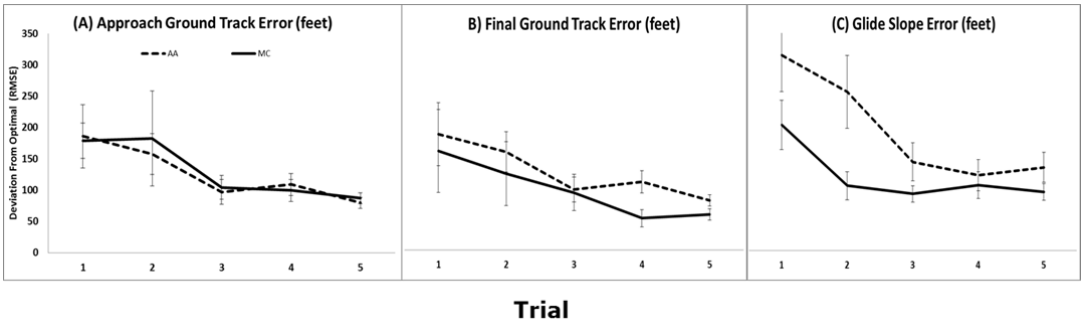

Figure 11: ***Experiment 3 Test Performance on Landing Task by Automation and Trial.*** **Deviation in root mean square error from optimal control in the Landing Task is shown for three metrics: Panel (A) Ground Track on Approach, Panel (B) Final Ground Track, and Panel (C) Glide Slope. Dashed lines show participants who received automation assistance (AA) from the autopilot activated in training block 2. Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

No main effect for group was observed in approach ground track $F(1,31)<1$, $\eta^2_p=.00$, or final ground track $F(1,31)<1$, $\eta^2_p=.01$, suggesting that the automation-assisted participants were able to recover from the heading control challenges they faced in training block three. No interaction of trial with group was observed in approach ground track $F(2.03,31)< 1$, $\eta^2_p=.00$, final ground track $F(2.08,31)<1$, $\eta^2_p=.01$, or glide slope $F(2.96,31)<1$, $\eta^2_p=.02$, which is consistent with the idea that the nature of the error fluctuations themselves showed similar patterns of improvement across time for both groups.

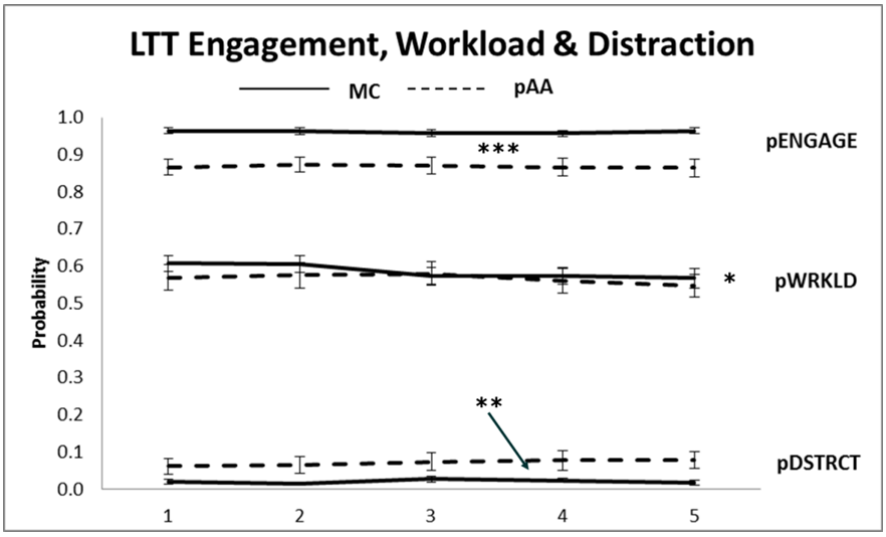

Figure 12: ***Experiment 3 EEG based probability measures of Engagement, Workload, and Distraction recorded by the B-Alert classifier across all five Landing Task Test trials.*** **Dashed lines show participants who received automation assistance (AA) from the autopilot activated in Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

As indicated in Figure 12, EEG data show a main effect for group in both engagement $F(1,31)=18.56$, $p<.01$, $\eta^2_p=.37$, and distraction $F(1,31)=4.96$, $p<.05$, $\eta^2_p=.14$ while no such difference was observed for workload $F(1,31)<1$, $\eta^2_p=.01$. A main effect for trial was evident in workload $F(1,31)=2.96$, $p<.01$, $\eta^2_p=.13$, but not engagement $F(2.13,31)<1$, $\eta^2_p=.01$ or distraction $F(2.03,31)=2.25$, $p>.05$, $\eta^2_p=.07$, providing evidence that the B-alert workload classifier was sensitive enough to show a decreasing workload trend over time during the five landing task test

trials. No interaction of trial with group was observed for workload $F(2.96,31)=1.56$, $p>.05$ $\eta^2_p=.05$, however, nor for engagement $F(2.13,31)<1$, $\eta^2_p=.01$ nor distraction $F(2.03,31)=1.12$, $p>.05$, $\eta^2_p=.04$.

Trial-by-trial Performance and EEG Analysis for Training Block One. A repeated-measures ANOVA conducted on performance recorded during training block one indicated a main effect for trial in altitude control error $F(5.91,31)=4.30$, $p<.01$, $\eta^2_p= .12$, airspeed control error $F(4.00,31)=11.39$, $p<.01$, $\eta^2_p=.27$), and criterion passed $F(2.38,31)=5.23$, $p<.01$, $\eta^2_p= 14$), but not heading control error $F(3.83,31)<1$, $\eta^2_p=.03$. This is consistent with the notion that training had an influence on trial-by-trial error with a general trend toward reduction, as indicated in Figure 13. The relatively flat learning profile for heading was expected given that directional control was held constant without the input tradeoffs required in the case of increased pitch for decreased airspeed in the altitude and airspeed metrics. No main effect for group or interaction of group with trial was found in any of the performance metrics during training block one which supports the assumption that random assignment of participants between odd (manual control) and even (automation-assisted) groups was successful in controlling for any significant individual differences in performance.

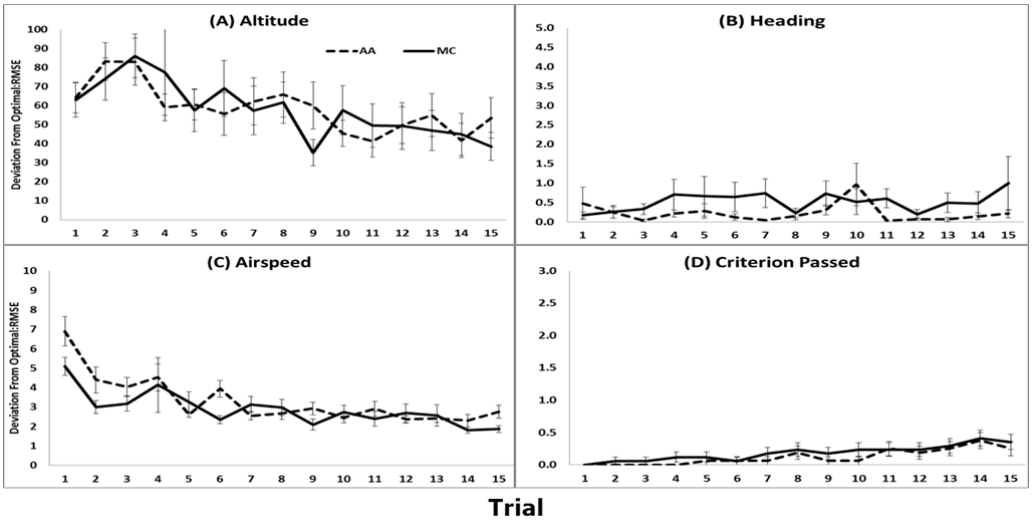

Figure 13: ***Experiment 3 Performance by Automation Group and Trial in Training Block 1.*** **Deviation in root mean square error from optimal control during training tasks is shown for three metrics: Panel (A) Altitude, Panel (B) Heading, and Panel (C) Indicated Airspeed (IAS). Panel D shows the average number of criterion passed per trial. Dashed lines show participants who received automation assistance (AA) from the autopilot activated in Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

EEG data for training block one showed no main effect for group in distraction $F(1,31)<1$, $\eta^2_p=.01$, engagement $F(1,31)=2.07$, $p>.05$, $\eta^2_p=.07$, or workload $F(1,31)<1$, $\eta^2_p=.01$ which is expected of pre-treatment cognitive profiles. No main effect of trial was recorded for distraction $F(4.37,31)<1$, $\eta^2_p=.03$, engagement $F(5.56,31)<1$, $\eta^2_p=.031$, or workload $F(7.73,31)=2.01$, $p>.05$, $\eta^2_p=.07$, nor was there any indication of an interaction between trial and group for distraction $F(4.37,31)<1$, $\eta^2_p=.03$, engagement $F(5.56,31)<1$, $\eta^2_p=.02$, or workload $F(7.73,31)<1$, $\eta^2_p=.01$, as indicated in Figure 14. These data also support the assumption that random assignment of participants was successful in controlling for individual differences in workload, engagement and distraction as well as performance.

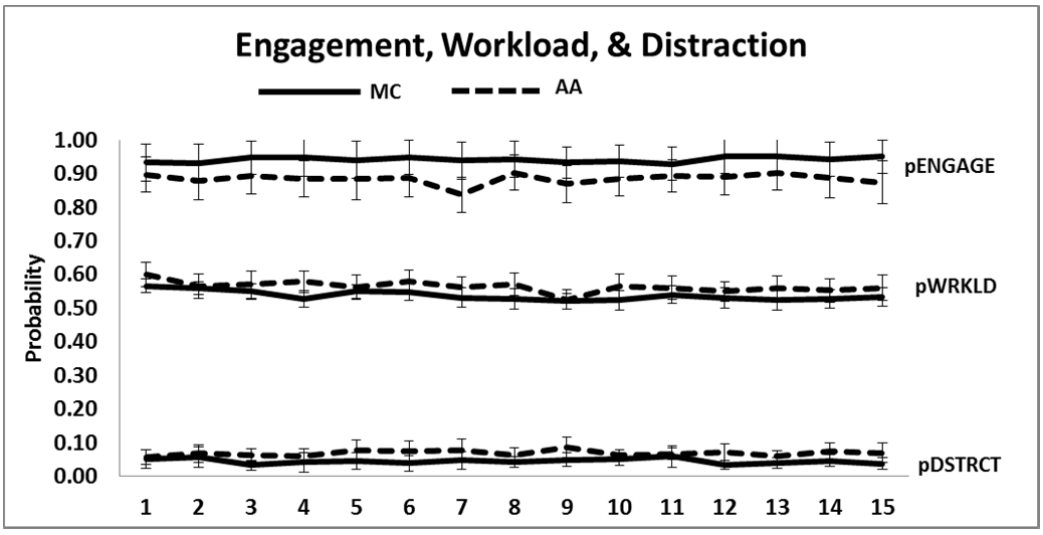

Figure 14: ***Experiment 3 EEG based probability measures of Engagement, Workload, and Distraction recorded by the B-Alert classifier across all fifteen trials in Training Block 1.*** **Dashed lines show participants who received automation assistance (AA) from the autopilot activated in Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

Trial-by-trial Performance and EEG Analysis for Training Block Two. Data collected during training block two can only be meaningfully compared for the heading performance metric since the autopilot was invoked for altitude and airspeed control in the automation-assisted group. These data indicate a main effect for trial $F(6.55,31)=17.06$, $p<.01$, $\eta^2_p=.36$, indicating that error fluctuated significantly over the course of training, but not for group $F(1,31)<1$, $\eta^2_p=.01$, which is surprising since manual control participants had to struggle with three aspects of control simultaneously while

the autopilot allowed the automation-assisted group to focus exclusively on bank angle and thereby have a higher likelihood of reducing their heading error. The observed interaction between group and trial $F(6.55,31)=3.24$, $p<.05$, $\eta^2_p=.10$ might provide a bit of insight regarding the surprising lack of an overall difference between groups, since it appears in Figure 15 that the automation-assisted group may actually have started out with a bit more error than the manual control group (perhaps because they have to adjust to autopilot influence) but were able to reduce their error more substantially over time in the form of a crossover trend.

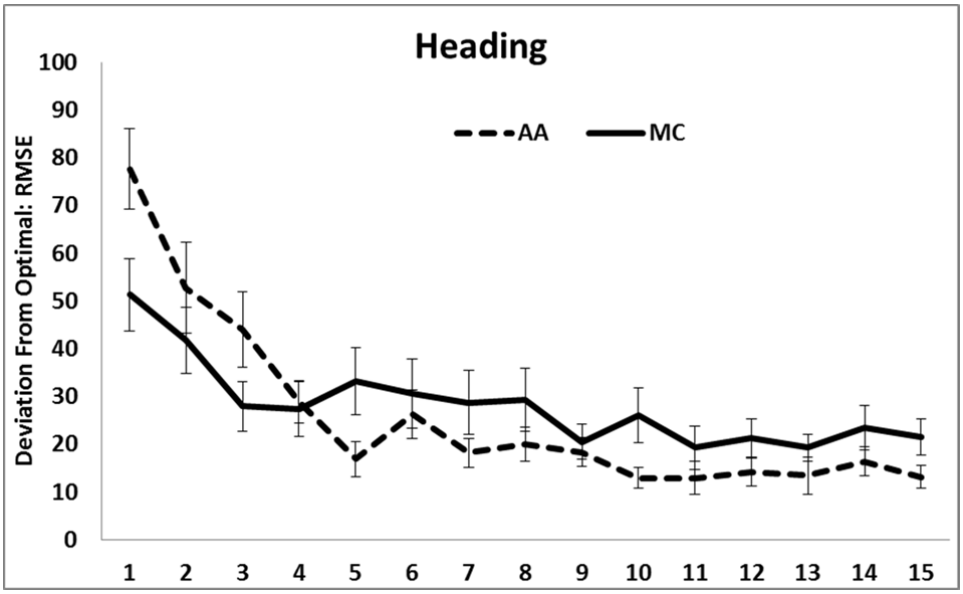

Figure 15: ***Experiment 3 Performance by Automation Group and Trial in Training Block 2.*** **Deviation in root mean square error from optimal Heading control during training is shown on the vertical axis. Dashed line shows participants who received automation assistance (AA) from the autopilot activated at the beginning of the training block. Solid line shows participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

EEG data for training block two once again showed no main effect for trial in distraction $F(5.39,31)=1.37$, $p>.05$, $\eta^2_p=.04$, engagement $F(6.38,31)=1.03$, $p>.05$, $\eta^2_p=.03$, or workload $F(6.14,31)<1$, $\eta^2_p=.03$ as indicated in Figure 16. These data do, however, show a significant main effect for group in both engagement $F(1,31)=13.7$, $p<.01$, $\eta^2_p=.31$, and distraction $F(1,31)=5.60$, $p<.05$, $\eta^2_p=.15$, suggesting the presence of substantial group differences in cognitive state despite the fact that no significant differences were noted between groups in performance. There was no group

effect observed for workload $F(1,31)<1$, $\eta^2_p=.00$, nor were there any interactions observed between trial and group for engagement $F(6.38,31)=1.22$, $p>.05$, $\eta^2_p=.04$, workload $F(6.14,31)=1.05$, $p>.05$, $\eta^2_p=.03$, or distraction $F(5.39,31)=1.12$, $p>.05$, $\eta^2_p=.04$.

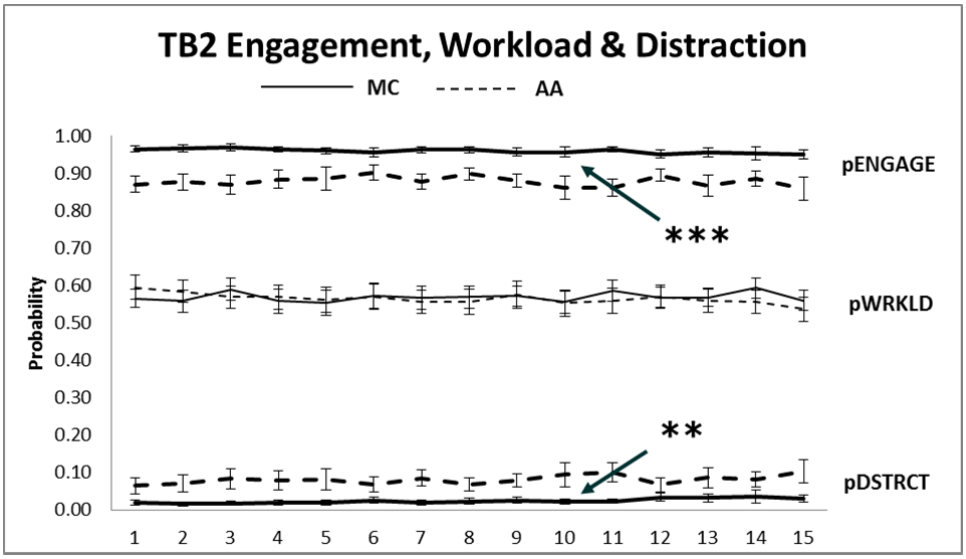

Figure 16: ***Experiment 3 EEG based probability measures of Engagement, Workload, and Distraction recorded by the B-Alert classifier across fifteen trials in Training Block 2.*** **Dashed lines show participants who received automation assistance (AA) from the autopilot activated at the beginning of the training block. Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

Trial-by-trial Performance and EEG Analysis for Training Block Three. Despite the fact that automation was removed during training block three the automation-assisted group recorded more control error across all four performance metrics as indicated by a main effect for group in altitude $F(1,31)=9.55$, $p<.01$, $\eta^2_p=.24$, heading $F(1,31)=4.38$, $p<.05$, $\eta^2_p=.12$, airspeed, $F(1,31)=8.37$, $p<.01$, $\eta^2_p=.21$, and criterion passed $F(1,31)=6.06$, $p<.05$, $\eta^2_p=.16$. The main effects observed for trial in altitude $F(5.91,31)=2.96$, $p<.01$, $\eta^2_p=.09$, airspeed $F(5.60,31)=7.28$, $p<.01$, $\eta^2_p=.19$, and criterion passed $F(8.64,31)=4.91$, $p<.01$, $\eta^2_p=.13$, provide evidence via significant error fluctuation that learning occurred for all participants on all performance metrics except heading control $F(2.25,31)=1.98$, $p>.05$, $\eta^2_p=.06$, as depicted in Figure 17. No interactions between trial and group occurred in altitude $F(5.91,31)<1$, $\eta^2_p=.02$, heading $F(2.25,31)=1.19$, $p>.05$, $\eta^2_p=.04$, airspeed $F(5.60,31)=1.03$, $p>.05$, $\eta^2_p=.03$, or criterion reached $F(8.64,31)=1.10$, $p>.05$, $\eta^2_p=.03$.

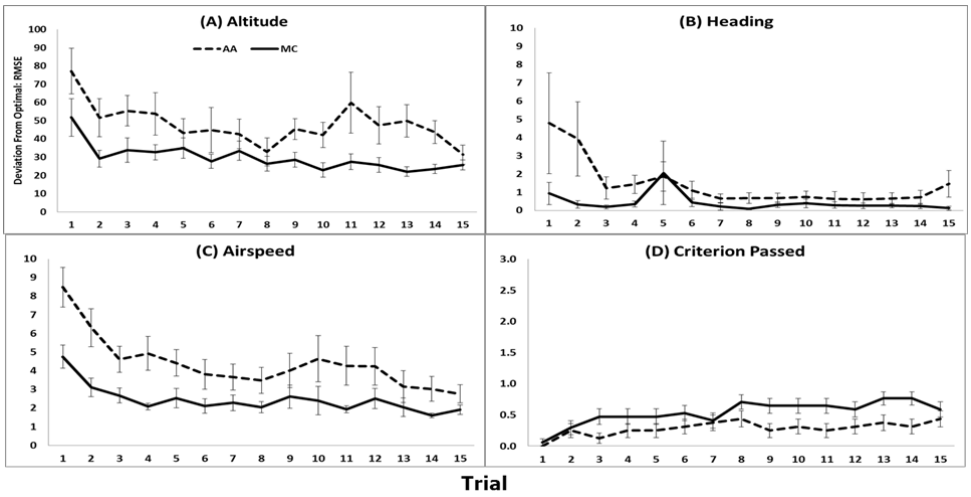

Figure 17: ***Experiment 3 Performance by Automation Group and Trial in Training Block 3.*** **Deviation in root mean square error from optimal control during training is shown for three metrics: Panel (A) Altitude, Panel (B) Heading, and Panel (C) Indicated Airspeed (IAS). Panel D shows the average number of criterion passed per trial. Dashed lines show participants who received automation assistance (AA) from the autopilot activated during Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

Main effects for group observed in EEG data displayed in Figure 18 show evidence of a continued disengagement $F(1,31)=7.92$, $p<.01$, $\eta^2_p=.304$ and distraction $F(1,31)= 7.92$, $p<.01$, $\eta^2_p=.17$ trend, despite the removal of automation before training began in block three. No group effect was observed for workload $F(1,31)<1$, $\eta^2_p=.01$, however, nor were there any trial effects recorded for engagement $F(6.43,31)=1.09$, $p>.05$, $\eta^2_p=.04$, workload $F(8.54,31)=1.64$, $p>.05$, $\eta^2_p=.05$, or distraction $F(5.26,31)<1$, $p>.05$, $\eta^2_p=.03$. No interaction of trial with group was observed for engagement $F(6.43,31)=1.15$, $p>.05$, $\eta^2_p=.04$, workload $F(8.54,31)<1$, $\eta^2_p=.03$, or distraction $F(5.26,31)<1$, $p>.05$, $\eta^2_p=.03$.

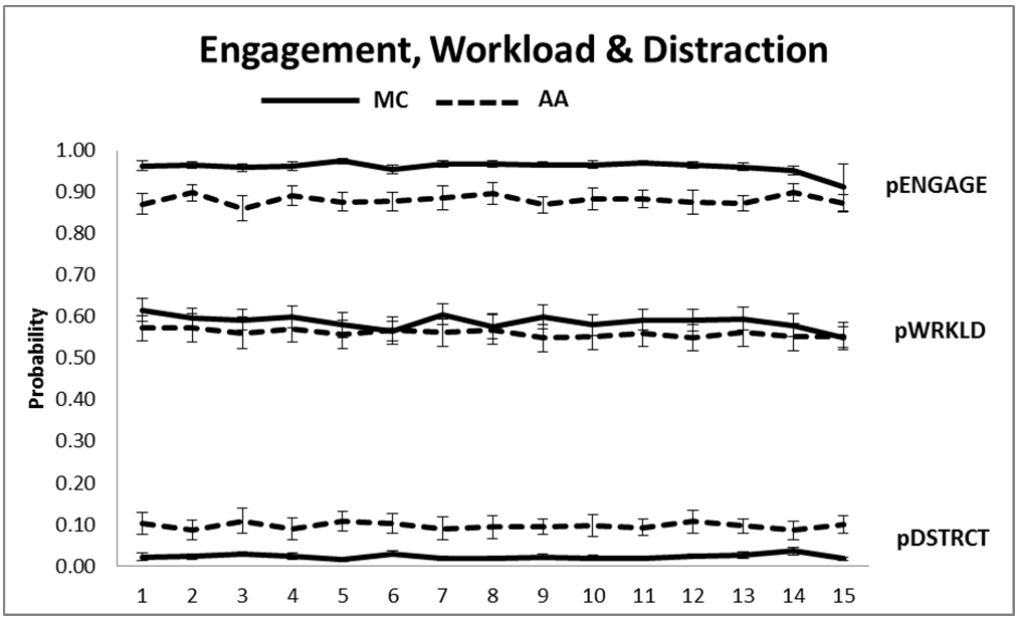

Figure 18: Experiment 3 EEG based probability measures of Engagement, Workload, and Distraction recorded by the B-Alert classifier across all fifteen trials in Training Block 3. **Dashed lines show participants who received automation assistance (AA) from the autopilot activated during Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

Study Wide EEG Analysis by Block. Commensurate with the block by block workload comparison between experiments one and two discussed in Chapter IV, a similar analysis was conducted on the EEG data indicated in Figure 19 using a repeated-measures ANOVA with Greenhouse Geisser corrections made where appropriate. These data also show a main effect for group in both engagement $F(1,31)=12.41$, $p<.01$, $\eta^2_p=.29$, and distraction $F(1,31)=4.08$, $p<.05$, $\eta^2_p=.12$ across training and test while no such difference was observed for workload $F(1,31)<1$, $\eta^2_p=.00$. No main effect for trial was observed in engagement $F(1.57,31)<1$, $\eta^2_p=.01$, workload $F(1.88,31)=1.57$, $p>.05$, $\eta^2_p=.00$ or distraction $F(1.42,31)<1$, $\eta^2_p=.02$, nor was any interaction between group and trial recorded for engagement $F(1.57,31)=2.28$, $p>.05$ $\eta^2_p=.07$, workload $F(1.88,31)=2.80$, $p>.05$, $\eta^2_p=.08$, nor distraction $F(1.42,31)=1.70$, $p>.05$, $\eta^2_p=.05$.

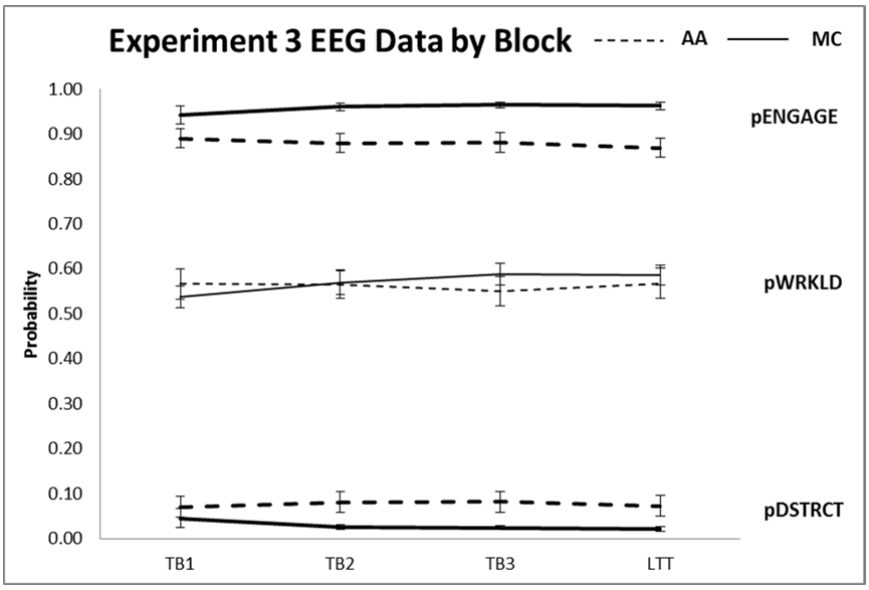

Figure 19: ***Experiment 3 average EEG based probability measures of Engagement, Workload, and Distraction recorded by the B-Alert classifier across all three training blocks and the Landing Task Test block.*** **Dashed lines show participants who received automation assistance (AA) from the autopilot activated during Training Block 2 (TB2). Solid lines show participants who performed in manual control mode (MC) throughout. Error bars show standard error.**

Experiment 3 Discussion

The landing task data analyzed above present a performance profile consistent with previous research on both an averaged and trial-by-trial basis. As such they suggest a successful replication of automation-induced performance deficits reported in previous research (Blitch & Clegg, 2011; Clegg et al., 2010). Given the lack of any group differences noted in the pre-treatment training block confirming the random assignment hypothesis (H6), and the large effect sizes reported in the same pattern as previous research (e.g. glide slope) these data provide strong support for hypothesis one despite having fewer trials during each training block than the procedure used by Blitch and Clegg (2011) or experiments one and two in this study.

EEG data collected during the landing task test, however, presented a surprising lack of support for the undesirable difficulty drop hypothesis (H2) in that workload was recorded at relatively steady levels across training with no substantial difference between groups. Moreover, the negligible effect sizes involved suggest that this is not due to a lack of statistical power. Even more

surprising, however, is the fact that the automation-assisted group recorded substantially less engagement and much higher distraction levels than the manual control group – a situation that is statistically backed once again by very large effect sizes. This not only fails to support the equivalent engagement hypothesis (H7), but it implicates the influence of an out of the loop mechanism as the source of automation-induced training deficits in lieu of an undesirable reduction in difficulty.

Looking across the trial-by-trial data as well as the EEG data averaged across each training block in Figure 19, this disengagement trend provides amplified evidence for the out of the loop account championed by Endsley (1995) and negligible support, if any, for the increased difficulty approach described by Wickens et al. (2012b). While the slight dip in EEG based workload plots between training block two and three combined with a moderate yet statistically insignificant effect size suggests the remote possibility of workload interaction between group and block, the substantial and highly significant performance and engagement deficits recorded by the automation-assisted group compared to the manual control group presents much more compelling evidence that participants had been drawn far out of the loop while automation was present in training block two and were struggling to reengage once it was removed in the session thereafter.

Despite the benefit shown with the part-task fractionalization approach reported by Mane, Coles, Wickens, and Donchin (1983), performance data show that automation invoked during training block two in this experimental paradigm did not provide any performance benefit for the automation-assisted group. The same EEG disengagement pattern discussed earlier was evident in training block two, however, despite the lack of any specific group difference in performance. Data recorded in training block one before automation was introduced show no significant difference in any performance or EEG metrics. This provides verification of the random assignment assumption and strong evidence of automation's complicity as the most likely source of disengagement effects and performance deficits observed after it was initially invoked, subsequently removed, and then withheld throughout the test period.

The goal of this experiment was to replicate automation-induced training deficits while monitoring workload and engagement with a neurophysiological instrument in lieu of subjective self-reporting. Although several hypotheses went unsupported via evidence provided in the exact opposite direction, this goal was achieved. The replication of these deficits in a nearly identical pattern with equivalent magnitude to those reported in previous research is quite interesting given that they were produced with 25% fewer trials. That fact, along with the identification of disengagement and not workload as their most likely source presents an intriguing challenge for theoretical accounts of automation's influence on training to be discussed in the next chapter along with findings from experiment one and two.

Chapter VI

Conclusions, Implications, and Future Work

Previous research exploring the role of workload in automation-assisted training has provided a wealth of empirical evidence identifying both its costs and benefits across a broad range of literature addressing whole-part techniques (Dattel, 2006; Fontana et al., 2009; Gutzwiller et al., 2013a; Wickens et al., 2012b), Cognitive Load Theory (Paas et al., 2003; Sweller et al., 1998; van Merriënboer et al., 2006), and desirable difficulty (Bjork & Bjork, 2006; Bowers et al., 1998; Healy et al., 2005a; Schneider, 1985). From an investment perspective on instructional design, however, it is not enough to simply record whether automation helps or hurts learning in the presence of selected pedagogical factors. In order to make wise decisions on how to design and/or restructure training programs for *future* tasks of unknown complexity with unclear consequences and uncertain resource constraints, it is vital that managers understand *why* automation has a particular positive or negative influence on learning. This is the first study we are aware of that clearly and empirically identifies disengagement, distraction, and excess introspection from subjective self-reporting as concurrent and prominent sources of cost associated with the injection of automation into training.

Conclusions

Disengagement v. Desirable Difficulty Drop. Perhaps the most surprising and interesting conclusion drawn from the work presented here is that automation's influence on training was not associated with workload fluctuation as often suspected in prior research. On the contrary, findings from three experiments using a partial autopilot treatment in this study provide compelling evidence for disengagement and distraction as a primary source of automation-induced training deficits in lieu of an assumed drop in otherwise desirable difficulty.

Once the pre-test and subjective measures were removed from protocols in experiments one and two, a nearly identical replication of the magnitude and pattern of automation's detrimental

influence established in previous research was achieved. While learning theory previously discussed in terms of skill acquisition for complex tasks (Bjork & Bjork, 2006; Healy, Ericsson, & Bourne Jr, 1999; Healy et al., 2005a; Schmidt & Wulf, 1997; Schneider et al., 2002; Wickens et al., 2012b; Wulf & Shea, 2002) provided ample reason to believe that an undesirable reduction in difficulty may have been the primary source of these effects, there is no evidence in this particular study that automation exerted that kind of influence.

Although self-reported workload in the first two experiments fluctuated significantly between tasks of varying difficulty, none of the measures employed in this study, subjective or neurophysiological, showed any evidence that the introduction or removal of automation had any significant influence on workload. This suggests that workload measurement may not provide the best method for understanding automation's influence during unmanned system operator training. It certainly proved to be inadequate in detecting the presence of automation in this particular experimental paradigm.

Against this backdrop of workload's insufficiency as a solitary indicator of automation's influence, neurophysiological measures of engagement fluctuated everywhere that performance between groups varied in comparison to pre-treatment conditions, and in one case (during training block two) actually provided advanced warning that differences in brain state had occurred long (fourteen trials) before performance differences appeared. This provided a consistent and powerful indication that participants who received automation assistance suffered an immediate and substantial engagement loss that was directly associated with performance deficits lasting long after automation was removed. Based on this evidence, it seems that instead of pushing novice trainees out of a sweet spot of desirable difficulty, automation pulled them out of the loop altogether through a disengagement and distraction mechanism that is more consistent with the voluminous amount of research exploring the relationship between automation and; situational awareness (Endsley, 1995; Kaber & Endsley, 2004) and responsibility offloading (Bainbridge, 1983; Cummings, 2006; Moray

& Inagaki, 1999) rather than complacency (Bailey & Scerbo, 2007; Parasuraman & Manzey, 2010; Singh, Molloy, & Parasuraman, 1993) and over-reliance (Dixon & Wickens, 2006; Wickens & Colcombe, 2007).

As Endsley (1995) points out in her discussion of situational awareness theory, both working memory and attention are primary factors associated with out of the loop effects. In a number of dissociations conducted via the application of neurophysiological methods, workload measurement has been shown to be more closely associated with working memory resources whereas cognitive states of engagement are tightly coupled with attention management (Berka et al., 2007; Marcotte, Meyer, Hendrix, & Johnson, 2013; Popovic et al., 2013; Stevens, 2012). Yet the vast majority of research associated with automation to date focuses on workload instead of engagement as a primary indicator of its influence (Guastello et al., 2013; Hutchins et al., 2013; Onnasch, Wickens, Li, & Manzey, 2013; Rusnock & Geiger, 2013).

Given the many factors that influence the degree of focus in attention (Wickens & McCarley, 2007; Wickens, McCarley, & Steelman-Allen, 2009), it would not be appropriate to equivocate engagement with awareness, so a substitution of all subjective instruments would be uncalled for. A consistent covariance in neurophysiological engagement, distraction, and subjective situational awareness, however, might allow for convincing conclusions to be made regarding the issue of whether cognitive resources liberated by automation have been reinvested back into the task at hand or allowed to wander to other extraneous activities.

Given the insufficient nature of workload as the sole indicator of cognitive activity, it is hoped, therefore, that this work will inspire consideration of neurophysiological measures such as distraction and engagement level as an augmentation of attentional factors to existing workload centric models of automation in future exploration of its influence on training.

Subjective Self-Reports Can Be Invasive. A second and perhaps equally surprising conclusion from the body of evidence discussed here is that the apparently common assumption that

subjective self-report measures are inherently non-invasive (Annett, 2002; de Guinea et al., 2012; Rubio et al., 2004) may not be valid when administered on a repeated-measures basis in automation research. The fact that the NASA-TLX proved to be insensitive to the invocation and removal of automation during training in two different experiments was a surprise as well. That observation, however, was substantially mitigated by the fact that neurophysiological metrics provided the same result – indicating that cognitive workload associated with working memory consumption actually remained constant throughout this particular training program – thereby restoring faith in the TLX's sensitivity to it.

The observation that the administration of the TLX in experiment two somehow prompted the manual control group to accumulate substantially more error than automation-assisted participants under identical conditions, however, points to an invasiveness factor that lies beyond instrument sensitivity. Because the large effect size and statistical significance of this surprising differential performance reversal occurred only in the with-TLX condition during experiment two, there is strong evidence for the presence of an invasive influence.

It should be noted, however, that nature of the TLX design might be considered to promote excess introspection and secondary task effects due to the number of times that its six scales have to be rated, and weighted, and compared against each other in order to arrive at calculated scores (Cao et al., 2009; Hart & Staveland, 1988). The multiple times that the TLX was repetitively administered during this investigation (eight times in experiment one and seven times in experiment two) also contribute to this explanation in a manner consistent with massed practice observations in discussed in context of resource allocation theory by Kanfer, Ackerman, Murtha, Dugdale, and Nelson (1994) where participants' cognitive resources never get a chance to be refreshed during training. It is doubtful, therefore, that a similar level of invasiveness would occur with relatively simple subjective self-report instruments involving a just a few modest questions, particularly if they are only administered once or twice during an investigation. In any case, ample motivation exists to further

explore this phenomenon in future work under the notion that excess introspection may be involved and/or that the TLX can act like a secondary task of extraneous nature and thereby drain resources otherwise devoted to learning (Horrey & Wickens, 2006; Mane & Wickens, 1986; Wickens, 2008a; Yeh & Wickens, 1988).

Implications

Mental Demand vs. Cognitive Investment. Taken together, the demonstrated sensitivity of the TLX to perceived difficulty in experiments one and two combined with disengagement data in experiment three suggest that while automation liberated cognitive resources in one manner or another during training block two, those resources were not reinvested back into the task by the automation-assisted group. If they had been, engagement levels would not have dropped in experiment three, nor would evidence of distraction have risen so substantially at the same time.

Given the reduction of complexity from three control inputs (throttle, roll & pitch axes) to only one (roll) for the automation-assisted group, there is little doubt that cognitive resources previously engaged in throttle and pitch control were liberated by automation. The question, however, is whether those liberated resources were reinvested back into the general mission of learning how to land the aircraft or reallocated instead to some other cognitive activity unrelated to that mission. The distraction components of the EEG data suggest the latter. Following this line of reasoning, the TLX instrument did what it was designed to do in the previous two experiments – measure perceived difficulty – which did not fluctuate between levels of automation despite the large differences indicated between task components.

The EEG data recorded here, by comparison, don't appear to measure mental demand per se, but rather the *cognitive investments* carried out to meet that demand. The EEG based workload metric implemented by the B-Alert system, after all, is derived from empirical assessments of working memory, whereas its engagement and distraction classifiers are more closely tied to perceptual processes associated with baseline vigilance tasks. It appears from the data presented here

that working memory investments remain generally constant throughout this particular variant of automation-assisted training, whereas the level of task engagement seems to fluctuate with level of automation. This might explain why the TLX failed to detect any fluctuations associated with level of automation changes despite demonstrating remarkable sensitivity to differences between task components themselves. The mental demand did not change, but investment strategy did.

What's perhaps most interesting about this influence is that disengagement and distraction persisted long after full manual control was returned, suggesting perhaps that participants no longer felt responsible for the error they accrued after automation was introduced. This is consistent with concerns for accountability for autonomous weapon use expressed by (Cummings, 2006; Grossman, 1996).

It may be that complacency and over-dependence dominate automation-induced performance deficits in operational settings where skill is already well established, while an apathetic tune out mechanism dominates that influence in novice operator training during which basic skill is initially acquired. As previously discussed, it is doubtful that novices acquired enough experience with the obviously imperfect automation used in this study to develop enough confidence to enable the over-reliance status implied by previous research (Dixon & Wickens, 2006; Madhavan & Wiegmann, 2007). Regardless of the implications of this for deadly force decisions, it's clear that just because a certain level of previously established demand is changed during training, a subsequent modification of cognitive resource investment will not necessarily be made to accompany that change.

Even though plots of perceived demand by participants in experiment one and two show a consistent and substantial fluctuation across training as indicated in Figure 10, performance and EEG data recorded when automation was invoked or removed provide clear evidence that investment of cognitive resources did not necessarily follow suit. It might be interesting to replicate these effects with alternative measures of attention such as eye tracking or pupilometry to determine whether the

distraction metric co-varies with other behavioral indicators. If so, the case for automation's influence on cognitive resource investment would be substantially strengthened.

Disengagement Guides Potential Mitigation. In consideration of the procedural differences between experiment one and two, it seems likely that the demand - investment disconnect noted above may have been mitigated in part by the presence of the pre-test. Consideration of pre-test influences on training manipulations have been expressed in previous literature (Gade & Chari, 2013; Goettl & Shute, 1996; Kromann et al., 2009; Randolph et al., 2013; Yildirim et al., 2013). Although participants were provided with instructional tutorials in both experiments one and two before hands on training began, the inevitable failure of novice participants to land the aircraft on their very first attempt during the pre-test may have somehow inspired the automation-assisted participants to reinvest liberated resources back into training during their second session, thereby improving performance and contributing to the washout of automation-induced deficits in experiment one. The concern for pre-testing bias in particular may serve as a sort of lemons to lemonade intervention with the potential to somehow inoculate automation-assisted trainees against disengagement and distraction influences noted in experiment three. Regardless of whether a pre-test results in replicating similar washout effects in future research, the neurophysiological measures used here as flags of disengagement effects may inspire the development of a variety of interventions to bring participants back *into* the loop during practice and thereby mitigate this kind of automation-induced training deficit.

Adaptive Automation Cueing: Reviews of adaptive automation historically discussed the use of workload and performance monitoring as a basis by which to allocate task functions between humans and machines or adjust levels of automation within overall system design (Inagaki, 2003; Kaber, 1996; Scerbo, 2007). Given that subjectively reported workload is difficult to monitor in real time without diverting substantial cognitive resources in an invasive manner, this approach is decidedly sub-optimal for dynamic high risk tasks such as flying an unmanned aircraft. Even if rapid

assessments of workload and performance are made via some sort of mission related secondary task performance, the adjustments to automation involved are bound to lag behind the need for intervention. It's not very useful for an autopilot to take over control of a high speed jet aircraft after the pilot has blacked out in a high speed dive and performance assessment methods only become aware of this fact *after* s/he has failed to pull back on the stick just before crashing. It's clear that this type of performance based approach to adaptive automation adjustment is destined to suffer from lag and latency effects that threaten its utility. As a result, neurophysiological measures such as the EEG disengagement metrics discussed above have the potential to present indications of impending performance deficits *before* they manifest in physical behavior with disastrous consequences.

Consider the data presented during experiment three discussed earlier. Participants showed no significant difference in performance between groups when automation was introduced in training block two. EEG data, however, revealed a reduced engagement profile (in terms of a reduced vigilance state and higher distraction probabilities), throughout this entire training block suggesting the presence of an emergent learning or attention deficit despite the absence of any such a pattern in performance data. This early detection factor from non-invasive neurophysiological monitoring enables the development of highly responsive interventions based on real time cueing sources such as the EEG measures used here. While this approach has already been demonstrated to enhance performance in uninhabited aerial vehicle operators (Wilson & Russell, 2007), it has the potential to benefit adaptive automation systems across the entire spectrum of human computer interaction in training – and potentially save the lives of blacked out air crews in operational environments as well.

Distraction & Engagement Enable Cognitive Load Distinctions. To the extent that neurophysiological measures collected by the B-Alert system are accurately assaying engagement, distraction and workload based on vigilance task performance and span task working memory profiles, the results from this study show promise for further isolated examination various load types proposed by Cognitive Load Theory. While proponents and supporters of Cognitive Load Theory

have recently welcomed the advent of neurophysiological monitoring within the behavioral sciences realm (Antonenko et al., 2010; Lee, 2013; Paas et al., 2003), the theory still lacks the wealth of crisp, empirically based distinctions between the various types of load necessary to enable the development of effective treatments and pedagogical manipulations used to pursue the increased effectiveness, efficiency, transfer, and durability goals required for training optimization (Bjork & Bjork, 2006; Healy et al., 2005c). Wickens and colleagues point out, for example, that despite the appreciable research progress made in cognitive resource characterization, the distinction between intrinsic and germane remains unclear (Wickens, Hutchins, Carolan, & Cumming, 2012a).

In the context of data recorded in this study, it may be that that the workload measures based on working memory like those provided by the B-Alert classifier present a reasonable approximation of intrinsic load associated with the need to maintain awareness of interactivity levels between multiple information or motor skill elements. Meanwhile the vigilance based cognitive state metrics might be considered to provide insight on how well intrinsic load levels monitored by the workload metric are balanced between extraneous (indicated by distraction) and germane (indicated by engagement) categories.

If we loosely apply those speculative associations to this particular study, it appears that intrinsic load did not fluctuate significantly throughout training whereas the distribution of germane vs. extraneous load changed substantially after the introduction of automation in training block two. Although participants were specifically advised to ignore altitude and airspeed feedback while the auto-pilot was partially engaged during that session, they were free to manipulate the throttle and stick in whatever manner they felt necessary to perform the turning task with minimum heading error. It may be that this degree of freedom in control surface manipulation served to counter act the goal of the auto-pilot manipulation in the first place – to reduce intrinsic load for novice operators.

Assuming that this explains why automation did not have a positive transfer value in accordance with Cognitive Load Theory, the distraction and engagement metrics might provide a

reasonable explanation for its negative influence as well. The EEG data in training block one show a relatively low probability of distraction and reasonably high engagement levels for all participants while performance error was steadily reduced - suggesting that an appreciable amount of learning occurred while germane and extraneous load was balanced in an appropriate high/low fashion. When automation was introduced, no appreciable fluctuation of intrinsic load was noticed, but perhaps the way it was previously distributed between germane and extraneous categories was perturbed in a negative fashion and learning suffered as a result. Regardless of whether these speculations hold up in future research, the crisp quantitative nature of neurophysiological measures used in this study are bound to be of value in future applications of Cognitive Load Theory within the unmanned systems realm and beyond.

Distraction & Engagement Provide Insight to Part-Task Training. While part-task training has been prominently featured in recent literature (Dattel, 2006; Gutzwiller et al., 2013a; Lim et al., 2009; McDermott, Carolan, & Wickens, 2012; Wickens et al., 2012b) following the nearly two decades since it was prominently examined in the context of aircraft carrier landings and complex computer simulations (Mane et al., 1983; Wightman & Lintern, 1985), the specific underlying sources of its mixed reputation for enhanced training transfer have yet to be identified. In consideration of that void, this is the first study we are aware of that involved all three kinds of PTT in a single experiment while monitoring training efficiency and workload with a combination of subjective and neurophysiological measures.

Since the segmentation and simplification forms of part task training were held constant between groups, it can be said that automation functioned as a "fractionalization agent" during all of the experiments discussed in this study. Commensurate with the replication objectives identified at the outset of this study, a backward form of segmentation was used in choosing which tasks were performed during hands on training sessions within the design of the Synthetic Training Environment. Simplification was used in a procedure roughly analogous to the "worked examples"

method from Cognitive Load Theory (Paas & Van Merriënboer, 1994; van Merriënboer et al., 2006) by essentially guiding the trainee into each scenario through shared control for the first ten seconds of each training trail. Fractionalization was implemented via the partial invocation of an auto-pilot during training block two which allowed half of the trainees to focus exclusively on the roll axis while others struggled with full manual control of pitch, roll, and throttle inputs.

In the same manner that the neurophysiological measures in this study may provide a fresh look at the concepts driving Cognitive Load Theory, we suspect that similar insight might be gained in regard to part-task training efficiency as well. We have seen that while automation proved to have no substantial influence on workload throughout this study, it had a negative and potentially disastrous influence on overall task engagement. We may be able to use this new knowledge in future examination of effects observed through part-task training enabled by or conducted with automated processing.

While recent reviews of the part-task training literature present a generally pessimistic view of its value for training transfer (Fontana et al., 2009; Gutzwiller et al., 2013a; Wickens et al., 2012b), it should not be forgotten that multiple cases exist where it has been shown to have a positive influence when used in backward chained segments (Goettl & Shute, 1996), and administered in an additive fractionalized fashion with tightly controlled variables in a complex computer simulation (Mane et al., 1983). It could be that the reason that these implementation of the part-task approach achieved positive training value because they were somehow able to inoculate trainees from becoming disengaged as hypothesized above in discussion of pre-testing effects. The bottom line here is that new measurement methods applied to previous research conducted without them present provides opportunities for greater understanding.

Limitations and Future Work

Low Statistical Power Invites Replication. As with many research projects conducted with limited resources, the experiments carried out in this study suffered from relatively small sample

sizes. Although effect sizes were reported for all experimental results with an eye toward statistical power potential, the conclusions here are summarily weakened. This situation invites replication with both subjective and neurophysiological measures within the confines of the Synthetic Training Environment or similar flight simulator. It is important to note, however, that accurate replications of these findings would require the use of a similar participant profile – undergraduate college students with no prior flight experience of any kind. That said, extension of this research into the skilled / expert realm would be of major value to the field, and should be conducted as soon as possible in order to establish a generalizable nature of the findings reported here.

Whole Task Control Condition in Part-Task Replication. In their seminal investigation of part-task training for aviators, (Wightman & Lintern, 1985) emphasize that any replication of their work ought to include a whole-task control condition for comparison of results. It is important to note, however, that while the current study makes many references to part-task effects in the context of automation design; it was never intended as an investigation of part-whole training per se. The apparatus made available for this study, the Synthetic Training Environment, happened to follow a part-task instructional paradigm that provided limited yet effective auto-pilot functionality that suited the goals of this investigation. Given the potential for additional insight to be gained from replication of the disengagement effects reported here and development of potential strategies to mitigate them (such as pre-testing), it would make sense to add a whole task control condition as well. It should be pointed out, however, that even more value might be gained by adding an additional test procedure (such as the cloud break task) to augment the part-task and whole-task procedures in order to provide a true test of transfer from both conditions to a novel task.

Measuring the Power of Automation. In discussion of experiment three, it is clear that automation-induced training deficits were replicated, albeit with different fluctuations in cognitive state than previously expected. What is perhaps most interesting is that such a large effect size for these deficits was produced after so little automation had been invoked. Given that the auto-pilot

manipulation was only applied to fifteen out of forty five total training trials, the maximum percentage that could be assigned to its influence is 33%. When one further considers that it was only applied to two of the three possible control modes during those trials (altitude & airspeed), the percentage of automation influence during this particular training program arguably drops to only 22% of all possible control opportunities. For such a small "amount" of automation to have such a large impact on learning (based on reported effect size) is worthy of special attention.

In their recent meta-analysis of automation's influence across the stages of cognitive processing previously discussed, Onnasch and colleagues make an attempt to quantify the "amount" of automation by essentially combining levels of automation observed across each of the four stages in an additive fashion (Onnasch et al., 2013). While this approach provides substantial insight in terms of a quantification model for automation influence, the manipulation used in this study shows this to be insufficient in that the number of opportunities for that influence to occur is not accounted for. Perhaps a measure which calculates the percentage of automation opportunities that were actually implemented might be of use as an additional factor in the model. Further dividing that implementation percentage by the effect size of the influence observed when automation was invoked might provide an even more useful indication of "automation power" in terms of the magnitude of return on an investment in automation expressed as the ratio of influence to opportunity.

Definition Upgrade. While the findings from this study support research literature that casts a generally pessimistic view toward the use of automation in training, it is important to consider this evidence in a more expansive context than what has been presented heretofore. Much of this pessimism stems from the limited and perhaps even myopic nature of the automation definition generally used throughout the human factors literature. Despite its oft-cited status and utility in laboratory experiments, the notion of automation as "execution by a machine (usually a computer) of a function that was previously carried out by a human" (Parasuraman & Riley, 1997) is woefully

inadequate in capturing its full potential to benefit humanity in both training and operational environments.

First and foremost, automation does not have to perform any task to be valuable. Consider the simulator used in these experiments – while it was designed with rudimentary autopilot functions, it didn't *have* to perform any task *for* the human to be of value. On the contrary, automation often presents humans with *new* tasks that they have never performed before (such as flying an aircraft for the first time without ever leaving the ground), or have performed inadequately (by crashing said aircraft on take-off or landing) – thereby dictating the need for initial training of novice operators or refresher training for those who have acquired an appreciable yet currently inadequate set of skills.

Secondly, this definition ignores one of the most important benefits of automation – the freedom to fail. There is enormous value in automation's presentation of new and horrifically challenging tasks to humans in the form of life threatening scenarios. Simulators used to be constructed for aircraft after they had originally been built in pursuit reduced training costs. They used to be evaluated on the basis of how "realistic" the system was, or how well skills acquired within them transferred to the "real world". Modern engineering and design efforts, however, often create fully functional simulators long before the experimental aircraft or vehicle get anywhere near the runway or test track. This is because they enable the development of new and potentially complex emergency procedures - without risking one's life or well-being in the process (Bell & Waag, 1998).

Thirdly, the unmanned systems and robotics field in particular typically endeavors to create machines that can perform *new* tasks that have never been performed by a human and *never will*. Consider the multitude of machines and automated devices that assemble microprocessors and other components of the advanced technologies that are becoming commonplace in today's society like smart phones and GPS watches. While some of these devices are still assembled by hand with skilled laborers peering through a large magnifying glass on their workbench, the vast majority of

mechanical assembly in the modern workplace requires one to build a machine to build another (product) machine (Rekiek, Dolgui, Delchambre, & Bratcu, 2002).

Independent of form factor and assembly line ergonomics, entire families of advanced robots are in development that unquestionably surpass human experiential precedent - not only in the traditional aspects of enhanced strength, precision and fatigue immunity, but also in the realm of polymorphic or "shape shifting" designs, hyper redundant limb structures, variable textured or adhesive skins, etc. (Laschi et al., 2012; Walker et al., 2005). Automation not only plays a role in the design of an artificial cephalopod limb or elephant trunk, they enable control of such enormously complicated mechanisms even though a human being has never coiled their arm around a tree trunk or controlled eight extendable tentacles in a collaborative fashion.

This issue represents the most profound and compelling return on any investment in human factors - the development of *new* capabilities and innovative coping strategies for *future* challenges that lay beyond the realm of previous human performance. It is with that optimistic perspective that we consider the results of this study under the powerful umbrella of their mitigation potential. It is hoped that these findings will enable instructors, program managers, and students alike to avoid the pitfalls associated with automation-induced training deficits inspired by hasty implementation amidst the crushing burden of increased demands dropped onto a floor of limited resources. It is only with further research into the relationship between human cognition and skill acquisition of this nature that maximum potential can be extracted from automation without incurring the veiled costs that can produce impoverished learning and training deficits like those observed here.

REFERENCES

Annett, J. (2002). Subjective rating scales: science or art? *Ergonomics, 45*(14), 966-987.

Antal, J. (2009). I Fight the Body Electric! *Military Technology, 33*(7), 22-30.

Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 1-14.

Ayres, P., & Paas, F. (2007). Can the cognitive load approach make instructional animations more effective? *Applied Cognitive Psychology, 21*(6), 811-820.

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science, 8*(4), 321-348. doi: 10.1080/14639220500535301

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775-779. doi: 10.1016/0005-1098(83)90046-8

Baldwin, C. L., & Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage, 59*(1), 48-56. doi: 10.1016/j.neuroimage.2011.07.047

Bell, H. H., & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology, 8*(3), 223-242.

Berka, C., Levendowski, D., Cvetinovic, M., Petrovic, M., Davis, G., Lumicao, M., . . . Olmstead, R. (2004). Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction, 17*(2), 151-170.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., . . . Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine, 78*(5), B231-B244.

Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K., . . . Stibler, K. (2005). *Evaluation of an EEG workload model in an Aegis simulation environment.* Paper presented at the Defense and Security.

Bjork, R. A., & Bjork, E. L. (2006). Optimizing treatment and instruction: Implications of a new theory of disuse. In L. G. O. Nilsson, Nobuo (Ed.), *Memory and society: Psychological perspectives.* (pp. 116-140). New York, NY US: Psychology Press.

Blitch, J. G. (2012). *Implications for automation assistance in unmanned aerial system operator training.* (Master of Science), Colorado State University, Fort Collins.

Blitch, J. G., Bauder, C. J., Gutzwiller, R. S., & Clegg, B. (2012). *Correlations of spatial orientation with simulation based robot operator training.* Paper presented at the 4th International Conference on Applied Human Factors and Ergonomics (AHFE), San Francisco CA.

Blitch, J. G., & Clegg, B. A. (2011). Automation influence on unmanned aerial system operator training. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55*(1), 143-146.

Bowers, C., Thornton, C., Braun, C., Morgan Jr, B. B., & Salas, E. (1998). Automation, task difficulty, and aircrew performance. *Military Psychology, 10*(4), 259-274.

Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: Software for assessing subjective mental workload. *Behavior Research Methods, 41*(1), 113-117. doi: 10.3758/brm.41.1.113

Clegg, B., & Heggestad, E. (2010). Automation and effective training: Colorado State University Technical Report for U.S. Army Research Office MURI Grant W911NF-05-1-0153.

Clegg, B., Heggestad, E., & Blalock, L. (2010). *The Influences of Automation and Trainee Aptitude on Training Effectiveness*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting.

Cowell, A., Hale, K., Berka, C., Fuchs, S., Baskin, A., Jones, D., . . . Fatch, R. (2007). Construction and validation of a neurophysio-technological framework for imagery analysis. In J. Jacko (Ed.), *Human-Computer Interaction. Interaction Platforms and Techniques* (Vol. 4551, pp. 1096-1105): Springer Berlin Heidelberg.

Cummings, M. (2004). *Automation bias in intelligent time critical decision support systems*.

Cummings, M. L. (2006). Automation and accountability in decision support system interface design.

Dattel, A. R. (2006). A comparison between a part-task cognitive training group and a part-task skill development group on aircraft performance. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 50*, 1986-1989.

de Guinea, A. O., Titah, R., Leger, P.-M., & Micheneau, T. (2012). *Neurophysiological Correlates of Information Systems Commonly Used Self-Reported Measures: A Multitrait Multimethod Study.* Paper presented at the System Science (HICSS), 2012 45th Hawaii International Conference on.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: a reliance-compliance model of automation dependence in high workload. *Human Factors, 48*(3), 474-486.

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: a workload analysis. *Human Factors, 47*(3), 479-487.

Dorneich, M. C., Whitlow, S. D., Mathan, S., Ververs, P. M., Erdogmus, D., Adami, A., . . . Lan, T. (2007). Supporting real-time cognitive state classification on a mobile individual. *Journal of Cognitive Engineering and Decision Making, 1*(3), 240-270.

Endsley, M., & Kaber, D. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics, 42*(3), 462-492.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32-64.

Fontana, F. E., Mazzardo, O., Furtado Jr, O., & Gallagher, J. D. (2009). Whole and part practice: a meta-analysis. *Perceptual and Motor Skills, 109*(2), 517-530.

Fournier, L., Wilson, G., & Swain, C. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training. *International Journal of Psychophysiology, 31*(2), 129-145.

Gade, S., & Chari, S. (2013). Case-based learning in endocrine physiology: an approach toward self-directed learning and the development of soft skills in medical students. *Advances in Physiology Education, 37*(4), 356-360.

Galán, F. C., & Beal, C. R. (2012). EEG estimates of engagement and cognitive workload predict math problem solving outcomes *User Modeling, Adaptation, and Personalization* (pp. 51-62): Springer.

Galster, S. M., Duley, J. A., Masalonis, A. J., & Parasuraman, R. (2001). Air traffic controller performance and workload under mature free flight: Conflict detection and resolution of aircraft self-separation. *The International Journal of Aviation Psychology, 11*(1), 71-93.

Goettl, B. P., & Shute, V. J. (1996). Analysis of part-task training using the backward-transfer technique. *Journal of Experimental Psychology: Applied, 2*(3), 227.

Gorman, J. C., Martin, M. J., Dunbar, T. A., Stevens, R. H., & Galloway, T. (2013). Analysis of semantic content and its relation to team neurophysiology during submarine crew training *Foundations of Augmented Cognition* (pp. 143-152): Springer.

Gramm, J., & Papp, S. (2009). *An insatiable demand:'manning'the US Air Force's unmanned aircraft systems with capable pilots.* (Masters Degree in Public Policy), Harvard Kennedy School of Government

Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., & John, M. S. (2008). *The red-line of workload: theory, research, and design.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Griffiths, P. G., & Gillespie, R. B. (2005). Sharing control between humans and automation using haptic interface: primary and secondary task performance benefits. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 47*(3), 574-590.

Grossman, D. (1996). On killing: The psychological cost of learning to kill in war and society. *No.: ISBN 0-316-33011-6*, 400.

Guastello, S. J., Shircel, A., Malon, M., & Timm, P. (2013). Individual differences in the experience of cognitive workload. *Theoretical Issues in Ergonomics Science*(ahead-of-print), 1-33.

Gutzwiller, R. S., Clegg, B. A., & Blitch, J. G. (2013a). Part-task training in the context of automation: current and future directions. *American Journal of Psychology, 126*(4), 417-432.

Gutzwiller, R. S., Clegg, B. A., Smith, C., Lewis, J. E., & Patterson, J. D. (2013b). *Predicted failure alerting in a supervisory control task does not always enhance performance.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Haider, H., & Frensch, P. A. (2002). Why aggregated learning follows the power law of practice when individual learning does not: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmieri (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(2), 392-406. doi: 10.1037/0278-7393.28.2.392

Hancock, P., & Meshkati, N. (1988). *Human mental workload*: Access Online via Elsevier.

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine, 69*(4), 360-367.

Hart, S. (2006). NASA-task load index (NASA-TLX); 20 years later. *Annual Meeting of the Human Factors and Ergonomics Society, 50*(9), 904-908.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload, 1*, 139–183.

Healy, A., Ericsson, K., & Bourne Jr, L. (1999). Optimizing the Long-Term Retention of Skills: Structural and Analytic Approaches to Skill Maintenance. Annual Report, 1991-1992 (ATTN:PERI-BR, Trans.) (Vol. 99-24): U.S. Army Research Institute.

Healy, A., Kole, J., Wohldmann, E., Buck-Gengler, C., Parker, J., & Bourne Jr, L. (2005a). *Optimizing the speed, durability, and transferability of training*.

Healy, A., Wohldmann, E., Parker, J., & Bourne, L. (2005b). Skill training, retention, and transfer: The effects of a concurrent secondary task. *Memory & Cognition, 33*(8), 1457.

Healy, A. F., Bourne Jr, L. E., Clegg, B., Fornberg, B., Gonzalez, C., Heggestad, E., . . . Buck-Gengler, C. J. (2010). Training knowledge and skills for the networked battlefield: DTIC Document.

Healy, A. F., Buck-Gengler, C. J., Barshi, I., Parker, J. T., Schneider, V. I., Raymond, W. D., . . . Bourne, L. E., Jr. (2002). Optimizing the durability and generalizability of knowledge and skills. In S. P. Shohov (Ed.), *Trends in cognitive psychology.* (pp. 123-192). Hauppauge, NY US: Nova Science Publishers.

Healy, A. F., Kole, J. A., Wohldmann, E. L., Buck-Gengler, C. J., Parker, J. T., & Bourne Jr, L. E. (2005c). Optimizing the speed, durability, and transferability of training (pp. 135): United States Army Research Institute for Behavioral and Social Sciences.

Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K., & Subbiah, I. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence, 30*(5), 425-447.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32*(2), 175-191.

Heggestad, E. D., Clegg, B. A., Goh, A., & Gutzwiller, R. S. (2012). How automation-based training aides and learner cognitive abilities impact training effectiveness *Principles of training: Theory and research*. New York, NY: Taylor & Francis.

Heuer, H., & Schmidtke, V. (1996). Secondary-task effects on sequence learning. *Psychological Research, 59*(2), 119-133.

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(1), 196.

Hutchins, S. D., Wickens, C. D., Carolan, T. F., & Cumming, J. M. (2013). The influence of cognitive load on transfer with error prevention training methods: a meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 55*(4), 854-874. doi: 10.1177/0018720812469985

Inagaki, T. (2003). Adaptive automation: sharing and trading of control *Handbook of cognitive task design* (pp. 147-169).

Jex, H. R. (1988). Measuring mental workload: Problems, progress, and promises. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (Vol. 52, pp. 5-39): Elsevier Science Publishers B.V.

Johnson, R. R., Popovic, D. P., Olmstead, R. E., Stikic, M., Levendowski, D. J., & Berka, C. (2011). Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological Psychology, 87*(2), 241-250.

Joseph, N. (2009). Metacognition needed: teaching middle and high school students to develop strategic learning skills. *Preventing School Failure, 54*(2), 99-103.

Kaber, D. B. (1996). *The effect of level of automation and adaptive automation on performance in dynamic control environments.*

Kaber, D. B., & Endsley, M. R. (1997). Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. *Process Safety Progress, 16*(3), 126-131.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science, 5*(2), 113-153.

Kanfer, R., Ackerman, P. L., Murtha, T. C., Dugdale, B., & Nelson, L. (1994). Goal setting, conditions of practice, and task performance: A resource allocation perspective. *Journal of Applied Psychology, 79*(6), 826.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology / Learning, Memory & Cognition, 31*(2), 187-194. doi: 10.1037/0278-7393.31.2.187

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a Mnemonic Debiasing Account of the Underconfidence-With-Practice Effect. *Journal of Experimental Psychology / Learning, Memory & Cognition, 32*(3), 595-608.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits - and costs - of dropping flashcards. *Memory, 16*(2), 125-136. doi: 10.1080/09658210701763899

Kornell, N., & Bjork, R. A. (2009). A Stability Bias in Human Memory: Overestimating Remembering and Underestimating Learning. *Journal of Experimental Psychology / General, 138*(4), 449-468. doi: 10.1037/a0017350

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition, 29*(5), 745-756.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education, 43*(1), 21-27. doi: 10.1111/j.1365-2923.2008.03245.x

Laschi, C., Cianchetti, M., Mazzolai, B., Margheri, L., Follador, M., & Dario, P. (2012). Soft robot arm inspired by the octopus. *Advanced Robotics, 26*(7), 709-727.

Lee, H. (2013). Measuring cognitive load with electroencephalography and self-report: focus on the effect of English-medium learning for Korean students. *Educational Psychology*(ahead-of-print), 1-11.

Lee, J. M., N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243-1270.

Lim, J., Reiser, R., & Olina, Z. (2009). The effects of part-task and whole-task instructional approaches on acquisition and transfer of a complex cognitive skill. *Educational Technology Research & Development, 57*(1), 61-77. doi: 10.1007/s11423-007-9085-y

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301. doi: 10.1080/14639220500337708

Mane, A., & Wickens, C. D. (1986). The effects of task difficulty and workload on training. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 30*(11), 1124-1127.

Mane, A. M., Coles, M. G., Wickens, C. D., & Donchin, E. (1983). The use of additive factors methodology in the analysis of a complex skill. *Proceedings of the 27th Annual Meeting of the Human Factors and Ergonomics Society*, 407-411.

Manzey, D., Bahner, J. E., & Hueper, A. D. (2006). *Misuse of automated aids in process control: Complacency, automation bias and possible training interventions.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Marcotte, T. D., Meyer, R. A., Hendrix, T., & Johnson, R. (2013). *The relationship between real-time EEG engagement, distraction and workload estimates and simulator-based driving performance*. Paper presented at the Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Bolton Landing, NY.

Marshall, S. P. (2002). *The index of cognitive activity: Measuring cognitive workload.* Paper presented at the Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on.

Martin, E., Lyon, D. R., & Schreiber, B. T. (1998). *Designing synthetic tasks for human factors research: An application to uninhabited air vehicles.* Paper presented at the Annual Meeting of the Human Factors Society.

McDermott, P. L., Carolan, T., & Wickens, C. D. (2012). *Part task training methods in simulated and realistic tasks.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control, 21*(4-5), 203.

Moroney, W. F., Biers, D. W., & Eggemeier, F. T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. *The International Journal of Aviation Psychology, 5*(1), 87-106.

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology, 8*(1), 47-63.

Newell, K. M. (1976). Knowledge of results and motor learning. *Exercise and Sports Science Review, 1*.

Newton, P., & Bristoll, H. (2010). Psychometric Success Spatial Ability Practice Test 1 (pp. 1-12).

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2013). Human performance consequences of stages and levels of automation: an integrated meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. doi: 10.1177/0018720813501549

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist, 38*(1), 63-71.

Paas, F., Tuovinen, J. E., van Merriënboer, J. J. G., & Darabi, A. A. (2005). A motivational perspective on the relation between mental effort and performance: optimizing learner involvement in instruction. *Educational Technology Research & Development, 53*(3), 25-34.

Paas, F., & Van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors, 52*(3), 381-410. doi: 10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2*(2), 140-160.

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

Parasuraman, R., & Wilson, G. F. (2008). Putting the brain to work: neuroergonomics past, present, and future. (Cover story). *Human Factors, 50*(3), 468-474.

Parasuraman, R. M., M; Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors, 38*(4).

Popovic, D., Stikic, M., Berka, C., Klyde, D., & Rosenthal, T. (2013). PHYSIOPRINT: a workload assessment tool based on physiological signals.

Randolph, J. J., Kangas, M., Ruokamo, H., & Hyvönen, P. (2013). Creative and playful learning on technology-enriched playgrounds: an international investigation. *Interactive Learning Environments*(ahead-of-print), 1-14.

Rekiek, B., Dolgui, A., Delchambre, A., & Bratcu, A. (2002). State of art of optimization methods for assembly line design. *Annual Reviews in Control, 26*(2), 163-174.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and Workload Profile methods. *Applied Psychology, 53*(1), 61-86.

Rusnock, C. F., & Geiger, C. D. (2013). *The impact of adaptive automation invoking thresholds on cognitive workload and situational awareness.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Russell, C. A., Wilson, G. F., Rizki, M. M., Webb, T. S., & Gustafson, S. C. (2005). *Comparing classifiers for real time estimation of cognitive workload.* Paper presented at the Proceedings of the 11th International Conference on Human-Computer Interaction.

Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A modern approach*. Englewood Cliffs NJ: Prentice-Hall.

Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics, 35*(11), 991-1009.

Scerbo, M. (2007). Adaptive automation. In R. R. Parasuraman, M. (Ed.), *Neuroergonomics: The brain at work* (pp. 239-252). 198 Madison Ave. New York NY 10016: Oxford University Press.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217.

Schmidt, R. A., & Wulf, G. (1997). Continuous concurrent feedback degrades skill learning: Implications for training and simulation. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*(4), 509-525.

Schneider, V., Healy, A., & Bourne Jr, L. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46*(2), 419-440.

Schneider, W. (1985). Training high-performance skills: Fallacies and guidelines. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 27*(3), 285-300.

Schooler, J., & Engstler-Schooler, T. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid* 1. *Cognitive Psychology, 22*(1), 36-71.

Sheridan, T. B., & Parasuraman, R. (2000). Human versus automation in responding to failures: an expected-value analysis. *Human Factors, 42*(3), 403-407. doi: 10.1518/001872000779698123

Shiffrin, R. M., & Schneider, W. (1984). Automatic and controlled processing revisited.

Singh, I., Molloy, R., Mouloua, M., Deaton, J., & Parasuraman, R. (1998). Cognitive ergonomics of cockpit automation. *Human cognition: A multidisciplinary perspective*, 242-253.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation- induced "complacency": development of the complacency-potential rating scale. *International Journal of Aviation Psychology, 3*(2), 111.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991-1006.

Smith, C., Fadden, S., & Boehm-Davis, D. (2005). *Use of a functional aviation display under varying workload conditions*.

Stevens, R., & Galloway, T. (2013). Towards the development of quantitative descriptions of the neurodynamic rhythms and organizations of teams. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*, 134-138.

Stevens, R., Galloway, T., & Berka, C. (2007). Integrating innovative neuro-educational technologies (I-Net) into K-12 science classrooms. *Foundations of Augmented Cognition*, 47-56.

Stevens, R., Galloway, T., Berka, C., & Sprang, M. (2009). *Neurophysiologic collaboration patterns during team problem solving.* Paper presented at the Human Factors and Ergonomics Society 53rd Annual Meeting, San Antonio TX.

Stevens, R. H. G., T.L.; Wang, P.; Berka, C. (2012). Cognitive neurophysiologic synchronies what can they contribute to the study of teamwork? *Human Factors: The Journal of the Human Factors and Ergonomics Society, 54*(4), 489-502.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review, 10*(3), 251-296.

Teo, G. W., Schmidt, T. N., Szalma, J. L., Hancock, G. M., & Hancock, P. A. (2013). *The effects of feedback in vigilance training on performance, workload, stress and coping.* Paper presented at the Human Factors and Ergonomics Society 57th Annual Meeting, San Diego CA.

USAF/AFRL. (2002). *Predator STE (Synthetic Training Environment) Installation Disk v 2.0*. U.S. Air Force Research Laboratory

van Merriënboer, J. J. G., Kester, L., & Paas, F. (2006). Teaching complex rather than simple tasks: balancing intrinsic and germane load to enhance transfer of learning. *Applied Cognitive Psychology, 20*(3), 343-352. doi: 10.1002/acp.1250

van Merriënboer, J. J. G. S., J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147-177.

Walker, I. D., Dawson, D. M., Flash, T., Grasso, F. W., Hanlon, R. T., Hochner, B., . . . Zhang, Q. M. (2005). *Continuum robot arms inspired by cephalopods.* Paper presented at the Defense and Security.

Wickens, C., & Colcombe, A. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors, 49*(5), 839-850. doi: Doi 10.1518/001872007x230217
Wickens, C., & McCarley, J. (2007). *Applied attention theory*: CRC.
Wickens, C., McCarley, J., & Steelman-Allen, K. (2009). NT-SEEV: A model of attention capture and noticing on the flight deck. *Human Factors and Ergonomics Society, 53*(12), 769-773.
Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177. doi: 10.1080/14639220210123806
Wickens, C. D. (2008a). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 449-455.
Wickens, C. D. (2008b). Multiple Resources and Mental Workload. *Human Factors, 50*(3), 449-455. doi: 10.1518/001872008x288394
Wickens, C. D., Dixon, S. R., & Ambinder, M. S. (2006). Workload and automation reliability in unmanned air vehicles. In N. J. P. Cooke, Heather L.; Pedersen, Harry K.; Connor, Olena (Ed.), *Human factors of remotely operated vehicles.* (pp. 209-222). Amsterdam Netherlands: Elsevier.
Wickens, C. D., Hutchins, S., Carolan, T., & Cumming, J. (2012a). Attention and cognitive resource Load in training strategies. In A. F. Healy & L. E. B. Jr. (Eds.), *Training cognition: optimizing efficiency, durability, and generalizability*. 711 Third Avenue, New York, NY 10017: Taylor and Francis, LLC.
Wickens, C. D., Hutchins, S., Carolan, T., & Cumming, J. (2012b). Effectiveness of part-task training and increasing-difficulty training strategies: a meta-analysis approach. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.
Wiener, E. L. (1988). Cockpit automation. *Human factors in aviation, (A 89-34431 14-54). San Diego, CA, Academic Press, Inc., 1988*, pp. 433-461.
Wightman, D. C., & Lintern, G. (1985). Part-task training for tracking and manual control. *Human Factors, 27*(3), 267-283.
Wilson, G., & Russell, C. (2003). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 45*(4), 635.
Wilson, G. F., Estepp, J., & Davis, I. (2009). *A comparison of performance and psychophysiological classification of complex task performance.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
Wilson, G. F., Russell, C., Monnin, J., Estepp, J., & Christensen, J. (2010). *How Does Day-to-Day Variability in Psychophysiological Data Affect Classifier Accuracy?* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
Wilson, G. F., & Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 49*(6), 1005.
Winston, P. H. (1984). *Artificial intelligence*. Reading, MA: Addison-Wesley.
Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic bulletin & review, 9*(2), 185-211.
Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 30*(1), 111-120.
Yildirim, N., Ayas, A., & Küçük, M. (2013). A comparison of effectiveness of analogy-based and laboratory-based instructions on students' achievement in chemical equilibrium. *Scholarly Journal of Education, 2*(6), 63-76.