UNDERSTANDING THE EFFECTS OF INTERRUPTIONS ON THE QUALITY OF TASK PERFORMANCE

by

David Michael Cades
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Psychology

Committee:

____ Director

____

____

____ Program Director

____ Dean, College of Humanities
and Social Sciences

Date: __________________________________ Spring Semester 2011
George Mason University
Fairfax, VA

Understanding the Effects of Interruptions on the Quality of Task Performance

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

David Michael Cades
Master of Arts
George Mason University, 2007

Bachelor of Science
Tufts University, 2003

Director: Deborah A. Boehm-Davis, University Professor and Chair
Department of Psychology

Spring Semester 2011
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to my amazing wife, Bess, my parents, Amy and Lonny, and my dogs past and present, Sarah and Bocephus.

Acknowledgments

First and foremost, I need to acknowledge Bess (and not only because she said I had to). Bess believed in me and supported me even when I didn't and most importantly tolerated me during some of the stressful periods periods and for that I am and will be eternally grateful. I love you Bess!

Working at home for the last few months of writing this dissertation was made warm and enjoyable because of my usually well-behaved companion, Bocephus. Thanks for being cute and cuddly Bo and even though you can't read this I'll be sure to give you an extra treat in appreciation.

Thank you to my parents and grandparents for believing in me and supporting me through my entire academic life from pre-school to graduate school and for helping me to keep a good balance of professional sports disappointments (Go Philadelphia!) and real life triumphs. I certainly could not have reached this point without your support and, only sometimes overbearing, love.

I also want to thank my Washington DC area family - Stevie and Ellen (the Godparents), Andy and Michelle (the cousins), and Ari, Noah, and Aviva (the Godchildren). You all treated Bess and I like your own children and siblings and it was a wonderful feeling to know that even away from my parents I was not without the love and support of my family. From receiving all of our wedding gifts to the more than occasional home cooked meal, I could not have asked for anything more! It meant so much to me to be able to watch the kids grow up from cute blobs to real people.

Next, I would like to thank those who helped me to grow as a person outside of school. Mainly in the areas of ice hockey and golf. These two activities, neither of which I had ever done before, have become and fun and important part of life helping me to maintain a balance in life that is as important to me as any professional milestone. Specifically, thanks to all my Bulldogs and other teammates and to those who tolerated my slightly too high handicap out on the course.

Lastly, I want to acknowledge all of the people who helped make my graduate school experience one of the greatest times of my entire life. First and foremost, no one could ask for a better advisor and mentor than Deborah Boehm-Davis. Debbie, thank you for accepting me into your family and pushing me when I needed it to achieve what you knew I could. I am constantly impressed with your ability to do so much at such a high level and always make time for your students, your family, and yourself. If I end up achieving even a fraction of what you have I will consider my career a success. I also want to thank Patrick McKnight for giving me a work home in the MRES lab for the last five years. No one got me to think and work outside of my comfort zone as much as you did, and as a result, I have been exposed to topics and approaches that I never would have considered; not to mention the fun we have had taking our work to Germany, Spain, Florida and even State College, PA. I want to thank my entire committee, Debbie, Chris Monk, and Kathy McKnight (I clearly chose the correct McKnight for this role) for making sure that I was

able to finish this dissertation at a high quality in a timely manner despite moving halfway across the country. It is not often that a student can have such a wonderful bond with all three members of his committee. Chris, the frequent gchat conversations to keep me on the ball; Kathy, the “working” dinners at Auld Shabeen to keep me fed and warm. This really has been an amazing experience for me that I sometimes wish did not have to end. I need to also thank Nicole Werner for keeping my work life organized and always being a good friend and colleague. Then there is David Kidd. If there was one person who on a day-to-day basis for the last five years has kept my nose to the grindstone it has been him. Writing a dissertation can be a lonely experience and David helped to ensure that mine was anything but. I consider it a privilege that he and I will be able to walk together at our graduation in May. Of course none of this work would have been possible without my sponsor from NASA, Kara Latorella, and the GSRP program, my data coders Holley and Maren, and my graders, David, Nicole, Erik Nelson, and Matt Kendra. Thanks to all of you for being such great friends and colleagues!

Table of Contents

List of Tables

List of Figures

Figure Page

Abstract

UNDERSTANDING THE EFFECTS OF INTERRUPTIONS ON THE QUALITY OF TASK PERFORMANCE

David Michael Cades, PhD

George Mason University, 2011

Dissertation Director: Deborah A. Boehm-Davis

The majority of previous research on interrupted task performance has primarily focused on the few actions surrounding the interruption (Trafton, Altmann, Brock, & Mintz, 2003; Ratwani, McCurry, & Trafton, 2008; Monk, Trafton, & Boehm-Davis, 2008) and most of these studies have been conducted in controlled laboratory environments (Eyrolle & Cellier, 2000; Oulasvirta & Saariluoma, 2004; Gillie & Broadbent, 1989). The purposes of the following studies are to expand the understanding of the disruptive effects of interruptions to include how interruptions impact the overall quality of the task being performed. Additionally, this research investigates how both quality and previously identified local effects generalize and can be understood in real-world environments, specifically the classroom and the flightdeck. Previous theoretical frameworks used to explain interrupted task performance in the lab (Altmann & Trafton, 2002, 2007; Oulasvirta & Saariluoma, 2004, 2006) were combined with theories of decision making (Brunswik, 1952, 1955; Adelman, Miller, Henderson, & Schoelles, 2003; Hogarth, 1987; Simon, 1955) to provide a more complete picture of how interruptions affect task quality in the real world. Two classroom studies and one flightdeck study were conducted to further our understanding of how interruptions actually affect performance in these environments at both a quality and local level. The

first classroom experiment showed that certain tasks can be performed with interruptions and distractions without showing a decrement in overall quality, although they might take longer. This study also showed that expertise was not necessary to avoid the disruptive effects of interruptions. The second classroom study showed that quality decrements due to interruptions are positively correlated with greater cognitive resource demand and that local disruptions have little influence on the overall quality of performance. The flightdeck experiment showed that sufficient expertise can help mitigate the disruptive effects of interruptions and that tools designed to facilitate resumption are generally liked by pilots, but that the information on them needs to be carefully designed. Taken as a whole, the results of these experiments suggest that theories of interrupted task performance need to be broadened and augmented with theories from other domains in order to provide a more complete view of how interruptions affect task performance in terms of both local metrics and more global measures such as quality.

Chapter 1: Executive Summary

This dissertation is comprised of an executive summary, two separate manuscripts, and a conclusions/lessons learned section. Chapter 2, titled “The Effects of Interruptions on the Quality of Task Performance”, describes two experiments exploring how interruptions and distractions affect the quality of task performance in a classroom setting. Chapter 3, titled “Examining Interruptions on the Flight Deck: can we Mitigate their Disruptive Effects?”, describes one experiment exploring the effects of interruptions on the flight deck, specifically focusing on the quality of decisions and possible solutions aimed at mitigating the disruptive effects of interruptions in this environment. This executive summary will provide an overview of the motivation (both theoretical and behavioral) for this work, a brief overview of the three experiments presented in the following two chapters, and some discussion of the findings. Chapter 4, the conclusions/lessons learned section will bring together the broader implications of all three experiments and discuss limitations and possible future directions of study in this area.

1.1 Motivation for the Current Studies

Cell phones, e-mail, instant messaging, knocks at your door — our lives are full of distractions and interruptions which disrupt our ability to complete tasks and get work done. Over the past 20 years or so, a significant body of research has emerged studying the effects of interruptions. The research has repeatedly shown that interruptions impair reaction time and accuracy following an interruption (Eyrolle & Cellier, 2000; Ziljstra, Roe, Leonora, & Krediet, 1999). Findings have been less clear, however, regarding how constant and repetitive interruptions and distractions affect the overall quality of the product produced or task being performed.

This question is garnering more attention in the popular media as our lives become inundated with sources of interruption and distraction. A recent article in Newsweek Magazine probes the possible disruptive effects of having a Blackberry in the Oval Office for the first time (Begley, 2009). Additionally, an article from the Washington Post examines the consequences that cell phone use (particularly text messaging among teens), might have on the productivity of today's youth (St. George, 2009). Although it is extremely unlikely that the President's Blackberry will lead to a national crisis, it is not unlikely that the mere presence of interruptions and distractions may be affecting the quality of the work we do. Work on interruptions has mainly focused on the micro level time and accuracy costs on the few steps directly following the interruption. It is not clear whether interruptions have a broader, or perhaps cumulative, effect on the overall quality of the task. Some might even argue that although certain distractions (e.g., listening to music or watching television) may slow progress on a task, distractions do not affect the quality of their work. Further, distractions could be argued to be beneficial by increasing arousal and enjoyment during monotonous or unenjoyable tasks (Speier, Vessey, & Valacich, 2003). This position is only partially supported by the literature. Two studies found that listening to music and music videos did not affect performance on homework tasks (Pool, Koolstra, & Voort, 2003), but having a spoken-word program on in the background led to slower and poorer performance (Pool, Voort, Beentjes, & Koolstra, 2000). The questions I will address here are how the effect of interruptions on the quality of our work can be understood and addressed from both a theoretical and applied perspective and how the type of task or environment affects the quality of performance.

1.1.1 Theories of Interrupted Task Performance

In general, theories used to explain interrupted task performance focus on effects at the micro action level and have not been, or cannot be, brought to bear on questions of overall quality and qualitative aspects of performance. Specifically, two memory-based theories — Memory for Goals (Altmann & Trafton, 2002, 2007) and Long-term Working Memory

(Oulasvirta & Saariluoma, 2004, 2006) — have been used successfully to explain some of the action-level disruptive effects of interruptions.

Memory for Goals

The Memory for Goals theory (Altmann & Trafton, 2002, 2007) posits that when performing a task, that task's goal drives behavior. Thus, the goal with the strongest activation level in memory will guide behavior. When an interruption occurs, the current task goal must be suspended and the interruption goal must be instantiated. While working on the interruption, the memorial representation, or activation level, of the main task decays, while that for the interruption grows. The longer the interruption, the greater the activation level of the primary task goal will have decayed and the longer it will take to retrieve this goal and resume working on it. Following the interruption, the primary task needs to be resumed and the goal retrieved from memory. The time it takes to retrieve this goal, and the probability that the correct goal is retrieved, is based on the current activation of the to-be-retrieved task or goal. Goals with lower activation levels will lead to a higher probability of the incorrect goal being retrieved upon resumption.

Altmann and Trafton (2002, 2007) identify two constraints that influence the activation level of the primary goal apart from the temporal-based activation decay function. The priming constraint suggests that cues, either environmental or mental, can help to boost activation for a suspended goal. The strengthening constraint suggests that activation levels of goals in memory may be boosted through rehearsal. The more frequently and recently a goal has been rehearsed, the higher its activation level will be.

Previous studies examining Memory for Goals (Altmann & Trafton, 2004; Cades, Trafton, Boehm-Davis, & Monk, 2007; Hodgetts & Jones, 2006a, 2006b; Monk, Boehm-Davis, & Trafton, 2004; Ratwani & Trafton, 2008; Trafton et al., 2003) have generally only focused on the few steps preceding and following the interruption to quantify the disruptiveness of various characteristics of interruptions. They do not provide a metric of overall performance or quality or make predictions about how or whether retrieval time has any influence on quality.

Given that interruptions impair both the accuracy and speed of goal retrieval, does this also suggest that the overall quality of the task will suffer? On one hand, Memory for Goals could predict that although a task performed with interruptions may take longer and have more errors on specific resumption steps than one performed without interruptions, the overall quality of that task will not suffer, especially if the resumption errors are corrected before the task is complete. Alternatively, Memory for Goals also allows that with enough interruptions and distractions over time, retrieval of the proper goal could become difficult and sufficiently error-prone to lead to overall quality decrements. For example, think of what happens with successive tellings of the same story as in the game whisper down the lane. Each time the story is passed on, the person passing it must retrieve the story from memory and retell it. After just a few retrievals, the story often bears little resemblance to the original. Interruptions could have the same type of effect on the overall quality of task performance. If each time the task is resumed, the actual goal that is retrieved from memory is in some way degraded or different than the original goal that was suspended, the quality of the end product certainly might suffer.

Long Term Working Memory

Long Term Working Memory theory (Ericsson & Kintsch, 1995) suggests that experts are able to encode information into a protected long term memory store and create specific retrieval structures for that information allowing faster and more accurate information retrieval compared to novices. Oulasvirta and Saariluoma (2004, 2006) applied this theory to the study of interruptions, conducting a series of experiments examining the influence of interruptions on performance in listening and reading comprehension tasks. These tasks were chosen because, according to Long Term Working Memory theory (Ericsson & Kintsch, 1995), people have sufficient expertise to use the protected memory stores. According to this theory, if people are using the protected retrieval structures associated with Long Term Working Memory, interruptions should not disrupt performance. If people are able to encode information about the primary tasks into Long Term Working Memory then later retrieval of that information following an interruption, should be fast and accurate. According to Long Term Working Memory, interruptions would not disrupt experts performing tasks within their expertise.

Although the work on interruptions and Long Term Working Memory does not specifically address overall quality, it would follow that if people have sufficient expertise to encode details of the primary task into Long Term Working Memory, the overall quality of that task performed with interruptions should not suffer. If people can quickly and accurately recall exactly what they were working on prior to an interruption, as Long Term Working Memory suggests, then there is no reason to believe that the quality of their work would decrease as a result of interruptions.

1.1.2 Other Theoretical Frameworks Related to Quality of Performance

Decision making and the cognitive mechanisms behind decision making might provide useful theoretical insights into how interruptions affect the quality of task performance. Where many of the theories of interrupted task performance focus only on the specific actions around the interruptions event, decision-making theories lend themselves to a more subjective interpretation of quality, which is also the way most of us perceive quality (grades in a class, performance reviews at work, etc.). Specifically, Brunswikian Correspondence Constancy (Brunswik, 1952, 1955), Cognitive Acceleration (Adelman et al., 2003), and Noncompensatory Strategies or Satisficing (Hogarth, 1987; Simon, 1955) theories may prove useful in helping to understand how interruptions may affect the quality of task performance.

Correspondence Constancy

Brunswik (1952, 1955) used the term Correspondence Constancy to refer to people's ability to maintain a constant level of performance across varying difficulties and levels of task performance. He found that people try to stabilize their performance of a task even as the environment and context of performing that task changes. Interruptions not only disrupt tasks, they also increase cognitive demands in the task environment. Correspondence Constancy would suggest that in the face of interruptions, people would try and maintain their level of performance until the processing or task demands from the interruptions became too much for them to handle. Anything past that point would lead to decreased task quality.

Cognitive Acceleration

Adelman et al. (2003) identify Cognitive Acceleration as one mechanism by which people try to maintain a constant level of performance in the face of increasing task demands. They specify two methods by which Cognitive Acceleration may occur: (1) increased processing per unit time and (2) filtered processing. In dealing with interruptions (or increased task demands), people may be able to increase their cognitive processing speed, (i.e., take up the cognitive slack in the system) to handle interruptions, up to a point, without sacrificing the quality of their performance (Edland & Svenson, 1993; Payne, Bettman, & Johnson, 1993). Performance would suffer only when interruptions caused demand increases beyond the ability of the performer to speed his or her processing. Filtered processing (Maule, Hockey, & Bdzola, 2000), on the other hand, would suggest that people may be able to maintain performance under the increased demands associated with interruptions by shedding non-critical judgments. However, as with Correspondance Constancy, if the demands become too great, and shedding non-critical judgments is insufficient to adapt to the increased demands, then the quality of performance may be adversely affected.

Noncompensatory Strategies - Satisficing

Satisficing occurs when a decision is made after a minimum acceptable criterion is met, regardless of whether better solutions exist (Hogarth, 1987; Simon, 1955). Jedetski, Adelman and Yeo (2002) found that when consumers were presented with an overwhelming number of options, they engaged in satisficing and did not evaluate all of the different options. Although performing a task with interruptions may not lead to a situation where a person has too many options, it may lead to situations that are overwhelming due to heightened demand. If they engage in satisficing to cope with increased demands, then the quality of their performance may suffer.

All three decision-making theories point to the same conclusion in how interruptions may affect the quality of task performance. If the presence of interruptions and distractions in addition to the primary task causes the overall demands to become too high for a person to handle, then the quality of task performance will decrease. It is not clear where this red line of performance is, and in fact it may be different for each task and for each individual, but these theories clearly suggest that there is a point at which interruptions and distractions will decrease the quality of task performance.

1.2 Investigating the Quality of Interrupted Task Performance

The experiments to follow investigate how interruptions affect the quality of task performance. It could be, as suggested by the memory-based theoretical frameworks presented above (Altmann & Trafton, 2002, 2007; Ericsson & Kintsch, 1995), that the locus of disruption associated with interruptions is confined to the first few actions upon resumption of the primary task and that the overall quality of task performance will not be impaired by interruptions. It could also be that the resumption errors predicted by the memory-based accounts have a cumulative effect and lead to a degradation in memorial representations that result in decreased quality of task performance.

Alternatively, if the presence of interruptions leads to overall increased task demands and participants are not able to adapt or do not have sufficient resources to handle the increased demands, decision-making theories would suggest (Brunswik, 1952, 1955; Edland & Svenson, 1993; Hogarth, 1987; Maule et al., 2000; Payne et al., 1993; Simon, 1955) that interruptions may lead to decreased quality of task performance. In this case, it will be important to supplement the memory-based theories of interrupted task performance with decision-making theories that reflect how interruptions affect decision making. This combination of memory-based and decision-making theories will allow for more complete explanations and more accurate predictions of interrupted task performance.

1.3 Overview of Experiments

The primary goal of the following three experiments is to explore how interruptions affect the overall quality of task performance in naturalistic environments. This work represents two new directions in the interruptions literature. Although it is widely accepted that interruptions are generally disruptive, the disruptions measured usually center around the few actions just prior to and just after the interruption (Altmann & Trafton, 2002; Trafton et al., 2003). These disruptions are usually quantified in terms of reduced accuracy (Eyrolle & Cellier, 2000; Gillie & Broadbent, 1989), task completion time increases (Ziljstra et al., 1999; Gillie & Broadbent, 1989) and increased errors and longer latencies when resuming the primary task (Altmann & Trafton, 2002; Hodgetts & Jones, 2006b). In addition to these more local measures, the following experiments will also look at how interruptions affect the entirety of the task by measuring the quality of performance.

The second new direction of this work involves examining interruptions in the real world. The majority of prior work looking at interrupted task performance has taken place in controlled laboratory settings (Monk, 2004; Altmann & Trafton, 2002; Trafton et al., 2003; Hodgetts & Jones, 2006b; Gillie & Broadbent, 1989). Comparatively less work has explored whether or not the effects observed in lab settings generalize to the real world. By carrying out studies both in classrooms (Chapter 2 - 2 studies) and on the flightdeck (Chapter 3 - 1 study) we seek to expand our understanding of how interruptions actually affect the way people work in real situations.

1.3.1 Classroom Experiments

Two exploratory experiments were carried out in a classroom environment to investigate how interruptions affect the quality of task performance. Both experiments required participants to complete two sessions: an interrupted session in which they completed the task while dealing with interruptions and distractions and a not-interrupted session in which they only worked on the primary task without any distractions or interruptions. In sessions where participants were not interrupted, they were required to work alone, in silence, and without using email, Internet or instant messaging. During interrupted sessions, participants were encouraged to talk to their classmates and instructor and use the Internet, email, instant messaging and any other device or program they wished. Additionally, specific external interruptions, such as people coming in to make announcements or having loud conversations just outside the classroom, were also used. All tasks completed in these experiments were graded by four independent graders. These grades served as the measure for quality in these experiments. The only difference between the two experiments was in the primary task used. In the first classroom experiment, the task was a graduate-level statistical problem set that included calculation and interpretation questions and an open-ended experimental design question. The second classroom experiment used a manuscript-editing task.

Classroom Experiment 1

The two different types of questions used in the first experiment were designed to address different theoretical explanations for potential quality decrements. Because none of the participants were experts in statistics, Long Term Working Memory (Ericsson & Kintsch, 1995) would predict that interruptions should cause quality decrements across both types of questions. Memory for Goals (Altmann & Trafton, 2002, 2007) does not make specific predictions about quality. This framework would, however, predict that interruptions should lead to resumption errors and, if the resumption errors go unnoticed or uncorrected, then quality would suffer. Memory for Goals does not inform whether resumptions errors will be corrected, but it makes sense that the more resumption errors that are made, the more likely some will go uncorrected. Decision-making theories (Brunswik, 1952, 1955; Edland & Svenson, 1993; Hogarth, 1987; Maule et al., 2000; Payne et al., 1993; Simon, 1955) would predict that the experimental design questions would show greater quality decrements due to interruptions than the calculation questions because the experimental design questions required more cognitive resources to complete.

The results of this first experiment showed that there were no quality differences on any of the questions between the interruption and no interruption conditions. While these results do not support a Long Term Working Memory explanation, they do not tease apart Memory for Goals or decision-making explanations.

Classroom Experiment 2

In order to differentiate these two possible explanations, we chose a more difficult and demanding task - document editing. The demands of each task type were estimated using a simplified version of a Natural Language Goals, Operators, Methods, and Selection Rules (NGOMSL) approach (Kieras & Polson, 1985). Unlike a traditional full NGOMSL analysis, which includes time for basic perception and movement, the analysis performed here identified only on the cognitive, or mental operation, aspects of the tasks performed. The analyses for the different task types are presented below.

Experiment 1 Tasks

- Calculation and interpretation tasks (3 Mental Operations)
    - Decide which analysis answers question
    - Recall how to perform chosen analysis
    - Decide how to interpret results
- Experimental design questions (at least 4 Mental Operations)
    - Recall results of current data
    - Decide on important next steps
    - Decide on writing approach
    - Decide on whether to review and/or edit writing

        - If task ends here then no more Mental Operations will be required
        - If reviewing or editing occurs then some more Mental Operations will be required

Experiment 2 Tasks

- Editing for local changes - spelling and grammar (2 Mental Operations)
    - Decide if spelling or grammar is correct
    - If wrong, recall and implement correction of spelling or grammar
- Editing for global changes - flow, content, style etc. (at least 6 Mental Operations)
    - Recall content of entire paper (not word for word, but enough to get the gist)
    - Decide where changes are needed
    - Recall and retain section that needs to be changed
    - Think about possible changes, while retaining context
    - Decide and implement change
    - Decide if change works
        - If task ends here then no more Mental Operations will be required
        - If more reviewing or editing occurs then some more Mental Operations will be required

The NGOMSL style analyses show that, in its simplest form, global editing requires greater cognitive resources than any of the other types of tasks that students had to perform in the two experiments. This is true even though the NGOMSL analyses do not take into account how much information must be stored and/or recalled in each task. NGOMSL treats all mental operations equally and therefore assumes that recalling the results of the current data in Experiment 1 requires the same amount of cognitive resources as recalling the entire paper to edit in Experiment 2. Assuming that it does require more cognitive resources to store and/or recall greater amounts of information, then another key distinction that sets the global editing apart from the experimental design questions is the need to retain much greater amounts of information while making decisions on how to proceed. For the experimental design questions, the only prior information that it is necessary to retain in memory are the results of the previous experiment. For global editing, on the other hand, it is imperative to retain knowledge about both the section being edited along with how it fits into the flow of the entire document. Combining this additional level of information retention with the larger number of Mental Operations required for global editing suggests that it uses the greatest amount of cognitive resources, and according to the decision-making theories, should be the most susceptible to the negative effects of interruptions.

According to Memory for Goals, as long as the primary tasks are visually similar (i.e., same program or screen layout) and the interruptions are the same, there should be no difference in the disruption caused by interruptions. If one of the tasks had large highlights on the last action prior to an interruption and the other task did not, Memory for Goals would predict that the large highlighted area would act as a resumption cue and facilitate faster and more accurate resumption (Altmann & Trafton, 2002). Because the tasks for both experiments used similar word processing documents, neither provided any better resumption cues than the other. A decision-making approach would predict greater quality decrements with interruptions in the document editing task because it required more cognitive resources. Consistent with the decision-making explanation, the second experiment showed that when trying to edit a document, the presence of interruptions and distractions did reduce the quality of task performance.

These two experiment together lend support for a decision-making explanation of how interruptions affect the quality of task performance. That is, when the task and interruptions require more cognitive resources, potentially more than a person has available, quality will suffer. Further, these results suggest that theories of interrupted task performance must be broadened to include more complete task metrics such as quality and that the disruptive effects of interruptions are not only contained in the first few actions prior to or just after an interruption. Lastly, these experiments suggest that there are some tasks that may not suffer in quality as a result of being performed with interruptions and distractions.

1.3.2 Flightdeck Experiment

In some settings, increased time to complete a task is not a problem as long as quality is maintained. In fact, in a classroom or office setting, even some loss in quality may be acceptable. However, in a safety-critical environment such as the flightdeck, even the smallest time increase or dip in performance is unacceptable as it may lead to catastrophic consequences. The presence and disruptiveness of interruptions on the flightdeck in terms of failures to return to a task and errors in resumption have been well documented (Dismukes, 2006; Latorella, 1998; LeGoullon, 2006). However, less focus has been given to how interruptions and distractions affect the quality of task performance and what methods can be used to help mitigate the disruptive effects of interruptions and help pilots resume.

Few tasks performed on the flightdeck lend themselves well to being measured in terms of quality. We chose one of them, a destination-choosing task. The scenario required pilots to deviate from their primary destination at the last second; they then had to choose a new destination based on a number of different attributes such as fuel, distance, airline operations, runway length, and weather. The potential alternate destination choices available were evaluated by a subject matter expert prior to the experiment and two of the destinations were chosen as optimal. This allowed us to measure the quality of performance based on whether pilots chose an optimal or non-optimal destination.

Forty pilots from a regional airline were randomly assigned to one of four conditions. In the uninterrupted condition, participants proceeded through the rerouting and destination choosing and the scenario ended when they chose a destination. The other three conditions had participants dealing with an interruption which required them to suspend the destination-choosing task and attend to a malfunctioning aircraft system. In all cases, the malfunctioning system was fixed and participants then resumed choosing a destination.

The three interruption conditions differed in what was presented to participants upon resumption. Participants either resumed to: (1) the same screen they were looking at prior to the interruption; (2) a simple resumption aid screen which reminded them to continue choosing a new destination; or (3) a detailed resumption aid screen which provided them with all of the information necessary for them to decide whether or not to land at each of the possible destinations.

The goal of the two resumption aids was to take advantage of the cueing constraint from the Memory for Goals framework (Altmann & Trafton, 2002, 2007), which suggests that providing reminders or pointers to an interrupted task helps to facilitate resumption of that task. The second resumption aid was designed to provide even more specific cueing and help lower pilot workload by providing them with information they would otherwise have to look up.

Interestingly, the only condition with a significant number of non-optimal decisions was the interruption with detailed resumption aid condition. This was most likely due to the fact that this aid contained non-commercial airports in the area that pilots would normally not consider as options unless they were in an emergency situation. Although these non-commercial choices did allow for safe landings, they did not have airline operations and would therefore cause significant problems in dealing with aircraft and passenger maintenance. All ten participants in the detailed resumption aid condition mentioned that they liked having this information on one screen and also said that they might have preferred a more limited field of choices (e.g., do not provide information on non-commercial airports unless emergency is declared). No workload differences were observed among the four conditions. The pilots were certainly experts at the tasks they were asked to complete and it could have been their training and expertise that helped to prevent any disruptive effects associated with interruptions. Although the quantitative data from this experiment may suggest that this particular type of interruption may not be problematic for pilots, the positive feedback on the resumption aids points to possible methods of improving the overall experience of pilots and help them deal with interruptions.

Chapter 2: Interruptions Effects on the Quality of Task Performance in a Classroom Setting

2.1 Introduction

Imagine a manager arriving to his office in the morning. The first item on his daily agenda is writing a letter to the company about a new policy change. The manager opens up a word processing program and begins to compose the letter. Shortly after beginning to type, there is a knock at his door. It's an employee who needs to have a brief meeting, so the manager stops typing the letter and speaks with the employee. After the conversation, the manager returns to the letter only to be interrupted again by a phone call. The person on the phone asks the manager to check something on the Internet. While doing this, the manager receives an instant message followed by an email and another employee comes by for a meeting. Fast forward to the end of this typical day. The manager is shutting down his computer and there is the letter he had started that morning - unfinished, forgotten, and now late. Does this scenario sound familiar? More and more, our lives are becoming inundated with multiple tasks and sources of information that are constantly vying for our attention, and all-too often, negatively affecting our work.

A large body of research exists examining the effects of interruptions and distraction (Eyrolle & Cellier, 2000; Gillie & Broadbent, 1989; Ziljstra et al., 1999), the mechanisms by which tasks are interrupted and resumed (Altmann & Trafton, 2002, 2007; Oulasvirta & Saariluoma, 2004, 2006) and methods by which the disruptive effects of interruptions may be mitigated (McFarlane, 2002; McFarlane & Latorella, 2002; Trafton, Altmann, & Brock, 2005; Cades, Boehm-Davis, & Smith, 2010). However, to date, most of this work has focused on the few actions just prior to and just after the interruption, quantifying the negative effects in terms of reduced task accuracy (Eyrolle & Cellier, 2000; Gillie &

Broadbent, 1989), task completion time increases (Ziljstra et al., 1999; Gillie & Broadbent, 1989), and increased errors on and longer latencies to resume the interrupted task (Altmann & Trafton, 2002; Hodgetts & Jones, 2006a, 2006b). Additionally, the bulk of this work has taken place in highly controlled laboratory settings (Monk et al., 2004; Altmann & Trafton, 2002; Trafton et al., 2003; Hodgetts & Jones, 2006a; Gillie & Broadbent, 1989) with less work being done in naturalistic settings (Czerwinski, Horvitz, & Wilhite, 2004; Cades, Werner, Boehm-Davis, & Arshad, 2010; Cades, Boehm-Davis, & Trafton, 2007).

The goal of the current work is to expand our understanding of the effects of interruptions and focus on the overall quality of task performance in an environment often inundated with interruptions and distractions — the classroom. It is one thing to know that interruptions cause resumption errors and longer task latencies, but how do these effects impact the students' tasks as a whole? Are they able to perform as well while dealing with interruptions and distractions as they would in a more sterile environment?

This question is garnering more attention in the popular media as our lives become inundated with sources of interruption and distraction. An article from the Washington Post examines the consequences that cell phone use (particularly text messaging among teens), might have on the productivity of today's youth (St. George, 2009). A recent article in Newsweek Magazine probes the possible disruptive effects of having a Blackberry in the Oval Office for the first time (Begley, 2009). Although it is extremely unlikely that the President's Blackberry will lead to a national crisis, it is not unlikely that the mere presence of interruptions and distractions may be affecting the quality of the work he does.

It is not clear whether interruptions have a broader, or perhaps cumulative, effect on the overall quality of the task. Some youth may argue that although certain distractions (e.g., listening to music or watching television) may slow progress on a task, distractions do not affect the quality of the work. Further, distractions could be argued to be beneficial by increasing arousal and enjoyment during monotonous or unenjoyable tasks (Speier et al., 2003). This position is only partially supported by the literature. Two studies found that listening to music and music videos did not affect performance on homework tasks (Pool et al., 2003), but having a spoken-word program on in the background led to slower and poorer performance (Pool et al., 2000). This work will examine the effects of interruptions on the quality of work from both a theoretical and applied perspective in a naturalistic classroom environment. This will allow us to see how and whether theories describing possible mechanisms of interrupted task performance, previously tested primarily in laboratory environments, make predictions for more realistic tasks and environments.

2.1.1 Theories of Interrupted Task Performance

In general, theories used to explain interrupted task performance focus on effects at the micro action level and have not been, or cannot be, brought to bear on questions of overall quality and qualitative aspects of performance. Specifically, two memory based theories — Memory for Goals (Altmann & Trafton, 2002, 2007) and Long-term Working Memory (Oulasvirta & Saariluoma, 2004, 2006) — have been used successfully to explain some of the action-level disruptive effects of interruptions.

Memory for Goals

The Memory for Goals theory (Altmann & Trafton, 2002, 2007) posits that when performing a task, that task's goal drives behavior. Thus, the goal with the strongest activation level in memory will guide behavior. When an interruption occurs, the current task goal must be suspended and the interruption goal must be instantiated. While working on the interruption, the memorial representation, or activation level, of the main task decays, while that for the interruption grows. The longer the interruption, the greater the activation level of the primary task goal will have decayed and the longer it will take to retrieve this goal and resume working on it. Following the interruption, the primary task needs to be resumed and the goal retrieved from memory. The time it takes to retrieve this goal, and the probability that the correct goal is retrieved, is based on the current activation of the to-be-retrieved task or goal. Goals with lower activation levels will lead to a higher probability of the incorrect goal being retrieved upon resumption.

Altmann and Trafton (2002, 2007) identify two constraints that influence the activation level of the primary goal apart from the temporal-based activation decay function. The priming constraint suggests that cues, either environmental or mental, can help to boost activation for a suspended goal. The strengthening constraint suggests that activation levels of goals in memory may be boosted through rehearsal. The more frequently and recently a goal has been rehearsed, the higher its activation level will be.

Previous studies examining Memory for Goals (Altmann & Trafton, 2004; Cades, Trafton, et al., 2007; Hodgetts & Jones, 2006a, 2006b; Monk et al., 2004; Ratwani & Trafton, 2008; Trafton et al., 2003) have generally only focused on the few steps preceding and following the interruption to quantify the disruptiveness of various characteristics of interruptions. They do not provide a metric of overall performance or quality or make predictions about how or whether retrieval time has any influence on quality.

Predictions from Memory for Goals

Given that interruptions impair both the accuracy and speed of goal retrieval, does this also suggest that the overall quality of the task will suffer? On one hand, Memory for Goals could predict that although a task performed with interruptions may take longer and have more errors on specific resumption steps than one performed without interruptions, the overall quality of that task will not suffer, especially if the resumption errors are corrected before the task is complete. Alternatively, Memory for Goals also allows that with enough interruptions and distractions over time, retrieval of the proper goal could become difficult and sufficiently error-prone to lead to overall quality decrements. For example, think of what happens with successive tellings of the same story as in the game whisper down the lane. Each time the story is passed on, the person passing it must retrieve the story from memory and retell it. After just a few retrievals, the story often bears little resemblance to the original. Interruptions could have the same type of effect on the overall quality of task performance. If each time the task is resumed, the actual goal that is retrieved from memory is in some way degraded or different than the original goal that was suspended, the quality of the end product certainly might suffer.

Unfortunately, Memory for Goals (Altmann & Trafton, 2002, 2007) does not make predictions about whether a resumption error will be corrected. For example, according to the theory, if a person was interrupted while adding 4 and 3, that person has a better chance of responding incorrectly upon resumption than if he or she was not interrupted. However, there is nothing in the Memory for Goals framework that predicts whether or not that person would later correct the error. If the errors are made, but then corrected, the overall quality of the finished task should not be impaired, even in the presence of interruptions.

Long Term Working Memory

Long Term Working Memory theory (Ericsson & Kintsch, 1995) suggests that experts are able to encode information into a protected long term memory store and create specific retrieval structures for that information allowing faster and more accurate information retrieval compared to novices. Oulasvirta and Saariluoma (2004, 2006) applied this theory to the study of interruptions, conducting a series of experiments examining the influence of interruptions on performance in listening and reading comprehension tasks. These tasks were chosen because, according to Long Term Working Memory theory (Ericsson & Kintsch, 1995), people have sufficient expertise to use the protected memory stores. According to this theory, if people are using the protected retrieval structures associated with Long Term Working Memory, interruptions should not disrupt performance. If people are able to encode information about the primary tasks into Long Term Working Memory then later retrieval of that information following an interruption, should be fast and accurate. According to Long Term Working Memory, interruptions would not disrupt experts performing tasks within their expertise.

Predictions from Long Term Working Memory

Although the work on interruptions and Long Term Working Memory does not specifically address overall quality, it would follow that if people have sufficient expertise to encode details of the primary task into Long Term Working Memory, the overall quality of that task performed with interruptions should not suffer. If people can quickly and accurately recall exactly what they were working on prior to an interruption, as Long Term Working Memory suggests, then there is no reason to believe that the quality of their work would decrease as a result of interruptions. This theory suggests that an expert's ability to perform a task with interruptions is virtually the same as without interruptions.

2.1.2 Other Theoretical Frameworks Related to Quality of Performance

Decision making and the cognitive mechanisms behind decision making might provide useful theoretical insights into how interruptions affect the quality of task performance. Where many of the theories of interrupted task performance focus only on the specific actions around the interruptions event, decision-making theories lend themselves to a more subjective interpretation of quality, which is also the way most of us perceive quality (grades in a class, performance reviews at work, etc.). Specifically, Brunswikian Correspondence Constancy (Brunswik, 1952, 1955), Cognitive Acceleration (Adelman et al., 2003), and Noncompensatory Strategies or Satisficing (Hogarth, 1987; Simon, 1955) theories may prove useful in helping to understand how interruptions may affect the quality of task performance.

Correspondence Constancy

Brunswik (1952, 1955) used the term Correspondence Constancy to refer to people's ability to maintain a constant level of performance across varying difficulties and levels of task performance. He found that people try to stabilize their performance of a task even as the environment and context of performing that task changes. Interruptions not only disrupt tasks, they also increase cognitive demands in the task environment. Correspondence Constancy would suggest that in the face of interruptions, people would try and maintain their level of performance until the processing or task demands from the interruptions became too much for them to handle. Anything past that point would lead to decreased task quality.

Cognitive Acceleration

Adelman et al. (2003) identify Cognitive Acceleration as one mechanism by which people try to maintain a constant level of performance in the face of increasing task demands. They specify two methods by which Cognitive Acceleration may occur: (1) increased processing per unit time and (2) filtered processing. In dealing with interruptions (or increased task demands), people may be able to increase their cognitive processing speed, (i.e., take up the cognitive slack in the system) to handle interruptions, up to a point, without sacrificing the quality of their performance (Edland & Svenson, 1993; Payne et al., 1993). Performance would suffer only when interruptions caused demand increases beyond the ability of the performer to speed his or her processing. Filtered processing (Maule et al., 2000), on the other hand, would suggest that people may be able to maintain performance under the increased demands associated with interruptions by shedding non-critical judgments. However, as with Correspondance Constancy, if the demands become too great, and shedding non-critical judgments is insufficient to adapt to the increased demands, then the quality of performance may be adversely affected.

Noncompensatory Strategies - Satisficing

Satisficing occurs when a decision is made after a minimum acceptable criterion is met, regardless of whether better solutions exist (Hogarth, 1987; Simon, 1955). Jedetski, Adelman and Yeo (2002) found that when consumers were presented with an overwhelming number of options, they engaged in satisficing and did not evaluate all of the different options. Although performing a task with interruptions may not lead to a situation where a person has too many options, it may lead to situations that are overwhelming due to heightened demand. If they engage in satisficing to cope with increased demands, then the quality of their performance may suffer.

Predictions from Decision-Making Theories

All three decision-making theories, though slightly different in their specific mechanisms, all basically would make the same prediction about whether and how interruptions would affect the overall quality of task performance. As people perform a task, they only use the cognitive resources they need to complete what they are doing. Often the required resources to perform a task are well below the maximum a person is able to call upon. As task characteristics and demands change, people innately attempt to maintain high levels of performance (Brunswik, 1952, 1955) by adjusting their approach and the amount of cognitive resources they use. Thus, people are often able to maintain a high level of performance even in the face of increased task demands, up to the point when the task requirements exceed the resources they can call upon. Interruptions are simply a change in the overall complexity of the task and require the person performing the task to use additional cognitive resources to remember the original task, switch to the interruption, and then resume the original task after the interruption. These theories would predict that tasks which require fewer cognitive resources will be less susceptible to the disruptive effects of interruptions than tasks which require greater cognitive resources. In fact, as long as the total cognitive resource requirement of the primary and interrupting tasks is less than the total available cognitive resources, these decision making theories would predict no impairments to the overall quality of the primary task. Estimates of the resources for given tasks can be made using various task analytic procedures.

Distinguishing between Theoretical Approaches

The key distinction among these theories is that only the Memory for Goals framework (Altmann & Trafton, 2002, 2007) does not take aspects of the primary task into account. Rather, when making predictions about performance, this framework suggests that it is only the ability to resume accurately or correct incorrect resumptions that will affect the quality of task performance. Long Term Working Memory theory (Ericsson & Kintsch, 1995) suggests that the primary task characteristics matter in terms of how familiar the person performing the task is with that task. That is, this theory predicts that if a person is performing a task with which he or she has sufficient expertise, interruptions will not affect the quality of task performance. Finally, the decision-making theories – Correspondence Constancy (Brunswik, 1952, 1955), Cognitive Acceleration (Adelman et al., 2003), and Noncompensatory Strategies (Hogarth, 1987; Simon, 1955) – predict that a person should be able to maintain a similar level of performance with interruptions as without until interruptions increase the demands of the tasks beyond what is available to that person. These decision-making theories suggest that it is the primary task resource demands which will affect whether or not interruptions will negatively affect the quality of task performance.

2.2 Classroom Experiment 1

The purpose of Experiment 1 is to conduct an exploratory study to investigate whether and how interruptions affect the quality of task performance in a naturalistic classroom setting. The classroom environment was chosen because it is prone to interruptions and distractions. Additionally, this environment provides a commonly-used and accepted metric of quality – student grades.

2.3 Method

2.3.1 Task

The task for this experiment was a graduate-level behavioral sciences statistical problem set. Two different problem sets were used, each with similar questions about different data sets. Each problem set involved statistical methods that the participants had already covered in their graduate statistics class and took no longer than 1 hour to complete (see Appendix A for actual problem sets used). Questions on the problem sets were in short-answer format and consisted of three or four calculation and interpretation questions and one experimental design question. The calculation and interpretation questions required participants to perform some analysis and provide justification for their answers. The last question in each problem set focused on experimental design; it required participants to design a follow-up study based on the data and analyses they had just completed.

Interruptions and Distractions

In the interrupted condition, interruptions and distractions came from two sources: preplanned external interruptions and participants' own computers and devices in the form of email, instant messages, phone calls, and other sources. During each interrupted session, three or four planned external interruptions occurred. Examples of these included, but were not limited to, people coming into the classroom to ask to borrow some office supplies, loud conversations directly outside the classroom, and conversations between participants and the teacher unrelated to the problem sets. It is important to note that all external interruptions were unrelated to the problem sets and most required participants to completely disengage from working on the problem sets. Additionally, in the interrupted condition, participants were free to interact with each other, their various electronic devices (e.g., cell phones, ipods, etc.) and the Internet, whereas in the no interruption condition, participants were asked to work alone and work only on the assigned problem sets.

2.3.2 Design

This experiment employed a 2 condition (interrupted: yes or no) within-subjects design performed across two sessions approximately one week apart. Participants were randomly assigned to being interrupted either in session 1 or session 2. This was done to help control for any effects of order of exposure to interruptions.

2.3.3 Measures

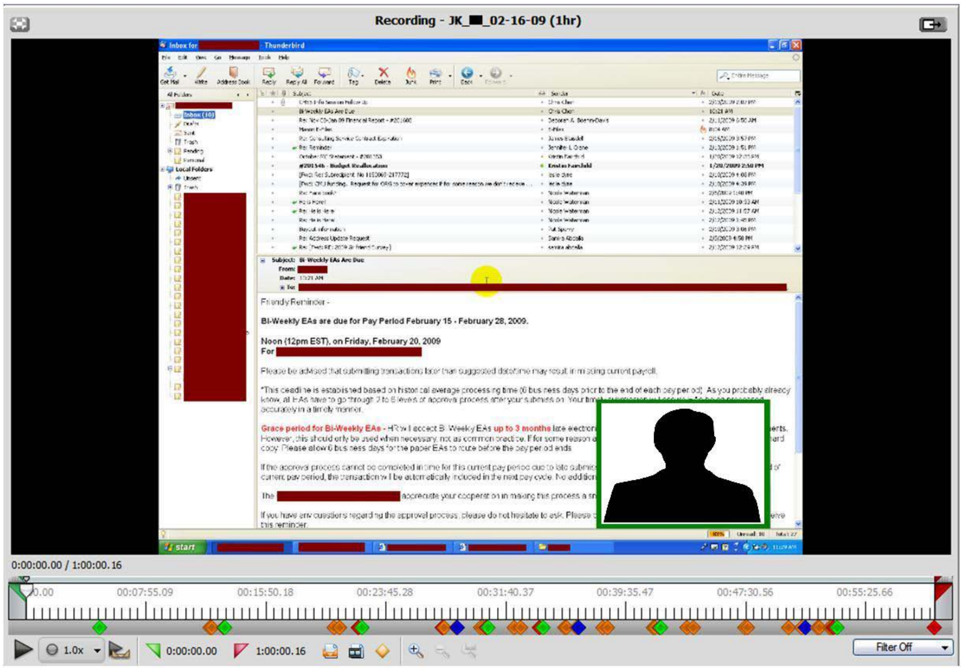
The primary dependent measure was the grades assigned to the students' work. The calculation and interpretation questions were discrete answer questions and were graded based on whether or not the participant got the correct numerical answer. All of the students' open-ended experimental design questions were distributed to four graders for more specific grading. Three of the four graders had completed at least two years of graduate level statistics classes and the fourth had completed one year. The graders were asked to grade the experimental design questions on the following attributes using an eight point Likert scale for each attribute:

- Spelling and grammar
- Flow of answer
- How well does the essay answer the question asked
- Overall grade

In addition to the grades, audio and video of the participants' computer screens was recorded. The recordings were used to calculate the time spent on the problem set, the number of interruptions, the time spent on the interruptions, and the resumption lags. The resumption lag is the time between the end of an interruption and the first action back on the primary task, in this case the problem set. This measure, taken from the Memory for Goals framework (Altmann & Trafton, 2002, 2007), has been used to quantify the disruptive effects of interruptions at the action level (Trafton et al., 2003; Hodgetts & Jones, 2006a, 2006b; Monk et al., 2008). Lastly, questionnaires were used to measure participants' ability to recall specific external interruptions and to gather information about their work habits and preferences.

2.3.4 Materials

The experiment was conducted on laboratory computers in a small classroom at George Mason University. The problem sets were created by the experimenters for this experiment. The data sets were taken from the textbook *Discovering Statistics Using SPSS* (Field, 2009) Participant interaction with the computers was recorded with Techsmith$^{\text{TM}}$Morae usability software (see Figure 2.1 for screen shot of recording). This software captured both audio and video of the computer screen in addition to timestamped key presses and mouse clicks. The demographics questionnaire, statistical screening test, and post-test questionnaire were administered using web-based survey software.



Recording - JK_[redacted]_02-16-09 (1hr)

Subject: Bi-Weekly EAs Are Due
From: [redacted]
Date: 11:21 AM
To: [redacted]

Friendly Reminder -

Bi-Weekly EAs are due for Pay Period February 15 - February 28, 2009.

Noon (12pm EST), on Friday, February 20, 2009
For [redacted]

Please be advised that submitting transactions later than suggested date/time may result in missing current payroll.

*This deadline is established based on historical average processing time (6 business days prior to the end of each pay period). As you probably already know, all EAs have to go through 2 to 6 levels of approval process after your submission. Your tim
accurately in a timely manner.

Grace period for Bi-Weekly EAs - HR will accept Bi-Weekly EAs up to 3 months late electron
However, this should only be used when necessary, not as common practice. If for some reason a
copy. Please allow 6 business days for the paper EAs to route before the pay period ends.

If the approval process cannot be completed in time for this current pay period due to late submis
current pay period, the transaction will be automatically included in the next pay cycle. No addition

The [redacted] appreciate your cooperation in making this process a sm

If you have any questions regarding the approval process, please do not hesitate to ask. Please
this reminder.

Figure 2.1: Screen shot of Techsmith^{TM}Morae usability software data collection setup.

2.3.5 Participants

Thirteen students (5 men, 8 women, average age 22.54 years) enrolled in the introductory graduate statistics class, Psychology 611/612 Introductory Graduate Statistics for the Behavioral Sciences, at George Mason University participated in this experiment. Participation was voluntary and participants received extra credit towards their final course grade in exchange for participation.

2.3.6 Procedure

Each participant completed two problem sets across two sessions separated by approximately one week. Upon arrival to the first session, participants filled out an informed consent form and a demographic questionnaire. All participants then completed a basic 12-question statistical screening test that was developed for use by the George Mason University Psychology Department. The purpose of the screening test was to ensure that all participants had at least a basic understanding of statistical concepts including measures of central tendency and correlation. Participants were not required to get all 12 questions correct; rather, it was important to see that they could understand terms such as mean, median, mode, and standard deviation. An additional purpose of the screening test was to provide a baseline measure of each participant's statistical ability. Once the screening test was complete, participants completed an additional pre-test questionnaire addressing their work habits, styles and preferences. These questions were designed to specifically address the types of environments in which the participants typically worked and liked to work, in addition to how many different projects they worked on at any given time. After completing the screening test and the pre-test questionnaire, participants were asked to complete the first problem set. The participants in the no interruption condition were not able to use the Internet, email, instant messaging, or their cell phone. They were also not allowed to talk to each other, with the exception of asking the instructor quietly if they had a question about the problem set itself. Participants in the interruption condition were specifically told that they could check their email, use the Internet and instant message clients, and use their cell phones if they so desired. In addition, they were told to freely talk among themselves and with the experimenter. Participants in both conditions were given one hour to work on the problem sets. After an hour, they had to stop working and turn in the assignment to the experimenter via email. After completing the problem sets, participants in the interruption condition completed a short questionnaire asking them to recall any external interruptions. Once the questionnaire was complete, the session was over.

The second session was completed within one week of the first and proceeded similarly, except that during the second session, participants did not complete the demographics and pre-test questionnaires or the statistical screening test. Those in the interruption condition during session 2 completed the interruption recall questionnaire after completing the problem set. The order in which participants completed the two problem sets was randomized such that half of the participants completed one of the problem sets during the interruption condition and the other half of the participants completed that same problem set in the no interruption condition.

After completing both sessions of this experiment, participants were debriefed as to the purpose of the experiment and given an opportunity to ask any questions or give any feedback.

2.4 Results and Discussion

2.4.1 Timing Data

One of the more robust effects of interruptions that has been shown in the past is that interruptions lead to increased time to complete a task (Gillie & Broadbent, 1989; Ziljstra et al., 1999). The data show that participants in the interruption condition experienced an average of 13.65 interruptions. To ensure that the interruptions used in this experiment were disruptive at this level, we examined the timing effects. A paired samples t-test revealed that participants in the interruption condition (M = 2131.73 sec, SE = 181.87) took significantly longer to complete the problem set than those in the no interruption condition (M = 1635.29 sec, SE = 212.74), $t(12) = 2.63$, *Cohen's* $d = .70$, $p < .05$ (Figure 2.2).

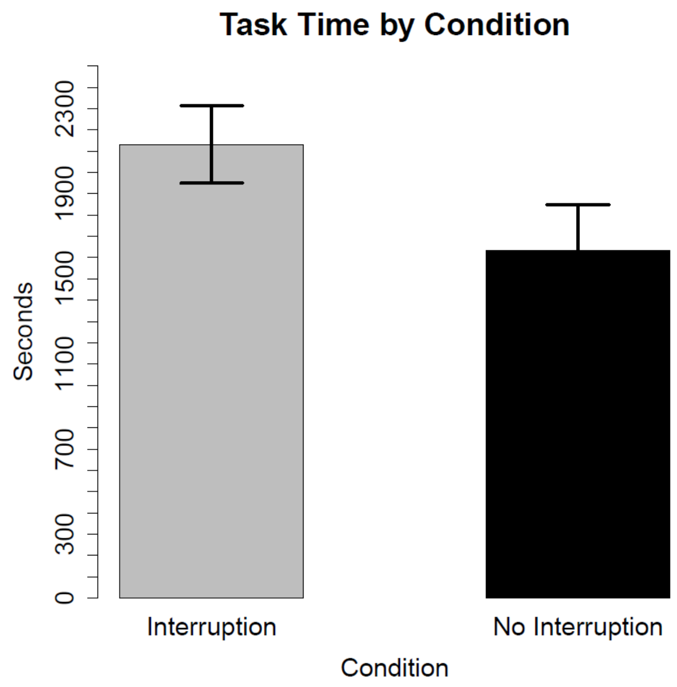


Figure 2.2: Total task time after interruption time was subtracted out by condition (Error bars represent standard error of the mean).

In addition to the statistical comparison of completion times in the interruption and no interruption sessions, we examined other timing metrics from the interruption sessions. Table 2.1 details these metrics observed in the interruption sessions. The interruptions referred to in this table include both the external planned interruptions as well as any interruption that was initiated by the participant, including but not limited to instant messages, email, web browsing, and phone calls.

Table 2.1: Average Timing Metrics (Standard error of the mean in parentheses).

Metric	Time (sec) or Frequency
Individual Interruption Length	36.73 sec (4.22)
Resumption Lag	6.85 sec (2.04)
Total Interruption Time per Session	440.48 sec (177.59)

What is interesting to note about the timing data is the large variability in all four metrics. This suggests that there are a lot of individual differences in how people work in distracted environments and cope with interruptions.

2.4.2 Quality of Performance - Grades

Calculation and Interpretation Questions

Each problem set had four calculation and interpretation questions. Out of a total of 52 responses, there were seven incorrect answers (86.5% correct). Only five participants made errors, each making at least two. Interruptions did not lead to an increase in errors on the calculation and interpretation questions, $F(1{,}12) = 1.00$, $MSE = .038$, $p = .34$, $\eta^2 = .08$. In fact more errors were made in the no interruption condition, $n = 4$, than in the interruption condition, $n = 3$.

According to Long Term Working Memory (Ericsson & Kintsch, 1995), the high levels of performance on these questions could be explained by arguing that the participants were experts in this content area (statistics) and were therefore able to take advantage of the protected memory stores of long term working memory to resume quickly and accurately. However, this explanation does not make sense in this context as the participants had completed at most two semesters of undergraduate statistics and the problem set was based on information covered in a graduate statistics class they were currently enrolled in. This certainly does not meet Ericsson's (1995) definition of expertise.

Alternatively, it could be that, in line with the Memory for Goals framework (Altmann & Trafton, 2002, 2007), any additional errors made upon resumption from an interruption were corrected at some point and the negative effects of interruptions for these questions were contained in the first action back on the primary task. However, observation of the video data show that participants were not observed making resumption errors.

According to theories of decision-making (Brunswik, 1952, 1955; Adelman et al., 2003; Hogarth, 1987; Simon, 1955), the high quality performance on the calculation and interpretation questions may be due to the fact that these questions with interruptions required fewer cognitive resources than the participants had available. If participants had more than sufficient cognitive slack to handle both the primary task and the interruptions, then perhaps the more resource demanding open-ended experimental design questions would be enough to cause quality decrements in the task performance.

Experimental Design Questions

If answering the experimental design questions in the interrupted condition required more cognitive resources than the participants could handle, we would expect to see lower overall grades in the interruption condition as compared to the no interruption condition. A paired samples t-test revealed no significant difference for the overall grades between the no interruption (M = 4.94, SE = .24) and the interruption (M = 4.97, SE = .22) conditions, $t(12)$ = -.170, *Cohen's d* = .03, p = .87.

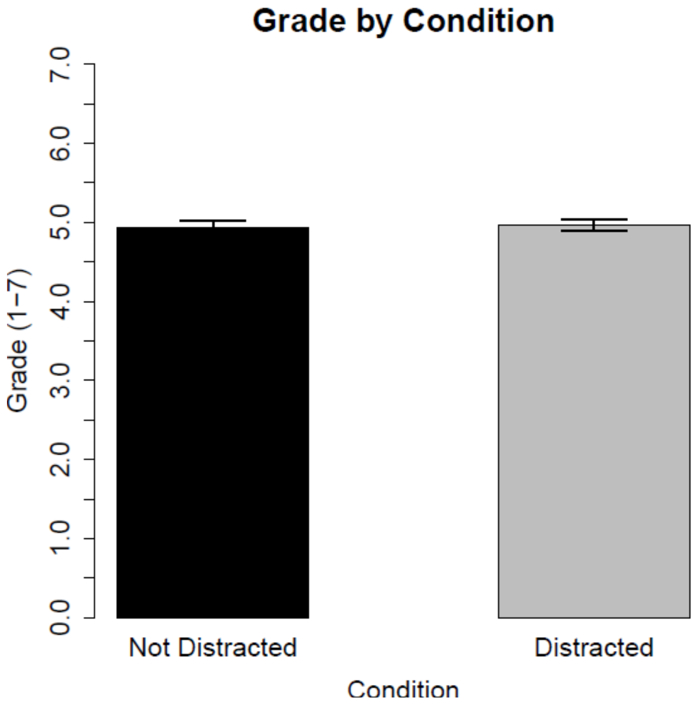


Figure 2.3: Mean overall grades for experimental design question in both conditions (Error bars represent standard error of the mean).

As Figure 2.3 shows, the overall grades for the two conditions were almost the same. Although the t-test revealed that there was not a significant difference between the overall grades for the two conditions, this type of test provides little support for the two conditions actually coming from the same population. This is because null hypothesis significance testing is sensitive to statistical power, sample size, and effect size. High power and an extremely large sample size would be necessary to be reasonably confident that a non-significant finding is actually no difference. In general, sample sizes large enough to satisfy this limitation are not feasible in these experimental settings. Bayesian statistical analyses offer a different, more subjective, approach to providing support for whether or not the two conditions are actually from the same population distribution.

In order to conduct the Bayesian analysis, a discrete prior distribution was created based on the hypothesis that, in general, interruptions are likely to impair the quality of task performance. It is important to note that the prior distribution is not based on any specific data, rather it is based on a hypothesized series of frequencies for the likelihood that interruptions will lead to lower quality performance. Figure 2.4 A shows the series of hypothesized values. Note that this prior distribution is positively skewed, suggesting that most people have a better than 0.5 probability that working with interruptions will lead to lower quality task performance. In fact, for this prior distribution, the area under the curve above a 0.75 probability is 43%. This means that, given this prior distribution, it would be expected that 43% of people would have at least a 0.75 probability of interruptions leading to lower quality task performance.

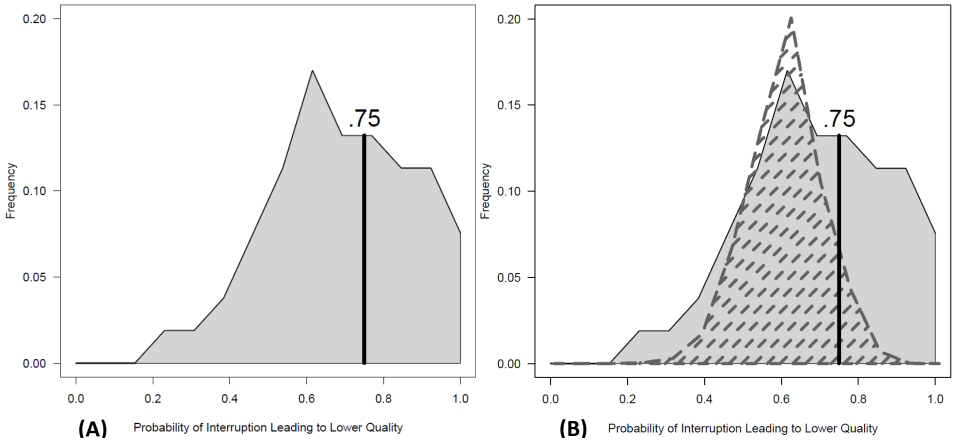


Figure 2.4: **(A)**. Prior distribution of likelihood that interruptions impair the quality of task performance. **(B)**. Dashed is the posterior distribution showing that, given the data, a greater portion of the population has a lower likelihood of showing lower quality task performance when working with interruptions.

The dashed, narrower distribution in Figure 2.4 B reveals the effect that the data from this experiment have on the prior distribution after applying the data using a Bayesian updating rule. The important aspects to note are that the posterior distribution has narrowed and shifted to a point where it is actually slightly negatively skewed. The posterior distribution (dashed distribution in 2.4 B) of a Bayesian analysis represents the effect of the actual data on the prior distribution. The narrowing of the posterior distribution reflects the fact that the data from the experiment have reduced the variability of the measurement. The fact that the posterior distribution is shifted to left, as compared to the prior distribution, reflects the data showing that, in general, there is a lower probability of interruptions leading to quality decrements than initially hypothesized. Most importantly, though, is that the area under the curve from a probability of 0.0 to 0.5 has greatly increased as compared to the prior distribution (shaded area). This shows that, given the data, we believe that there is a greater than initially hypothesized percentage of people (almost half) who have a less than 50% chance of showing lower quality task performance when working with interruptions. Recall that for the prior distribution, it was expected that 43% percent of people would have at least a 0.75 probability of interruptions leading to lower quality performance. For the posterior distribution, the area under the curve above a probability of 0.75 is only 5%. This suggest that, given the data, 95% of people have less than a 0.75 probability of interruptions leading to lower quality performance. According to this analysis, interruptions, for this set of tasks, do not seem to increase a person's chances of producing lower quality work. This provides support that the two groups (interruption and no interruption) are actually being drawn from the same population and that, for these tasks, interruptions do not have a high likelihood of leading to quality decrements.

These first analyses reflect only the overall grades assigned to the experimental design questions by the graders. These questions were also graded by the four graders on four attributes (spelling and grammar, flow, how well the answer addressed the question, and overall grade). To investigate whether or not interruptions had any effect on the grades for these attributes or whether individual graders differed in their grading a 2 (condition: interrupted or not interrupted) x 4 (question: spelling and grammar, flow, how well the answer addressed the question, and overall grade) x 4 (grader: grader 1, grader 2, grader 3, and grader 4) repeated measures ANOVA was conducted. This analyses revealed significant main effects of grader (2.5 A), $F(3,36) = 4.45$, $MSE = .84$, $p < .01$, $\eta^2 = .27$, attribute (2.5 B), $F(3,36) = 6.26$, $MSE = 1.46$, $p = .01$, $\eta^2 = .34$, and a significant grader by attribute interaction, $F(9,108) = 3.21$, $MSE = 1.03$, $p = .01$, $\eta^2 = .21$. Least significant difference post-hoc analyses showed that grader 1 is ($M = 4.82$, $SE = .19$) and 3's ($M = 4.77$, $SE = .23$) grades across all attributes were significantly lower than grader 2's ($M = 5.16$, $SE = .24$), $p < .05$, and grader 3's grades were significantly lower than grader 4's ($M = 5.06$, $SE = .25$), $p < .01$. The significant grader by attribute interaction is not a concern because, even though some of the graders differed in their grading, they did not differ between the interruption and no interruption condition as evidenced by the non-significant interaction between condition and grader. Graders 1 and 3, who graded on the lower side overall, graded lower in both conditions whereas graders 2 and 4 graded on the higher side in both conditions.

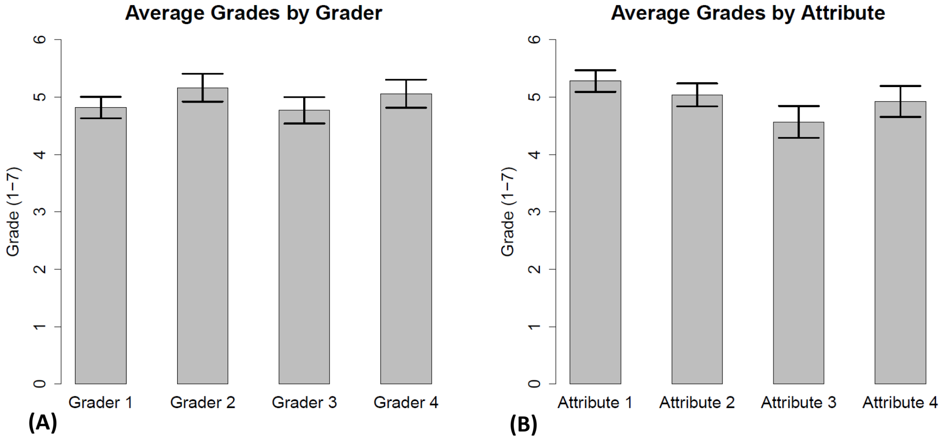


Figure 2.5: **(A)**. Grades by grader across attributes and conditions. **(B)**. Grades by attribute across graders and conditions. 1 = Spelling and Grammar, 2 = Flow, 3 = Appropriateness of Response, 4 = Overall Grade.

Least significant difference post-hoc tests showed that the main effect of attribute was due to fact that the grade for how well the response answered the questions asked ($M = 4.57$, $SE = .28$) was significantly lower than the grade for spelling and grammar ($M = 5.28$, $SE = .19$), $p < .05$, the grade for the flow of the response ($M = 5.04$, $SE = .20$), $p < .05$, and the overall grade for the response ($M = 4.92$, $SE = .27$), $p < .01$. As with the difference in graders, the difference in attribute grades did not differ by condition. The grades on these attributes were consistent across both the interruption and no interruption conditions.

Lastly, a bivariate regression showed that the number of interruptions significantly predicted overall grades, $b = .03$, $t(12) = 3.25$, $p < .01$, and number of interruptions also accounted for a significant proportion of variance in overall grades, $R^2 = .49$, $F(1, 11) = 10.54$, $MSE = .35$, $p < .01$. Interestingly this relationship suggests that, when interrupted, more interruptions actually lead to better performance.

2.4.3 External Interruption Recall

All of the analyses to this point have focused on metrics related to the primary task (i.e., the problem set). If the combination of the primary task and the interruptions required more resources than the participants had available, the decision-making theories would predict that performance would suffer. We hypothesized that this would result in lower quality task performance on the primary task. However, these theories do not specify where decrements might be observed. It could be that participants could show performance decrements in dealing with the interruptions themselves. In each interruption session, participants were exposed to at least 4 external interruptions. After the session, they were asked to recall all of the specific external interruptions they could. Decrements in this task would be evidenced by incorrect recall or inability to recall the interruptions they had just experienced. Out of 52 possible interruptions across all participants, there were only 3 instances where participants were unable to recall all of the interruptions they experienced. Three participants were able to recall only 3 of 4 interruptions. This 94% correct recall rate does not appear to reflect a decrement in performance. These data further support the finding that the interruptions in this experiment did not affect the quality of task performance.

2.5 Work Life Questionnaire

Prior to completing the problem sets, participants answered questions about their work habits, work settings and sources of distraction. Responses reported below were indicated by at least half of the participants, except where noted. In terms of work setting (i.e., amount of noise and distractions present), participants were asked in what setting they typically worked, in what setting they preferred to work, and in what setting they would work for an important assignment. They indicated that they typically work in settings with a little bit of noise, $n = 9$, (as opposed to silence or a lot of noise), they preferred to work in silence $n = 7$, and for important work, they would work in silence $n = 9$. Taken as a whole, this suggests that people generally try to avoid unnecessary distractions and noise in order to get their work done.

In terms of work habits, participants were asked about how long a typical work session would be for them, how many projects they generally work on in each session, and how many projects they would prefer to work on in a given session. Participants indicated that they usually work for at least 2 hours at a time, $n = 8$ (as opposed to ”1-2 hours” or “less than 1 hour”). They usually work on 1 project at a time, but sometimes include a 2nd or 3rd in a typical work session, $n = 9$ (as opposed to “only 1 project” or “typically 2 or more projects” in a given session). In terms of preference, participants indicated that they also prefer to work on 1 project at a time, but sometimes include a 2nd or a 3rd, $n = 6$, however preferring to work on 1 project at a time was only endorsed by slightly less than half of the participants. These answers suggest that people typically try to concentrate on one project at a time, but there are frequent occasions where it makes sense to work on additional projects at the same time or when working on additional projects cannot be avoided.

In terms of distractions, participants were given a list of 10 possible sources of distractions: phone call, text message, instant message, music with no words, music with words, television, email, web surfing, face to face conversations, and social networking. They were asked to indicate all of the potential distractions that they would not want to experience while working, actually do experience while working, help them with their work, and hurt their work. Participants indicated that they would prefer not to receive phone calls, $n = 8$, or instant messages, $n = 8$, watch television, $n = 9$, engage in face to face conversations, $n = 7$ or use social networking tools, $n = 9$. However, they indicated that when working they actually do not receive text messaging, $n = 7$, email, $n = 10$, and do not surf the Internet, $n = 8$. They indicated that only music without words, $n = 7$, helps them work and that phone calls, $n = 7$, television, $n = 8$, and social networking, $n = 9$, hurt their work. Interestingly, there were no items, positively or negatively, that were endorsed by more than 10 of the 13 participants. A few other potential distractions received at least four endorsements, suggesting that there was a lack of consensus among participants for this question. Additionally, 9 of the 10 possible distractions were endorsed by at least one person for all of the categories (i.e., distractions they would not want to experience, distractions they actually do experience, distractions they would prefer not to experience, and distractions that could help them work). These answers seem to suggest that although people know that there are many distractions vying for our attention while we try and work there is little consensus about how they affect our work and much individual variation in how people use these tools while they work.

The two different types of questions used in the first experiment were designed to address different theoretical explanations for potential quality decrements. Because none of the participants were experts in statistics, Long Term Working Memory (Ericsson & Kintsch, 1995) would predict that interruptions should cause quality decrements across both types of questions. Memory for Goals (Altmann & Trafton, 2002, 2007) does not make specific predictions about quality. This framework would, however, predict that interruptions should lead to resumption errors and, if the resumption errors go unnoticed or uncorrected, then quality would suffer. Memory for Goals does not inform whether resumptions errors will be corrected, but it makes sense that the more resumption errors that are made, the more likely some will go uncorrected. Decision-making theories (Brunswik, 1952, 1955; Edland & Svenson, 1993; Hogarth, 1987; Maule et al., 2000; Payne et al., 1993; Simon, 1955) would predict that the experimental design questions would show greater quality decrements due to interruptions than the calculation questions because the experimental design questions required more cognitive resources to complete.

The results of this first experiment showed that there were no quality differences on any of the questions between the interruption and no interruption conditions. While these results do not support a Long Term Working Memory explanation, they do not tease apart Memory for Goals or decision-making explanations.

2.6 Classroom Experiment 2

Experiment 2 was designed to help tease apart whether the Memory for Goals or a decision-making approach is more appropriate when trying to explain how interruptions affect the quality of task performance. In order to differentiate these two possible explanations, we chose a more difficult and demanding task - document editing. The demands of each task type were estimated using a simplified version of a Natural Language Goals, Operators, Methods, and Selection Rules (NGOMSL) approach (Kieras & Polson, 1985). Unlike a traditional full NGOMSL analysis, which includes time for basic perception and movement, the analysis performed here identified only on the cognitive, or mental operation, aspects of the tasks performed. The analyses for the different task types are presented below.

Experiment 1 Tasks

- Calculation and interpretation tasks (3 Mental Operations)
  - Decide which analysis answers question
  - Recall how to perform chosen analysis
  - Decide how to interpret results
- Experimental design questions (at least 4 Mental Operations)
  - Recall results of current data
  - Decide on important next steps
  - Decide on writing approach
  - Decide on whether to review and/or edit writing
    - If task ends here then no more Mental Operations will be required
    - If reviewing or editing occurs then some more Mental Operations will be required

Experiment 2 Tasks

- Editing for local changes - spelling and grammar (2 Mental Operations)
    - Decide if spelling or grammar is correct
    - If wrong, recall and implement correction of spelling or grammar
- Editing for global changes - flow, content, style etc. (at least 6 Mental Operations)
    - Recall content of entire paper (not word for word, but enough to get the gist)
    - Decide where changes are needed
    - Recall and retain section that needs to be changed
    - Think about possible changes, while retaining context
    - Decide and implement change
    - Decide if change works
        - If task ends here then no more Mental Operations will be required
        - If more reviewing or editing occurs then some more Mental Operations will be required

The NGOMSL style analyses show that, in its simplest form, global editing requires greater cognitive resources than any of the other types of tasks that students had to perform in the two experiments. This is true even though the NGOMSL analyses do not take into account how much information must be stored and/or recalled in each task. NGOMSL treats all mental operations equally and therefore assumes that recalling the results of the current data in Experiment 1 requires the same amount of cognitive resources as recalling the entire paper to edit in Experiment 2. Assuming that it does require more cognitive resources to store and/or recall greater amounts of information, then another key distinction that sets the global editing apart from the experimental design questions is the need to retain much greater amounts of information while making decisions on how to proceed. For the experimental design questions, the only prior pieces information that are necessary to retain in memory are the results of the previous experiment. For global editing, on the other hand, it is imperative to retain knowledge about both the section being edited along with how it fits into the flow of the entire document. Combining this additional level of information retention with the larger number of Mental Operations required for global editing suggests that it uses the greatest amount of cognitive resources, and according to the decision-making theories, should be the most susceptible to the negative effects of interruptions.

2.6.1 Hypotheses

According to the Memory for Goals framework, the content of the primary task should not make a difference in how people resume it following an interruption. If this is the case, then switching to a harder primary task with the same types of interruptions should show the same lack of effect on quality that was observed in Experiment 1. Alternatively, in line with a decision-making approach, if the lack of a difference in Experiment 1 was actually due to the participants having more available cognitive resources than the tasks demanded, increasing the amount of resources demanded by the tasks should lead to a decrease in the quality of performance, assuming that the increase is sufficient to exceed the capabilities of the participants.

2.7 Method

2.7.1 Task

Manuscript editing was chosen as the task for Experiment 2. Editing a document is most similar to the experimental design questions from Experiment 1 in that it has no discrete answer and requires a more subjective and creative response. However, unlike the experimental design questions, which did not require participants to keep any information in memory, high quality document editing requires a person to keep large amounts of information in memory in order to remember the context and flow of the document. Certain types of low-level editing (i.e., spelling, grammar, formatting) can be accomplished with little effort, but editing the content, flow, and organization of a manuscript requires concentration and focus.

Two first drafts of manuscripts of approximately equal length were used in this experiment. First drafts were chosen to ensure that there was sufficient need for editing. Both documents were stripped of any identifying remarks to ensure author anonymity.

Interruptions and Distractions

The interruptions and distractions used in the interrupted condition of this experiment were the same as those used in Experiment 1.

2.7.2 Design

The design was also the same as Experiment 1. It was a 2 condition (interrupted: yes or no) within-subjects design completed across two sessions, approximately one week apart.

2.7.3 Measures

All scores were based on a 10-point Likert scale. The number of values available to graders was increased to allow for increased variability in the scores. As with Experiment 1, the primary dependent measure was the overall grade assigned to each edited manuscript. In addition to overall grade, the manuscripts were scored on how well the editor improved the following attributes:

- Spelling and grammar
- Flow
- Elegance
- Structure
- Redundancies reduced

- Succinctness
- Clarity
- Insightfulness of the suggestions or comments made by the editor
- Detail of the suggestions or comments made by the editor
- Quality of rewritten portions of the document
- Degree to which edits were local (i.e., minor word or spelling and grammar changes) versus global (i.e., higher level content or structure issues)
- Overall grade

The same four graders from Experiment 1 were given all of the edited manuscripts and asked to rate each of the twelve attributes only for the portion of the document that the participant completed. This resulted in twelve grades for each document. Participants were not penalized for portions of the document that they were not able to complete.

In addition to the quality metrics, frequency, timing, and progress measures were calculated using the Techsmith^{TM}Morae screen and audio recordings. Specifically, time on task, total time on interruptions, number of interruptions, average interruption length, resumption lag, percentage of document completed (based on number of lines edited), and editing rate (lines/minute) were measured.

2.7.4 Materials and Equipment

The materials and equipment used in Experiment 2 were the same as those used in Experiment 1, with the exceptions that manuscripts were used instead of problem sets and no statistical screening test was used. The Techsmith^{TM}Morae usability suite was again used to capture audio, video, and keystrokes. The same demographics questionnaire from Experiment 1 was used in Experiment 2, but the work habits and interruption recall questionnaires were not used.

2.7.5 Participants

Sixteen students (6 men, 10 women, average age 27.94) enrolled in the introductory graduate statistics class, Psychology 611/612 Introductory Graduate Statistics for the Behavioral Sciences, at George Mason University, who did not participate in experiment 1, participated in this experiment. Participation was voluntary and participants received $5.00 in exchange for participation.

2.7.6 Procedure

The procedure was similar to the one used in Experiment 1. After giving informed consent and completing the demographics questionnaire, participants in both conditions were instructed to edit the manuscript given. They were told to edit for spelling, grammar, format, content, flow, overall structure, and any other relevant changes they saw fit to make. In the no interruption condition, participants had to work alone in silence without using the Internet, email, instant messaging, or their phones. In the interruption condition, participants were free to use whatever additional programs, devices, or tools they wanted to. Participants were given one hour to work on the edits and were told to make as much progress as possible and that it was alright if they were not able to finish. After working for an hour, participants in the interruption condition completed the interruption recall questionnaire. Session 2 proceeded exactly as session 1 did, except that participants did not fill out the demographics questionnaire in session 2.

2.8 Results and Discussion

2.8.1 Quality of Performance - Grades

The grade data for each of the attributes in Experiment 2 were collapsed across graders due to the fact that, as shown in Experiment 1, grader performance was consistent across interruption and no interruption conditions.

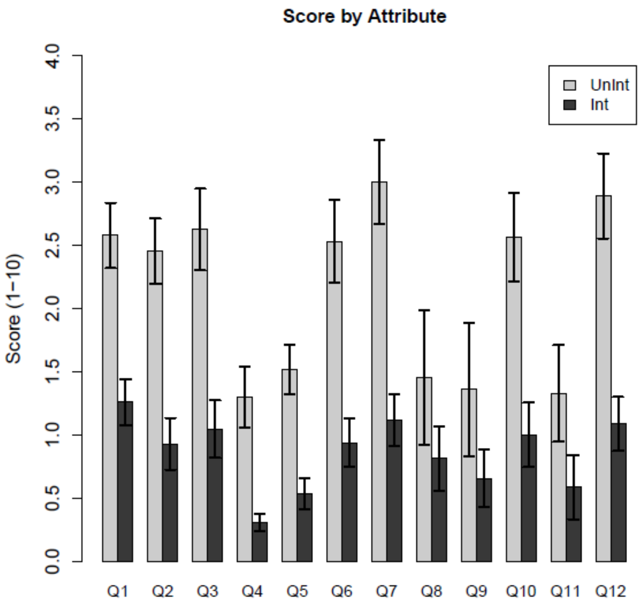


Figure 2.6: Grade by attribute. Attributes are: A1-Spelling and Grammar, A2-Flow, A3-Elegance, A4-Structure, A5-Redundancies, A6-Succinctness, A7-Clarity, A8-Quality of Comments, A9-Detail of Comments, A10-Rewrites, A11-Local(1) vs. Global(10), A12-Overall Grade (Error bars represent standard error of the mean).

The graders were instructed to assign scores for the 12 attributes reflecting only the portion of the documents that participants were able to edit. This was done to ensure that the scores reflected the work that the participants did and would not be negatively biased by the tendency for tasks to take longer when performed with interruptions and distractions (timing effects are discussed more thoroughly in the Timing Data section of the results). Figure 2.6 shows the average scores across graders of all 12 attributes for both the interruption and no interruption conditions. The grades for every attribute are lower in the interrupted condition than in the not interrupted condition. A 2 factor (Interruption vs. No Interruption) repeated measures MANOVA with the 12 attributes as the multiple measures, revealed that all of the attribute grades were significantly lower in the interruption condition except for how good or insightful (A8) and how detailed (A9) the editor's comments were. Table 2.2 shows the MANOVA statistics for each attribute.

Table 2.2: 2 factor (Interruption vs. No Interruption) MANOVA statistics for 12 graded attributes.

Attribute	*F*	*p*	*MSE*	η^2
Spelling and Grammar	18.60(1,15)	<.01	.75	.55
Flow	42.02(1,15)	<.001	.44	.74
Elegance	31.73(1,15)	<.001	.63	.68
Structure	20.52(1,15)	<.001	.38	.58
Redundancies Reduced	32.37(1,15)	<.001	.24	.68
Succinctness	33.20(1,15)	<.001	.61	.69
Clarity	52.61(1,15)	<.001	.54	.78
Insightfulness of Comments	2.31(1,15)	=.15	1.42	.13
Detail of Comments	3.12(1,15)	=.10	1.27	.17
Quality of Rewritten Portions	28.32(1,15)	<.001	.69	.65
Local vs. Global Edits	6.91(1,15)	<.05	.64	.32
Overall Grade	43.96(1,15)	<.001	.59	.75

Further, when all twelve attributes were averaged together, a paired samples t-test revealed that the grades for the interruption condition ($M = .86$, $SE = .15$) were significantly lower than the grades for the no interruption condition ($M = 2.13$, $SE = .27$), $t(15) = 5.69$, *Cohen's* $d = 1.48$, $p < .001$. Taken as a whole, the grading data show that, for this set of tasks, interruptions did indeed reduce the quality of task performance.

The task analyses suggest that the document editing task required more cognitive resources than the statistical problem set tasks used in Experiment 1, and seemingly more total cognitive resources than the participants had available. That data seem to support the task analyses and the results are consistent with a decision-making approach prediction that suggests that quality decrements would be observed if the cognitive resources demanded by the tasks and interruptions exceeded the available resources of the participant. Further, these results neither support nor refute the predictions made by the Memory for Goals framework, which predicts that the effects on quality should be realized in whether or not resumption errors are made and not corrected. As in Experiment 1, participants were not observed making resumption errors. The lack of observed resumption errors with the presence of interruptions is, however, contrary to what would be predicted by the Memory for Goals Framework.

2.8.2 Timing Data

All timing data are based on the Techsmith^{TM}Morae recordings. The recordings for four of the interruption condition sessions were corrupted; therefore, the data only reflect the remaining 12 interruption sessions.

In terms of the raw time actually spent working on the editing task, a paired samples t-test, with mean imputation for the four missing data points, revealed that participants spent significantly more time editing the document in the no interruption condition ($M =$ 2690.48 seconds, $SE = 5.14$) than in the interruption condition ($M = 2042.53$ seconds, $SE =$ 96.47), $t(15) = 6.62$, *Cohen's D* $= 2.37$, $p < .001$ (results of analysis with mean imputation were the same as a similar analysis with unequal n). Participants were able to edit more of the document, in terms of number of lines, in the no interruption condition. This is not surprising because the total time for the session was capped at 1 hour. In order to better understand how interruptions affected the speed at which people were able to work we also investigated the rates at which they performed the editing task.

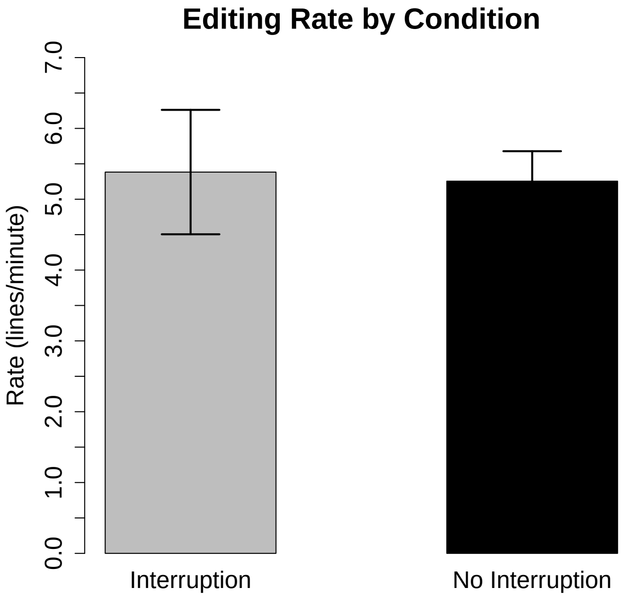


Figure 2.7: Editing rate (lines/minute) by condition (Error bars represent standard error of the mean).

In Experiment 1, even though no quality difference were found between interruption and no interruption conditions, we did find that interruptions caused the task to take longer to complete. It could be that quality decrements were observed in this experiment due to a speed accuracy trade off (i.e., participants worked faster, but less well in the interruption condition). In order to investigate this possibility, an editing rate was calculated for all participants based on the number of lines of the document they completed in the time they spent working on the editing task. A paired samples t-test, imputing the mean for the 4 missing data points, showed no significant difference between the editing rate in the interruption condition ($M = 5.38$ lines/minute, $SE = .88$) and the no interruption condition ($M = 5.26$ lines/minute, $SE = .42$), $t(15) = .13$, *Cohen's* $d = .05$, $p = .90$ (see Figure 2.7) (results of analysis with mean imputation were the same as a similar analysis with unequal n). This finding suggest that the lower quality performance was not due to a speed accuracy trade off. Additional support for the lack of a speed accuracy trade off was shown by a paired samples t-test comparing the number of corrections made per line, imputing the mean for the 4 missing data points which revealed a non-significant difference between the interruption condition (M = .32 corrections/line, SE = .07) and the no interruption condition (M = .43 corrections/line, SE = .06), $t(15)$ = 1.71, *Cohen's d* = .41, p = .11.

Table 2.3 details the additional frequency and timing metrics observed in Experiment 2.

Table 2.3: Average Timing and Frequency Metrics (Standard error of the mean in parentheses).

Metric	Time (sec) or Frequency
Number of Interruptions	14.33 (2.84)
Individual Interruption Length	51.89 sec (7.89)
Resumption Lag	7.32 sec (1.39)
Total Interruption Time per Session	655.71 sec (128.55)

For the most part, these values are similar to those obtained in Experiment 1. The largest disparity between the two experiments is in the total interruption time (overall time in Experiment 1 was shorter than in Experiment 2), however, an independent samples t-test found no significant difference between these two times, $t(24)$ = .98, *Cohen's d* = .39, p = .34. No significant differences were found between the two experiments for any of the timing and frequency metrics.

Overall, the results of Experiment 2 lend support for a decision-making explanation for how interruptions and distractions affect task performance. When interruptions and distractions occurred during a more cognitively demanding task, the quality of task performance suffered. Although participants in this experiment were able to work at the same rate regardless of the presence of interruptions, they were not able to maintain high quality task performance with the addition of interruptions and distractions. These results suggest that interruptions can negatively impact the quality of task performance as the resources demanded by the task increase.

2.9 General Discussion

The environments in which students perform their work today are vastly different from even five or ten years ago. The list of possible sources of interruptions and distractions is ever growing and students are being forced to handle these multiple inputs and still complete their work. While we know from the interruptions domain that interruptions and distractions are generally disruptive at the local or action level, causing longer resumption latencies (Trafton et al., 2003), task times (Gillie & Broadbent, 1989), and poorer task accuracies (Eyrolle & Cellier, 2000), there has not been much work focused on the global effects of these interruptions and distractions on the quality or outcome of task performance.

These two experiments were designed to begin to explore how interruptions and distractions affect the quality of task performance in naturalistic settings, specifically classrooms. An additional goal of this work was to see whether theoretical frameworks used to explain performance at an action level with interruptions in the laboratory experiments could be used to predict and explain the effects on task quality in more realistic settings. Lastly, we were interested to see whether frameworks from other areas, specifically decision making, could be used to add to or more comprehensively explain how interruptions affect the quality of task performance in the classroom

Both experiments found that interruptions disrupted performance, but in different ways. In Experiment 1, interruptions and distractions did not negatively impact the quality of task performance, but did lead to longer task completion times. Experiment 2, on the other hand, showed that when participants completed a more demanding task, the presence of interruptions led to lower quality task performance.

Theoretically, the results from Experiment 1 seem to suggest that it is possible to reduce the negative effects of interruptions without expertise in the task, as would be suggested by Long Term Working Memory. Further, these data do not differentiate between the theoretical predictions of the Memory for Goals or the decision-making theories. The use of a more cognitively demanding task in Experiment 2 was designed as a critical test to rule out either the Memory for Goals or decision-making explanations for this finding. Memory for Goals predicts that any differences in performance would be based on whether or not people make and then correct resumption errors. This framework does not make differential predictions based on different primary tasks. However, decision-making theories predict that if the task demands are increased to a point beyond a person's available resources, then the quality of that task performance should suffer. While the results of Experiment 2 do not necessarily allow us to rule out a Memory for Goals explanation, they do lend stronger support for the decision-making theories' explanation. With a more demanding task, as shown by the NGOMSL task analyses, participants in the interruption condition showed decrements in the quality of their task performance as compared to those in the no interruption condition.

Taken as a whole, these results suggest that there are some tasks that people may be able to perform in distracted and interrupted settings without sacrificing quality, although they may take longer to complete. It seems that a key component in predicting whether or not quality will be reduced is how cognitively demanding a task is, with greater demanding tasks being more likely to show quality decrements when performed with interruptions and distractions.

It is important to understand that these two experiments represent a first exploratory investigation into the effects of interruptions on the quality of task performance. Both experiments were conducted in relatively uncontrolled naturalistic settings. Interruptions and distractions, for the most part were controlled completely by the participants and there was great variability in the number, frequency, length, and type of interruptions in which people engaged. Future work investigating the quality of task performance with interruptions should work to evaluate how these, and other, task characteristics affect quality. Additionally, these experiments did not compare performance to any type of personal baseline for the participants. Although their performance in the no interruption condition can be viewed as one type of baseline, future work should incorporate comparisons of an individual's performance to a more broad baseline (e.g., grades in school, prior performance on similar tasks). This would allow more direct attribution of any differences observed to the presence of interruptions and distractions.

2.10 Conclusions and Practical Implications

With our lives and work settings becoming increasingly filled with distractions and interruptions, it is essential that we continue to work towards understanding what we can or should do versus what we cannot or should not do in distracted settings. In the classroom, quality decrements could result in lower grades; in the office, maybe a poorly worded report or confusing memo, but what about other, potentially safety critical, environments? The implications of interruptions and distractions have the potential to have large and serious consequences. What these two experiments show most clearly is that we are just beginning to understand some of the broader effects of interruptions and distractions.

From a theoretical perspective, this work shows that the current theories of interrupted task performance (Altmann & Trafton, 2002, 2007; Oulasvirta & Saariluoma, 2004, 2004) are not sufficient to explain all of the effects of interruptions and must either be changed or augmented with other theoretical frameworks to provide a more comprehensive picture of performance with interruptions. For this work, we showed that general theories of decision making (Brunswik, 1952, 1955; Adelman et al., 2003; Hogarth, 1987; Simon, 1955) can be used to better explain some of the global effects of interruptions. By broadening and supplementing current theories on interruptions with frameworks and theories from other domains, we can work towards providing a more complete picture of how people deal with interruptions and distractions and we can more accurately predict when and how they will be disruptive.

From an applied perspective, this work shows that interruptions and distractions will not always reduce the overall quality of the final product. There are certain tasks, or types of tasks, that people can complete just as well in distracted environments. This suggests that working alone in silence may not always be necessary for optimal performance. Moving forward, it is essential that we explore the effects of interruptions at both the local/action and global levels in many different environments with many different tasks.

Chapter 3: Examining Interruptions on the Flight Deck: Can we Mitigate their Disruptive Effects?

3.1 Introduction

February 1, 1991 - a 10-passenger Fairchild Metroliner commuter flight had been cleared to hold in take-off position on runway 24L by the tower at Los Angeles International Airport (LAX). Just after sending the hold clearance, the tower controller, who had been communicating with the Metroliner, was interrupted by a request to reestablish contact with another aircraft that had somehow gotten off of the tower frequency. Once the controller got this aircraft back on frequency, he was contacted by a 737 on final approach to LAX asking for landing clearance. The controller issued the 737 landing clearance on runway 24L - the same runway the Metroliner was holding on. By the time the controller realized he had cleared the 737 to land on the same runway that the Metroliner was about to take off on, it was too late. The 737 touched down on top of the smaller aircraft. Thirty-four people lost their lives (NTSB, 1991).

Although no accident can be linked to only one cause, it is clear that the interruption of the tower controller contributed significantly to this incident. Interruptions have been listed as one of many contributing factors to numerous incidents reported to the FAA. Across numerous domains, studies on interrupted task performance have consistently found that interruptions lead to both time (Eyrolle & Cellier, 2000; Trafton et al., 2003) and accuracy costs (Cutrell, Czerwinski, & Horvitz, 2001; Edwards & Gronlund, 1998). Given the prevalence of interruptions in aviation operations and the fact that even a single interruption can lead to disastrous consequences, it is imperative to understand and then mitigate their disruptive effects.

In the aviation domain Dismukes, (2006), Latorella (1998), and LeGoullon (2006) have documented both the frequency and negative effects of interruptions on the flight deck, showing how interruptions can lead to errors and failures to return to interrupted tasks. Dismukes and colleagues (1998; 2001; 2006; 2007) have approached the problem of interruptions on the flight deck as one of Prospective Memory, that is remembering to remember (Brandimonte, Einstein, & McDaniel, 1996). In examining aviation accident reports citing crew error as a primary cause, Dismukes et al. (1998) found that about half were in some way related to interruptions, distractions, or performance of other tasks leading to failures or lapses in prospective memory. Loukopoulos et al. (2001) found observational evidence that when pilots were interrupted or distracted during highly learned and proceduralized tasks, they were more likely to make errors in returning to the initial task, usually by omitting the next relevant step in the task. They proposed the pilots had a difficult time retrieving memory of where they were in the task. As proceduralized tasks are learned, episodic memory traces are formed whereby each step helps to trigger the next step in the sequence. When the chain of steps is broken by interruptions and distractions, the person performing the task can no longer rely on the previous step to trigger the next step in the sequence. Rather, the person must use prospective memory to remember the next step that must be performed when the task is resumed following the distraction. Once a task is broken in the middle by an interruption or distraction, it can often be difficult for the person performing the task to remember the correct next step and errors become more likely.

It is essential to use our knowledge of why interruptions are problematic to work towards mitigating these negative effects. In terms of prospective memory failures, Dismukes (2006) suggests that creating salient reminder cues, using checklists where possible, and speaking aloud next action steps or periodically reviewing equipment status out loud could possibly be used on the flight deck to combat common problems caused by interruptions and distractions.

Latorella (1998) investigated whether the disruptive effects of interruptions on the flight deck could be mitigated by taking advantage of cross-modal (e.g., visual and aural) task combinations as suggested by Multiple Resource Theory (Wickens, 2002). She found that although pilots committed slightly fewer errors later on in the resumed task in cross-modal as compared to single mode pairs, cross-modality in general did not help them perform faster and in certain cases, actually led to worse performance (i.e., more errors on the interrupting task) than single-mode combinations. These findings suggest that interruptions' negative effects cannot be alleviated by simply displaying information on another channel.

Finally, LeGoullon (2006) investigated the role of expertise in mitigating the disruptive effects of interruptions on the flight deck. In comparing untrained undergraduate students to commercially licensed airline pilots, she found that there was no benefit in terms of time to resume a task following an interruption. Consistent with Long Term Working Memory theory (Ericsson & Kintsch, 1995), which has been applied to show that expertise can help reduce, or even eliminate, the negative effects of interruptions (Oulasvirta & Saariluoma, 2004, 2006), LeGoullon did find that the pilots made fewer errors upon resumption than the novices, suggesting that expertise may help to reduce the disruptive effects of interruptions.

3.2 Approach

Previous work looking at interruptions on the flight deck has documented their disruptive nature and frequent occurrence in this safety critical environment. However, most of this is focused locally, that is, on the few steps directly prior to or following task interruption. The goal of this work is to extend our understanding of how interruptions affect the quality of task performance on the flight deck in terms of an entire task. Specifically we are interested in task quality, which reflects a more global view of the entire task. That is, do interruptions and distractions lead to poorer quality decisions? This effect cannot readily be measured with the local cost measures of previous work. A second goal is to test possible methods aimed at mitigating the disruptive effects of interruptions at both local (time) and global (quality) levels and to see how these mitigation strategies affect the overall workload of the scenario.

In this experiment, we examine the effects of interruptions on the quality of pilots' decision making while also evaluating the effectiveness of two possible flight deck tools aimed at facilitating both resumption and accurate decision making following interruptions. We expect that interruptions will lead to longer task times (even after removing the time dealing with the interruptions) and lower quality decisions. Further, we expect that the use of resumption aids will lead to better decision quality without further increasing the time it takes to complete the task. Lastly, we expect that overall workload will be highest for the interrupted condition without resumption aids and lowest for the non-interrupted condition. We expect that workload will be intermediate for the interruption conditions with resumption aids.

3.3 Method

3.3.1 Task and Scenario

The task consisted of a simulated flight. For this scenario the pilots were told they were flying a regional jet, similar to the type on which they were rated, from Denver (DEN) to Wichita (ICT). To achieve this, we used a story board approach, developed with the help of a Subject Matter Expert who is a current commercial airline pilot. For example, Figure 3.1 shows a story board slide for the aircraft state shortly after take off. The scenario was presented by showing pilots static images of what their Flight Management System would look like over time. We also provided them with all of the relevant information that would be included in a flight bag (e.g., release, route, fuel information, destination information, airport maps, etc.) or that they would receive from control and dispatch along the way (e.g., weather, frequency changes, updated airport information, etc.).

As pilots went through the scenario, they had to speak aloud how they would change the aircraft's settings (e.g., speed, altitude, heading, frequency, etc.) and respond to clearances given by the experimenter. Prior to final approach into Wichita, they were informed that all runways at Wichita were reporting nil braking and were closed due to snow. Further, they were informed that the time required to clear the runways was longer than the amount of fuel they had on board. This scenario was set up to force them to choose a new destination airport at this point. Given that no alternate flight destination was provided in the initial flight plan, pilots were required to engage in decision making at this point in the scenario.

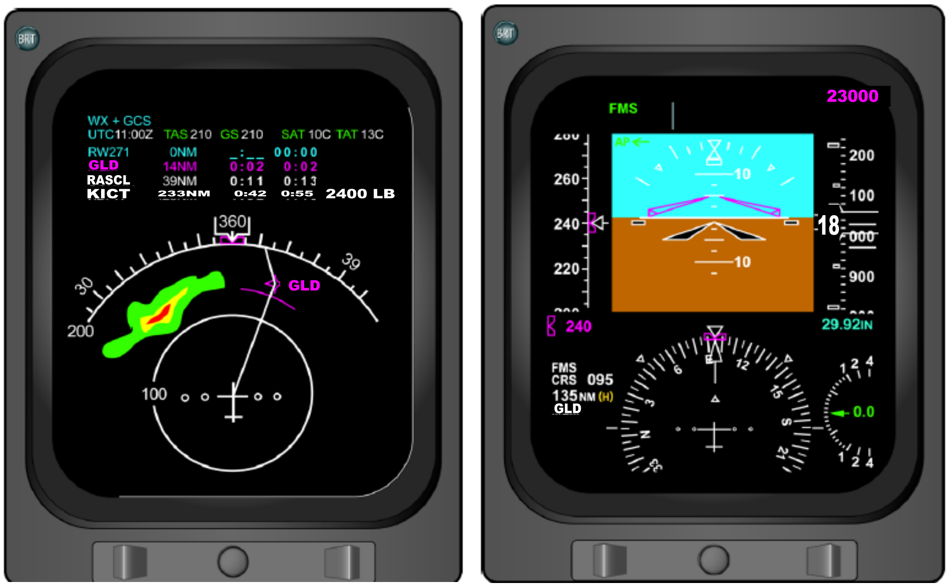


Figure 3.1: Example screen from the story board showing general aircraft status and position information.

Choosing a New Destination

In choosing an alternate airport, a number of considerations need to be taken into account. Here we provide one example of the decision making process that might be used to choose a new destination when it is not preprogrammed into the Flight Management Computer (FMC). It should be noted that not all of these tasks must be performed; rather, this is an example of how this task might be performed.

- Deviate from the current path towards a preprogrammed holding fix
- Enter into a holding pattern

- Check aircraft status (i.e., fuel, flaps, landing gear etc.)
- Determine current location
- Locate nearest suitable airports (suitability may be determined by distance, runway length, airline operations, airport type etc.)
- Choose a few candidate locations (if possible) and evaluate each based on the suitability criteria above (these may be identified using paper maps, airline presence maps, or the FMC)
- Perform fuel calculations to ensure sufficient fuel to make it safely to new destinations
- Consult co-pilot, airline operations, and controllers for more detailed information about suitable destinations
- Captain makes ultimate decision on where to reroute to with safety as primary concern

When choosing a new flight destination, a flight crew must consider many different factors including fuel, distance, weather, runway length, airport type, and airline operations. Taking these factors into account, the constraints and factors mentioned above were managed in such a way that pilots could safely land at a number of airports. Tulsa (commercial) and a number of non-commercial (i.e., general aviation or military) airports were actually closer, but between the bad weather at Tulsa and the repercussions of landing at a non-commercial airport (i.e., no airline services, only acceptable in true emergency situations, more unwanted publicity for pilot and airline, etc.), these alternatives were less appealing. Oklahoma City and Kansas City were the two optimal decisions, with both having good weather, long runways, and airline operations, even though they would result in going slightly into the fuel reserves. It is not unsafe or uncommon for a flight to go into the reserve fuel under the circumstances outlined in the scenario.

Experimental Conditions

Pilots were randomly assigned to one of four experimental conditions: uninterrupted, interrupted with no resumption aid, interrupted with a simple resumption aid, and interrupted with a detailed resumption aid. In the uninterrupted condition, pilots worked through the scenario outlined above and the session ended once they made their decision about the new destination airport. In all of the interrupted conditions, just after beginning the reroute decision-making process, the pilots were given a screen showing a flap failure (Figure 3.2) that required them to run through a portion of their Quick Reference Handbook (QRH), which we provided for them.

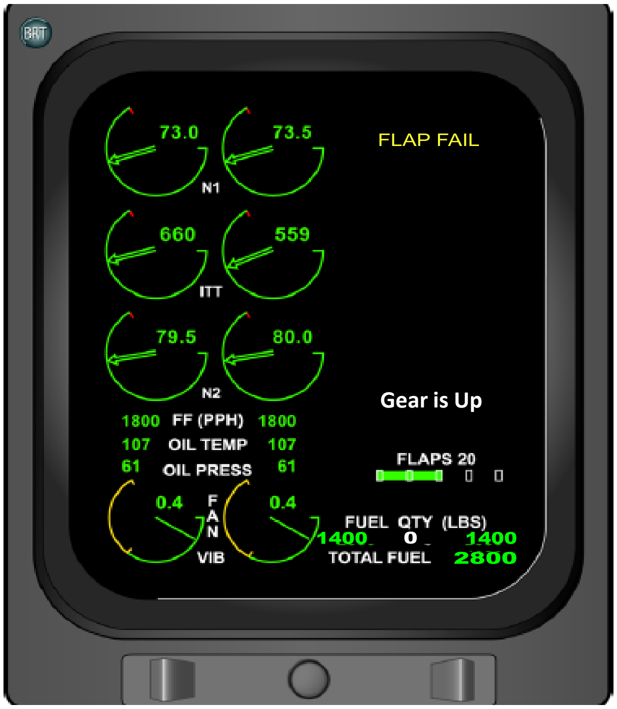
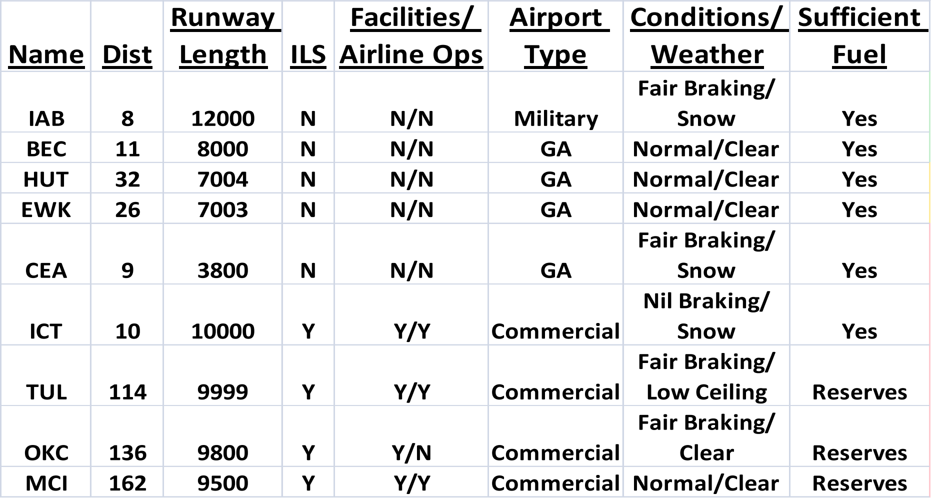


Figure 3.2: Flap failure notification.

Once they completed the QRH for flap failure, they were told that the flap failure was actually due to a problem with the circuit breakers, which were reset, making the flaps fully operational. It is not uncommon for alarms to go off on the flight deck or for pilots to resolve them. Had the flaps remained in a failed state, the pilots would have been forced to declare an emergency and land at the nearest airport regardless of considerations for airline operations and airport type. For this reason, we ensured that when the pilots were considering where to go, they had a fully functional aircraft.

When the interruption was over, pilots returned to the task of choosing a destination under one of three conditions: no resumption aid, simple resumption aid, or detailed resumption aid. In the no aid condition, they were not given any additional instructions and had to return to the destination choosing task on their own. In the simple aid condition, they were presented with a simple screen that read “Don’t forget to choose a new destination!” This simple aid was meant to serve as a sort of post-it-note type reminder. The detailed resumption aid provided the pilots with a table that contained a list of all available airports sorted by distance with name, runway length, distance, facilities and operations information, current weather, and airport type information for the closest commercial and non-commercial airports (Figure 3.3). The detailed resumption aid was designed with the help of the Subject Matter Expert and informal interviews with commercial airline pilots. The two resumption aids differed in the amount and usefulness of information each provided the pilots and the amount of effort the pilots needed to use to understand them (i.e., we suspected that more effort might be required to take in the information in the detailed aid). As with the uninterrupted condition, the experiment ended once the pilots chose a new destination.



Name	Dist	Runway Length	ILS	Facilities/ Airline Ops	Airport Type	Conditions/ Weather	Sufficient Fuel
IAB	8	12000	N	N/N	Military	Fair Braking/ Snow	Yes
BEC	11	8000	N	N/N	GA	Normal/Clear	Yes
HUT	32	7004	N	N/N	GA	Normal/Clear	Yes
EWK	26	7003	N	N/N	GA	Normal/Clear	Yes
CEA	9	3800	N	N/N	GA	Fair Braking/ Snow	Yes
ICT	10	10000	Y	Y/Y	Commercial	Nil Braking/ Snow	Yes
TUL	114	9999	Y	Y/Y	Commercial	Fair Braking/ Low Ceiling	Reserves
OKC	136	9800	Y	Y/N	Commercial	Fair Braking/ Clear	Reserves
MCI	162	9500	Y	Y/Y	Commercial	Normal/Clear	Reserves

Figure 3.3: Detailed resumption aid.

3.3.2 Design

This experiment used a four group, between-subjects design: uninterrupted, interrupted with no resumption aid, interrupted with a simple resumption aid, interrupted with a detailed resumption aid. Condition order between the participants was counterbalanced using a Latin Square.

3.3.3 Measures

Demographic data including age, pilot role (Captain or First Officer), and hours of flight experience were collected prior to the session. Quality of the decision was measured as either "Optimal" (Oklahoma City or Kansas City) or "Non-optimal" (any other airport), based on input from the Subject Matter Expert.

The sessions were recorded using Techsmith Morae^{TM}Usability software. These recordings were used to calculate all of the relevant timing data including total task time, interruption length, and resumption time following the interruption. Lastly, the perceived difficulty of each condition was measured at the conclusion of each session using the NASA-TLX Workload Scale (Hart & Staveland, 1988).

3.3.4 Materials

The story boards were developed with the help of the Subject Matter Expert in Microsoft Powerpoint and were presented on a 13” laptop computer. Paper copies of additional flight bag materials included a flight information sheet, a QRH, weather information, and flight maps.

3.3.5 Participants

Forty participants (39 men, 1 woman, 19 Captains, 21 First Officers), 10 in each of the four conditions, participated in this experiment in exchange for a $25.00 gift card. They had an average age of 29.3 years ($SD = 4.1$) and an average of 4564 ($SD = 1830$) hours of flight experience. All participants were pilots for the same regional airline.

3.3.6 Procedure

Upon arrival, participants filled out an informed consent form and answered demographic questions. They were then given a preflight briefing on the flight plan (i.e., aircraft type, fuel amount, etc.) for the session and given all relevant flight information (i.e., routes, weather maps). Participants also were told that at various times throughout the scenario the experimenter would act as pilot flying, pilot not flying, or the air traffic controller, as needed.

Participants were asked to act as if they were actually on the flight deck for the hypothetical flight and to read back clearances as appropriate and pretend to adjust the proper hardware (e.g., turn the altitude bug to climb or descend). In addition, participants were instructed to use a speak-aloud protocol throughout the scenario.

The scenario proceeded through slides (e.g., Figure 3.1) updating position and aircraft status including the reroute from Wichita and the flap failure interruption for the proper conditions. All resumption aids were displayed on a Powerpoint slide that directly followed the flap failure slide. After the participants verbally confirmed that their decision for a new destination was final, they were asked to rate how confident they were in their decision on a 10-point scale. Laslty, they were asked to complete the NASA-TLX Workload Scale.

3.4 Results and Discussion

All timing data were calculated using the session screen recordings from Techsmith Morae$^{\text{TM}}$Usability software. The videos were time coded at the relevant points (i.e., interruption start, interruption end, resumption point) and these times were used in subsequent analyses. Due to lack of variance in the confidence ratings (i.e., all responses were 7 or above, with 10 being most confident) these data are not discussed further other than to point out that in all cases when pilots made their decisions, they were extremely confident.

3.4.1 Decision Quality

Table 3.1 shows the distribution of non-optimal decisions out of a possible 10 in each condition.

Table 3.1: Number of non-optimal decisions by condition.

Condition	Number of Non-Optimal Decisions
No Interruption	0
No Resumption Aid	1
Simple Resumption Aid	1
Detailed Resumption Aid	5

First, it should be noted that non-optimal decisions were only made in the presence of interruptions. In order to statistically evaluate whether interruptions or the presence of resumption aids affected the distribution of non-optimal decisions, we used a non-parametric binomial test using the total number of errors across all four conditions as the expected distribution. Across the four conditions, there were seven non-optimal decisions out of forty total decisions. This yielded a non-optimal decision rate of 17.5%. Using this rate as the expected non-optimal decision rate, we found that both the no interruption condition, $p < .001$, and the interruption with detailed resumption aid condition, $p < .05$ were significantly different from the expected distribution. However, it is important to point out that this effect is driven by the presence of the detailed resumption aid condition. Without this condition, it is unlikely that we would have observed any significant effects.

Of the seven non-optimal decisions that were made, three were made by captains and four by first officers with a combined average of 4657 hours of flight experience, slightly more than the average for the total sample. This suggests that neither flight role nor experience affected the likelihood of a non-optimal decision.

Only two of the seven non-optimal decisions were to a commercial airport (Tulsa), one in the simple resumption aid condition and one in the detailed resumption aid condition. Tulsa was considered non-optimal because of deteriorating weather there. Although it was the closest commercial option to Wichita, getting to Tulsa still required pilots to go into their reserve fuel and there was a chance that, given the deteriorating weather, landing at Tulsa would be impossible by the time they arrived. The other five non-optimal decisions were to non-commercial airports - 4 to a local military base and 1 to a nearby general aviation airport. The one pilot in the no resumption aid condition who made a non-optimal decision had extensive personal knowledge of all the airports in the area. He chose to reroute to the military base because he was concerned about getting the aircraft on the ground as soon as possible and he was aware that this airport was the closest option with a long runway. The three pilots in the detailed resumption aid condition who also chose the military base cited the same concern for getting the aircraft down as soon as possible; they also were slightly concerned that the flap failure they had just recovered from may be a sign of other mechanical problems. The other pilot in the detailed resumption aid condition who made a non-optimal decision chose to reroute to a nearby general aviation airport as opposed to the military base because he thought there would be a better chance of getting services (i.e., fuel, cleaning, equipment, etc.) at the general aviation facility; he also was interested in getting the aircraft on the ground as soon as possible. It is important to point out that by choosing to land at an airport that we determined to be non-optimal does not necessarily suggest that it is a wrong decision. It merely suggests that taking into account all of the relevant factors, including airline operations and airport type, there were better options that could have been chosen without sacrificing safety.

An interesting theme that emerged from this pattern of non-optimal decisions is that five out of the seven chose a "non-optimal" destination based on information to which they typically would not have access. At no point in this scenario was it necessary for them to declare an emergency, and without doing so or specifically requesting non-commercial landing options, the pilots would not have been presented with them. The only reason they were aware of these options was due to the detailed resumption aid that was provided to them or by extensive local knowledge. In current operations, pilots would combine information from their paper maps, airline service maps, FMS, dispatch, airline operations and air traffic control, all of which would highlight and suggest the commercial options first and foremost. Only if none of these were possible would they be offered non-commercial options. The detailed resumption aid presented the non-commercial options along with standard commercial options. Given this new type of information, almost half of the pilots in this condition chose a non-commercial option. It should be noted that the pilots most likely would have been able to land safely at both the military base and the general aviation airport, but these were deemed non-optimal due to the lack of airline operations and services combined with the ability to reach a commercial airport safely.

3.4.2 Workload

Workload was measured using the paper version of the NASA-TLX Workload Scale (Hart & Staveland, 1988). This workload metric is a weighted score based on subjective ratings of mental demand, physical demand, temporal demand, performance, effort, and frustration. The scores can range from 0-100. No significant differences in overall workload were found between any of the four conditions, $F < 1$ (Figure 3.4).

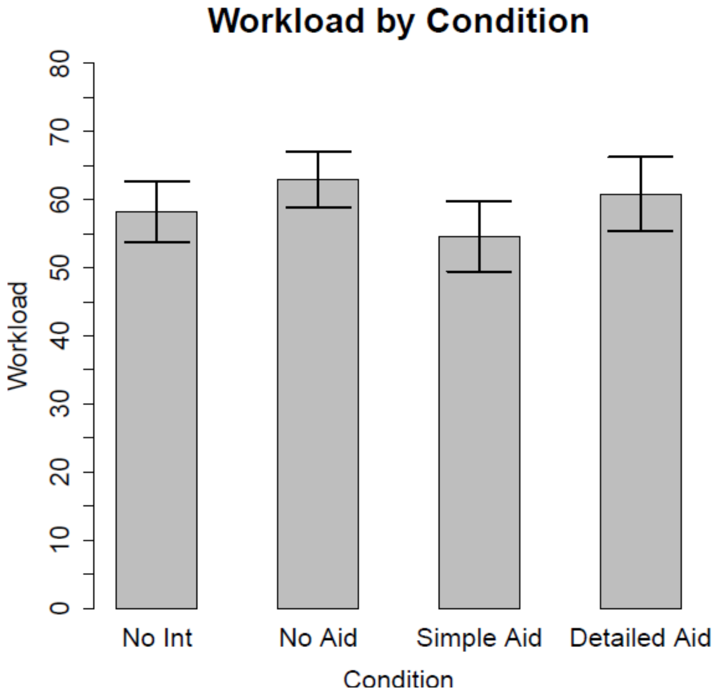


Figure 3.4: Workload across conditions.

The lack of workload differences could be due either to the small effect size or low statistical power for this effect ($Power$ = .11). A post-hoc power analysis revealed that approximately 560 participants would have been required to reliably observe a difference in workload ratings between these conditions. One reason for this is that pilots are specifically trained to deal with abnormal and stressful events on the flight deck. Their level of training qualifies them as experts in dealing with the adverse events presented to them in this experiment. Additionally, this was not a "real" scenario and the simulation was of a relatively low fidelity, which may have added to equal workload levels across conditions. The experimenter observed each and every pilot react calmly and methodically to both the flap failure and need to choose a new destination. At no point did any of the pilots seem overwhelmed or flustered by the potentially dangerous and stressful situation the experiment created. It could be that this training allows them to handle the added stress without affecting their perceived workload. Workload could have also been rated similarly due to the use of a simulation. Were this scenario a real situation, there would be actual potential danger and there would be real lives at stake. Perhaps the lack of realism in our simulation led to lower workload scores in all of the conditions.

3.4.3 Timing Data

The following timing measures were calculated using the Techsmith Morae Usability software recordings: total task time (including the interruption time), total interruption time, and time to resume following an interruption. The total time on task (without the interruption time) was also calculated from the measures listed above. Task time was measured from the reading of the first clearance to the pilots' vocal confirmation of their new destination as their final decision. Interruption time was measured from the flap failure warning tone to the request to put the Quick Reference Handbook away. The time to resume was calculated from the time that the pilot put the Quick Reference Handbook away to the first verbal remark referencing the destination selection task. As this was a self-paced task, there was some individual variation in how fast each pilot went through different sections of the scenario, but the benchmark points used for the timing calculations were common to all participants. In addition, one participant's video data in the simple resumption aid condition was corrupted; thus no timing data could be calculated for this participant.

There were no significant differences in either the total task time (including interruption time), $F(3,35) = 1.76$, $MSE = 33394.64$, $p = .17$, $\eta^2 = .13$, or the total task time (without interruption time), $F < 1$ (Figure 3.5). It was surprising that the no interruption condition did not take less time than the three interruption conditions when the interruption time was included in the measurement, but this lack of a difference may be due to the large variance in task completion times across pilots. It was not surprising that there were no total time differences when the interruptions were taken out. Once the interruption time was removed, the only additional time in the interruption conditions was the resumption time itself. Relative to the total task time, this reflected an extremely small percentage (2.2%) of the total task time, not enough to lead to differences in the total time across the four conditions.

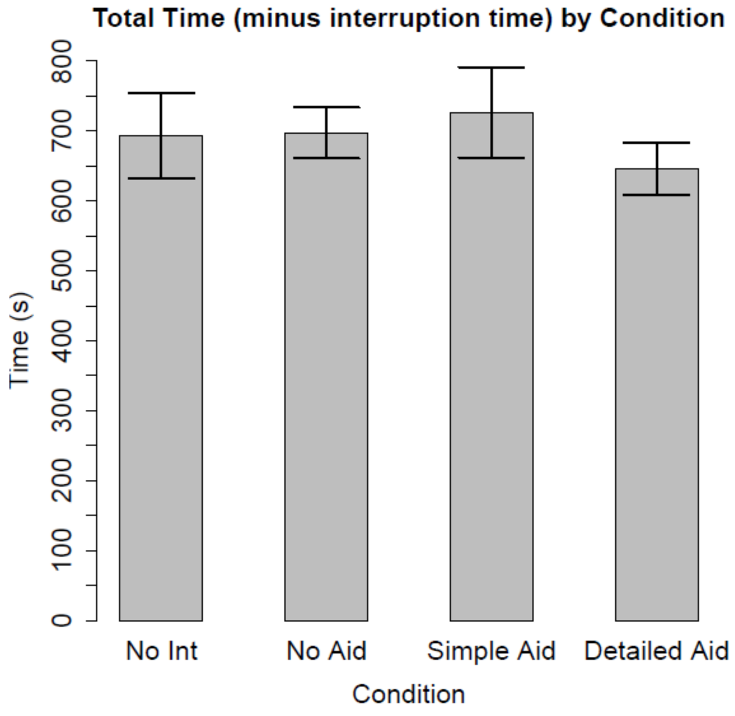


Figure 3.5: Total task time without interruption time.

Analyses of interruption time and time to resume the decision making task were limited to the three interruption conditions. Not surprisingly, there were no differences in the interruption time across the three interruption conditions, $F < 1$. This was expected as the interruption in each of the three conditions was identical. A rather large effect was observed in the difference in time to resume the decision-making task following the interruption, $F(2,26) = 8.01$, $MSE = 106.60$, $p < .01$, $\eta^2 = .38$ (Figure 3.6). Bonferroni post-hoc analyses show that it took pilots longer to resume the decision making task when they had the detailed resumption aid than in either the no aid, $p < .01$, or simple aid, $p < .01$, conditions. There was no difference between the no aid and simple aid conditions, $p = .99$.

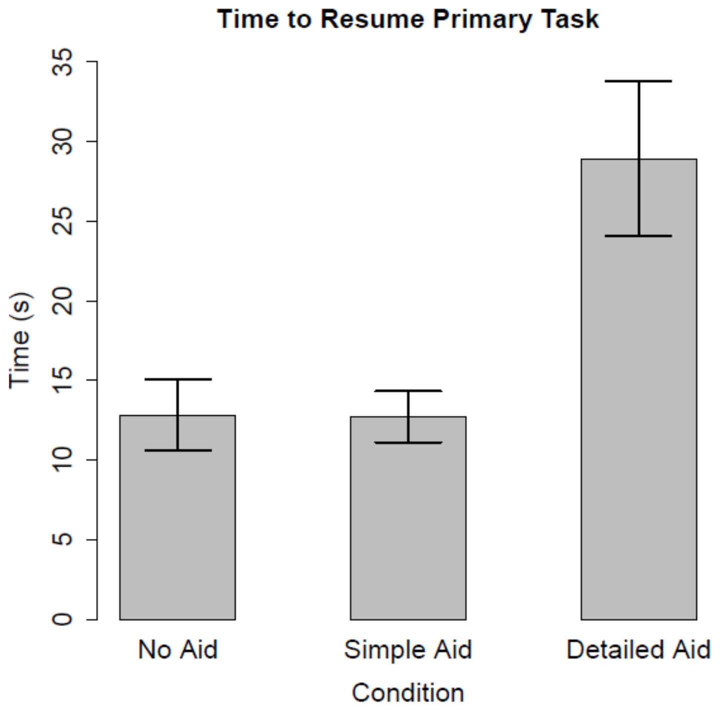


Figure 3.6: Time to resume decision making task following an interruption.

The extra time pilots took to resume the decision-making task with the detailed resumption aid can most likely be attributed to reading or familiarization time. When pilots were first shown the detailed resumption aid (see Figure 3.3), it was the first time they had ever seen it or the information contained within it. All ten pilots in this condition, when first presented with the aid, spent some time reading over the screen and trying to understand what they were looking at. It remains unclear whether this extra time would be needed if pilots had previous exposure to this type of screen and information.

3.5 General Discussion

The goal of this experiment was to investigate how interruptions affect the quality of decision making on the flight deck. In line with previous research documenting the negative effects in this domain (Dismukes, 2006; Latorella, 1998; LeGoullon, 2006), this experiment showed that non-optimal decisions (i.e., lower quality performance) were only made in the presence of interruptions, and were significantly more prevalent when pilots were given additional and novel information intended to help them recover from the interruption. Although confined to only one task, it does suggest that it is possible that interruptions may lead to lower quality performance on the flight deck. It is important to remember that the decisions were classified by one SME and may not necessarily reflect bad decisions. Clearly, further investigation incorporating more types of tasks and outcomes would be necessary to further support or refute this idea.

A secondary goal of this research was to see whether providing information relevant to an interrupted task upon resumption could help pilots resume more quickly or lead to better decisions. Surprisingly, the more information that was provided upon resumption, the more non-optimal decisions pilots made and the longer it took them to resume. It is important to note that when resuming with the detailed resumption aid, pilots were observed taking a long time reading the information in the aid before they made any resumption action. As the resumption time was calculated based on this action, the longer resumption times associated with the detailed resumption aid may be at least partially attributable to the additional reading time. Further, it appears that this may be a result of the amount and type of information provided in the detailed resumption aid. Of the five pilots who made non-optimal decisions with the detailed resumption aid, four of them chose to go to a non-commercial airport they would not necessarily have known about had it not been identified in the detailed resumption aid. All ten pilots in the detailed resumption aid condition commented that they liked the idea of this type of aid. Specifically, they liked having all of the information they would have had to gather from multiple sources on one page. That being said, these results clearly indicate that while the idea behind presenting a detailed resumption aid is sound and appreciated, further investigation is necessary into the amount, the type, and the format of the information presented.

3.6 Conclusions and Practical Implications

The data suggest that interruptions may lead to poorer quality decisions on the flight deck. They also indicate that pilots are interested in tools that may help them recover from interruptions. The simple resumption aid in this task was designed to support a cognitive cuing mechanism (Altmann & Trafton, 2002), which suggests that any type of information that directly points to or is associated with a to-be-remembered task will facilitate the retrieval of that task and help prevent errors commonly seen in prospective memory tasks (Brandimonte et al., 1996). However, the data also make it clear that when designing and implementing these types of tools, it is important to be careful about what information is included. For example, with this type of task and resumption aid, participants indicated that it would be important to them to be able to sort the information and allow different levels of filtering of the information included. Alternatively, it was suggested that the system could be designed to prioritize the list for the pilots based on criteria from either the airline or the Federal Aviation Administration. Additionally, it is important to remember that this was the first exposure pilots had to the detailed resumption aid. They would certainly need to be provided with exposure to any new system or tool prior to implementation. Even though the detailed resumption aid used in this experiment provided pilots with information they would have gathered on their own in order to make a decision, they spent a significant amount of time, when initially shown the aid, looking it over to figure out both what information was in the tool and how this information was arranged. As suggested by Long Term Working Memory theory (Ericsson & Kintsch, 1995), if the pilots in this experiment had sufficient training and exposure to develop expertise with the detailed resumption aid, they likely would have been able to resume much more quickly following the interruption.

When interpreting these results, it is also important to keep in mind that this experiment only examined one specific flight scenario and one specific task that the detailed resumption aid was designed to support. A challenge moving forward will be to identify the characteristics of the tools that help to support task resumption following an interruption and then generalize them to other tasks on the flight deck and other environments (e.g., surface transportation, office work, emergency room, and hospitals). Some type of resumption aid may have prevented the accident detailed in the beginning of this article by reminding the controller of the aircraft waiting on the runway(NTSB, 1991).

With the constant addition of new and more systems on the flight deck, the number and frequency of interruptions will surely increase, as will the incidence of errors resulting from these interruptions. In an environment with zero tolerance for accidents, this will only make it more important to come up with solutions aimed at helping pilots recover from these unavoidable interruptions quickly, accurately, and without reducing the quality of their performance.

Chapter 4: Summary, Conclusions, and Lessons Learned

The present series of studies begins to bring the focal point of interruptions research back to higher-order measures of performance, which were used in early interruptions studies (e.g., Gillie & Broadbent, 1989). After such measures failed to yield meaningful results, theories were developed that localized the effects of interruptions at or near the resumption actions. Having established those local effects of interruptions, this set of experiments represents a return to higher-order performance measures. In an attempt to bridge the established theoretical accounts of interrupted task performance with naturalistic tasks, this body of research aims to test various theories of interrupted task performance, previously evaluated almost entirely in laboratory settings, in dynamic naturalistic environments. Additionally, the three experiments in this dissertation work to examine the effects interruptions have on the overall quality of task performance, whereas most previous research focused only on the few steps immediately following resumption of the primary task.

4.1 Broader Implications

Taken as a whole, the three experiments show that in order to fully understand the disruptive effects of interruptions we must expand our focus from the few steps surrounding the interruption point to include other metrics reflecting the task as a whole, using quality as an example. The findings suggest that some tasks show quality decrements when performed in a distracted and interrupted environment while others do not. Preliminary analysis suggests that tasks that are more cognitively demanding may be more susceptible to quality decrements in the presence of interruptions. Current theories of interrupted task performance may be insufficient to explain these effects as these theories primarily have been used to explain findings in laboratory settings. Additionally, these theories have mainly been tested with non-realistic tasks and have not been adequately evaluated in dynamic naturalistic environments. It is not even clear that current theories are capable of predicting and explaining effects in real-world environments.

4.1.1 Classroom Experiments

The classroom experiments demonstrated that interruptions may or may not negatively affect the overall quality of task performance. In the first experiment interruptions had no effect on overall measures of quality. Long Term Working Memory (Ericsson & Kintsch, 1995) would predict that having expertise would provide protection from the negative effects of interruptions. Thus, Long Term Working Memory does not provide a prediction for non-experts, other than the implication that they would perform more poorly without access to the protected memory stores. Participants in the classroom studies were not experts in the tasks they performed, but were still always able to perform the tasks without quality decrements. This suggests that expertise is not necessary to prevent quality decrements caused by interruptions.

Additionally, these results make it difficult to support or refute the Memory for Goals (Altmann & Trafton, 2002, 2007) framework or decision-making theories (Brunswik, 1952, 1955; Adelman et al., 2003; Hogarth, 1987; Simon, 1955). In line with the Memory for Goals framework (Altmann & Trafton, 2002, 2007), these null findings could be due to participants correcting resumptions errors. However, participants were not observed making or correcting many resumption errors, therefore, there is not evidence to support or refute this explanation. Alternatively, the null findings could have been a result of the tasks demanding relatively few cognitive resources. That is, the resources required may have been within the capacity of the participants to perform the primary and interruption tasks without sacrificing quality. The second classroom experiment was designed to tease apart the Memory for Goals and decision-making explanations.

In the second study interruptions led to performance decrements. We hypothesized that this was due to the use of a more complex task that taxed the participants' available resources. The cognitive task analyses indicated that the document editing task used in the second classroom experiment required the greatest number of mental operations and showed quality decrements when performed with interruptions. The discrete or creative tasks used in the first classroom experiment and the local editing task in the second experiment required fewer mental operations than the global editing task and showed no quality decrements when performed with interruptions.

Theoretically, these results lend support to a decision-making explanation of performance. It would make sense that with a more cognitively demanding task, participants did not have sufficient cognitive resources to cope with the increased demand. According to this explanation, quality suffered because the demand for cognitive resources was higher than the available supply that participants could utilize. Although these results do not lend direct support to a Memory for Goals explanation of task decrements (i.e., the locus of disruption is in the first few actions immediately after resumption), they also do not provide evidence to refute such an explanation.

4.1.2 Flightdeck Experiment

One goal of the flightdeck experiment was to examine a situation in which expertise may have helped mitigate some of the negative effects associated with interruptions. No workload or timing differences were observed across the four conditions, suggesting that the pilots were able to handle the destination-choosing task with an interruption without significant difficulty. This may indicate that their expertise with flight operations helped to protect them from any negative effects of interruptions.

An additional goal of the flightdeck study was to evaluate the effectiveness of potential task resumption aids, designed to help mitigate the disruptive effects of interruptions in a safety-critical environment. Interestingly, five of the ten participants exposed to the detailed resumption aid made non-optimal decisions. However, it is important to point out that although the decisions were rated as non-optimal, this does not mean that they were wrong or unsafe. It only means that there were other decisions that the pilots could have made that would have been better when considering criteria including safety, airline operations, facilities, and other destination aspects. Although all of the pilots in the detailed resumption aid condition commented that they liked the information provided by the aid, the results clearly indicated that there was a disconnect between the intention of the designer and the use of the tool.

Two possible explanations for the failure of the detailed resumption aid were that (a) the aid provided too much information and (b) lack of training. The detailed resumption aid contained information on military and general aviation airports for which the pilots would not generally have access unless they declared an emergency. The scenario in this experiment was designed specifically to not require that an emergency be declared. This was done to increase the list of potential acceptable solutions. Future iterations of this type of tool should consider limiting the options and information presented. The other possible explanation for the high number of non-optimal decisions associated with use of the detailed resumption aid is lack of training. There are few professions which provide and require as much training as commercial aviation. Before pilots ever fly a real aircraft, they have already logged hundreds, if not thousands, of hours in high-fidelity simulators, exposing them to all types of flights and both normal and emergency conditions. Pilots would almost certainly not encounter a new piece of technology for the first time on the flightdeck during an actual flight.

In addition to having the greatest number of non-optimal decisions, the detailed resumption aid also led to longer resumption times than either the simple or no aid conditions. This could be, at least in part, due to the lack of training and experience with the detailed aid, but it may also be due to the time participants spent reading the information on the detailed aid upon resumption. The resumption time was calculated from the end of the interruption to the first verbal or mouse action back on the task. Many of the pilots were observed reading through the detailed aid before making a resumption action, increasing the calculated resumption time. Although total task time in this experiment was the same, with more exposure to this tool pilots might be able to actually complete the task more quickly.

4.2 Experimental Limitations

When interpreting the results of this research and assessing their impact on our understanding of interrupted task performance, it is important to keep in mind certain limitations. In order to accurately understand how interruptions affect task performance in naturalistic environments, participants must perform the tasks as they would in those environments. To this end, the classroom studies were designed to try and ensure that the participants were motivated to complete the task to the best of their ability (i.e., extra credit in experiment 1 and payment in experiment 2). Even with these incentives, it is possible that the participants were not taking the tasks as seriously as they would if it were their own work. One way to address this in the classroom is to employ a similar type of experimental design in actual classes or lab settings, taking care to not knowingly place any students in a situation that might impair their learning experience. For the flightdeck experiment, this issue becomes a little more complicated. The best way to ensure high motivation of the participants would be to introduce similar scenarios into actual flight; however, for clear passenger safety reasons that is probably not possible. The next best solution would be to increase the fidelity of the flight simulation. Using a static story board approach may not have allowed the pilots to completely immerse themselves in the task and it certainly did not take into account any actual interactions with the flight systems.

These experiments were conducted with two specific groups of participants. A broader sample is necessary to generalize these findings beyond George Mason psychology graduate students and regional jet commercial airline pilots from one airline. This design cannot account for any differences among students from different programs, schools, or at different levels of education nor can it account for differences between pilots at different airlines or pilots who are rated on different types of aircraft. Lastly, from an overall impact perspective, these studies were limited in that they concentrated on one global metric (i.e., quality) and used a limited number and type of external interruptions and primary tasks. Given these limitations, the results of this line of research should be viewed as a strong first step towards investigating global metrics of interruptions in naturalistic environments.

Another goal of this line of research was to evaluate interrupted task performance in real world settings. In order to truly accomplish that goal, every aspect of the study needs to be as close to the real world as possible. Within the clear limitations of participants knowing they were part of a study or having to use a relatively low-fidelity simulator, care was taken to choose realistic tasks. Completing problem sets and editing documents are certainly tasks that a psychology graduate student would experience and the flightdeck scenario was designed with the help of a Subject Matter Expert to ensure that it was realistic. The interruptions used in these scenarios must also be realistic. While the use of the Internet, instant messaging, email, cell phones and other distractions by the participants in the classroom studies was certainly realistic, there may be some concern that some of the external interruptions may have appeared planned or contrived. For the most part, the external interruptions (e.g., someone coming in to the room to ask a question, loud conversations outside, etc.) seemed to come across as natural; however, the experimenter observed several instances where the participants acknowledged that they believed the distraction to be part of the experiment. Similarly, with the flightdeck study, upon being interrupted some of the participants responded that they expected more uncommon events because they were part of a research study. While it is certainly difficult to completely address the realism of every aspect of any study, it is important to understand what some of these limitations may be.

Certain metrics are easier to measure and define than others. For instance, reaction time is very clearly the time it takes someone to react to something and can be measured at varying levels of specificity with any timekeeping device. Other metrics, such as quality, can be much more ambiguous. For the purposes of this work, quality in the classroom was defined as a grade given by a grader and quality on the flightdeck was based on an optimal or non-optimal decision rating determined by a subject matter expert. A limitation of this study is that there may be other ways to more objectively define and measure quality.

Three potential methods to make the measure of quality more objective are the use of a true gold standard, individual baselines, or validated rubrics. In the classroom experiments there was not a gold standard, or already established high quality example of either the problem sets in the first experiment or the manuscripts in the second experiment. This issue could be addressed by taking a published work, using it as the gold standard, purposely perturbing it and then asking participants to edit it. Comparing the participants' edited version of the perturbed document to the gold standard would provide a more objective and direct measure of the quality of task performance. The use of personal baselines could also make the quality measurements more objective. If each participant's abilities and performance were known prior to the experiment, then we could directly compare the effects of the experiment to each participant's baseline. One potential way to do this in the classroom is to use participants' actual grades from relevant coursework. On the flighdeck, factoring in flight hours and any information from pilots' records with regards to performance issues (e.g., missed approaches, near-misses, other incidents, etc.) could be used to establish a baseline which could be compared to their performance in the experiment. The use of a baseline would provide a more direct measure of the impact of interruptions on quality. Another approach to more objectively assess quality would be to establish evaluation rubrics for the tasks being performed. At least in the classroom, many of the tasks that teachers ask students to do are tasks that have been around for a long time and that the teachers are extremely familiar with. Some teachers have already established rubrics to guide them through evaluation of the tasks. The rubrics can be developed by experts on a task by task basis to reflect the most important features of each task and can then be employed by the evaluators to provide more reliable and objective measurements of quality.

4.3 Theoretical Limitations

One aim of this work was to evaluate theories, primarily developed in and based on laboratory work, in dynamic naturalistic environments. The goal was to have these experiments, as much as possible, serve as critical tests of the theories, helping to firmly support or refute each one. We certainly understand more now about how accurate and effective the theories evaluated (i.e., Long Term Working Memory, Memory for Goals, and the decision-making frameworks) are in informing and predicting interruptions' effects on task quality in naturalistic environments. However, these experiments fell short of being able to directly support or refute any of the theoretical frameworks discussed.

The first classroom experiment showed that expertise is not necessary to avoid quality decrements with interruptions, but the flightdeck study showed that expertise may be helpful in avoiding these negative effects. Together, these findings certainly cannot rule out the role Long Term Working Memory (Ericsson & Kintsch, 1995) may have in mitigating the disruptive effects of interruptions. A better test of this theory would be to have people perform a task with which half of them have expertise and the other half are novices. Both groups would perform the task with and without interruptions and if the experts were able to take advantage of Long Term Working Memory encoding and retrieval, their performance should be similar in both conditions. Additionally, the novices should show decrements in performance when they performed the task with interruptions.

Similar to the lack of direct support or refutation for Long Term Working Memory, both the flightdeck and classroom studies do little to directly support for or refute the Memory for Goals framework (Altmann & Trafton, 2002, 2007). This framework does not make specific direct predictions in regards to how interruptions should affect the quality of task performance. It would predict that interruptions may lead to more resumption errors which, if uncorrected, could negatively affect quality. However, as participants in all three experiments were not observed making resumption errors, it was difficult to evaluate the predictions associated with Memory for Goals. Perhaps the most direct support for any theory was the support for the decision-making theories (Brunswik, 1952, 1955; Adelman et al., 2003; Hogarth, 1987; Simon, 1955) from the second classroom experiment. The critical manipulation in this experiment was using a cognitive task analysis to measure the cognitive resources required by the tasks, and choosing tasks that, according to the tasks analyses, required more cognitive resources than any of the tasks used in the first experiment. The tasks used in the second classroom experiment, which required the greatest amount of cognitive resources according to the task analyses, did show quality decrements when performed with interruptions. However, there is no way from the current experiments to determine if this the quality decrements are directly due to the increased cognitive demand or if they are actually due to some other difference between the tasks used in the first and second classroom experiment.

These experiments' lack in providing critical tests of the theories is mostly due to two factors: loss of experimental control in naturalistic environments and not enough experiments. Evaluating the real world impact of what we learn in the laboratory is a key part of a successful Human Factors research initiative. However, anytime we leave the lab we almost always sacrifice experimental control. In laboratory interruptions research, we can control exactly when and how people are interrupted and exactly where how and how they resume following that interruption. We can control the length, the difficulty, the modality, and many other aspects of either the interruption or the primary task in order to address specific research questions. As soon as we begin to gather data and perform research in real-world environments we sacrifice a lot of this control. Looking at the two classroom experiments, we, for the most part, could not control when, how, or how often people were interrupted or what they were interrupted with.

Additionally, in lab settings we can carefully manipulate one or two variables between experiments or conditions while holding everything else constant, and then directly attribute any observed effects to those manipulations. In real-world environments we can manipulate different aspects of the experimental conditions, but it becomes much more difficult to hold all other variable constant, or to even identify all of the variables that could potentially affect the results. From the first to the second classroom experiment we quantified the change in cognitive resources required using a cognitive task analysis and we know that quality suffered with interruptions in this second experiment. While that finding is in line with decision-making theories, we do not know if that is the only reason that a quality decrement was observed. It could be that the tasks changed, or maybe that participants were using different types of interruptions for some reason. This all points the difficulty of designing critical tests of theories in naturalistic environments.

One way to overcome the difficulty of critically testing theories in the real-world is to cast a wider net in terms of the environments and tasks tested. We ran three experiments across two environments using three types of tasks. Any conclusions drawn from these findings could potentially only apply to these specific situations. In order to determine if and how well these effects generalize it would be necessary to conduct similar experiments using different tasks in the same and other domains. If, for instance, we conducted another ten experiments in the classroom using different types of tasks, and we found that the tasks that were more cognitively demanding reliably led to greater quality decrements, it would lend much stronger support to the decision-making explanation of task quality with interruptions. We were limited in this regard by the scope of this dissertation, but moving forward it is perfectly reasonable to begin to conduct other similar experiments and work towards building a database of different tasks that do and do not show quality decrements with interruptions. These tasks can then be compared across multiple attributes (e.g., cognitive resources, modality, complexity, etc.) to better determine what is actually causing the quality decrements.

4.4 Next Steps

Many of the limitations mentioned speak to logical future directions to gain a more complete understanding of the role of interruptions in the real world. Future studies should incorporate a wider array of tasks for both the primary and interruption content. More objective or quantitative metrics of workload may help further tease apart the decision-making and memory-based explanations of interrupted task performance. Another important next step is to develop, refine, and implement mitigation strategies and tools. While the flightdeck study is an example of an attempt at implementing a mitigation tool, the results clearly suggest that further iterations or other tools are necessary. In all environments it is important to consider how the mitigation tools and strategies developed will generalize to other tasks and settings. In an ideal situation, the development of the tools and strategies will be followed with experimental evaluation of the tools and strategies. Employing an iterative design process will allow the best tools and training programs to be developed and implemented. As a Human Factors practitioner it is essential to not only add knowledge, but to also use that knowledge to improve how we interact with our world.

4.5 Final Thoughts

Moving forward, it will be important to expand both the breadth and interaction of the theories from various domains (e.g., expertise, memory, decision-making) that may bear on interrupted task performance. Furthermore, we must continue to explore interruptions' effects in other naturalistic environments, for other task sets, and on other metrics. This work suggests that a complete understanding of interrupted task performance may come through integration of theories from various domains, incorporating both local (i.e., timing) and global (i.e., quality) metrics. Additionally, it shows the both the value and compromises that must be considered when pushing outside of the laboratory into the real world. Studies in the real world are essential in order to truly understand effects observed in the lab, but attention must be given to the controls that are sacrificed as we increase mundane realism. What we can really take away from this research is that a complete understanding of interrupted task performance will likely rely on more than a single theory and will be achieved only by integrating work from other domains and evaluating performance in both laboratory and naturalistic environments.

Appendix A: Problem Sets from Classroom Experiment 1

PROBLEM SET 1

Please complete all questions below to the best of your ability. If you are not familiar how to perform any of these functions in SPSS please ask the instructor. You have 30 minutes to complete this task.

Using the **ChickFlick.sav** data:

1. Determine if there is a relationship between gender and arousal?

2. Determine if there is a relationship between film watched and arousal?

3. Are the arousal levels significantly different between the two films viewed?

4. Are the arousal levels significantly different between male and female viewers?

5. Design a follow up experiment/data collection to investigate other factors that may affect the arousal level of movie viewers? Be as detailed as possible.

PROBLEM SET 2

Please complete all questions below to the best of your ability. If you are not familiar how to perform any of these functions in SPSS please ask the instructor.

A student was interested in whether there was a positive relationship between the time spent doing an essay and the mark received. He got 45 of his friends and timed how long they spent writing an essay (**hours**) and the percentage they got in the essay (**essay**). He also translated the grades into their degree classifications (grade): first, upper second, lower second, and third class. Using the data in the file **EssayMarks.sav**:

1. Find out what the relationship was between the time spent doing an essay and the eventual mark in terms of percentage and degree class.

2. Determine whether the differences between the classifications were significant.

3. Outline the design of a follow up experiment/data collection to determine what other factors might affect performance on the essay task. Be as specific as possible.

References

References

Adelman, L., Miller, S. L., Henderson, D. V., & Schoelles, M. (2003). Using Brunswikian theory and a longitudinal design to study how hierarchical teams adapt to increasing levels of time pressure. *Acta psychologica*, *112*(2), 181–206.

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*, *26*, 39–83.

Altmann, E. M., & Trafton, J. G. (2004). Task interruption: Resumption lab and the role of cues. In *Proceedings of the 26th annual conference of the cognitive science society.* Erlbaum.

Altmann, E. M., & Trafton, J. G. (2007). Timecourse of recovery from task interruption: Data and a model. *Psychonomics Bulletin and Review*.

Begley, S. (2009). Will the blackberry sink the presidency? *Newsweek*.

Brandimonte, M., Einstein, G. O., & McDaniel, M. A. (1996). *Prospective memory: Theory and applications.* Lawrence Erlbaum Associates New Jersey, USA.

Brunswik, E. (1952). *The conceptual framework of psychology, international encyclopedia of unified science, vol1, no. 10.* University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193.

Cades, D. M., Boehm-Davis, D. A., & Smith, M. B. (2010). Mitigating the disruptive effects of interruptions on the flight deck. In *Hci-aero 2010.*

Cades, D. M., Boehm-Davis, D. A., & Trafton, J. G. (2007). Interruptions in the office: An observational field study. In *2007 american psychological association division 21, division 19 and human factors and ergonomics society potomac chapter annual* *symposium on applied experimental research.* Fairfax: George Mason University.

Cades, D. M., Trafton, J. G., Boehm-Davis, D. A., & Monk, C. A. (2007). Does the difficulty of an interruption affect our ability to resume? In *51st annual human factors and ergonomics society conference* (pp. 234–238). Baltimore, Maryland: Human Factors and Ergonomics Society.

Cades, D. M., Werner, N. E., Boehm-Davis, D. A., & Arshad, Z. (2010). Interruptions are disruptive in the real world: Evidence from an office setting. In *54th annual human factors and ergonomics society conference.* San Francisco, California: Human Factors and Ergonomics Society.

Cutrell, E., Czerwinski, M., & Horvitz, E. (2001). Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In M. Hirose (Ed.), *Human-Computer interaction- INTERACT 2001 conference proceedings* (pp. 263–269). Amsterdam: IOS Press.

Czerwinski, M., Horvitz, E., & Wilhite, S. (2004). A diary study of task switching and interruptions. In *Human factors in computing systems: Proceedings of CHI'04* (pp. 175–182). New York: ACM Press.

Dismukes, R. K. (2006). Concurrent task management and prospective memory: pilot error as a model for the vulnerability of experts. In *50th annual human factors and ergonomics society conference* (Vol. 50, p. 909-913). San Francisco, California: Human Factors and Ergonomics Society.

Dismukes, R. K., Berman, B. A., & Loukopoulos, L. D. (2007). *The limits of expertise: Rethinking pilot error and the causes of airline accidents* (1st ed.). Burlington, VT: Ashgate.

Dismukes, R. K., & Young, G. (1998). Cockpit interruptions and distractions: Effective management requires a careful balancing act. *ASRS Directline*(10), 4–9.

Edland, A., & Svenson, O. (1993). Judgment and decision making under time pressure: Studies and findings. *Time pressure and stress in human judgment and decision making*, 27–40.

Edwards, M. B., & Gronlund, S. D. (1998). Task interruption and its effects on memory. *Memory*, *6*(6), 665–687. (133FQ Times Cited:8 Cited References Count:24)

Ericsson, K. A., & Kintsch, W. (1995). Long-Term Working-Memory. *Psychological Review*, *102*(2), 211–245.

Eyrolle, H., & Cellier, J. M. (2000). The effects of interruptions in work activity: Field and laboratory results. *Applied Ergonomics*, *31*(5), 537–543.

Field, A. P. (2009). *Discovering statistics using SPSS.* SAGE publications Ltd.

Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive? a study of length, similarity, and complexity. *Psychological Research*, *50*, 243–250.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, the Netherlands: Elsevier Science.

Hodgetts, H. M., & Jones, D. M. (2006a). Contextual cues aid recovery from interruption: The role of associative activation. *Journal of Experimental Psychology-Learning Memory and Cognition*, *32*(5), 1120–1132.

Hodgetts, H. M., & Jones, D. M. (2006b). Interruption of the tower of london task: Support for a goal-activation approach. *Journal of Experimental Psychology-General*, *135*(1), 103–115.

Hogarth, R. M. (1987). *Judgment and choice.* New York: Wiley.

Jedetski, J., Adelman, L., & Yeo, C. (2002). How web site decision technology affects consumers. *IEEE Internet Computing*(March-April), 72–79.

Kieras, D., & Polson, P. G. (1985). An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies*, *22*(4), 365–394.

Latorella, K. A. (1998). Effects of modality on interrupted flight deck performance: Implications for data link. In *Proceedings of the human factors and ergonomics society 38th annual meeting.* Santa Monica: Human Factors and Ergonomics Society.

LeGoullon, M. D. (2006). Spring ahead or fall back? exploring the nature of resumption errors. In *50th annual meeting of the human factors and ergonomics society* (Vol. 50, pp. 363–367). San Francisco, CA: Human Factors and Ergonomics Society.

Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2001). Cockpit interruptions and distractions: A line observation study. In R. Jensen (Ed.), *11th international symposium on aviation psychology.* Columbus, OH: Ohio State University.

Maule, A. J., Hockey, G. R. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta Psychologica*, *104*(3), 283–301.

McFarlane, D. C. (2002). Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human Computer Interaction*, *17*, 63–139.

McFarlane, D. C., & Latorella, K. A. (2002). The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction*, *17*(1), 1–61.

Monk, C. A. (2004). The effect of frequent versus infrequent interruptions on primary task resumption. In *48th annual meeting of the human factors and ergonomics society* (pp. 295–299). Santa Monica: Human Factors and Ergonomics Society.

Monk, C. A., Boehm-Davis, D. A., & Trafton, J. G. (2004). Recovering from interruptions: Implications for driver distraction research. *Human Factors*, *46*(4), 650–663.

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, *14*(4), 299–313.

NTSB. (1991). *Runway collision of USAir flight 1493, boeing 737 and SkyWest flight 5569 fairchild metroliner los angeles international airport* (Tech. Rep.). National Transportation Safety Board.

Oulasvirta, A., & Saariluoma, P. (2004). Long-term working memory and interrupting messages in human-computer interaction. *Behaviour & Information Technology*, *23*(1), 53–64.

Oulasvirta, A., & Saariluoma, P. (2006). Surviving task interruptions: Investigating the implications of long-term working memory theory. *International Journal of Human-Computer Studies*, *64*(10), 941–961.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive judgment maker.* New York: Cambridge University Press.

Pool, M. M., Koolstra, C. M., & Voort, T. H. A. (2003). The impact of background radio and television on high school students' homework performance. *The Journal of Communication*, *53*(1), 74–87.

Pool, M. M., Voort, T. H. A., Beentjes, J. W. J., & Koolstra, C. M. (2000). Background television as an inhibitor of performance on easy and difficult homework assignments. *Communication Research*, *27*(3), 293–326.

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2008). Predicting post completion errors using eye movements. In *Computer human interaction 2008 (CHI 2008).* Florence, Italy.

Ratwani, R. M., & Trafton, J. G. (2008). Spatial memory guides task resumption. *Visual Cognition*, *16*(8), 1001–1010.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 99–118.

Speier, C., Vessey, I., & Valacich, J. S. (2003). The effects of interruptions, task complexity, and information presentation on computer-supported decision-making performance. *Decision Sciences*, *34*(4), 771–797.

St. George, D. (2009). 6,473 texts a month, but at what cost? *Washington Post*, A01.

Trafton, J. G., Altmann, E. M., & Brock, D. P. (2005). Huh, what was i doing? how people use environmental cues after an interruption. In *Human factors and ergonomics society* (pp. 468–472). Orlando, FL.

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human Computer Studies*, *58*(5), 583–603.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177.

Ziljstra, F. R. H., Roe, R. A., Leonora, A. B., & Krediet, I. (1999). Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology*, *72*(2), 164–185.

Curriculum Vitae

David M. Cades was born in Abington, Pennsylvania on September 23, 1981. He received the degree of Bachelor of Science in Engineering Psychology from Tufts University in 2003. He received his Masters of Arts in Psychology in 2007 from George Mason University. He has worked as a human factors professional at Verizon Laboratories, Electronic Ink, Inc., the American Institutes for Research, and Lighthouse International. He is currently a scientist at Exponent Failure Analysis Associates.