Memory Aids to Improve Follow-Through on Intentions in Complex Task Environments

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Stephen Whitlow

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Caroline Hayes, Advisor

December 2015

Acknowledgements

First and foremost, I would like to thank Dr. Caroline Hayes for her guidance and wisdom as my adviser. Despite her becoming a department chair at Iowa State, she graciously offered to continue on as my adviser. I appreciate how generously engaged she was in this thesis, especially in light of the heavy workload of a new department chair. Her support, patience and humor have made this an enjoyable and enriching journey.

To my current and former committee members, I thank you for your guidance and feedback. Your multi-disciplinary perspectives informed and guided the execution of this thesis. A special thanks to Dr. Tom Stoffregen whose insights inspired the prospective memory aid design. Our discussions on Ecological Interface Design were enriching and enjoyable.

I would also like to thank my managers at Honeywell, Rose Mae Richardson and Olu Olofinboba, for your unwavering supporting of my goal. You not only helped me navigate the bureaucracy to secure tuition reimbursement, your genuine interest was much appreciated. I would also like to thank my current and former colleagues at Honeywell, Drs. Bill Rogers, Trish Ververs, Santosh Mathan, and Michael Dorneich, who encouraged me to take the leap and wrote reference letters on my behalf. To Trent Reusser, for his valuable software development consultations; they always helped me resolve a problem and never made me feel like a novice. You inspired me to be a better programmer and I always struggled to resolve some coding impasse on my own, so I was only bringing you the hard problems.

To my colleagues in the Human First lab, thanks for allowing me to borrow some lab space at a very convenient location on the East Bank campus, which certainly helped the attendance rate of recruited participants. And a special thanks to Janet Creaser, now Dr. Janet Creaser, who inspired me to pursue this goal and generously provided the benefit of her experience in becoming the first Ph.D. from the Human Factors and Ergonomic program.

And a 20-year belated thanks to my Masters graduate advisor at the University of Illinois at Urbana-Champaign, Dr. Neal Cohen, who inspired a lifelong interest in the study of human memory.

Dedication

I dedicate this thesis to my family whose support made it possible and who inspire me to be a better person. To my wife Lisa-- thanks for your unconditional, unwavering support of this goal. I would never have embarked on this journey if I did not know that I had your total support and blessing. To my children, Quinn and Auden, you inspire me to better myself and hope that you see me as a good role model. Thank you all for your patience and understanding when "dad was in the thesis room". I worked very hard to be present with all of you and do the lion's share of work after everyone was in bed. While I enjoyed the journey, I do understand that it came at a cost which you all bore graciously.

I would like to recognize my Mom and Dad, Miriam Whitlow and Dr. Roger Whitlow, who provided me excellent role models of lifelong learning and self-fulfillment. As a husband and parent, I fully appreciate the sacrifices you made to complete advanced degrees with small children. Dad--I still fondly remember our trips to St. Louis University as you finished your Ph. D-- stops at Stuckeys for chili dogs and trinkets, hanging at the library, and the modest hotel accommodations. Mom--I also vividly remember your Masters graduation and how proud I was of you and how happy I was to see you so excited and proud of yourself.

Abstract

The purpose of this research was to assess whether an adaptive prospective memory (PM) aid could benefit PM performance while minimizing costs such as interfering with primary tasks, user annoyance, and potential for complacency. This was investigated in a series of computer-based experiments that involved dynamic flight scenarios, multiple primary tasks, and 12 unique, embedded PM tasks. The cues to trigger PM tasks were presented in the simulated flight deck environment, such as "call Air Traffic Control at 10000 feet altitude". There were two independent variables (IV): multiple PM aid types were investigated across two primary task workload levels. PM task difficulty was fixed across IV levels such that there were a consistent number of “easy” and “hard” PM tasks across all conditions. Dependent variables included PM measures, such as PM performance and PM reaction time (RT), primary task measures, such as percent correct and reaction time, and subjective impression rating for PM aids and perceived workload. The potential benefits and costs of two different adaptive PM aids modes were investigated: one based on PM task difficulty and the other primary task workload. While PM aids supported a greater benefit for “hard” PM tasks performance compared to “easy” ones, the practical impact was modest and did not justify costs. In a follow-up experiment, there was both a statistical and substantive practical PM performance benefit found across primary task workload levels. Based on the benefits to PM performance and the actual and likely costs of each aid type, we concluded that an adaptive PM aid based on primary task load has the most advantageous cost/benefit ratio in a challenging real-world task environment.

Table of Contents

Contents

List of Tables

List of Figures

Introduction

Prospective Memory (PM) is remembering to do something at a later time, such as calling a colleague when you get to work the following day. There are two basic types of PM, *time-based* e.g. remembering to call someone in 45 minutes or at noon, and *event-based* e.g. picking up milk when you drive by a convenience store. PM tasks are usually "background" tasks (e.g., remembering to pick up milk) while concurrently carrying out one or more primary tasks (e.g., driving home, listening to news on radio, etc.). PM is a complex cognitive process that relies on multiple component processes such as encoding (committing PM task to memory), working memory (keeping PM task in working memory), persistent attention (monitoring external or internal cues that indicate the time and situation to retrieve and execute PM task), and to a lesser degree retrieval of retrospective memory (retaining basic details of PM task and when to execute it) (Einstein & McDaniel, 1990). Compromise of any of the component processes can result in the failure to retrieve and successfully execute the PM task at the right time.

PM is critical for safe and effective operations in dynamic task environments where operators are required to execute delayed intentions while concurrently performing multiple primary tasks. In flight operations, pilots are presented with new tasks that must be delayed due to ongoing higher priority tasks or until the situation is appropriate for action. For example, pilots need to remember to resume performing a pre-flight checklist if interrupted, to contact Air Traffic Control (ATC) when entering a new sector, and update the landing runway in the flight management computer during the heavy workload of final approach. PM is often compromised in such high workload, high tempo operational environments due to operator working memory limitations, inattention and distraction. Anecdotal and experimental evidence highlight how regularly PM fails—with varying degrees of impact depending on the situation and domain. For example, Nowinski et al. determined that of the 75 retrospective and prospective memory errors identified in an analysis of a random sample of 20% of commercial aviation incidents (n= 1299) reported to NASA's Aviation Safety Reporting System in 2001 (ASRS), 74 were prospective memory errors where pilots failed to execute a delayed intention (2003).

Everyone has had the experience of knowing that there is something that they intended to do but they cannot remember what that is—this is a PM failure. An example would be telling a co-worker that you will bring in a book for them to borrow the next day. When the opportunity arises, usually when leaving the house for work, this intention is often hard to remember the next

day in a different context and in the midst of the many competing tasks that must be done in the rush to get out the door. This is also seen in laboratory studies where changing contexts and high workload at retrieval worsens PM performance (Burgess & Shallice, 1997). We fail at other PM tasks, such as remembering to transfer clothes from the washer to the dryer. This situation is complicated since laundry rooms are typically out of the way so people cannot rely on walking by the laundry room to trigger remembering to transfer the clothes. Another interesting situation is trying to remember to NOT flush the toilet due to some plumbing issue. This intention frequently fails since this bathroom task is a highly scripted routine where each action cues the subsequent action, so before you know it you are flushing the toilet in spite of reminding yourself just minutes or seconds prior. This is a less serious example of an habit-capture failure which is common in aviation incidents (Nowinski et al., 2003).

Various memory aids are often used to help people reduce PM failures, such as post-it notes and location-aware mobile applications. These can help people to remember when to perform tasks, but this support comes with the following potential costs:

- Annoyance to users
- Interruption of or distraction from ongoing tasks
- Over-reliance on the aid
- PM skill degradation

For example, a software dialog box that "pops-up" and grabs attention can annoy and distract users. Such intrusive aiding not only interferes with users' workflow but can also impact their acceptance of such a system. Other examples of costs are overreliance and skill degradation, like forgetting a meeting when pop-up reminding is not enabled in a calendar application. People develop inherent strategies and cognitive process to support PM performance, henceforth referred to as native PM skills.

Over time users can also become over-reliant on an aiding system such that they don't bother to remember things without it. Systems should not discourage maintaining native PM skills since one can never guarantee the availability and infallibility of an aiding system. An aid that minimally disrupts ongoing tasks and does not compromise natural memory processes would be considered non-intrusive. The challenge is to design a memory aid that is informative enough to

support remembering but is also non-intrusive in that it neither interferes with ongoing tasks or natural remembering processes nor induces over-reliance.

There are two facets of PM aids that will be explored in this thesis:

1. The benefit of non-intrusive PM aids compared to no aids.
2. Adaptive aids that provide help under certain situations, such as for only difficult PM tasks or under high primary task workload

The results of this work will help developers of PM aids to understand what features will make their aids most effective while reducing negative impacts on users.

As is the case for most computer-mediated aiding/tutoring applications, there are persistent and challenging design trade-offs between short-term performance and long-term skill development and retention. We could design a highly intrusive aid that could almost guarantee perfect PM performance, but it would inevitably compromise both primary task performance and native PM skills retention. Likewise, we could design a very subtle, non-intrusive aid that preserves primary task performance and native PM process development, but does not adequately support PM performance. A possible design solution to this tradeoff is providing subtle aiding for PM tasks only under certain circumstances, or adaptive aiding. Two possible triggers for selection are "high" PM task difficulty, when factors such as long duration and subtle or no cues conspire against PM performance and "high" primary task workload when ongoing tasks recruit high levels of cognitive resources that cannot support PM performance. The argument in favor of this would be as follows:

- PM aiding should be provided when PM task difficulty is "high", since users will likely benefit more from aiding when task factors increase the likelihood of PM failure, as compared to "low" task difficulty where users are more likely to maintain high levels of PM performance without aiding.
- PM aiding should be provided when primary task workload is high, since they will likely benefit more from aiding since they should have sufficient cognitive resources under low levels of workload to support high level of PM performance without aiding.

Such adaptively guided provision of aids would lessen the overall impact of PM aiding, reducing primary task impact and fostering less over-reliance. In a series of experiments, the work

reported here explored this tradeoff culminating in an investigation of adaptive memory aiding with a primary task manipulation that simulates adaptive aiding. This research targeted aiding solutions with applicability to flight operations, since there are many documented aviation PM failures and current flight deck designs and procedures provide no explicit PM support.

To explore this trade-off, the research addressed a series of three research questions, with corresponding experiments. First, can a non-intrusive PM aid be effective in supporting PM performance, like an intrusive aid can, without interfering with primary task performance? This would determine if a non-intrusive aid, one that supports all PM tasks, supports improved performance compared to no-aiding. Next, we investigated possible adaptive aiding that could reduce some side-effects of consistent aiding. To investigate adaptive aiding, we addressed the second research question: Is there a difference in beneficial impact of PM aiding across levels of PM task difficulty? It is possible that "easy" PM tasks do not benefit enough to offset the known costs of always-ON or consistent aiding. If, as in prior work, the "hard" PM task benefits more from aiding, then PM task difficulty is a candidate dimension to trigger selecting aiding. Depending on the outcome of the second question, this could lead to a third and final question: Is there a differential benefit of aiding across primary task workload to support adaptive aiding only under high primary task workload? Would it be sufficient to provide aiding only when users' cognitive resources are most taxed by primary task performance? This would establish whether adaptive aiding could be an acceptable design trade-off to reduce impact on primary task and likelihood of inducing over-reliance in users.

General Approach

These questions were investigated in a series of computer-based experiments that involved dynamic flight scenarios, multiple primary tasks, and 12 unique, embedded PM tasks. Each PM task was introduced with instructions that included the action to be performed and the triggering cue for when to perform it, such as "Call Air Traffic Control at 10000 feet altitude". The potential benefits and costs of two different adaptive PM aids modes were investigated: one based on PM task difficulty and the other primary task workload. Accordingly, three primary research questions were addressed across experiments:

1. Can non-intrusive PM aids improve PM performance compared to no-aiding?
2. Does the performance benefit of PM aiding across PM task difficulty levels justify the costs?

3. Does the performance benefit of PM aiding across primary task workload levels justify the costs?

The proposed research will contribute to our understanding of how effective non-intrusive memory aids can be in supporting PM performance in complex task environments. This will help designers of PM aids in future complex safety-critical environments consider the trade-offs between memory aid effectiveness and their intrusiveness. It will also inform their understanding of over-reliance on aiding that could have long-term consequences on user's memory skills and operational performance. Considering the importance of PM in many safety critical domains, design knowledge from this research could improve operational safety by better supporting PM tasks while minimizing the negative impacts of aiding.

Literature Review

PM is a complex cognitive phenomenon that involves the encoding of some delayed intention, retaining the intention in memory, monitoring for a situation in which to perform action, retrieving and executing the action, and finally evaluating the outcome of the action. This phenomenon was described by Ellis as having the following phases (1996):

1. Encoding
2. Retention & Monitoring
3. Retrieval
4. Execution
5. Evaluation

The two basic types of PM are event-based and time-based. An event-based PM task would be stopping at a grocery store that you drive by on the way to work to pick up milk; a time-based task would be remembering to call your doctor at 10 am. McDaniel and colleagues further sub-divided event-based PM into immediate-execute tasks, to be done as soon as the situation arises, and delayed-execute tasks, to be done after some time interval after the situation arises (2004). Delayed execute tasks occur more frequently in real life since ongoing tasks usually prevent immediate execution. PM studies either use event-based cues, such as presentation of a target word, such as "cow" within verbal tasks, or time-based cues, such as a specific time interval (e.g.,

"in 15 minutes") or at a fixed time (e.g., "at noon"). These require different combinations of cognitive processes since event-based requires monitoring the environment for an external event, whereas the time-based memory requires actively maintaining the intention in working memory to support recognition of the time to execute.

In most PM experimental paradigms, participants are presented with a PM task which includes a cue to monitor for and an action to perform upon encountering the cue. The PM task is done concurrently with an ongoing task in which the cue is encountered. This ongoing task is referred to as a cover task, concurrent task or primary task. A real-world example of a PM task would be to remember to stop by the grocery to pick up milk while the ongoing task(s) would be driving home from work; the cue would be the grocery store sign. For simplicity, ongoing task(s) will be referred to as primary task(s).

How PM is Measured

As a complex cognitive phenomenon, various methodologies are required to investigate all aspects. It is difficult to perform experiments within realistic task environments due to the challenge of exerting control.

Research studies where the task environment resembles real-life operations are considered ecologically valid.

Accordingly, PM has been explored with self report methodologies that rely on participants' reporting PM failures in everyday life, simplified experimental paradigms involving simple primary tasks and artificial PM tasks with simple cues, and experimenter introduced longer-duration PM tasks.

Self Report Methodologies

To quantify PM performance in everyday life, researchers will ask participants to record PM failures in a log over an extended period of time. Eldridge et al. conducted a diary study to capture a memory problem corpus from researchers during their workday (1992). They found that 52 of 182 memory issues logged were PM errors and that these represent a variety of different problems that rely on a variety of contextual cues to trigger retrieval, such as one participant forgetting to call their doctor back. They concluded that awareness of a user's work context and their plans, such as a meeting with a given co-worker, could enable a technology support for PM.

PM is a phenomenon inherently situated in complex real situations, making it challenging to study in the lab. Czerwinski et al. asked information workers to document the impact of interruptions on PM (2004). Participants recorded all of the following details, an indication of the complexity of studying PM failures:

- Time of task start
- Difficulty switching to the task
- What documents were included in the task
- What was forgotten if anything
- Number of interruptions experienced
- Any additional comments

Participants reported a variety of strategies to support PM performance such as emailing themselves or creating a dynamic web page of reminders to complete some task. Other users suggested tools that would recreate computer-mediated work contexts, such as open documents and applications, which could facilitate PM recall.

Simplified Dual Task Experimental Paradigms

To date, most laboratory studies of PM use artificial primary tasks and simplified PM tasks compared to those found in real task environments. For example, Ellis et al. deployed two different primary tasks--prose reading (excerpts from Edgar Allen Poe) and semantic judgments (responding to series of statements such as "Doctors undergo a long training" that tap general knowledge) (1999). For each primary task, the prospective memory task was to respond when they encounter a target word, either "prefect" during the reading task or "ship" in the judgment task (1999).

These aforementioned paradigms are variants inspired by earlier work by Einstein and McDaniel in which participants press a key when they encounter a target word (PM cue) within primary tasks such as prose reading, short-term memory tasks, etc. (Einstein, Holland, McDaniel, & Guynn, 1992; Einstein & McDaniel, 1990; McDaniel & Einstein, 1993). For example, Guynn, McDaniel, & Einstein used a word fragment completion primary task in which participants wrote down the first word that came to mind based on the presented word fragment (Guynn et al.,

1998). The concurrent PM task was to circle any of 3 target words learned previously (e.g., school, unicorn, celery). Another variant required participants to delay their response until a task change following the presentation of a cue (screen turns red) while performing a series of cognitively demanding 1 minute tasks (Einstein et al., 2003).

Interruption Paradigm

Another means to experimentally manipulate PM is with a task interruption paradigm. When a primary task is interrupted, remembering to resume it following some distraction task is a PM task (Dodhia & Dismukes, 2009). The interruption paradigm supports multiple experimental manipulations, such as timing of interruption and interrupting task length and type, and multiple measures like resumption rate and resumption delay.

Ecologically-valid Experimental Measure

In addition to the dual-task paradigm, researchers often introduced a single ecologically valid PM task during the test session such as asking participants to call them one week later or remembering to ask the researcher to return a personal item at the end of session (Kvavilashvili & Fisher, 2007; Adda et al., 2008). While ecologically valid, this paradigm produces very sparse data and is insufficient for evaluations of PM aiding solutions.

Sellen et al. performed a naturalistic study in which participants used electronic badges to perform event and time-based PM tasks over a 2-week period (1997). The event-based task was to respond when they were in a particular room while the time-based task was to respond every 2 hours. Participants were also asked to indicate with the badge whenever they thought about the PM task.

Craik and Bialystok used a computer simulation to collect richer experimental data within an ecologically valid task, preparing breakfast for four people (2006). The prospective memory component of the task involved remembering when to start and stop the cooking of different foods at the right time to achieve the overall breakfast plan. Cook time varied between 1 and 4.5 minutes. This experiment also included a distracter task, setting the virtual table, to increase the cognitive demands on planning and prospective memory.

PM Impact on Primary Task Performance

In most cases, there are ongoing primary tasks during the PM retention period. Previous work has established costs to primary tasks from PM tasks. Smith (2003) found a cost to a low-level cognitive task, lexical decision, which is deciding whether a letter string is a word or not. The concurrent PM task was embedded with a lexical decision task such that participants were required to press a key when seeing one of 6 "critical" words that they were presented earlier. Reaction time measures to lexical decision have been shown to be sensitive to changes in cognitive resources. Participants had slower lexical decision times when performing concurrent PM tasks than on no PM task trials; this demonstrated that there was a primary task performance cost that was not associated with performing primary task actions. These results suggest that simply maintaining a PM task requires some cognitive resources.

While lexical decision was a very simple task, Loft and colleagues demonstrated a cost in a more complex task environment (Loft & Remington, 2010). They evaluated performance on an air traffic control (ATC) primary task, which was detecting and preventing conflicts of aircraft defined as violating 5 mile lateral and 1000 feet vertical separation. The PM task was to remember to press a different key for aircraft with altitude indicator flashing than the routine key. They found participants missed more conflicts, a lowering of primary task performance, when performing concurrent PM tasks; again this impact was not related to overlapping tasks actions, but it was a performance cost incurred from simply keeping the PM intention in memory.

PM Failures and Causes

Anecdotal and experimental evidence highlight how regularly PM fails across work domains and in everyday life—with varying degrees of impact depending on the situation and domain. Doctors make PM errors in the high stress, high tempo world of modern medicine. For example, surgeons often fail to remember to remove foreign objects such as surgical sponges and instruments from patients' body cavity. This was estimated to have occurred 1500 times in 2003 alone, despite the best intentions of competent surgical teams (Patient Safety Monitor Alert, 2003). Gawadne et al. found that surgical incidents with PM errors, operationalized as leaving foreign objects in the body, were 9x more likely to be emergency situations and 4x more likely to involve unexpected procedural changes, as compared to control incidents that were the same type of surgery

performed without incident (2003). These PM failures often result in post operative injury and illness including infection, perforated bowels, sepsis and even death. All researchers agreed that the actual rate of this type of memory failure is under-reported due to the sensitivity to medical malpractice lawsuits. Undoubtedly there are substantial costs incurred by lawsuits, second surgeries, deaths, and infections from retained objects.

PM is also critical for safe and effective operations in dynamic task environments where operators are required to execute delayed intentions along with immediate actions and long-term monitoring, such as modern flight operations; however, PM is often compromised in such high workload, high tempo operational environments due to operator memory limitations, habit capture, and distraction. Habit capture is a phenomenon where people are unable to execute a deviation to an over-trained complex task since each subtask cues the next subtask—leading them to forget the deviation. Nowinski et al. determined that of the 75 retrospective and prospective memory errors identified in an analysis of an aviation incident reporting system (ASRS), 74 were prospective memory errors where pilots failed to execute a delayed intention (2003). Retrospective memory refers to the remembering of content from past experience; a retrospective memory error would be failing to accurately recall something from the past. The low rate of retrospective memory errors was expected since retrospective memory has been traditionally well-supported by pilot training regimes and flight deck user interfaces like checklists.

Nowinski et al. determined the sources of PM errors included:

- Encoding Failure: Poor encoding of initial intention (14, 19%)
- Monitoring Failure: Failure to monitor for window of opportunity to execute delayed intention (19, 26%)
- Retrieval Failure: Lack of salient cues to help retrieve delayed intention (27, 36%)
- Execution Failure: Habit capture by ongoing, common tasks (execution failure) (14, 19%)

In interviews, corporate test pilots reported persistent PM problems in many different aircraft related to the cross-feed function (Whitlow, 2015). Most multi-engine aircraft include a fuel cross-feed system to balance fuel across left and right wing fuel tanks and accommodating engine failure. The PM challenge is that most cross-feed systems do not have an automated shutoff but

rely on pilots to remember that it is on and to turn it off once fuel is balanced; however, cross-feed indications are not very salient and the procedure can be lengthy such that pilots can get distracted and fail to shut-off the cross-feed, which has lead to an engine failure due to fuel exhaustion, flameout, on the side from which fuel was being cross-fed. Pilot reports were confirmed by reviewing ASRS reports in which a quick search revealed at least two incidents that were reported since 2009. In Report # 894615 from 2010, a pilot reported how distraction contributed to PM failure:

- "We then again started a series of deviations to avoid weather and I lost track of my cross feed situation. Good 40 or more minutes passed until we finally re-established a clear flight path on course. It was then that I discovered I had a serious imbalance of fuel to deal with. My right tank was near 2000 LBS and my left tank was at 8000 LBS."

In Report # 983453 from 2011, a pilot reported how the subtle cross feed indication contributed to imbalance:

- "Never seeing the cross feeds open, we closed the cross-feed and the imbalance stopped"

For the proposed research, I will focus on facilitating monitoring and retrieval cues in a complex task environment that resembles modern flight operations. All of the sources of PM failures are discussed below, as well as others such as Systems and Organization factors.

Encoding Failure

Complex sociotechnical systems, such as nursing care, impose a heavy PM load on the nurses who need to remember to order medications, communicate new information to attending physicians about certain patients, and order special meals---all of which are most likely to occur when they are in patient rooms and away from computer-mediated means to execute these intentions. Furthermore, nurses will frequently change work contexts that results in inconsistent opportunities for task support or stable contexts to enable successful retrieval of their delayed intentions (Fink et al., 2010). While doing rounds, nurses encounter many new tasks that distract them from adequately encoding these intentions.

Monitoring Failure

It is generally agreed that PM tasks and most primary tasks compete for monitoring resources (Guynn, 2003; Smith, 2003). When participants are required to delay PM task execution following cue presentation, as compared to immediate execution, their performance degrades due to the increased monitoring demands (McDaniel et al., 2004). Other research indicated that monitoring performance degrades over time. Kiddler et al. required participants to remember to make a response at either 1-minute or 2-minute intervals while performing the primary working memory task (1997). Despite being allowed to check elapsed time during the intervals, participant PM performance was worse with 2-minute intervals, suggesting difficulty maintaining the task in working memory over the longer interval.

Delayed execute tasks occur more frequently in real life since ongoing tasks require some delay prior to executing PM tasks. For example, pilots encounter frequent distractions and interruptions to highly-trained task flows and checklists that frequently occur while taxiing away from the gate—all competing with monitoring for the opportunity to realize the delayed intention. This was confirmed by incidents and accidents analyses that concluded that 23% of errors and 38% of safety threats happen prior to take-off (Helmriech et al., 2001). Of the 27 airline accidents from 1987 to 2001 where crew error was a causal factor, five involved a procedural omission which is a PM error (Dismukes, 2006). In an infamous 1987 example, an aircrew was interrupted in the middle of a checklist while taxiing out, leading them to forget to set the flaps to the take-off position. That error, along with the failure of a warning circuit, resulted in a crash that killed all but one onboard (Holbrook et al., 2005).

Retrieval Failure

In the absence of salient external retrieval cues, PM performance can degrade since it then relies on internal cognitive processes , i.e. self-generated cues, that are more easily compromised by fatigue, workload, and aging than are environmental cues (Craik, 1986). The availability of external cues, which effectively externalizes the memory requirement and reduces controlled processing requirement, is why event-based cues generally support higher PM performance (Sellen et al., 1997). This is also why event-based PM Is more robust in the face of task demands, age-related declines, and distraction. A previous example, forgetting to move laundered clothes to the dryer, illustrates the impact of retrieval cues. This PM task is "out of sight, out of mind" since the laundry room is frequently out of the way, so people are not reminded of the outstanding

task by walking by the laundry room. To address this common problem, most manufactures include an auditory indicator to alert people that the clothes are ready to be transferred.

Execution Failure

In addition to anecdotal reports of PM execution failures in domains such as aviation and medicine, controlled experiments have also investigated PM errors with doctors working on patient simulators. In one study, participants failed to execute 20% of "important" intentions, which would have impacted patient safety in a real world setting (Dieckman et al., 2006).

Studies in office domains have also revealed the prevalence and seriousness of PM issues. Researchers who volunteered as participants logged 182 total memory failures over a six week period, of which 53 were PM failures (Eldridge et al., 1992). Similar results were reported from the diary results from information workers (Czerwinski et al., 2004). A common situation is office workers forgetting to attach a document to an email before sending, even when that is often the primary point of the task. This is another example of people being distracted from executing their original intention, in this case a habit capture failure specifically. The normal progression of actions—adding email addresses, subject, salutation, body, closing, then hitting send—progressively cue the next action so that upon closing people tend to hit send without adding an attachment, since it is a deviation from the norm. Possibly to address this, email clients have recently added menu commands in productivity software to directly send the current file as an attachment, though this provides another means to send a file and does not fundamentally address the habit capture failure.

Systems and Organizational Factors

Operators of complex systems such as flight operations are generally provided with modern, advanced user interfaces that support elegant and effective execution of the four primary aviation tasks: aviate, navigate, communicate, and manage systems (Schutte & Trujillo, 1996); however, they do not support the executive processes which is the cognitive "overhead" in managing these tasks and associated prospective memory requirements within dynamic, complex operations. Observational studies have identified the substantial PM demands placed on flight crews (Loukopoulos et al., 2001); however, unlike retrospective memory, which is extensively and formally trained and is supported by paper and electronic checklists in flight operations, PM is afforded no such support. One challenge is that prospective memory tasks cannot be predicted a

priori given the dynamic nature of flight operations, so flight crews cannot be briefed on them and prepare ahead of time. Second, flight decks do not allow flight crew to modify the interfaces for dynamic task support, requiring flight crew to solely depend on internal cognitive processes for maintaining PM tasks. Not only does the flight deck provide no explicit PM support, the nature of their task environment conspires against successful PM performance by overloading, distracting, and imposing substantial monitoring and working memory burden on flight crews. A meta-analysis of the previously mentioned findings identified factors that compromise PM performance on the modern flight deck. These factors are categorized by causal categories below:

- Encoding Failure
  - High workload compromises both encoding and retrieval
- Monitoring Failure
  - Static flight deck designs do not have flexibility to support sufficient cueing
  - Active, conscious monitoring is costly in terms of mental workload—automatic cueing is much more efficient but it is susceptible to PM failure during disruption to normal task flow
- Retrieval Failure
  - External cues are more effective in supporting PM than internal cues—and there are often no salient external cues available
  - High workload compromises both encoding and retrieval
- Execution Failure
  - Pilots must rely on highly trained task flow to manage workload; however, if PM is deviation of habitual task flow it will likely be forgotten resulting in error of omission.
- Systems and Organization Factors
  - No training in PM strategies
  - Little or no explicit user interface support for task management
  - Cannot anticipate or prepare for PM tasks —dynamic, variable requirements
  - Over-reliance on happenstance retrieval
  - Pilots cannot dynamically create persistent visual cues to remind them

Approaches to Aiding

Prior work has explored PM aiding solutions that address the failure causes mentioned above. All PM aids incur some kind of cost which needs to be weighed against expected benefit of improved PM performance. The classes of PM aids and their associated trade-offs will be discussed below. Developing a further understanding of the likely benefits and costs of PM aids motivate the research questions of this work.

Facilitate Encoding

A common finding is that primary task demands during encoding impair PM performance (Einstein, et al., 1997; McGann, Ellis, & Milne, 2002). Since participants' attention is divided, they have fewer resources to encode a robust intention. This basic finding also holds when participants' attention is divided at retrieval as well. For example, if participants are required to monitor a series of voiced digits for a series of odd digits, PM performance is compromised (Einstein et al., 1998; McDaniel, Robinson-Riegler, &Einstein, 1998). To facilitate encoding of the current task to be resumed in an interruption paradigm, Dodhia & Dismukes introduced a long delay prior to an interrupting task so participants had the resources to adequately encode the intention to resume the current task (2005). While this aiding solution did enhance PM performance, most operational environments could not accommodate such delays; for example, when asked about ways to improve prospective memory, airline pilots responded that they would repeat air traffic control (ATC) instructions aloud to enhance encoding of the intention (Dodhia et al., 2001).

Facilitate Monitoring

For those PM tasks with long delays between encoding and retrieval, people need to monitor for the cue to execute the task; such monitoring is known to tap executive processing. Marsh and Hick investigated the impact of executive processing load on performance of a PM task, pressing a key when the presented word was a fruit (1998). During retention period between PM instructions and cue to respond, participants performed the Star Counting Test, a complex working memory task with known executive processing requirements; this task required participants to continuously increment or decrement a running count based on a complex visual

array of "stars" (asterisks), plus signs, and minus signs. Executive processing load was manipulated by having random increments and decrements in high load compared to a fixed number of two for low load. They found that PM performance was significantly lower under high executive processing load (.50) compared to low load (.75). Presumably the primary task uses the same pool of resources that supports monitoring for the window of opportunity to execute the PM intention.

In a variant of the interruption paradigm, McDaniel et al. assessed whether a simple visual reminder, a small blue dot, could facilitate monitoring for PM task cues over an interruption period (2004). Participants performed a series of primary tasks for 1 minute such as simple math problems, rating pleasantness of words, and judging which of two lines was longer. In addition, a persistent digit monitoring task was included to increase multi-tasking requirements. A series of single digits was presented auditorily every 2 seconds. Participants were required to press a button whenever two consecutive odd numbers were presented. At random intervals a red screen was presented to indicate a PM task which required them to press a key once the current primary task ended. For some of the PM trials, participants were presented with an interruption task after the current primary task ended. They found a 10 second interruption produced significantly lower PM (.60) compared to no-interruption (.80). In a follow-up experiment, they assessed the impact of a visual reminder (blue dot) of PM tasks, and found that this significantly improved PM performance across the interruption (.96) compared to no reminder (.79).

This aid had PM performance benefits even though it provided no information about the to-be-performed task or the time to perform it, only reminding participants that there was an outstanding delayed task to resume; moreover, it supported improved PM performance without impacting primary task performance. This external aid helped maintain the intention in working memory, thus facilitating monitoring despite the ongoing executive processing requirements.

Facilitate Retrieval

PM often fails due to the inadequacy of available cues to trigger retrieval. When external cues are neither salient nor related to PM tasks, PM performance is compromised especially under divided attention conditions (McDaniel & Einstein, 1993). External cues support PM that is robust to

many challenges since it reduces the controlled processing required to maintain the intention. Another common finding is that the distinctiveness of the PM cue has a substantial impact on performance. For example, when the cue word was common, participants remembered to perform PM tasks only 31% of the time compared to uncommon cues that supported 100% PM performance (McDaniel & Einstein, 1993).

While McDaniel et al. (2004) did find PM performance benefit from a cue, the small blue dot, that did not correspond to the primary task, other research found that cues that corresponded with the primary task facilitated PM retrieval effectiveness. Non focal cues, those that are neither specific nor related to the processing of the primary task, resulted in decreased PM performance when compared with focal cues that are specific and the processing of the primary tasks actually aids in detecting them. Non-focal cues are target syllables such as "tor" when the primary task was semantic word processing (Einstein et al., 2005); whereas cues like "deer" are considered focal cues when the primary task is animal categorization (Marsh & Meeks, 2007). Focal cues supported superior PM retrieval performance and lower reaction time than non-focal cues (Marsh & Meeks, 2007). Cues can also be overlooked if a new task occurs coincidently with the opportunity to perform some PM tasks. Dodhia and Dismukes found that a new task interferes with resuming an interrupted task (2009).

Atance and O'Neil proposed that "future thinking", or a detailed imagining of a future event, could facilitate successful PM performance (2001). Specifically, "future thinking" would support the selection of a more effective mnemonic for retrieving a delayed intention. By "pre-experiencing" events, one could assess effectiveness of mnemonics by determining how salient they will be in the future context and how likely one is to encounter them. This supposition was later confirmed by many researchers, including McDaniel, Howard, and Butler, who investigated the impact of asking participants to imagine themselves performing the PM task in the future (2008). They found imagining their implementation intention allowed participants to maintain the level of PM task performance even under high attentional demands, unlike the standard PM instructions.

Facilitate Execution

In everyday life, PM often fails since people generate general intentions that do not include a

specific mapping between the cue and the delayed intention (Gollwitzer, 1999). Gollwitzer and Sheeran concluded that people focus on the action to be performed at the expense of encoding and retrieving triggering event details (2006). PM can also fail because the triggering cues are not specific enough for people to recognize them. Researchers also determined that reminders that only included the intended action supported lower PM improvement compared to reminders that included details of target event and the intended action (Guynn et al., 1998). Similarly, Einstein & McDaniel concluded that only those reminders that strengthen the associative link between target cues and target action would improve PM performance (1996). For example, when driving to work, it may occur to you that you need to send an email to a colleague about a proposal. In general, people do not imagine the specific circumstances in which delayed intentions can be realized—such as when you are seated at your computer and reading email. Without this direct mapping, people need to continually remind themselves (e.g. rehearsal and active monitoring) which is easily compromised by distraction and delays.

Trade-offs

While memory aids can improve PM performance, they can incur short-term costs to primary task performance and/or longer-term costs by inducing user over-reliance which can discourage users' native PM skills. Memory aids could be very intrusive and "capture" attention-- an involuntarily re-direction from ongoing tasks to the memory aid (Yantis, 1993). This can interfere with how people direct their attention to primary tasks as well as to monitoring for cues to perform a PM task. Intrusive memory aids that employ attention capture could disrupt primary tasks by re-directing participants' attention from their performance.

Prior work has also had success with inserting artificial delays to facilitate encoding of, and then retrieval of PM tasks (Dodhia & Dismukes, 2005). While highly effective in reducing PM errors, this solution is not practical for many operational environments since it would disrupt task flow and introduce unacceptable delays in the tempo of operations. Disruptions to primary task could induce more errors and slower response times, referred to as primary task performance cost. Non-intrusive memory aids are subtle, peripheral cues that do not capture attention and should interfere less with primary tasks than intrusive aids.

In the short-term, memory aids that leverage attention-capture could be effective in supporting PM performance since they insure participants attend to the PM aid; however, there would likely

be short-term costs to primary task performance and long-term costs in terms of over-reliance (Bailey et al., 2001; Parasuraman et al., 1993). While non-intrusive aids are desirable due to minimizing their negative impact, **there is the question of how effective a non-intrusive memory aid can be in supporting PM performance.** The aiding solution should also minimize operator complacency and over-reliance that is common with aiding support systems. Consistent, reliable aids can lead participants to overly rely on them, essentially offloading or supplanting their own native memory processes. For example, Parasuraman et al. indicated that consistent automated monitoring support resulted in significantly more complacency than variable support where automation was not consistently reliable (1993). Skitka et al. also demonstrated that participants accustomed to an automated monitoring aid that was redundant with 100% reliable gauges, had lower detection rates of some abnormal event, 57% compared to 97%, than those participants in a non-aiding condition (1999). Smith, McCoy & Layton also demonstrated how automated support changed the cognitive processes and performance of dispatchers and pilots in a flight planning task (1997).

This pattern of automated aiding-induced complacency has also been shown with domain experts, and not just student participants (Mosier et al., 1998; Parasuraman, Molloy & Singh, 1993). In a follow-up study, Singh et al. found that subsequent monitoring performance was worse when participants became accustomed to constantly reliable automation (87.5% automation accuracy), compared to those participants that used inconsistently reliable automation (alternating blocks of 56.25% and 87.5% accuracy, respectively) (1997). Interestingly, in another study that compared pilots and non-pilots in an automation support experiment unrelated to aviation, pilots were significantly more apt to leave an automated monitor on, even after it had failed, than were non pilots (Riley, 1996).

This prior work indicates the proclivity of users to become over-reliant on automated monitoring if it is reliable and always present; however, in most operational settings, memory aids would either be fallible or not always available. Accordingly, any aiding solution should not interfere with or discourage users' natural memory processes. One possible solution to the challenge of supporting PM performance without inducing over-reliance and discouraging native memory processes is to adaptively aid only those PM tasks most in need of support. This notion is supported by prior work indicating that PM aids differentially improve performance under demanding conditions, but provide no benefit if the demands are reduced (Dodhia & Dismukes,

2009). They found that a reminder at encoding benefits PM performance when participants have only 2.5 seconds to retrieve the intention, but does not when they are given 10 seconds (2009).

Adaptive Aiding

To achieve a more satisfactory balance of trade-off than for aids that were always "on", adaptive aiding was considered. This approach is consistent with Rouse's (1988) conclusion that aiding/automation solutions should only be applied when human performance needs it (1988). Prior work has validated that aiding solutions realize the greatest benefits under high workload. Dorneich et al. found that a communication scheduling aid supported significantly higher primary task performance under high workload only but not under low workload (2006).

Prior work has also identified many factors that compromise PM performance and could be used to define situations for adaptive aiding. This approach was supported by McDaniel & Einstein who suggested researchers target those situations most likely to challenge PM with aiding (2007). The following factors were considered to trigger adaptive PM aids:

- Encoding Failure
    - High workload during encoding
    - High working memory load during encoding
    - Less specific mapping between retrieval cues and intended action
- Monitoring Failure
    - Longer delay between retrieval cue and time to execute
    - Participant has low working memory span
    - Longer duration of retention interval
    - High workload during retention period
    - Ongoing tasks that interfere with rehearsal of PM task
- Retrieval/Execution Failure
    - PM task involves a deviation from a habitual, serial task
    - New task introduced during window of opportunity
    - Lack of salient retrieval cues
    - Less specific mapping between retrieval cues and intended action
    - Little or poor alignment of PM cues with primary tasks
    - Relative frequency of retrieval cue (infrequent cues facilitate PM retrieval)

Based on analysis of prior work, PM task difficulty was selected as trigger for initial exploration of adaptive aiding. The following dimensions were selected to manipulate PM task difficulty ("easy", "hard"):

- Duration of retention interval
- Salience of PM retrieval cues
- Alignment of PM cues with primary task

First, this requires an evaluation of the differential impact of PM aiding across two levels of PM task difficulty: **will "hard" PM task benefit more from aiding than "easy" PM tasks?** Initial experiments also served to validate the definition of "easy" and "hard" PM task within these experimental task environments. A subsequent experiment investigated the differential benefit of PM aiding across levels of primary task load. Consideration of the performance benefit and the observed costs and likely costs would determine the mode of adaptation, either PM task difficulty or primary task load, based on which has the most advantageous cost/benefit ratio.

Experimental Testbed

The research questions were evaluated in three controlled experiments in which participants performed 2-3 primary and 12 PM tasks concurrently. All tasks in all experiments were performed within a dynamic multi-window, multi-tasking experimental environment that is similar to that of a modern aircraft. This experimental testbed created different primary task workload and PM aiding conditions by changing user interface elements and task parameters, to be described subsequently. The primary tasks were variants of common flight deck tasks that are simplified so as not to require particular expertise to successfully execute. While performing the primary tasks to be described below, participants also concurrently performed a sequence of 12 PM tasks that required them to remember to do something at a later time when the situation arises, as described in the PM Tasks section below. As in many operational environments, participants performed multiple primary tasks while monitoring for the opportunity to complete the PM tasks.

Scenarios

For all three experiments, six scenarios were generated with Python scripts to simulate flight performance data and provide all information to populate an Experiment Task Interface, including:

- Time of Day
- Aircraft Speed
- Aircraft Altitude
- Aircraft Location
- Schedule speed
- Progress in terms of distance and time to next waypoint
- Time/distance since departure airport
- Time/distance to arrival airport

A C# Windows application was developed to depict the graphical and information elements of these scenarios. A text-to-speech capability was implemented within the application to provide primary task feedback and present PM tasks to participants. The C# application stepped through scenario data at regular intervals to update the flight and information elements to simulate the task environment. The aircraft location was depicted as a blue circle that moved along a flight path over a green "Map" display, as depicted in Figure 1.

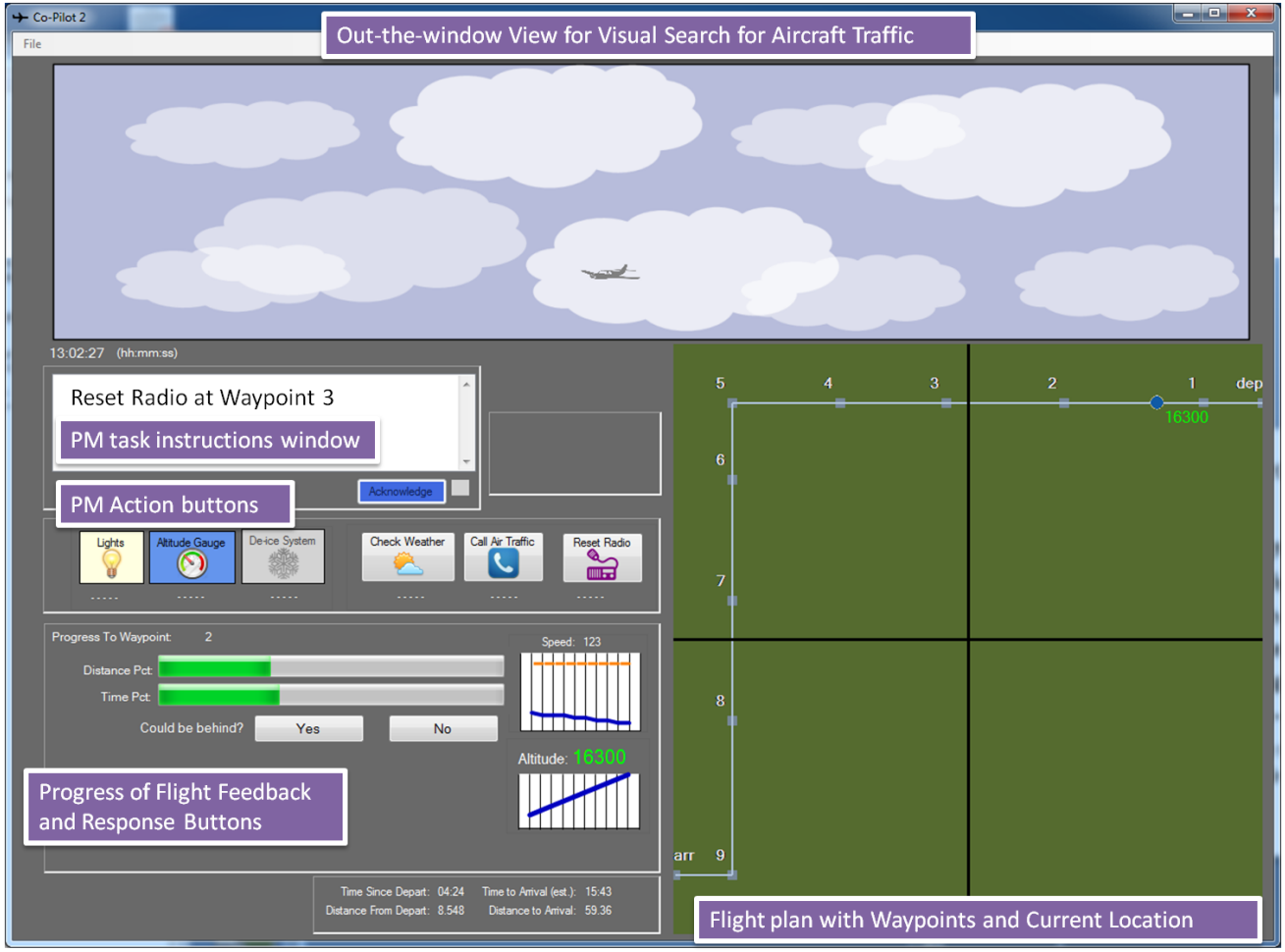


Figure 1: Experimental Task Interface

The update interval varied based on the speed of the aircraft but averaged to be approximately once per second. A sample of scenario data can be seen in Appendix B. All six scenarios included 1078 samples of simulation data which resulted in scenarios averaging just less than 18 minutes; all scenarios started at 1000 feet altitude and 2:00 into the flight and 13:00 time of day.

Scenarios were designed to be equivalent in the following dimensions:

- Length: all scenarios were between 17:56 and 18:00 in duration
- Speed variations were counterbalanced across 6 scenarios such that the deviation to waypoints had equal occurrences of being behind, on time, and ahead for each scenario

Flight scenarios varied along the following dimensions to minimize possible learning effects across conditions:

- General direction of flight on 2D map:
    - 2 scenarios went bottom to top
    - 2 scenarios went left to right
    - 2 scenarios went right to left
- Waypoints: all scenarios had flights that progressed through 10 waypoints
    - Waypoint names varied across 6 scenarios
- Altitude profile variation was balanced across 6 scenarios such that absolute vertical change was equivalent across entire flight but varied across 7 segments
- Flight schedule deviation (behind, on time, ahead) order was counterbalanced across 6 scenarios

Primary Tasks

Participants assumed the role of a co-pilot on very short flights that have just taken off. All flight simulation scenarios started two minutes after takeoff, two miles from the departure airport, at 1000 feet altitude, and at 13:00 time of day. Of the two primary tasks, Visual Search and Progress Assessment, only Progress Assessment required participants to consider simulation data such as aircraft location relative to waypoints and schedule status. The Visual Search task presented aircraft targets on a random schedule that was independent of simulation data. The primary tasks were introduced to tax participant cognitive resources so they had fewer resources

to perform PM tasks, thus increasing likelihood of PM task errors and supporting an evaluation of impact of PM aids.

Visual Search

First, participants continuously searched a simulated out-the-window view for other aircraft, as depicted in Figure 2. This visual search task required participants to monitor their "out-the-window" view for an aircraft. Once detected, participants moved a mouse cursor over the aircraft and clicked on it. Aircraft appeared at a randomized interval, location, and duration; random intervals, durations, and locations were generated from a range of values.

Figure 2: Visual Search Task

Event frequency and duration were used to create low and high workload levels, as depicted in Table 1.

	Frequency (per minute)	Average Duration
Low Workload	2.5	3.5 seconds
High Workload	4	2.75 seconds

Table 1: Visual Search Workload Levels

The aircraft disappeared after participants clicked on them. If participants did not click on an aircraft before it disappeared, they heard "traffic missed" in an American female voice as a reminder to perform the visual search task.

Progress Assessment

For the second concurrent primary task, participants were asked to monitor the flight's progress to the next waypoint; waypoints are a navigation reference point along a flight plan between departure and arrival airports. Participants were instructed to assess schedule and respond whether they "Could be behind?" by pressing either the "Yes" or "No" button. All flights included nine (9) waypoints with their corresponding segments. Each flight had a timetable for each segment and participants assessed progress by comparing the percentage of actual distance traveled to the percentage of the estimate time to the next waypoint. Participants also needed to consider whether the current speed was lower or higher than scheduled speed; scheduled speed is the speed required to maintain schedule for that segment. Participants were trained on the response heuristics and were given sufficient time during training to become proficient. Response heuristics are depicted in Table 2.

Could be behind?	Current Speed **Lower** than Schedule Speed	Current Speed **Higher** than Schedule Speed
Distance > Time	No	No
Distance < Time	Yes	No

Table 2: Progress Assessment Response Heuristics

Participants continuously assessed the flight's progress and responded based on the aforementioned heuristics. The simulated flight speed varied within each flight segment, and schedule speed was varied across segments to require that participants regularly attended this task. Furthermore, participants were instructed to regularly review and update their assessment. The task software reinforced this by resetting their response and issuing a voiced reminder to "track progress" in an American female voice. The resetting/reminder interval varied across workload condition as follows:

- Low workload—30 seconds
- High workload—15 seconds

Progress and speed data were depicted in different formats across workload conditions based on likely information processing demand differences. For the "low" workload condition, progress data and speeds were depicted graphically, as seen in Figure 3. Current speed and schedule speed trends were depicted as blue line and orange line respectively. The rightmost point on line was current value, and remainder consisted of previous nine values going back in time.

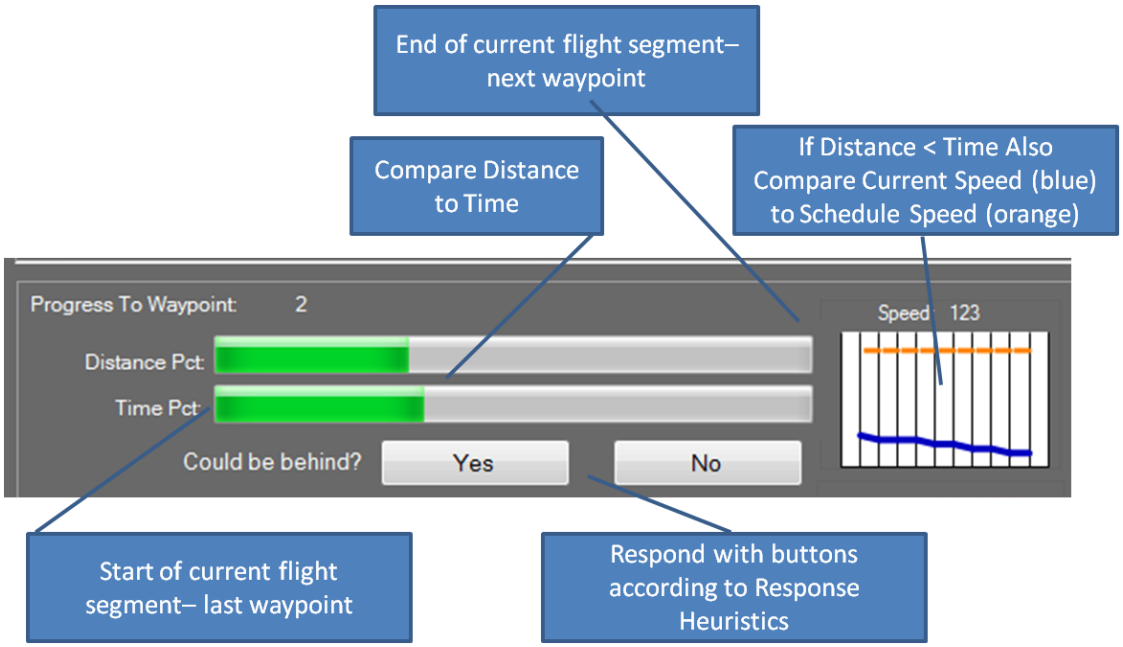


Figure 3: Progress Assessment Low Workload Task Interface

For the "high" workload, progress data and speeds were displayed as numeric values as depicted in Figure 4. The assumption underlying the two versions of the task display is that it requires more attentional and cognitive effort to visually parse and compare numeric values than processing graphical elements. All other task parameters were identical to the "low" workload version.

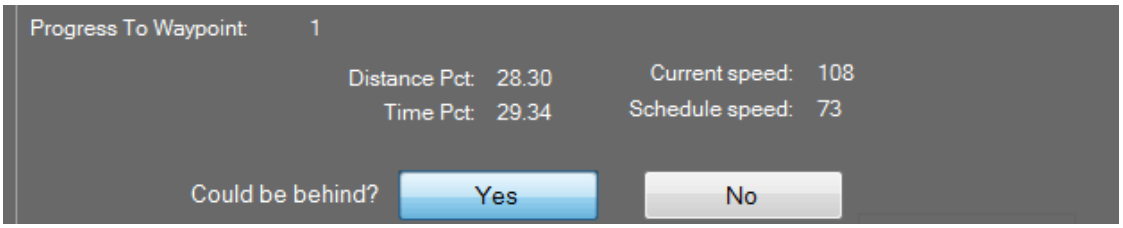


Figure 4: Progress Assessment High Workload Task Interface

Table 3 summarizes the workload variations for the primary tasks.

Task	Dimension	Low Mental Workload	High Mental Workload
Traffic Search	Duration	Longer (3.5 seconds)	Shorter (2.75 seconds)
	Targets Per Minute	Fewer (2. 5 per minute)	More (3.5 per minute)
Progress Assessment	Information processing requirements	Graphical	Numeric
	Reminder/Reset	Less frequent	More frequent

Table 3: Primary Task Workload Levels

Participants were required to balance the competing demands of these tasks that varied over time. The cognitive overhead of multi-tasking environments is known as task management, and is critical across many dynamic operational domains. Table 4 summarizes the experimental primary tasks and how they mapped to corresponding flight deck tasks and cognitive processes.

Experimental Primary Tasks	Analogous Flight deck Task	Cognitive Processes
Visual Search: looking for traffic targets in simulated out-the-window (OTW) view	Searching OTW for other airborne traffic	Visual Search Visual Attention
Progress Assessment: assess progress of simulated flight	Assessing progress of flight schedule to next waypoint	Continuous Monitoring Focused Attention Mental Calculation
Task Management: balancing demands of multiple concurrent tasks including PM tasks	Task Management: balancing demands of multiple concurrent tasks including PM tasks	Executive Control Working Memory

Table 4: Primary Task Details

Prospective Memory Task

In addition to two primary tasks, participants were required to perform twelve (12) prospective memory (PM) tasks throughout each flight. Each PM task was introduced by a simulated voice

message from the "pilot", in a British female voice, that asked the participant to do something in the future under specified conditions. For example, participants could hear "Reset Radio at Waypoint 3" or "Check Weather when entering Next Sector". Each new PM task message was also displayed in a dedicated "PM task instruction" window for eight seconds to provide participants further opportunity to encode the task, as seen in Figure 5. Each PM task instruction from the pilot consisted of a triggering condition, like "Waypoint 3", and an action, like "Reset Radio".

Triggering Conditions

All conditions that triggered PM task actions were displayed at a variety of locations on the task interface. They were:

- Waypoint
- Next Sector
- Altitude
- Time (of day)
- Elapsed time (e.g. 3 minutes)
- Flight time since departing
- Flight distance from departure airport
- Flight time to arrival airport
- Flight distance to arrival airport

Actions

The delayed actions for all PM tasks were executed by clicking one of a row of buttons in the task interface and included:

- Turn Lights On/Off
- Reset Altitude Gauge
- Turn De-Ice System Off
- Check Weather
- Call Air Traffic (ATC)
- Reset Radio

The triggering conditions and action buttons are labeled and highlighted in Figure 5 below.

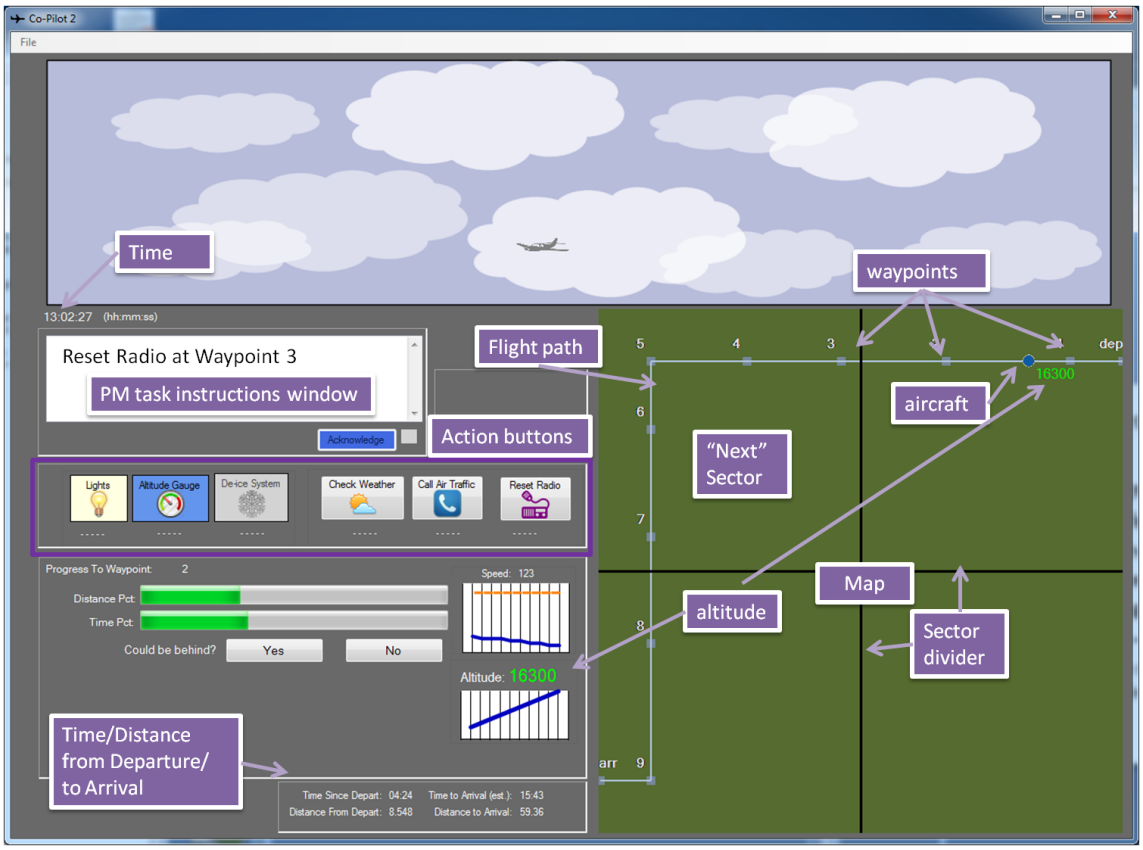


Figure 5: PM Task Details

For PM tasks, participants monitored conditions to determine when to perform the action, such as when the flight enters a new sector as indicated on the map display. At this point participants were required to retrieve the appropriate action to perform, such as "Check Weather", which involved clicking on the "Check Weather" button.

PM tasks were designed to be "easy" or "hard" based on the following dimensions identified in prior PM work: duration, salience of cues, and alignment with primary tasks, as depicted in Table 5:

Dimension	Easy	Hard
Duration	~ 70 seconds	~ 155 seconds
Salience of Cues	High: Altitude indication and Sector line—center of display, large font and line width	Low: Time/Distance Measures—bottom of display in small font; time interval—time of day indication only
Alignment with Primary Task (progress assessment)	High: Waypoints and Altitude indication were aligned with progress assessment tasks	Low: Absolute Time/Distance measures used as cues were not associated with progress assessment

Table 5: PM Difficulty Dimensions

Each scenario had an equal number of "easy" and "hard" PM tasks to assess PM performance. This was a fixed factor across all conditions, unlike primary task workload which varied across conditions. Across scenarios, PM tasks were equated for PM task total duration, overlapping tasks, schedule, type and difficulty as follows:

- Number of difficulty and easy PM tasks was equal within and across scenarios
  - 6 "easy", 6 "hard"
- There were never more than two overlapping PM tasks at a given time
- Duration (seconds)
  - "Easy" tasks average time: 67.8 (min) – 74.5 (max)
  - "Hard" tasks average times: 151.7 (min) – 161 (max)
  - Average all PM tasks: 113 (min)- 115.08 (max)
  - Total time of all PM tasks : 1356 (min) – 1383 (max)
- PM cue types—same number of each type across scenarios
  - 3 Waypoints
  - 2 Time/Distance (from departure or to arrival)
  - 2 Altitude
  - 1 Sector
  - 3 time interval
  - 1 Time of Day

PM tasks for one scenario are depicted in Table 6.

PM Item	Difficulty
Call ATC at 17500 feet Altitude	Easy
De-Ice is ON. Turn OFF in 3 minutes	Hard
Reset Radio at Waypoint U	Easy
De-Ice is ON. Turn OFF in 3 minutes	Hard
Turn Lights OFF at Next Sector	Easy
Reset Altitude Gauge at 29500 feet Altitude	Easy
Reset Radio at Waypoint X	Easy
Check Weather 13 minutes after Departure	Hard
De-Ice is ON. Turn OFF in 3 minutes	Hard
Turn Lights ON at Waypoint Z	Easy
Call ATC 1 minute before Arrival	Hard
Check Weather 1.5 miles from Arrival	Hard

Table 6: PM Tasks

The schedule for these PM tasks is depicted in Figure 6 with elapsed time in minutes indicated in top row. Each row corresponds to a separate PM task. The blue rectangles depict the duration of the PM tasks between when participants received PM instructions from pilot (leftmost edge) and when the triggering situation occurred (rightmost edge).

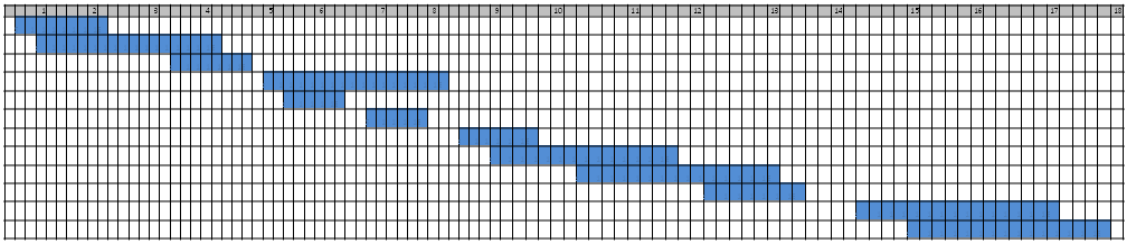


Figure 6: PM Task Schedule for One Scenario

Flight scenarios were also balanced across scenarios for PM tasks for the dimensions:

- PM cue: sector line
  - Half of scenarios had PM cues that were vertical sector lines
  - Half of scenarios had PM cues that were horizontal sector lines
- Point in scenario when trigger sector line encountered:
  - 2 scenarios had sector line cues encountered in first third of scenario
  - 2 scenarios had sector line cues encountered in middle third of scenario
  - 2 scenarios had sector line cues encountered in last third of scenario

After each trial, a dialog box (see Figure 7) presented participant performance feedback on primary and PM tasks in terms of percent correct. Primary and PM task performance was logged by the C# application to support analyses to answer the research questions.

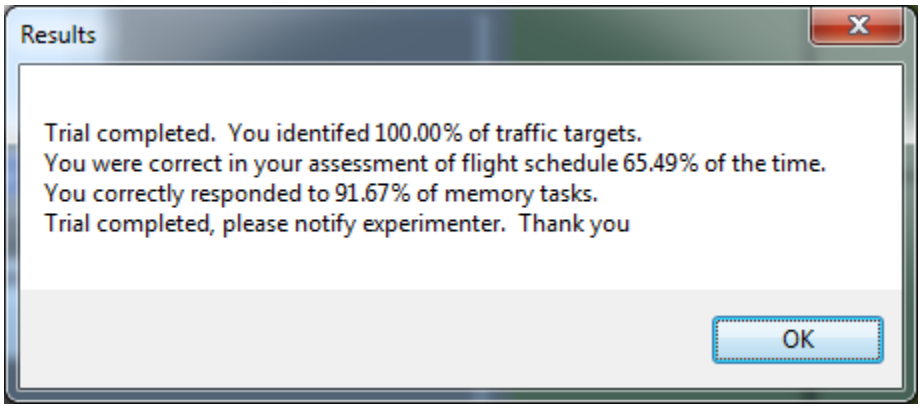
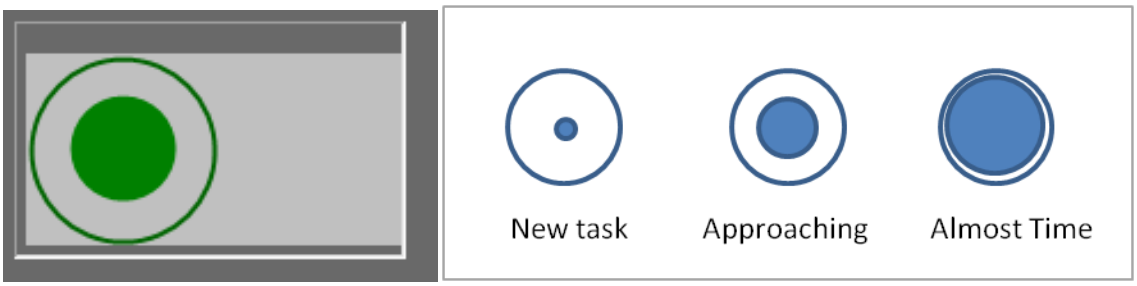


Figure 7: Participant Performance Feedback Dialog Box

Memory Aids

Participants were provided three (3) different levels of memory aiding (no-aiding, non-intrusive aiding, and intrusive aiding) to help them remember PM tasks. In the no-aiding condition which served as a baseline, participants received no software-based PM aid and had to rely on their own native PM skills. In the non-intrusive aiding condition, participants were provided with a peripheral graphical cue that there is an outstanding PM task.

The non-intrusive aid leveraged Ecological Interface Design (EID) principles to design a graphical form that facilitates processing. EID was selected because it can convey dynamic relationships with minimal cognitive processing required, thus minimizing attention capture (Burns & Hajdukiewicz, 2004). During discussion of the proposed aids, it was pointed out that most of the PM tasks are related to the continuously changing relationships over time (T. Stoffregen, personal communication, May 2, 2012). For example, if the PM task was to "Reset Radio when the aircraft reaches waypoint 3", the PM aid can estimate, based on current speed, how long it would take to reach ABC, the time to perform the PM task. This changing relationship was depicted as a small graphical form, a circle that progressively filled up from a dot in the center to the outline to reflect how much time is remaining until expected executions; this graphical timer is depicted in Figure 8.



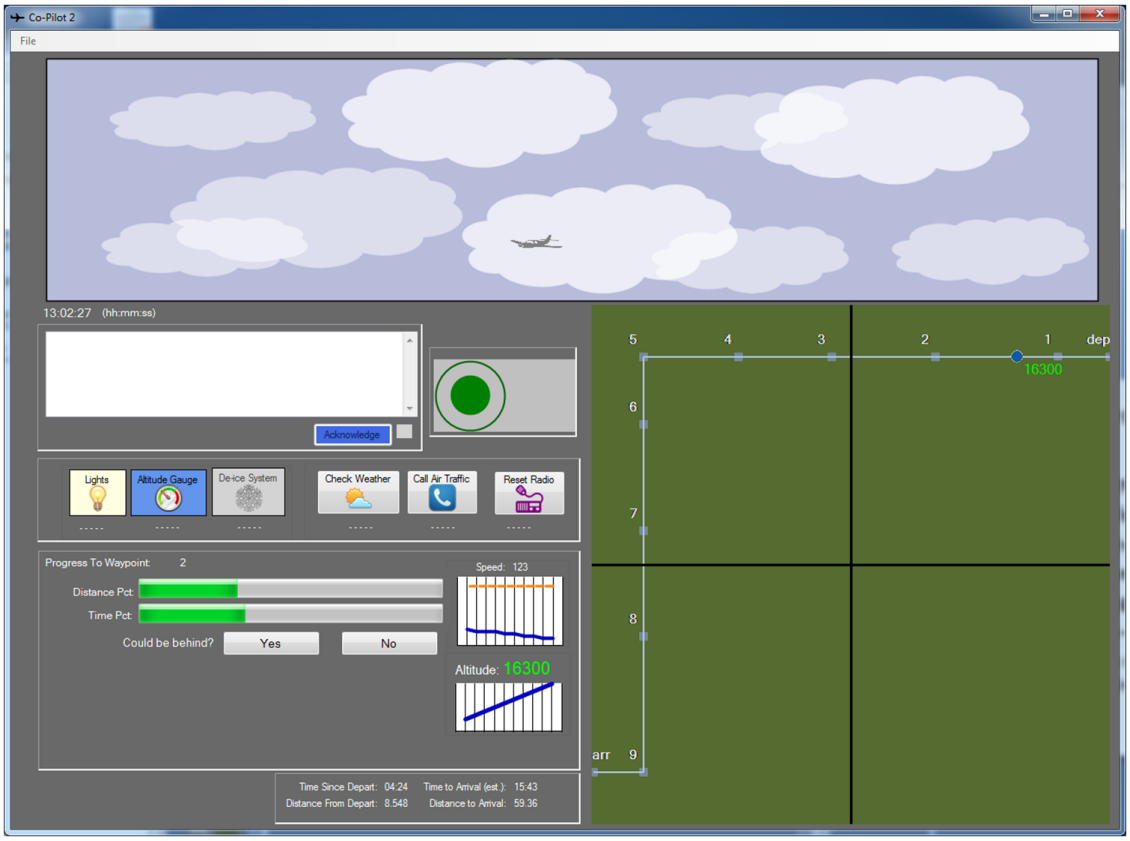


Figure 8: Non-intrusive PM Aid

There could be up to two non-intrusive aids visible at one time supporting two concurrent PM tasks. This aid was placed in the periphery of the experimental task interface, as seen above, when there was an active PM task(s). After the PM task is due, the graphical cue is removed.

For the intrusive aiding condition, participants received periodic pop-up dialog boxes that reminded them of the number of outstanding PM tasks (0, 1 or 2), as depicted below. The intrusive aid was presented on a pseudo-random schedule that distributed ten (10) presentations

across the trial. The proposed intrusive aid was a modal dialog that "pops-up" and grabbed the participant's attention to remind them of an active PM task, as seen in Figure 9.

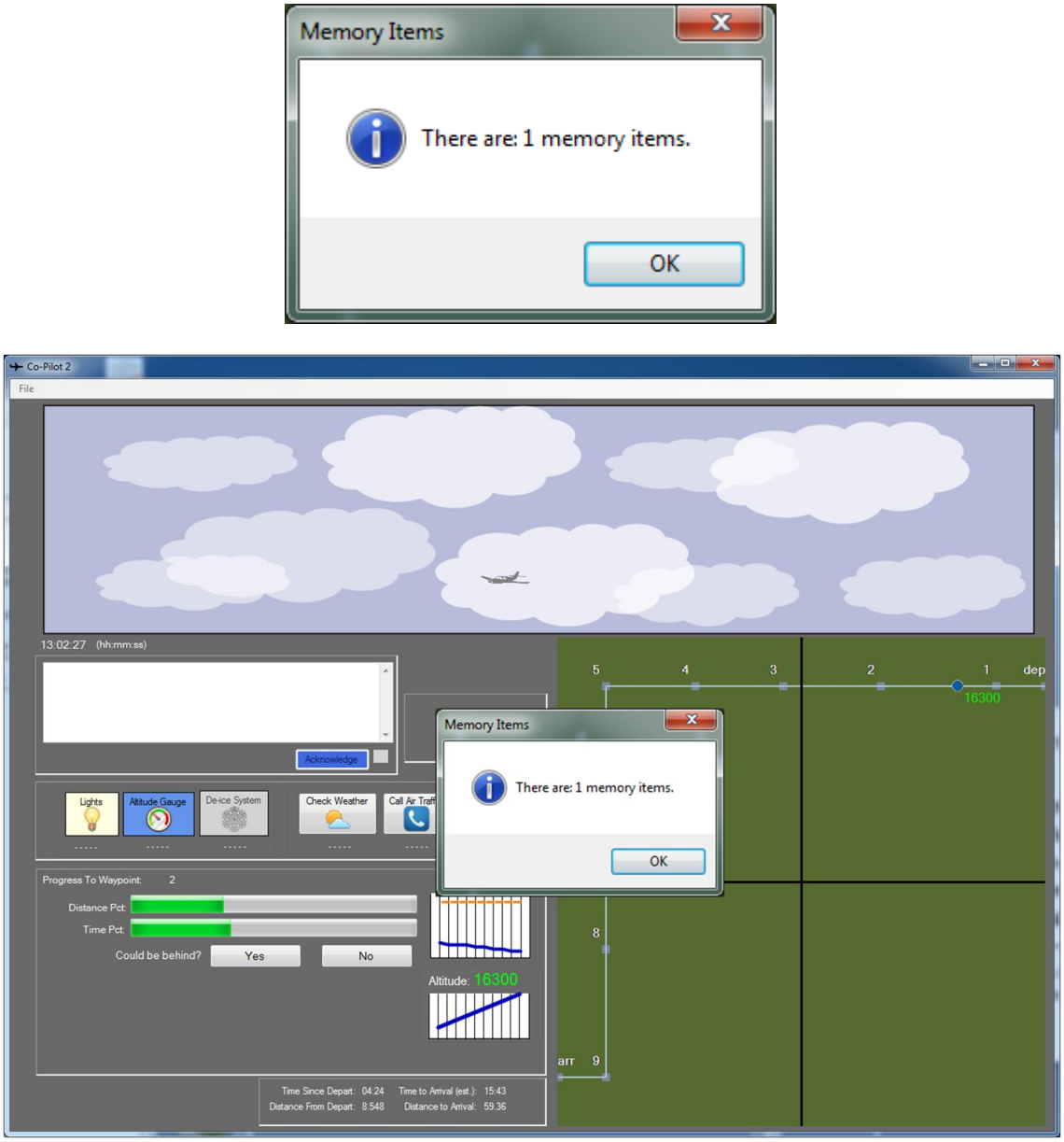


Figure 9: Intrusive PM Aid

Modal dialog boxes "grab" the focus of a user interface which means they are displayed on top of all active Windows and require user response to dismiss them. This is similar to common calendar software such as Microsoft Outlook which "pops-up" appointment reminders. This reminding paradigm was similar to that used by Guynn et al. (1998).

General Approach

The potential benefits and costs of two different adaptive PM aids modes, PM task difficulty and primary task load, were investigated in a series of computer-based experiments that involved dynamic flight scenarios, multiple primary tasks, and 12 unique, embedded PM tasks. There were two independent variables (IV): multiple PM aid types were investigated across two primary task levels. PM task difficulty was fixed across IV levels such that there were a fixed number of "easy" and "hard" PM tasks. Dependent variables included PM measures, such as PM performance and PM reaction time (RT), primary task measures, such as percent correct and reaction time, subjective impression rating for PM aids and subjective workload.

Experimental Design and Research Questions

The research program investigated the PM performance benefits, the potential costs to primary task performance, and subjective impression of different memory aid designs across a series of controlled, repeated-measures experiments. In addition to the aforementioned three primary research questions, there were also four secondary questions that are summarized below.

Primary Research Questions

1. Can non-intrusive PM aids improve PM performance compared to no-aiding?
2. Does the performance benefit of PM aiding across PM task difficulty levels justify the costs?
3. Does the performance benefit of PM aiding across primary task workload levels justify the costs?

Secondary Research Questions

4. Did the operational definition of "easy" and "hard" PM tasks, based on prior work, induce different PM performance in a complex task environment?
5. Can non-intrusive aiding support equivalent PM performance to intrusive aiding?
6. Did participants experience a difference in subjective intrusiveness between intrusive and non-intrusive PM aids?
7. What would the impact of aiding be on primary task performance?

Experiment 1A investigated different PM aids across primary task loading conditions to address Research Questions: 1, 4, 5, & 6. Experiment 1B investigated different PM aids under high primary task load to address Research Questions: 1, 2, 4, 5, &7. Finally, Experiment 2

investigated PM aids across primary task loading conditions to investigate Research Questions: 3&7.

Institutional Review Board Approval

A social/behavioral study was submitted to and approved by the University of Minnesota Institution Review Board (IRB) on August 5, 2014; the study was assigned IRB code number 1407P52522 and entitled "Memory Aids to Improve Follow-Through on Intentions in Complex Task Environments". A Change in Protocol Request was submitted and approved on December 19, 2014. The change request was submitted to permit recruitment via ResearchMatch (https://www.researchmatch.org/about/) and email. ResearchMatch is an online recruitment tool that connects researchers with volunteers. The University of Minnesota is a participating academic institution in this national volunteer registry supported by the U.S. National Institutes of Health as part of the Clinical Translational Science Award (CTSA) program. A Continuing Review was submitted on June 3, 2015 that detailed participant enrollment at 7 male and 11 female for a total of 18 participants. There was one participant withdrawal. This participant withdrew during training explaining that he found the multi-tasking to be challenging and somewhat stressful. Data from one participant were excluded due to experimenter error. There were no adverse events.

Participant Inclusion and Exclusion Criteria

The recruitment details and criteria described here apply for all three experiments. Participants were recruited from the University of Minnesota community. Participants self-selected based on two inclusion criteria. The first inclusion criterion was that they were between the ages of 18 and 65. Participants over 65 years of age were excluded since age-related deficits in PM have been shown to be particularly sensitive to task difficulty and could confound the results (Einstein et al., 1992). Second, participants self-selected if they had not been diagnosed nor are being treated for any neuropsychological disorder, such as attention deficit hyperactivity disorder (ADHD), or neuropsychiatric disorder, such as depression and schizophrenia. Participants with these classes of disorders were excluded since they often involve various cognitive deficits that could confound the results. For example, prior work has identified PM deficits in persons with ADHD (Altgassen et al., 2014).

Experiment 1A

Research Questions

To review, the following Research Questions and associated hypotheses were addressed in Experiment 1A:

- Primary Question
    - 1. Can non-intrusive PM aids improve PM performance compared to no-aiding?
        - Hypothesis: Non-intrusive aiding conditions will support superior PM performance compared to the no-aiding.
- Secondary Questions
    - 4. Did the operational definition of "easy" and "hard" PM tasks, based on prior work, induce different PM performance in a complex task environment?
        - Hypothesis: Participants will commit more PM errors on "hard" PM tasks than "easy" PM tasks.
    - 5. Can non-intrusive aiding support equivalent PM performance to intrusive aiding?
        - Hypothesis: Given its support of monitoring for PM task triggering condition, non-intrusive aiding will support an equivalent PM performance to intrusive aiding.
    - 6. Did participants experience a difference in subjective intrusiveness between intrusive and non-intrusive PM aids?
        - Hypothesis: Participants will rate non-intrusive aids as less negatively impactful than intrusive aids.

Experimental design was a 3 (aiding) x 2 (workload) repeated measures design (within subjects). The levels of the two independent variables were as follows.

Independent Variables

1. Aiding:
    a. No aid (Figure 5) baseline display
    b. Non-intrusive aid (Figure 8) baseline display with graphical timer
    c. Intrusive aid (Figure 9) baseline display with pop-up dialog reminders
2. Primary Task Workload
    a. Low (low workload variants of Visual Search and Progress Assessment (see Table 3))

b. High (high workload variants of Visual Search and Progress Assessment (see Table 3))

The experimental conditions are outlined in Table 7.

Workload/Aid	None (no aiding)	Non-intrusive	Intrusive
Low	Low Workload No Aid	Low Workload Non-intrusive Aid	Low Workload Intrusive Aid
High	High Workload No Aid	High Workload Non-intrusive Aid	High Workload Intrusive Aid

Table 7: Experiment 1A Design

With a 3x2 design (3 levels of aiding, 2 levels of workload =6 conditions) the number of required participants in a complete counterbalance would be a multiple of 720 (6x5x4x3x2x1). A practical alternative is using an incomplete counterbalance design such as a balanced Latin Square design that ensures that each condition follows all other conditions once- thus lowering the risk of carryover effects. See table below for an example of a Latin Square for a 6 condition experiment requiring multiples of 6 participants. The conditions are coded as follows in Table 8: aiding—intrusive (Int), non-intrusive (Non), no-aiding (No); workload—low (L) and high (H);

	Order					
Group	1	2	3	4	5	6
A	Int-L	Int-H	No-H	Non-L	No-L	Non-H
B	Int-H	Non-L	Int-L	Non-H	No-H	No-L
C	Non-L	Non-H	Int-H	No-L	Int-L	No-H
D	Non-H	No-L	Non-L	No-H	Int-H	Int-L
E	No-L	No-H	Non-H	Int-L	Non-L	Int-H
F	No-H	Int-L	No-L	Int-H	Non-H	Non-L

Table 8: 6-Condition Latin Square

Dependent Variables

To assess the effectiveness of the different PM aids, participant PM performance was measured by:

- PM task percent correct (% of PM tasks successfully executed)
- PM task reaction time (elapsed time from triggering situation to execution of action)
- Total PM errors (see PM task response categories below)

These measures were selected since they provide insight into how quickly and accurately PM tasks are executed, both important dimensions for successful PM performance. In this series of experiments, participant PM task responses were categorized and logged into one of the following 4 categories:

- Too Early, Right Action: participant executed the correct action prior to the correct time
- Right Time, Wrong Action: participant executed an incorrect action at the correct time
- Miss: participant did not execute the correct action within 20-second grace period of triggering situation
- Correct: participant executed the correct action within 20-second grace period after triggering situation

A 20-second grace period was selected based on prior PM research with a commensurately complex task environment (Altgassen et al., 2014). For analysis purposes, PM errors were the

total of the first three categories. PM task percent correct was defined as Correct/Total Tasks (12).

Primary task performance was measured by multiple measures, as described above, to assess impact of the different PM aiding conditions. To summarize:

- Visual Search
    - Percent correct (% of targets detected and clicked before disappearing)
    - Reaction Time (elapsed time from onset of target)
- Progress Assessment
    - Percent correct (% of total trial time with correct assessment)

These measures were selected since they are sensitive to effects of attention-capture, such as delayed response and detection failure (McDaniel & Einstein, 2007); attention capture is a potential downside to intrusive PM aids within a multi-tasking environment.

Subjective impression of the memory aids was assessed with a custom survey described above that included items pertaining to aid intrusiveness, impact on primary task performance, and interference with native PM skills. The measures were:

- Rating responses for each survey item

All participants received a baseline condition with no-aiding against which to compare the aiding conditions and anchor their responses.

Participants

Six participants (4 female, 2 male) were recruited from the University of Minnesota community and were compensated between $35 and $40 depending on the total duration of experimental session. Participants' age ranged from 20 to 47 years with a mean age of 27.66 years. Participants completed one 6-condition Latin Square counterbalance.

Procedure

Upon arriving for experiment, all participants read and signed an informed consent form and were asked if they had any questions or concerns about the experimental procedure. After any questions were answered, participants then received a PowerPoint briefing that described the experimental protocol, experimental user interface, survey, primary tasks, and PM task. All

training was individualized to each participant such that training would not progress until the participant was comfortable with the current task or tasks. Following the briefing, participants completed training scenarios during which they familiarized themselves with the Progress Assessment (PA) task first; then once they were comfortable with the PA task, they were asked to perform the Visual Search (VS) task concurrently with the PA task. Once comfortable with both Primary tasks, participants executed a second training scenario where they performed both primary tasks and PM tasks without any aiding. Next participants performed a training scenario with intrusive aid followed by a scenario with non-intrusive aids. Total training duration varied across participants between 35 and 60 minutes.

Before each trial, participants were informed about the aiding and workload level; for example, they were told that the upcoming trial would be a low workload trial with no aiding. After completing the trial, participants were provided feedback on primary and PM task performance via dialog in Figure 7. Participants were then asked to rate their subjective impression of the memory aids with a custom survey that included items pertaining to their experience with PM tasks, impact of aid on primary task performance, and how aid impacted their native PM skills. The survey was designed to assess perceived intrusiveness of the different aiding conditions. Intrusiveness of the PM aids was operationalized in terms of negative impact on primary task performance and interference with native PM skills. Participants rated their experience on a 7-point Likert scale that was anchored at 1 with "strongly disagree" and 7 with "strongly agree." After aiding conditions (intrusive and non-intrusive), participants responded to all 8 items listed below. After no-aiding conditions, participants only responded to the first 3 items.

1. It was easy to develop a rhythm or flow.
2. It was easy to remember memory tasks.
3. I was often distracted from the main tasks: visual search or progress assessment.
4. It was easy to integrate the memory aid into strategies for remembering memory tasks.
5. The memory aid distracted me from main tasks.
6. I lost my concentration after paying attention to the memory aid.
7. The PM aid hurt my performance on the visual search task.
8. The PM aid hurt my performance on the progress assessment task.

After the completion of all experimental trials, participants were thanked for their participation and received compensation.

Results

Graphs include standard error bars to enable visual comparison of condition means. Tables with condition means include overall mean for each level in "Totals" column for workload levels and row for aiding levels, respectively. Alpha (α) level of 0.05 was used for all analyses unless otherwise stated. Full ANOVA tables are in Appendix D. For post hoc t-tests, details are presented in a Comparisons table where p-value is bolded for significant results.

PM Performance by Aid Type

When assessing PM performance with a 2 factor repeated-measures analysis of variance (ANOVA), aiding ($F(2)= .03$, $p= .37$) nor workload ($F(1) = 0$, $p =1.0$) main effects were not significant, as indicated in Table 10. This is evident in comparing condition means in Table 9, especially for workload differences within aiding condition. The Bonferroni correction was applied for the following 4 comparisons such that the significance level was ((α) /4 or 0.05/4) 0.0125 for following t-test analyses. Consistent with statistical conventions, paired t-tests were run as post hoc comparisons with Bonferroni correction for repeated-measures design. For the critical comparison between non-intrusive aiding and no-aiding, neither difference was significant for either low workload (.85 vs. .79; $t(5)= .57$, $p = 0.594$, paired t-Test, two-tailed test), nor high workload (.81 vs. .82; $t(5)= -.24$, $p= 0$ .822, paired t-Test, two-tailed test). For the comparison between non-intrusive aiding and intrusive aiding, neither difference was significant for either low workload (.85 vs. .72; $t(5)= 1.0$,$p= 0.363$, paired t-Test, two-tailed test), nor high workload (.81 vs. .74; $t(5)= 1.05$, $p=0.341$, paired t-Test, two-tailed test). It is clear from these results that the primary task workload manipulation did not produce a sufficiently difficult "high" workload condition to drive differentiation in PM task performance.

	Intrusive	None	Non-intrusive	Totals
Low Workload	0.722222	0.791667	0.847222	0.787037
High Workload	0.736111	0.819444	0.805556	0.787037
Totals	0.729167	0.805556	0.826389	0.787037

Table 9: PM Percent Correct by Aiding and Workload Levels

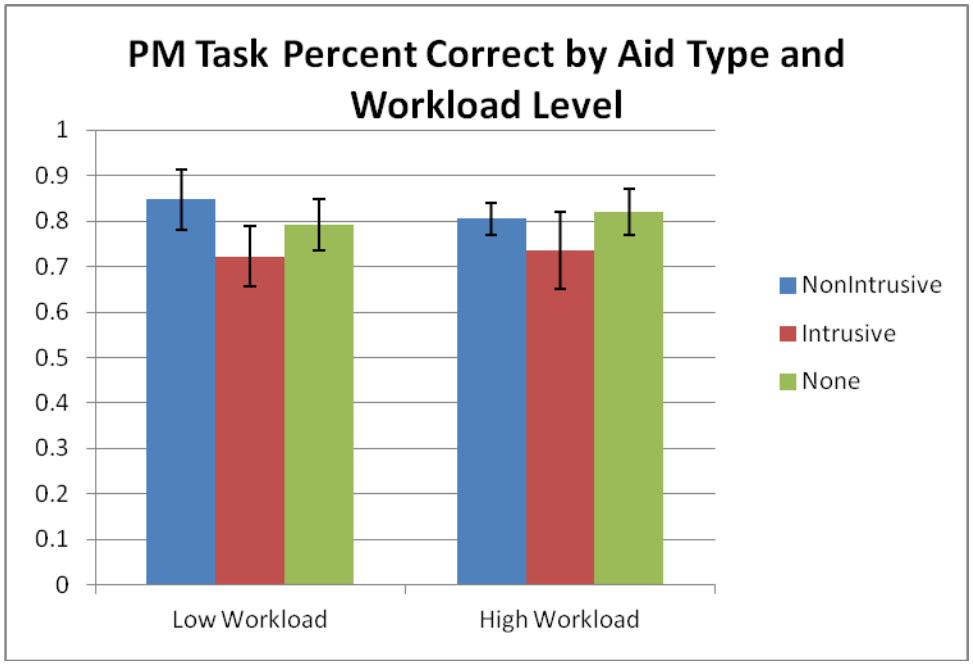


Figure 10: PM Task Percent Correct by Aiding and Workload Levels

Experiment 1A PM Task Percent Correct 2 Factor ANOVA Workload x Aid					
Source	SS	df	MS	F	P
Workload	0	1	0	0	1.000000
Aid	0.0629	2	0.0314	1.0865	0.374121
Workload x Aid	0.0081	2	0.0041	0.1547	0.858683

Table 10: PM Task Percent Correct 2 Factor ANOVA

PM Pct Correct Comparisons					
Condition 1	Mean 1	Condition 2	Mean 2	p-value of t-test	Significance level
Non-intrusive Low Workload	.85	None Low Workload	.79	0.594	.0125
Non-intrusive High Workload	.81	None High Workload	.82	0.822	.0125
Non-intrusive Low Workload	.85	Intrusive Low Workload	.72	0.363	0125
Non-intrusive High Workload	.81	Intrusive High Workload	.74	0.341	0125

Table 11: PM Pct Correct Comparisons

PM Performance by PM Difficulty

Across conditions there were 432 PM tasks (6 participants x 6 conditions x 12 PM tasks per trial). When collapsing across aiding and workload conditions, participants committed 92 total errors for an effective error percentage of .21 (92/432). All 6 experimental conditions had 6 "hard" and 6 "easy" PM tasks, as defined above, for a total of 36 of each. When PM performance was broken out by PM task difficulty, participants made significantly more total errors on "hard" tasks (9.5 out of 36) compared to "easy" tasks (5.33 out of 36) across 6 conditions, as depicted in Table 12. This difference was found to be statistically significant ($t(5) = -3.76$, $p = 0.013$, paired t-Test, two-tailed test).

	Easy	Hard
Average Errors	5.33333333	9.5

Table 12: PM Errors by PM Task Difficulty

To illustrate that this was not simply an aggregated effect driven by a few participants, error data were plotted by participant. As is clear from Figure 11, all participants exhibited this pattern, making more errors on "hard" PM tasks compared to "easy" PM tasks.

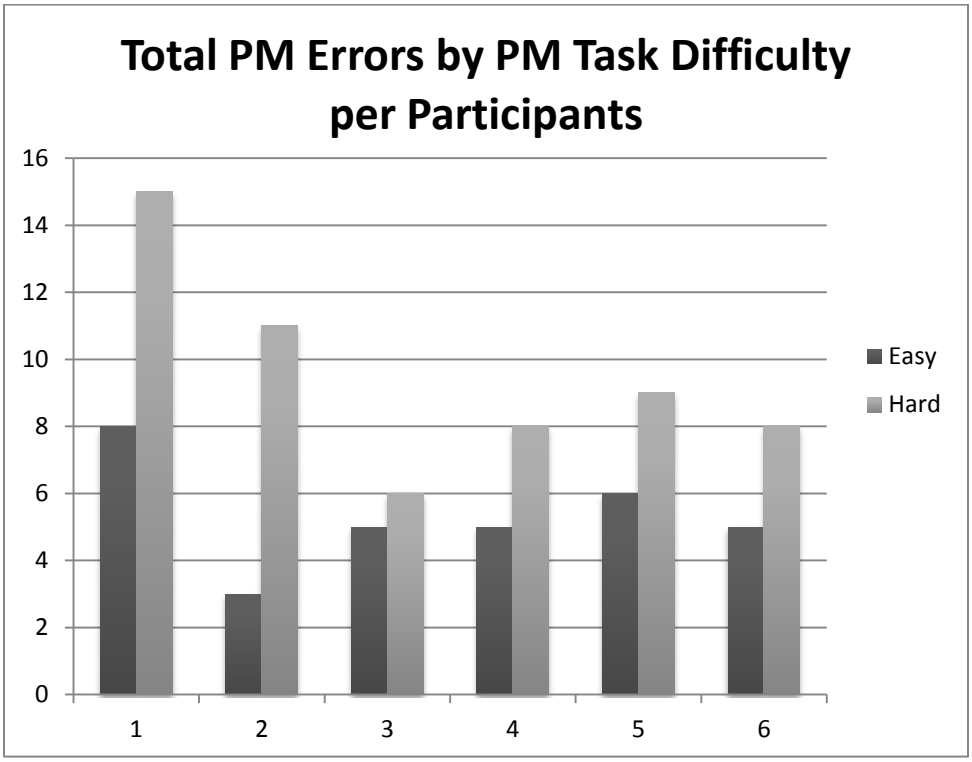


Figure 11: PM Errors by Difficulty per Participant

Subjective Impression—Intrusiveness Survey

For subjective impression findings, median values will be presented and analyzed; as a measure of central tendency, median is less sensitive to the skewed and sometimes biased nature of subjective response data including possible outliers. Participants responded with the following rating scale and median responses are depicted in Figure 12:

- 1 = Strongly disagree
- 2 = Disagree
- 3 = Somewhat disagree
- 4 = Neither agree or disagree
- 5 = Somewhat agree
- 6 = Agree
- 7 = Strongly agree

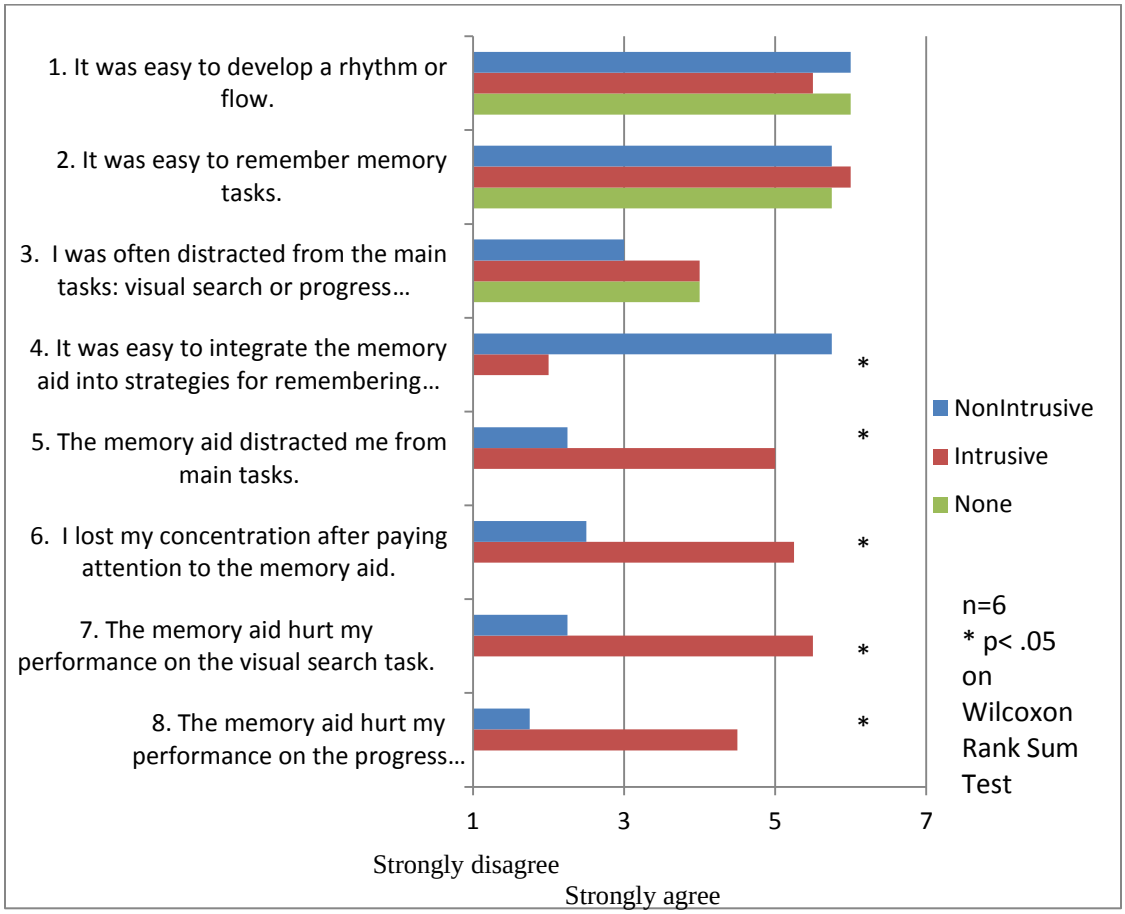


Figure 12: Survey Item Median Results by Intrusiveness of Aid

Items 4-8 asked about their impression of the different memory aids, intrusive and non-intrusive. The following comparisons used the Wilcoxon Rank Sum Test for non-parametric data (such as subjective survey responses)(Higgins, 2003); all comparisons were significant ($W(6) = 0$, $p < .05$, two-tailed test. Significant comparisons for individual survey items are identified with asterisks ("*").

For item 4, which asked about the ease of integrating the aid into a memory strategy, participants responded with a significantly higher level of agreement for the non-intrusive aid (5.75, 6= Agree) compared to intrusive aid (2, 2= Disagree). Items 5-8 asked about negative impact of aid on primary tasks and participants responded with a significantly higher level of agreement for intrusive aid compared to non-intrusive aid as depicted in Table 13 :

Item	Intrusive	Non-intrusive
5	5, 5= somewhat agree	2.25, 2= Disagree
6	5.25, 5= somewhat agree	2.5, 3= somewhat disagree
7	5.5, 5= somewhat agree	2.25, 2= disagree
8	4.5, 4= neither agree nor disagree	1.75, 2= disagree

Table 13: Survey Items Results

Primary Task Performance

When reviewing primary task percent correct performance across workload levels with standard error bars in Figure 13, it is evident that the workload manipulation was not effective. For Visual Search, performance across workload was essentially equivalent (Low = .95, High = .95; F(1) = 0, p= 1.0), as indicated in Table 16 . For Progress Assessment, the modest difference was not statistically significant (Low = .81, High = .77; F(1) = .61, p = .47), as indicated in Table 17.

	Intrusive	None	Non-intrusive	Totals
Low Workload	0.951917	0.926893	0.96553	0.948113
High Workload	0.956769	0.932758	0.955746	0.948424
Totals	0.954343	0.929825	0.960638	0.948269

Table 14: Visual Search Percent Correct

	Intrusive	None	Non-intrusive	Totals
Low Workload	0.821119	0.820501	0.798701	0.813441
High Workload	0.771336	0.782468	0.759895	0.771233
Totals	0.796228	0.801484	0.779298	0.792337

Table 15: Progress Assessment Percent Correct

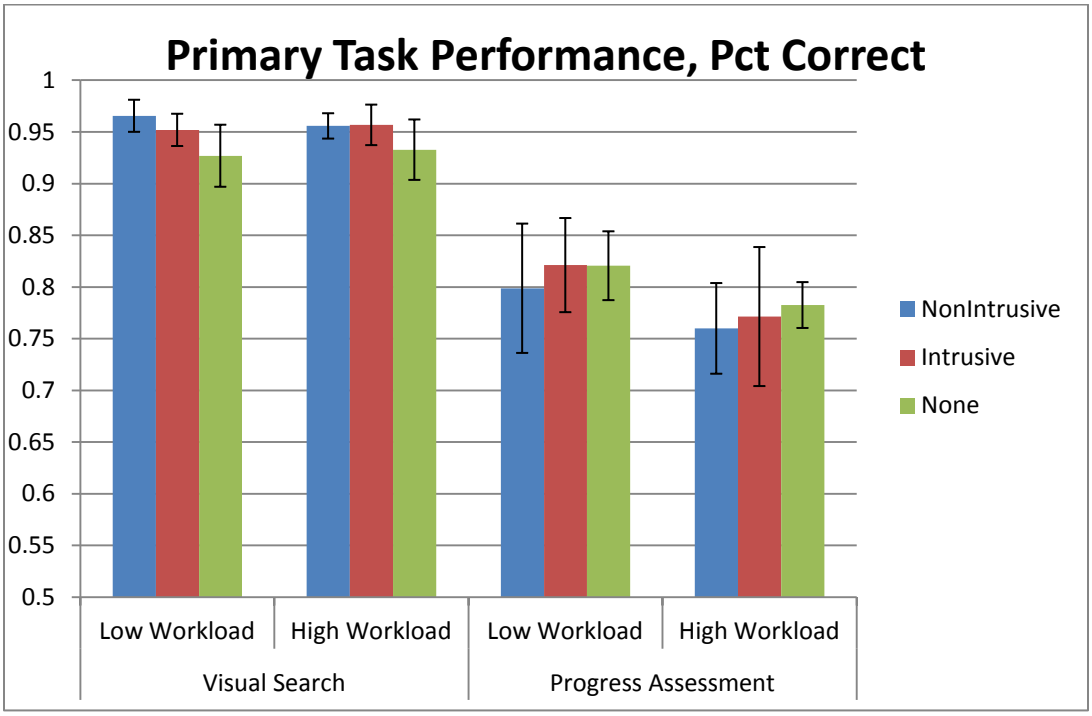


Figure 13: Primary Task Percent Correct Performance across Aiding Levels

Experiment 1A Visual Search Percent Correct 2 Factor ANOVA Workload x Aid					
Source	SS	df	MS	F	P
Workload	0	1	0	0	1.000000
Aid	0.0064	2	0.0032	2.9091	0.100975
Workload x Aid	0.0005	2	0.0002	1	0.401878

Table 16: Visual Search Percent Correct 2 Factor ANOVA

Experiment 1A Progress Assessment Percent Correct 2 Factor ANOVA Workload x Aid					
Source	SS	df	MS	F	P
Workload	0.016	1	0.016	0.6154	0.468281
Aid	0.0032	2	0.0016	0.1739	0.842869
Workload x Aid	0.0003	2	0.0001	0.0118	0.988283

Table 17: Progress Assessment Percent Correct 2 Factor ANOVA

Discussion

After running 6 participants through one complete Latin Square counterbalance, all participants were able to complete the study with no breaks and no degradation of performance over the 2 hour experimental session. Many reported that the time went fast and that it was engaging and interesting, but "not that hard" in the words of several participants. These self-reports dove-tailed with the minimal impact of "high" workload on primary and PM task performance. Visual Search performance was equivalent across workload levels. Likewise, when grouped across aiding levels, PM performance was equivalent (.79) for both low workload and high workload. This suggests that the primary tasks under "high" workload did not sufficiently tax participants' cognitive resources to induce errors. Accordingly, there was essentially no difference between aiding conditions due to this ceiling effect, thus **Hypothesis from Research Question 1 & 5** were not confirmed:

- Non-intrusive aiding conditions did not support superior PM performance compared to the no-aiding.
- Due to insufficiently high primary task workload, could not reliably assess whether non-intrusive aiding performance was equivalent to intrusive aiding.

However, of the errors committed, participants did commit significantly more errors on "hard" PM tasks (9.5) compared to "easy" PM tasks (5.33). This indicates that the manipulation of PM tasks difficulty was effective in inducing a differential error rate. This is noteworthy since difficulty was defined by dimensions that have been traditionally validated within simple task environments with simple PM tasks, and it was unclear how well it would translate to complicated PM tasks in a complex task environment. Accordingly, **Hypothesis from Research Question 4** was confirmed:

- Participants committed more PM errors on "hard" tasks than "easy" tasks.

Finally, on those survey items that asked about negative impact of aid, participants agreed at a significantly higher level for intrusive aid compared to non-intrusive aid. This suggests that the peripheral, graphical Non-intrusive aid design was perceived as less impactful on primary task performance than the intrusive aid. Thus, **Hypothesis 6** was confirmed:

- Participants rated Non-intrusive aids as less negatively impactful than intrusive aids.

After the experiment, many participants reported there were many lulls in the primary task load that allowed them to rehearse the PM tasks. Participants indicated that they had the time and cognitive resources to maintain PM tasks in memory. This effectively limited the beneficial impact of aiding since participants had the time and cognitive resources to maintain and perform PM tasks without aiding. We could not investigate the impact of different PM aids because the primary tasks were too easy. The participants had sufficient cognitive resources to perform PM and did not need the PM aids, as indicted by high levels of PM performance across aiding conditions. Harder primary tasks were required before participants would start needing the PM aids to support PM performance. Based on this feedback, an additional primary task was added in an attempt to tax participant resources sufficiently to induce more errors and possibly realize aiding benefits in Experiment 1B.

Experiment 1B

Primary task load in Experiment 1A was not sufficient to tax participant resources. Experiment 1B was made harder. Specifically, to disrupt and interfere with retention and retrieval, Experiment 1B introduced a third primary task which simulated radio queries from Air Traffic Control (ATC) about the state of the flight. This new task required participants to assess the veracity of compound statements about their current flight situation. Compound statements consisted of two statements connected by logical "and" operator. If both statements were true participants responded by clicking the "Yes" button, if one or both of the statements was false, participants responded with the "No" button, as depicted below. All statements from ATC were presented aurally, via text-to-speech functionality, in an American female voice. This was a different text-to-speech voice than the English female "pilot" that presented the PM tasks. Statements were presented approximately every 28 seconds. Participants were given a 10 second window to respond after the application completed "speaking" the statement. If they did not respond within 10 seconds, they received voiced feedback "Message missed". Statements asked participants about the following dimensions:

- Speed (greater than, less than, between)
- Speed Trend (stable, increasing, decreasing)
- Altitude (greater than, less than, between)
- Altitude Trend (stable, increasing, decreasing)
- Direction of Flight (Northbound, Southbound, Eastbound, or Westbound)
- Location relative to Waypoints– before, after, between
- Sector Location (1,2,3,4)

In addition to "Yes" and "No" response buttons for the Radio Query tasks, the experimental task interface also included reminders regarding the directional alignment of the map (North, South, East, and West) as well as the Sector layout (1,2,3,4), as seen highlighted in purple in Figure 14.

.

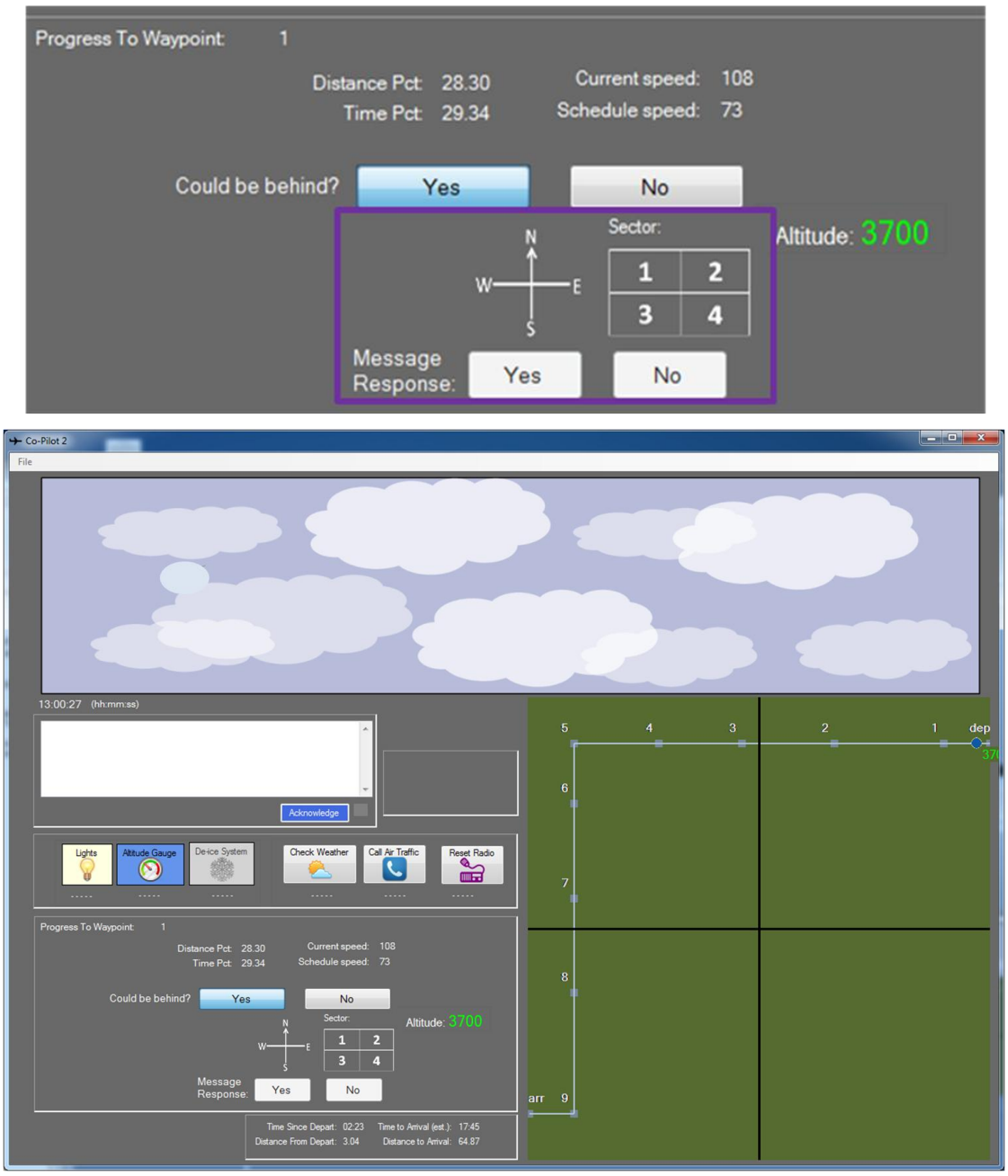


Figure 14: Radio Query Task Interface

Python scripts generated pseudo-random queries that were unique for each experimental scenario, insuring**:**

- An equal number of queries across trials (36)
- 50% of statements were True and 50% were False
- Of False statements, the first statement was True 50% of the time (if first statements were overwhelmingly false, participants could ignore most of the second statements since they

could deduce that statement was false without having to listen to the second statement. This would have decreased attention and workload demands).

The following are examples of Radio Queries:

- Speed Trend is Increasing and Located Between Waypoints Departure and 7
- In Sector 3 and Not Flying Westbound
- Speed is less than 140 and Located After Waypoint 1
- Speed Trend is Not Decreasing and Located After Waypoint 1
- Not in Sector 1 and Altitude Trend is Not Decreasing

Research Questions

To review, the following Research Questions were addressed in Experiment 1B:

- Primary Question
    - 1. Can non-intrusive PM aids improve PM performance compared to no-aiding?
        - Hypothesis: Non-intrusive aiding conditions will support superior PM performance compared to the no-aiding.
    - 2. Does the performance benefit of PM aiding across PM task difficulty levels justify the costs?
        - Hypothesis: No specific hypothesis, this is an outstanding empirical and theoretical question.
- Secondary Questions
    - 4. Did the operational definition of "easy" and "hard" PM tasks, based on prior work, induce different PM performance in a complex task environment?
        - Hypothesis: Participants will commit more PM errors on "hard" PM tasks than "easy" PM tasks.
    - 5. Can non-intrusive aiding support equivalent PM performance to intrusive aiding?
        - Hypothesis: Given its support of monitoring for PM task triggering condition, non-intrusive aiding will support an equivalent PM performance to intrusive aiding.
    - 7. What would the impact of aiding be on primary task performance?

- Hypothesis: There will be no differential impact of primary task performance across aiding conditions.

Experimental design was a single factor (3 levels of aiding) repeated measures design (within subjects). The levels of the one independent variable were as follows.

Independent Variables

1. Aiding:
    a. No aid (Figure 5) baseline display
    b. Non-intrusive aid (Figure 8) baseline display with graphical timer
    c. Intrusive aid (Figure 9) baseline display with pop-up dialog reminders

Unlike in Experiment 1A, primary task workload was held constant at high workload across the three levels of aiding in Experiment 1B (high workload variants of Visual Search and Progress Assessment (Table 3) plus Radio Query task (Figure 14). The experimental conditions are outlined in Table 18:

No aiding	Non-intrusive	Intrusive

Table 18: Experiment 1B Design

Three flight scenarios from Experiment 1A were randomly selected for use in Experiment 1B. Aside from the introduction of Radio Query task, all other primary task and PM task details were the same as in Experiment 1A. Presentation order was counterbalanced with a 3-condition Latin Square design. See below for an example of a Latin Square for a 3 condition experiment requiring multiples of 3 participants. The conditions are coded as follows in Table 19: aiding—intrusive (Int), non-intrusive (Non), no-aiding (No); workload: low (L) and high (H).

	Order		
Group	1	2	3
A	No-H	Non-H	Int-H
B	Int-H	No-H	Non-H
C	Non-H	Int-H	No-H

Table 19: 3-Condition Latin Square

Dependent Variables

The dependent variables were identical to those described in detail in Experiment 1A, aside from the addition of a Radio Query percent correct measure. To review, variables are listed below:

- PM performance:
  - PM task percent correct (% of PM tasks successfully executed)
  - PM task reaction time (elapsed time from triggering situation to execution of action).
  - Total PM errors (see PM task response categories below)
- Primary task performance:
  - Visual Search
    - Percent correct (% of targets detected and clicked before disappearing)
    - Reaction Time (elapsed time from onset of target)
  - Progress Assessment
    - Percent correct (% of total trial time with correct assessment)
  - Radio Query
    - Percent correct (% of correct responses to queries)

Participants

Six participants (4 female, 2 male) were recruited from the University of Minnesota community and were compensated between $30 and $40 depending on the total duration of experimental session. Participants' age ranged from 22- 61 years with a mean age of 44.5 years. Participants completed two 3-condition Latin Square counterbalances.

Procedure

The procedure was identical to Experiment 1A except for the following differences. Participants received the same training regime as Experiment 1A with the addition of Radio Query task training as well. After participants were comfortable with performing Progress Assessment and Visual Search concurrently, the Radio Query task was introduced and participants performed all three primary tasks until they were comfortable. Total training duration varied across participants between 45 and 75 minutes.

Results

Graphs include standard error bars to enable visual comparison of condition means. Alpha (α) level of 0.05 was used for all analyses unless otherwise stated. Full ANOVA tables are in Appendix D. For post hoc t-tests, details are presented in a Comparisons table where p-value is bolded for significant results.

PM Performance by Aid Type

While a single factor repeated-measures ANOVA (Aid) did not reach significance ($F(2,15) = 1.05$, $p = .378$), non-intrusive PM performance (.74) was found to be significantly higher, when compared to both intrusive (.53) ($t(5) = -3.48$, $p = 0.017$, paired t-Test, two-tailed test) and None (.54) ($t(5) = -3.26$, $p = 0.022$, two-tailed test). The Bonferroni correction was applied for these two comparisons such that the significance level was ((α) /2 or 0.05/2) .025 for following t-test analyses. When looking at the critical comparison between non-intrusive and no-aiding (None), it is worth noting that all participants exhibited the same pattern, with higher Non-intrusive performance. In fact, the effect size for this analysis ($d = 1.40$) was found to exceed Cohen's (1988) convention for a large effect ($d = .80$).

Graphs include standard error bars to enable visual comparison of condition means and significantly different values are identified with asterisks (*). PM task percent correct by aid type are presented in Figure 15.

	Intrusive	None	Non-intrusive	Total
High Workload	0.527778	0.541667	0.736111	0.601852

Table 20: PM Pct Correct by Aiding

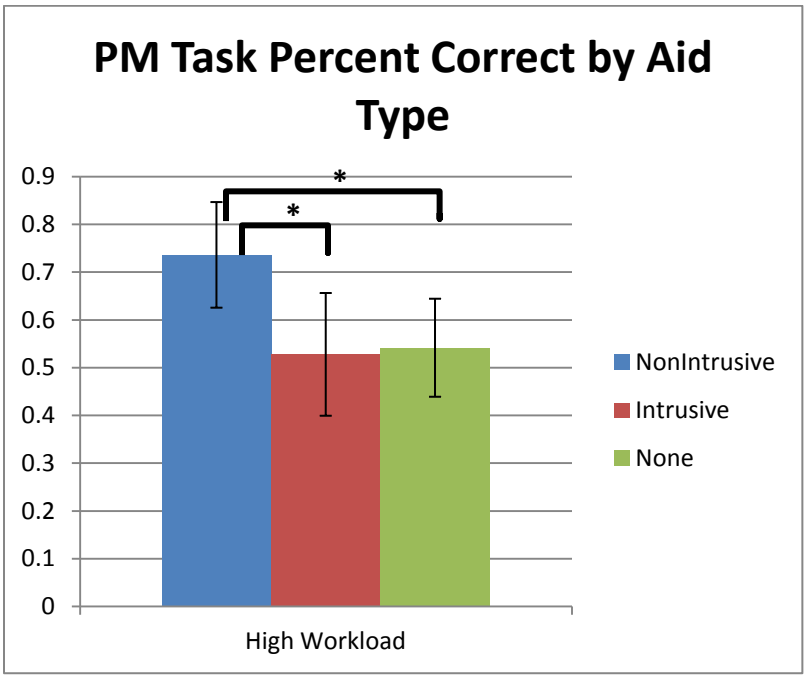


Figure 15: PM Task Percent Correct by Aiding

PM Pct Correct Comparisons					
Condition 1	Mean 1	Condition 2	Mean 2	p-value of t-test	Significance level
Non-intrusive	.74	Intrusive	.53	**0.017**	.025
Non-intrusive	.74	None	.54	**0.022**	.025

Table 21: PM Pct Correct Comparisons

Another way to evaluate PM performance is reaction time. Within this PM paradigm, reaction time is defined as the elapsed time from the onset of the triggering situation (e.g., at 14000 feet altitude) to when participant clicks the action button ("Check Weather"). In a single factor repeated-measures ANOVA on PM Reaction Time for aiding Condition, there was a significant main effect of aiding ($F(2)= 7.03$, $p < .010$), as depicted in Table 23

For the critical comparison, a t-Test was performed following the single factor ANOVA. Again, the Bonferroni correction was applied for the following two comparisons such that significance level was 0.025 (($\alpha= 0.05$) /2). As is evident in Figure 16, participants performed delayed action significantly sooner after triggering situation with non-intrusive aiding (2258 msec.) when

compared to both intrusive (6557 msec.) (t(5) = 4.13, p=0.009, paired t-Test, two-tailed test) and marginally significantly sooner than None (5060 msec.)(t(5) = 2.76, p= 0.039, two-tailed test).

	Intrusive	None	Non-intrusive	Totals
High Workload	6556.522	5060.262	2258.26	4625.015

Table 22: PM RT 1 Factor ANOVA

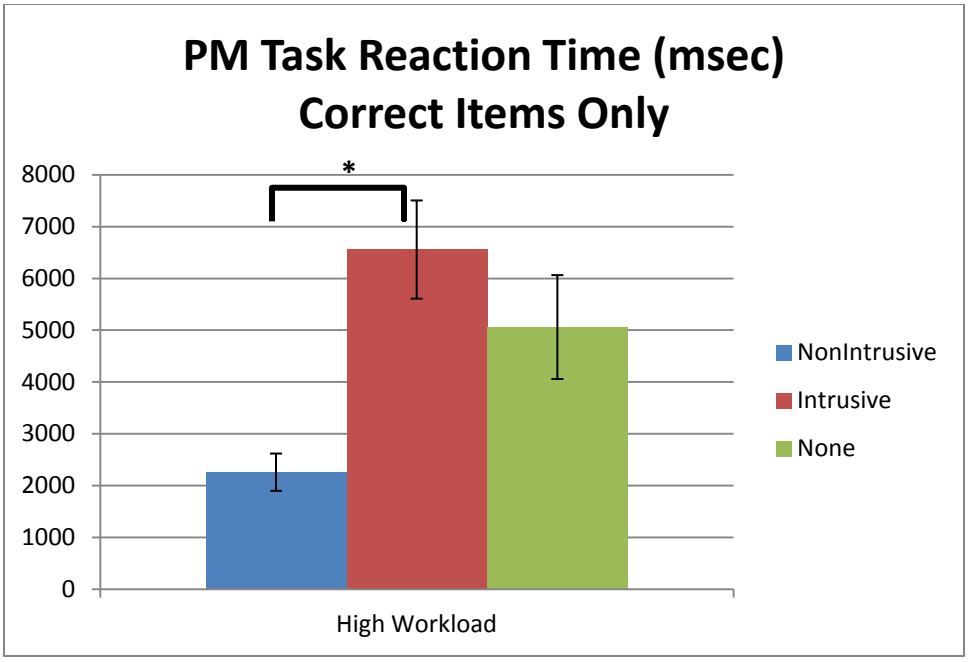


Figure 16: PM Task RT by Aiding

Experiment 1B PM Task Percent Correct 1 Factor ANOVA Aid					
Source	SS	df	MS	F	P
Aid	57130130	2	28565065	7.027108	0.007025

Table 23: PM RT 1 Factor ANOVA Aid

PM RT Comparisons					
Condition 1	Mean 1	Condition 2	Mean 2	p-value of t-test	Significance level
Non-intrusive	2258	Intrusive	6557	**0.009**	.025
Non-intrusive	2258	None	5060	0.039	.025

Table 24: PM RT Comparisons

PM Performance by PM Difficulty

Across conditions there were 216 PM tasks (6 participants x 3 conditions x 12 PM tasks per trial.) When collapsing across aiding conditions, participants committed 86 errors for an effective error percentage of .40 (86/216). All 3 experimental conditions had 6 "hard" and 6 "easy" PM tasks for a total of 18 of each. As in Experiment 1A, participants committed significantly more errors on "hard" PM tasks (9.5 out of 18 total "hard" tasks) when compared to "easy" PM tasks (4.83 out of 18 total "easy" tasks) ($t(5) = -5.08$, $p= 0.0038$, paired t-Test, two-tailed test). The effect size for this analysis ($d= 2.17$) was found to exceed Cohen's (1988) convention for a large effect ($d= .80$). As in Experiment 1A, all participants exhibited the same pattern, with more errors on "hard" PM tasks compared to "easy" tasks, as illustrated in Figure 17.

	Easy	Hard
Average Errors	4.833333333	9.5

Table 25: Total PM Errors by PM Task Difficulty

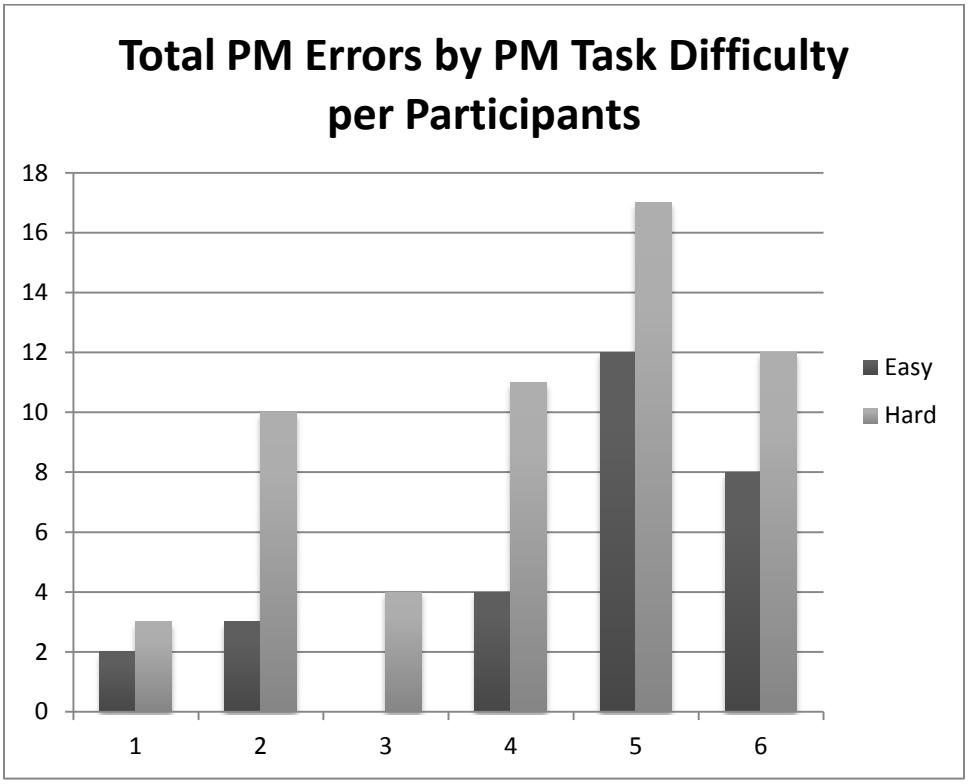


Figure 17: PM Errors by Difficulty per Participant

Differential Impact of Aiding across PM Task Difficulty

At the aggregated level, all aiding conditions exhibited the same pattern as global findings for error frequency. There were more errors on "hard" tasks compared to "easy" tasks across all aiding conditions, as seen in Figure 18.

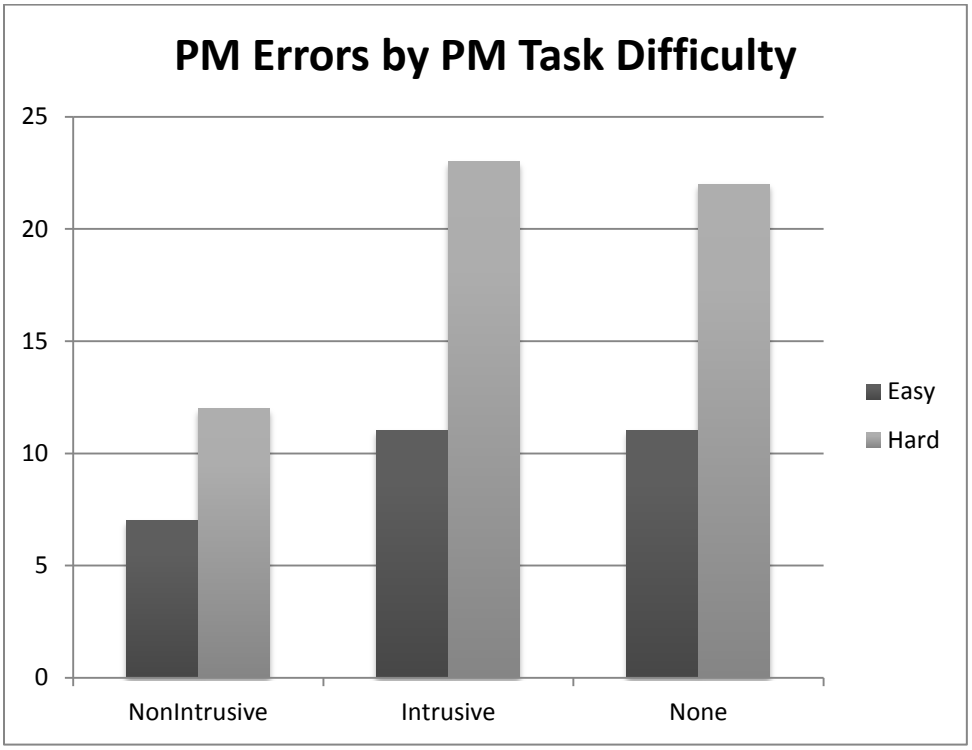


Figure 18: PM Errors by Difficulty across Aiding Levels

To look at differential impact of aiding across PM task difficulty, a relative difference score was calculated for all trials. Relative difference was calculated by subtracting "hard" PM task percent correct from "easy" PM task percent correct then dividing by "easy" percent correct for each participant, as depicted in Table 26. Relative differences scores for all participants across aiding conditions are presented in Figure 19. When looking across aiding conditions, the aid with the lower relative score could be presumably said to provide better support on "hard" tasks. Again, the Bonferroni correction was applied for the following two comparisons such that significance level was 0.025 ((α= 0.05) /2). The relative difference was significantly less for non-intrusive aiding (.20) compared to intrusive (.56)($t(5)= - 4.34$, $p= 0.007$, paired t-Test, two-tailed test), and also to None (.51) ($t(5) = 3.40$, $p= 0.019$, paired t-Test, two-tailed test).

Participant	Intrusive	None	Non-intrusive
1	0	0.25	0
2	0.8	0.4	0.2
3	0.5	0.166667	0
4	0.75	0.5	0.333333333
5	1	1	0.666666667
6	0.33333333	0.75	0
average	**0.56388889**	**0.511111**	**0.2**
std error	**0.14784982**	**0.1285**	**0.108866211**

Table 26: Relative Difference Values across Participants and Aiding

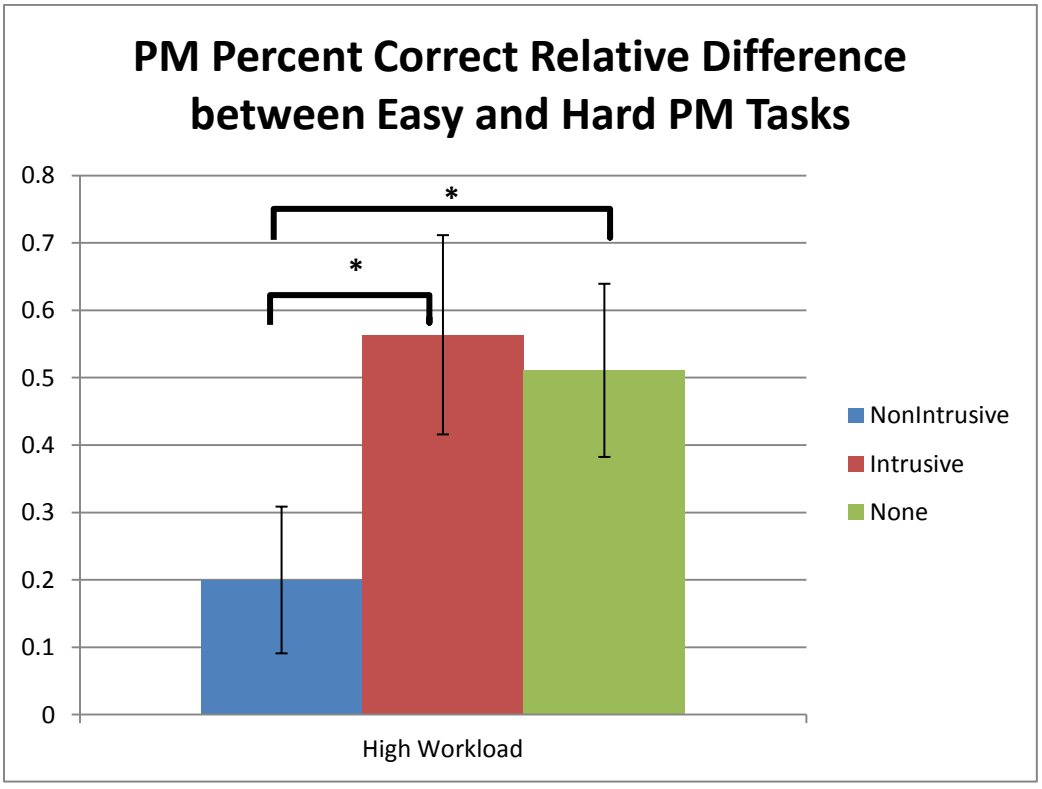


Figure 19: Relative Difference across Aiding Levels

PM Pct Correct Relative Difference Comparisons					
Condition 1	Mean 1	Condition 2	Mean 2	p-value of t-test	Significance level
Non-intrusive	.20	Intrusive	.56	**0.007**	.025
Non-intrusive	.20	None	.51	**0.019**	.025

Table 27: PM Pct Correct Relative Difference Comparisons

However, the practical impact of this was relatively minor—half of the participants had the same number of errors for "easy" and "hard" PM tasks and no participant had a difference greater than 2, as depicted in Table 28.

Participant	Easy	Hard	Difference
1	0	0	0
2	1	2	1
3	0	0	0
4	0	2	2
5	3	5	2
6	3	3	0
Total	**7**	**12**	**5**

Table 28: PM Errors by PM Task Difficulty

Primary Task Performance

Next, the impact of aiding on primary task was considered. Average percent correct performance for all three primary tasks were presented across aiding condition in Figure 20 and Table 29. Separate single factor repeated-measures ANOVAs (Aiding) indicated that differences in percent correct performance for neither Visual Search ($F(2) = .10$, $p = .91$) nor Progress Assessment ($F(2) = .53$, $p = .60$) reached statistical significance, as depicted in Table 30; however, there was a significant result for Radio Query Performance ($F(2) = 4.67$, $p < .05$). Follow-up comparisons, with Bonferroni corrected significance level of 0.025, indicated that Radio Query performance was significantly lower for non-intrusive condition (.63) compared to none (.75; $t(5) = -10.51$, $p = 0.00013$, paired t-Test, two-tailed test) but not intrusive (.71; $t(5) = -1.61$, $p = .168$, paired t-Test, two-tailed test).

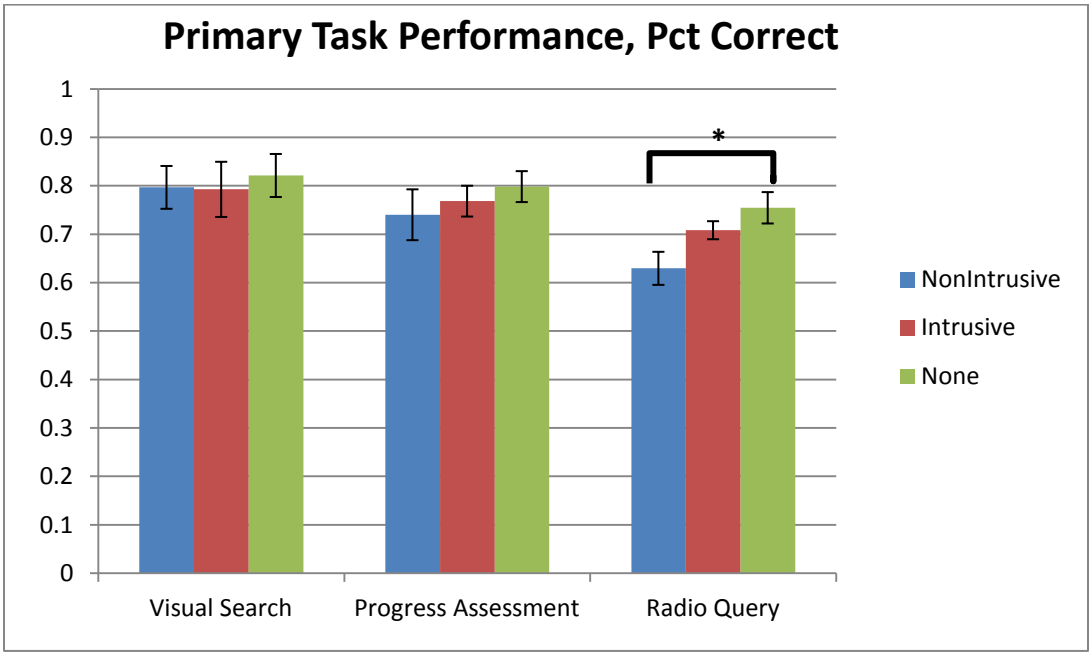


Figure 20: Primary Task Percent Correct Performance across Aiding Levels

Percent Correct Performance by Aiding Levels				
	Intrusive	None	Non-intrusive	Totals
Visual Search	0.792697528	0.82133861	0.796777993	0.803605
Progress Assessment	0.768398268	0.798392084	0.74025974	0.769017
Radio Query	0.708333333	0.75462963	0.62962963	0.697531

Table 29: Primary Task Percent Correct across Aiding Levels

Experiment 1B Visual Search Percent Correct 1 Factor ANOVA Aid					
Source	SS	df	MS	F	P
Aid	0.002880371	2	0.00144	0.100413	0.905067
Experiment 1B Progress Assessment Percent Correct 1 Factor ANOVA Aid					
Source	SS	df	MS	F	P
Aid	0.01014155	2	0.005071	0.529298	0.599622
Experiment 1B Radio Query Percent Correct 1 Factor ANOVA Aid					
Source	SS	df	MS	F	P
Aid	0.04792524	2	0.023963	4.673913	0.026438

Table 30: Visual Search, Progress Assessment, and Radio Query Percent Correct 1 Factor ANOVAs

Radio Query Pct Correct Comparisons					
Condition 1	Mean 1	Condition 2	Mean 2	p-value of t-test	Significance level
Non-intrusive	.63	Intrusive	.71	0.168	.025
Non-intrusive	.63	None	.75	**0.00013**	.025

Table 31: Radio Query Pct Correct Comparisons

In addition to the PM performance differences from Experiment 1B, primary task performance was compared with Experiment 1A to assess the impact of introducing a third primary task. While Progress Assessment performance was roughly equivalent (.77 in 1B compared to .79 in 1A), there was a marked reduction in Visual Search performance across all aiding levels in Experiment 1B (.80) as compared to Experiment 1A (.95), as indicated in Table 32.

Experiment 1A Totals	Intrusive	None	Non-intrusive	Totals
Visual Search	0.954343	0.929825	0.960638	0.948269
Progress Assessment	0.796228	0.801484	0.779298	0.792337

Table 32: Experiment 1A Primary Task Performance

Another measure of primary task performance was Visual Search reaction time (RT). RT data was analyzed for those periods when there were active PM tasks; this provides insight into potential performance costs to primary tasks of maintaining active PM tasks (Loft & Remington, 2010; Loft, Smith, & Bhaskara, 2011). While RT under non-intrusive condition was higher than no aiding (none) (1662 msec vs. 1571 msec), a single factor repeated-measures ANOVA indicated that there was no main effect of aiding types ($F(2)= .31$ $p = .74$), as indicated in Figure 21, Table 33, and Table 34.

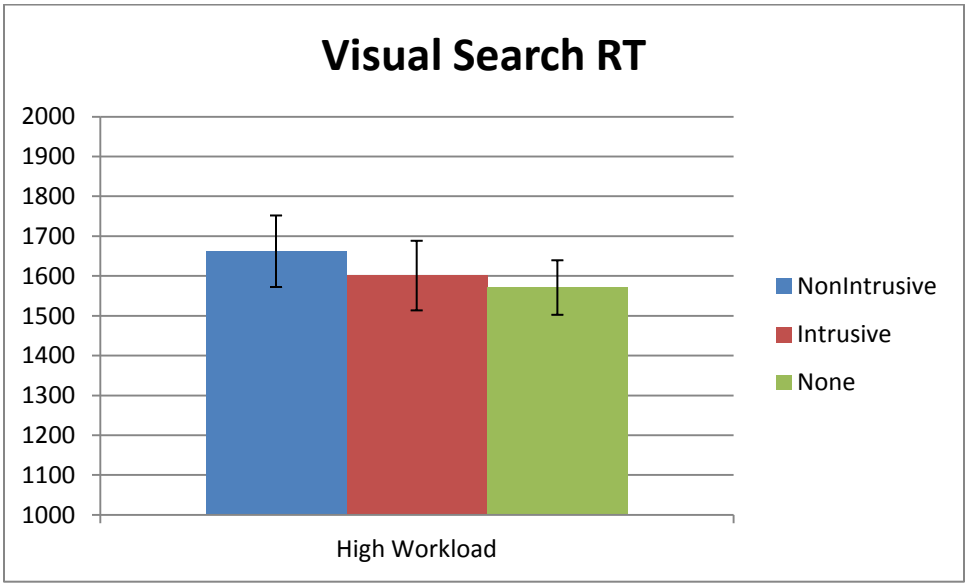


Figure 21: Visual Search RT

	Intrusive	None	Non-intrusive	Totals
Visual Search RT	1608.533	1571.016	1662.198217	1613.916028

Table 33: Visual Search RT across Aiding Levels

Experiment 1B Visual Search Reaction Time 1 Factor ANOVA Aid					
Source	SS	df	MS	F	P
Aid	25203.1	2	12601.55186	0.308926553	0.738796

Table 34: Visual Search RT Single Factor ANOVA

Discussion

There are three basic components to PM—encoding item, retention/monitoring, and retrieval/execution. From Experiment 1A participants' performance and self-reports, they had time and available mental resources to retain, monitor and retrieve PM tasks within the primary task dynamic of Experiment 1A. To fairly assess any memory aid, disrupting encoding was not appropriate. Starting in Experiment 1B, a third primary task, Radio Query, was added to increase workload in order to reduce participant cognitive resources for supporting the PM task. The task was designed to be time-consuming and effortful to occupy cognitive resources so they could not be applied to retaining and monitoring for the PM task. The premise was that the aid should reduce the retention and monitoring overhead to facilitate PM performance when aided, but that the increase in primary task workload would compromise PM performance when un-aided.

The Radio Query task involved new queries every approximately 28 seconds; the complex, compound statements required participants to dedicate full attention to listen to them, assess the truth via information presented in the experimental interface, then respond. Given the time to attend to and respond to these messages, the availability of time and cognitive resources for retaining and monitoring PM tasks was reduced. This premise was validated with impact on primary task performance, a substantial decrease in Visual Search task from Experiment 1A to 1B: Visual Search .80 total average percent correct compared to .95 in Experiment 1A

Substantial decline in Visual Search suggested the updated primary task design impacted participant cognitive resources. This is encouraging since primary task workload levels were not sufficiently high to induce an aiding benefit in Experiment 1A; however, after the introduction of a third primary task, participants committed a substantially higher percentage of errors in Experiment 1B (.40) compared to Experiment 1A (.21). The impact of PM task performance was also validated by significantly better performance for the non-intrusive aiding conditions than no aiding condition in Experiment 1B. For the critical comparison, non-intrusive aiding supported a

significantly higher PM performance (.74) compared to no aiding (.54). Accordingly, **Hypothesis from Research Question 1** was confirmed:

- Non-intrusive aiding supported superior PM performance, significantly higher PM task percent correct, compared to the no-aiding

Non-intrusive aiding also supported a significantly faster response time to PM tasks. This was expected given that the aid provides feedback about the timing of PM tasks. Also, by offloading the active monitoring for a triggering situation to the aid, participants presumably have more mental resources to monitor the timing feedback and respond faster.

The intrusive aid design, a pop-up dialog box, is the most common computer-based reminder and thus could be considered a de facto standard; to be considered as a viable alternative, the non-intrusive aid should support equivalent level of PM performance to the intrusive aid. In fact, non-intrusive aiding supported a significantly higher PM performance (.74) compared to intrusive aiding (.53). Accordingly, **Hypothesis from Research Question 5** was confirmed:

- Non-intrusive aiding supported equivalent PM performance to intrusive aiding

Experiment 1A results were replicated in that participants committed significantly more errors on **"hard"** PM tasks compared to "easy" tasks, thus **Hypothesis from Research Question 4** was re-confirmed:

- Participants committed more errors on "hard" PM tasks than "easy" PM tasks.

Primary task performance was assessed to measure potential costs across aiding conditions. While there were no differences across aiding conditions for Visual Search and Progress Assessment, performance on Radio Query was significantly different across aiding conditions with lower performance on non-intrusive trials. It is not clear why this primary task was impacted and not the others; it is not altogether surprising to have a modest performance cost for such an aid since it introduces some cognitive overhead with remembering to attend to and process the aid periodically. Accordingly, **Hypothesis from Research Question 7** was confirmed for Visual Search and Progress Assessment tasks, but not for Radio Query task:

- No differential impact of primary task performance across aiding conditions for Visual Search and Progress Assessment tasks
- There was a differential negative impact of non-intrusive aiding on Radio Query task. This will be re-assessed in the next experiment in the series.

In Experiment 1B, the benefit of adaptive aiding based on PM task was assessed by comparing performance improvement when comparing "easy" and "hard" PM tasks across aiding conditions;(e.g., relative difference). The lower the relative score, the better the support provided by the aid since it has equalized the PM difficulty. The performance results relevant to **Research Question 2** were as follows:

- The PM task performance difference between no-aiding and non-intrusive aiding conditions was greater for "hard" PM tasks compared to "easy" PM tasks;
- Specifically, the relative difference for non-intrusive aiding was significantly less than other aiding conditions

Despite this statistical difference, the practical impact of this was relatively modest; in fact, 3 of 6 subjects had the same number of errors on "hard" and "easy" tasks. Given the relatively modest practical impact of aiding of PM task difficulty, the benefit may not be sufficient given the potential cost in terms of operator confusion for an adaptation that could be simultaneously "on" for a "hard" PM task and "off" for a concurrent "easy" PM task. This could make it more difficult to integrate aiding with their workflow and native PM skills. This would also impose a burden for operators to understand some system definition of "hard" and "easy" PM tasks for them to understand and predict the memory aid. Thus, **Research Question 2** was not affirmatively answered:

- The modest practical benefits were not sufficient to justify the likely costs.

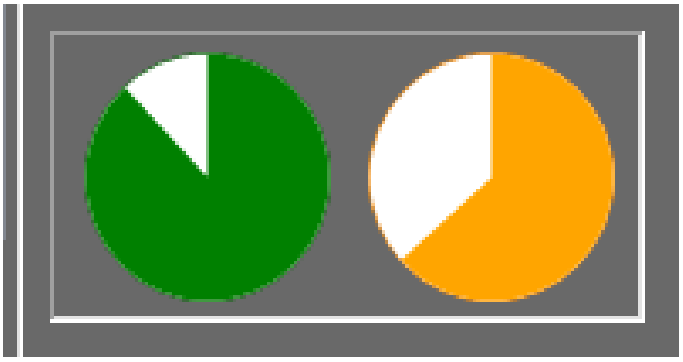
Accordingly, Experiment 2 explored aiding benefit across disparate workload levels. Prior work had established the differential benefit of aiding solutions under higher cognitive workload when compared to lower cognitive workload (Dorneich et al., 2006). Unlike PM task difficulty, adapting on current workload level would either be on or off over a given time period, possibly ameliorating some confusion. This adapting scheme would likely enable operators to more easily understand and predict aiding behavior, making it easier to integrate with their workflow and native PM skills. Given the poor PM performance and subjective rating results for the intrusive aid compared to non-intrusive Aid, it was eliminated as a level of aiding in Experiment 2.

Experiment 2:

After determining that the benefits of adapting based on PM task difficulty did not justify the costs, it was decided to investigate a more predictable adaptation for Experiment 2, one based on primary task loading; furthermore, having established superior effectiveness and subjective impression of non-intrusive aid over the intrusive aid, the intrusive aid was eliminated from further investigation. Experiment 2 was designed to evaluate the differential benefit of PM aids across disparate primary task workload conditions. Experiment IB established a more reliably high workload Primary task environment, with the addition of Radio Queries to Visual Search and Progress Assessment tasks; likewise, the low workload condition from Experiment 1A was repeated. By manipulating primary task workload within each trial, Experiment 2 simulated an adaptive system that was "on" (with non-intrusive aid) for high workload and "off" (no-aiding) for low workload.

To further increase potential aid benefits, PM task difficulty was also increased by increasing the number of "hard" PM tasks. This finding was consistent with McDaniel & Einstein's contention that aiding is most beneficial when targeting situations that most challenge PM (2007). The question is would this be exhibited for PM aid within a complex task environment.

With regard to primary task impact, Experiment 1B results indicated non-intrusive aiding condition produced a significantly lower performance on Radio Query task; to try to minimize impact of aid on Primary Task Performance, the Non-intrusive aid was re-designed to reduce information processing demands, while retaining feedback that supported PM performance benefit. The new design, as seen in Figure 22, resembled a circular timer that is progressively filling like a clockwise movement of a seconds hand on a clock. This design also resembled the progress indicator for downloading attachments on Apple iDevices, as seen in Figure 23.



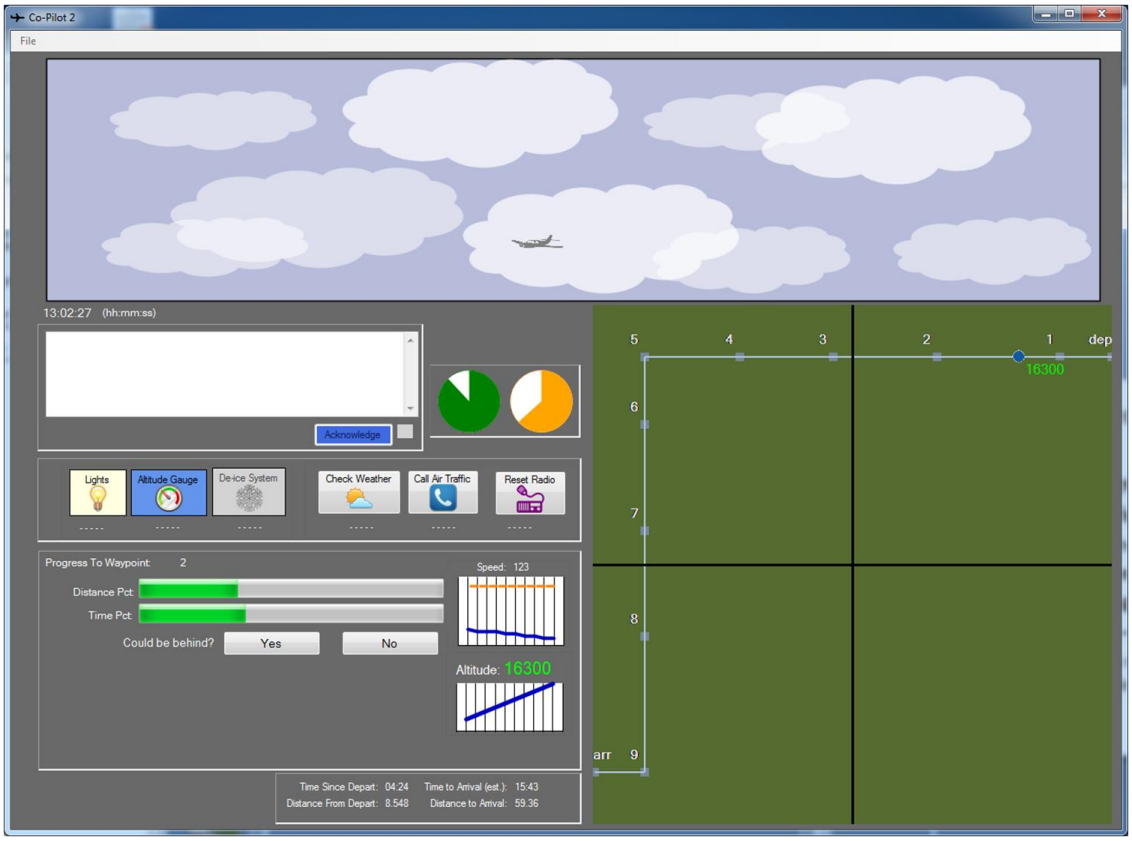


Figure 22: Non-intrusive Aid Updated Design

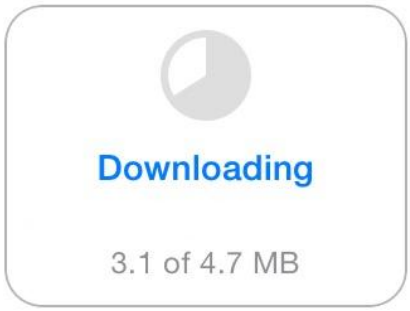


Figure 23: Apple iDevice Download Progress Indicator

The premise was that increased familiarity and intuitive presentation could reduce processing demands compared to the concentric circle graphical aid used in Experiments 1A and 1B.

To summarize, Experiment 2 differed from the prior experiments in the following ways:

1. The three primary task High workload conditions from Experiment 1B (Visual Search, Progress Assessment & Radio Query) was combined with the Low workload primary task condition from Experiment 1A (Visual Search & Progress Assessment).
2. PM difficulty was increased by including more "hard" PM tasks (8 of 12) for each scenario as compared to Experiments 1A and 1B (6 of 12)
3. PM non-intrusive aid was re-designed to reduce information processing demands that could have contributed to significantly lower primary task performance on Radio Query task in Experiment 1B.
4. Given the poor PM performance and subjective rating results for intrusive aid compared to non-intrusive Aid, it was eliminated as a level of Aiding in Experiment 2.

To review, the following Research Questions were addressed in Experiment 2:

- Primary Question
    - 3. Does the performance benefit of PM aiding across primary task workload levels justify the costs?
        - Hypothesis: No specific hypothesis, this is an outstanding empirical and theoretical question.
- Secondary Questions
    - 7. What would the impact of aiding be on primary task performance?

- Hypothesis: There will be no differential impact of primary task performance across aiding conditions.

Answering **Research Questions 3 & 7** also required answering the following supporting questions:

1. Did participants experience different levels of subjective workload between low and high primary task workload conditions within aiding condition, as measured by NASA TLX responses?
2. With the new aid design, was there a difference between Primary Task performance across aiding condition within workload condition (e.g., high workload: no-aiding compared to non-intrusive aid; low workload: no-aiding compared to non-intrusive aid)?

Experimental design was a 2 (aiding) x 2 (workload) repeated measures design (within subjects). The levels of the two independent variables were as follows.

Independent Variables

1. Aiding:
    a. No aid (Figure 5) baseline display
    b. Non-intrusive aid (Figure 8) baseline display with graphical timer
2. Primary Task Workload
    a. Low (low workload variants of Visual Search and Progress Assessment from Experiment 1A (see Table 3)
    b. High (high workload variants of Visual Search and Progress Assessment (see Table 3) plus Radio Query task from Experiment 1B (Figure 14) from Experiment 1B)

The experimental conditions are outlined as follows:

Workload/Aid	No aiding	Non-intrusive
Low	Low Workload No-Aid	Low Workload Non-intrusive Aid
High	High Workload No-Aid	High Workload Non-intrusive Aid

Table 35: Experiment 2 Design

Four flight scenarios from Experiment 1A were randomly selected for use in Experiment 2. Low workload primary tasks setup was identical to Experiment 1A and high workload primary tasks setup was identical to Experiment 1B. To assess subjective workload, participants completed the following select subscales of the NASA Task Load Index (TLX) subjective workload assessment tool after each trial (Hart and Staveland, 1988):

- Mental Demand – How mentally demanding was the task? (e.g., thinking, attending, remembering)
- Temporal Demand – How much time pressure did you feel in completing the task (e.g., relaxed pace or fast and furious?)
- Performance – How successful were you in performing the task?
- Effort – How hard did you have to work to accomplish your level of performance?
- Frustration – How irritated and stressed versus content and relaxed did you feel during the task?

Presentation order was counterbalanced with a 4-condition Latin Square design. See Table 36 for an example of a Latin Square for a 4 condition experiment requiring multiples of 4 participants. The conditions are coded as follows: aiding—intrusive (Int), non-intrusive (Non), no- aiding (No); workload—low (L) and high (H);

	Order			
Group	1	2	3	4
A	No-Low	NonInt-High	No-High	NonInt-Low
B	NonInt-High	No-Low	NonInt-Low	No-High
C	No-High	NonInt-Low	No-Low	NonInt-High
D	NonInt-Low	No-High	NonInt-High	No-Low

Table 36: 4-Condition Latin Square

Dependent Variables

The dependent variables were identical to those described in detail in Experiment 1A and 1B, aside from the addition of the NASA Task Load Index (TLX) subjective workload assessment. To review, variables are listed below:

- PM performance:
  - PM task percent correct (% of PM tasks successfully executed)
  - PM task reaction time (elapsed time from triggering situation to execution of action)
  - Total PM errors (see PM task response categories below)
- Primary task performance:
  - Visual Search
    - Percent correct (% of targets detected and clicked before disappearing)
    - Reaction Time (elapsed time from onset of target)
  - Progress Assessment
    - Percent correct (% of total trial time with correct assessment)
  - Radio Query
    - Percent correct (% of correct responses to queries)
- Subjective impression of the memory aids
  - Rating on custom survey described above that included items pertaining to aid intrusiveness, impact on primary task performance, and interference with native PM skills.
- Subjective Workload
    - Ratings on select subscales of the NASA Task Load Index (TLX) subjective workload assessment.

Participants

Four participants (2 female, 2 male) were recruited from the University of Minnesota community and were compensated $30. Participants' age ranged from 19 to 56 years with a mean age of 32.75 years. Participants completed 1 4-condition Latin Square counterbalance.

Procedure

The procedure was identical to Experiment 1A & 1B except for the following differences. Participant training included all elements in Experiment 1B training. In addition, participants performed low workload training scenarios similar to those described in Experiment 1A. Total training duration varied across participants between 40 and 70 minutes.

After each trial, participants rated their workload on the selected NASA TLX subscales along an incremented continuum from "Low" to "High" for all subscales except Performance which was anchored by "Good" and "Poor". Surveys were presented in paper form and participants circled the increment that corresponded to their perceived subjective workload. The NASA TLX administered to participants can be seen in the Appendix A.

Results

Graphs include standard error bars to enable visual comparison of condition means. Alpha (α) level of 0.05 was used for all analyses unless otherwise stated. Full ANOVA tables are in Appendix D. For post hoc t-tests, details are presented in a Comparisons table where p-value is bolded for significant results.

Subjective Workload Assessment

Subjective workload was assessed with the NASA TLX survey instrument. Participant selections were converted to an integer value between 0 ("Low", "Good") and 20 ("High", "Poor"). Median values were used for analyses and graphs. The small sample size (n=4) precluded the use of Wilcoxon Rank Sum Test for non-parametric data. When comparing the different levels of workload within aiding Level, standard error bars provide an approximation of statistical significance, provided they are non-overlapping. The critical subscale for subjective workload was Mental Demand. As illustrated by Figure 24 with non-overlapping standard error bars, participants rated the Mental Demand of both aiding conditions commensurately with the

workload manipulation of the primary tasks such that High workload conditions were rated higher than low workload conditions. This was not just an aggregated effect; within aiding condition, all participants rated High workload scenarios higher than low workload scenarios.

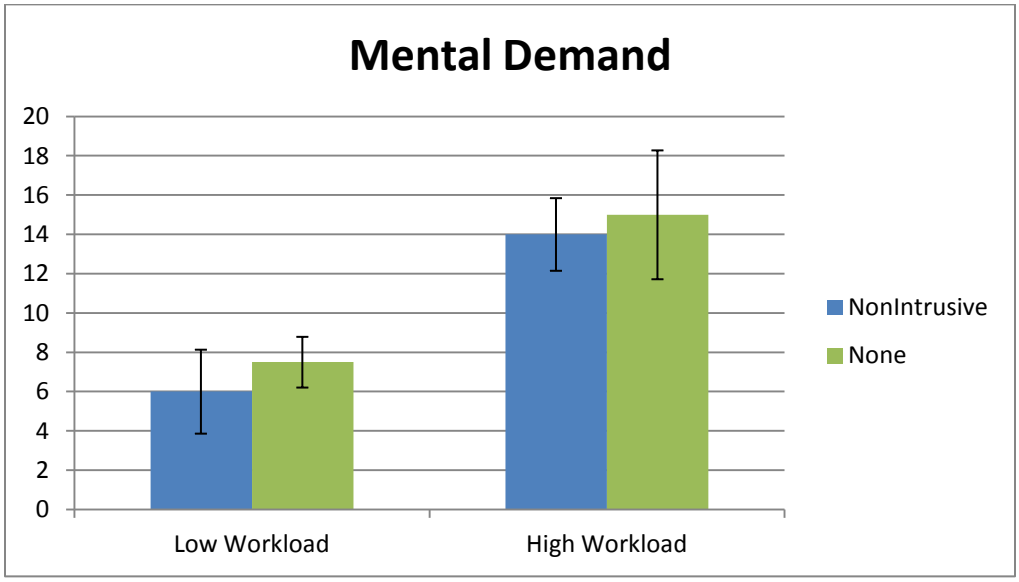


Figure 24: NASA TLX Mental Demand Ratings across Aiding and Workload Levels

A related measure is Effort, on which participants rated how hard they had to work to achieve their performance. While there is some overlap across workload levels for no aiding (none) condition, non-intrusive conditions error bars do not overlap across workload levels, such that high workload trials were rated as requiring higher effort than low workload trials, as depicted in Figure 25.

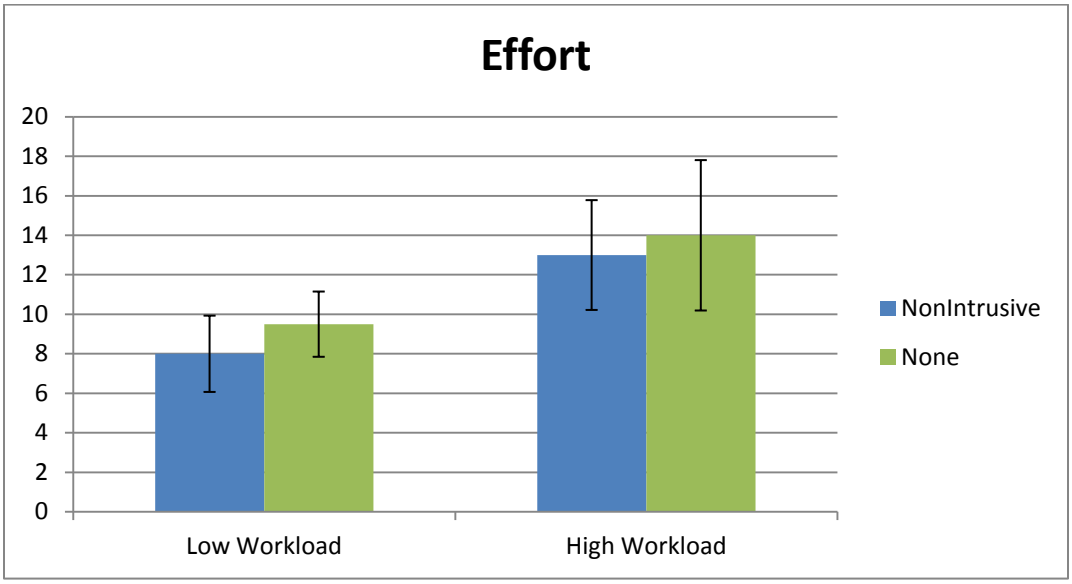


Figure 25: NASA TLX Effort Ratings across Aiding and Workload Levels

Considering the dynamic nature of the multitasking environment, Temporal Demand was also relevant. While there is some overlap of standard error bars within the no aiding condition, there is no overlap for across workload level for Non-intrusive conditions, such that participants rated higher Temporal Demand on high workload trials compared to low workload, as depicted in Figure 26.

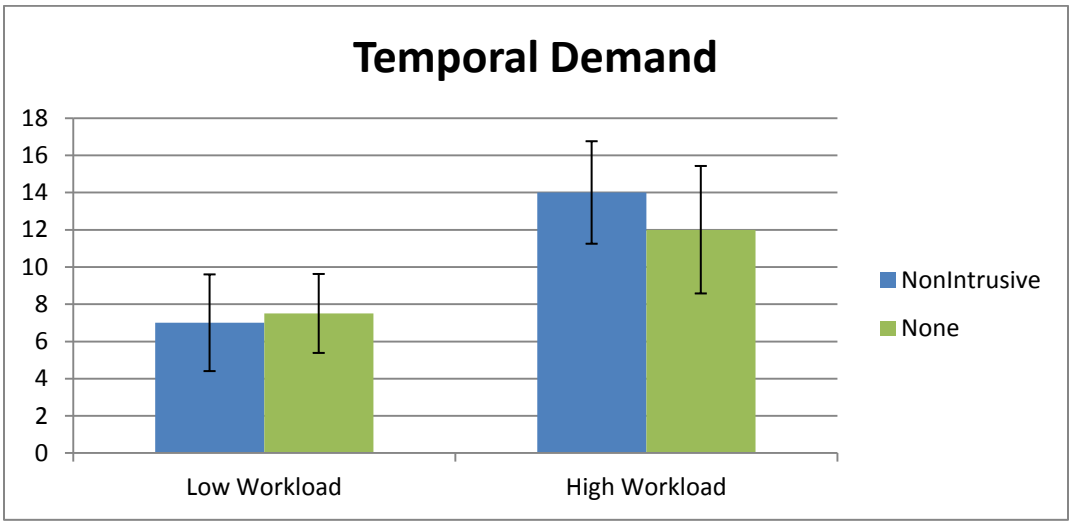


Figure 26: NASA TLX Temporal Demand Ratings across Aiding and Workload Levels

Results from other NASA TLX subscales will be discussed in a later section.

Differential Impact of Aiding Across Workload Conditions

Differential impact was operationalized as the percentage increase in correct PM tasks within aiding conditions between low and high workload. When investigating the differential impact of aiding across workload, PM performance improvement was as follows:

- Average of 95% more correct PM tasks with aiding under high workload
- Average of 24% more correct PM tasks with aiding under low workload

Figure 27 illustrates the differential benefit of aiding under high workload as compared to low workload.

Note: This is represented as error percentage as compared to error counts as in Experiment 1A and 1B since there were more "hard" PM tasks (8) than "easy" PM tasks (4) for Experiment 2; whereas there was an equal number of "hard" and "easy" PM tasks in Experiment 1A and 1B.

PM performance data was subjected to a 2 factor workload (2 levels) x aiding (2 levels) repeated-measures ANOVA. There was a marginally significant main effect of workload ($F(1) = 8.77$, $p = .059$), a significant main effect of aiding ($F(1) = 10.76$, $p < .05$), and a marginally significant workload x aiding interaction ($F(1) = 8.64$, $p = .060$), as illustrated in Table 38.

Again, the Bonferroni correction was applied for the following three comparisons such that significance level was 0.016 (($\alpha = 0.05$) /3). Analyses within workload level yielded significantly higher PM performance for non-intrusive aiding than no-aiding (None) (.94 compared to .52) under high workload ($t(3) = -5.48$, $p = 0.012$, paired T-test, two-tailed test); the effect size for this analysis ($d = 1.66$) was found to exceed Cohen's (1988) convention for a large effect ($d = .80$). Under low workload, non-intrusive aiding was not significantly higher (98% compared to 83%) ($t(3) = -1.27$, $p = .293$, paired T-test, two-tailed test). In fact, aiding supported PM performance under high workload that was statistically equivalent to low workload (94% compared to 98%) ($t(3) = -1.73$, $p = .181$, paired t-Test, two-tailed test). In terms of practical performance, participants successfully executed nearly twice the number of PM tasks while aided under high workload.

	None	Non-intrusive	Totals
Low Workload	0.833333	0.979167	0.90625
High Workload	0.520833	0.9375	0.729167
Totals	0.677083	0.958333	0.817708

Table 37: PM Percent Correct by Aiding and Workload Levels

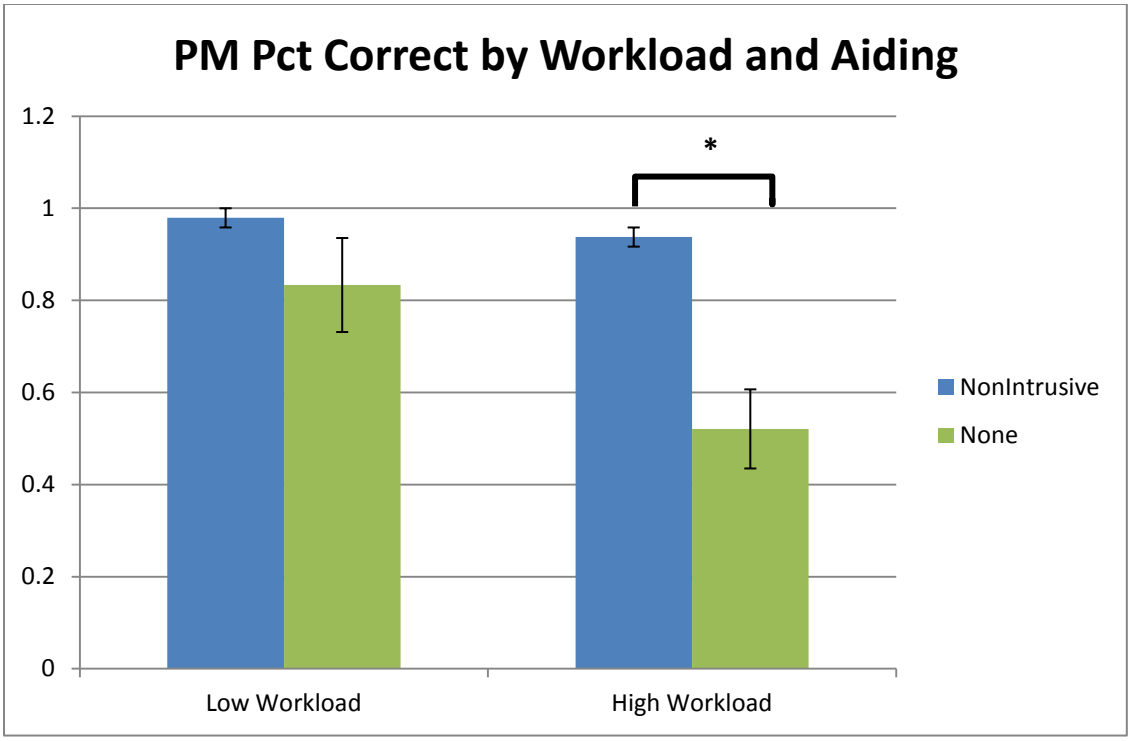


Figure 27: PM Percent Correct across Aiding and Workload Levels

Experiment 2 PM Task Percent Correct 2 Factor ANOVA Workload x Aid					
Source	SS	df	MS	F	P
Workload	0.1254	1	0.1254	8.7692	0.059483
Aid	0.3164	1	0.3164	10.7619	0.046407
Workload x Aid	0.0734	1	0.0734	8.6353	0.060579

Table 38: PM Percent Correct 2 Factor (Aiding, Workload) ANOVA

PM Task Pct Correct Comparisons					
Condition 1	Mean 1	Condition 2	Mean 2	p-value of t-test	Significance level
Non-intrusive High Workload	.94	None High Workload	.52	**0.012**	.016
Non-intrusive Low Workload	.98	None Low Workload	.83	0.293	.016
Non-intrusive High Workload	.94	Non-intrusive Low Workload	.98	0.181	.016

Table 39: PM Pct Correct Comparisons

Unlike in Experiment 1B, there was a substantial practical impact of primary task workload adaptation

All participants showed a benefit of aiding in high workload. There was an average of 5 fewer errors, a 42% reduction, with aiding compared to no aiding.

PM Errors	High Workload			
Participant	None	Non-intrusive	Difference	% Decrease in Errors
1	7	1	6	0.5
2	4	1	3	0.25
3	8	1	7	0.583333333
4	4	0	4	0.333333333
Total	**23**	**3**	**20**	
Average	**5.75**	**0.75**	**5**	**0.416666667**

Table 40: PM Errors

Next, participant reaction time to performing delayed action was assessed. A 2 factor workload (2 levels) x aiding (2 levels) repeated-measures ANOVA yielded a marginally significant main effect of aiding ($F(1) = 7.82$, $p = .068$), with non-intrusive levels (3178 msec) supporting faster reaction time than no aiding (6231 msec), as supported by Figure 28, Table 40, and Table 41.

PM RT	None	Non-intrusive	Totals
Low Workload	6115.638889	2961.715909	4538.677399
High Workload	6346.0625	3393.543561	4869.80303
Totals	6230.850695	3177.629735	4704.240215

Table 41: PM Task RT by Aiding and Workload Levels

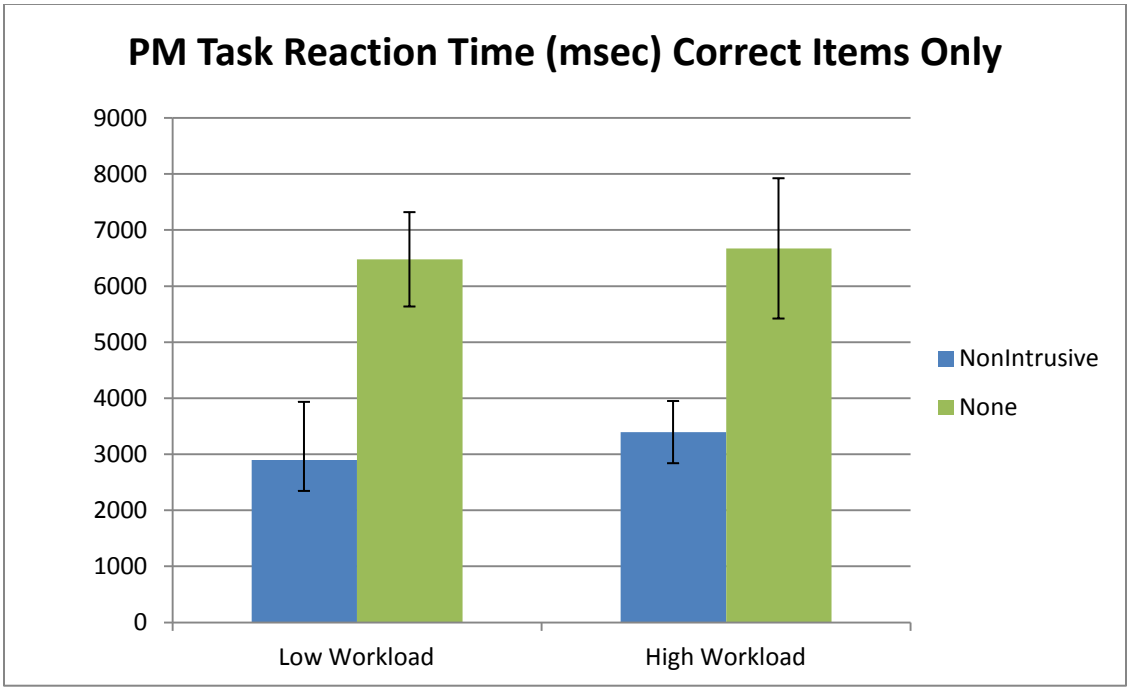


Figure 28: PM Task RT by Aiding and Workload Levels

Experiment 2 PM Task Reaction Time 2 Factor ANOVA Workload x Aid					
Source	SS	df	MS	F	P
Workload	438576.7344	1	438576.7344	0.2387	0.658633
Aid	37288632.9132	1	37288632.9132	7.8217	0.068037
Workload x Aid	40563.5874	1	40563.5874	0.1	0.772555

Table 42: PM Task RT 2 Factor (Aiding, Workload) ANOVA

Primary Task Impact

When comparing primary task performance across aiding conditions, it was evident that there was no systematic cost to aiding; in fact, non-intrusive condition performance on primary task within workload level is either higher than (Visual Search both workload Levels, as illustrated by Figure 29, Progress Assessment high workload), equivalent to (Radio Query), or is within .01 of no aiding performance.

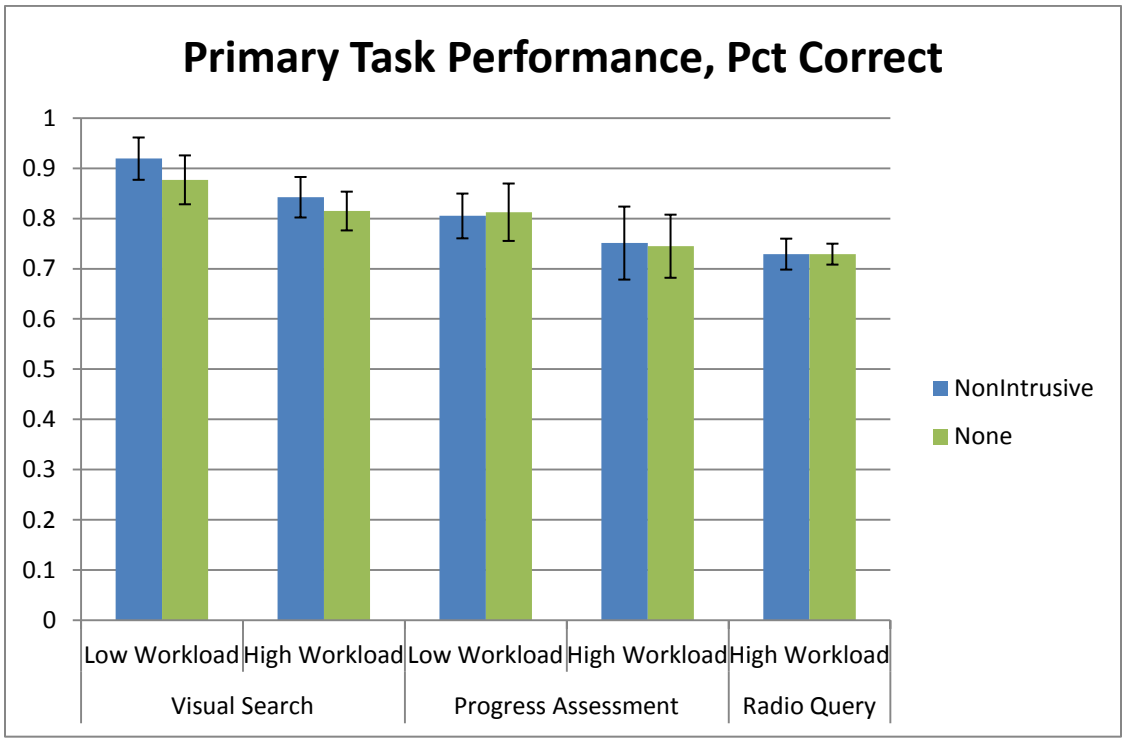


Figure 29: Primary Task Performance by Aiding and Workload Levels

To look at potential primary task performance cost of aiding, Visual Search RT data was collapsed across workload conditions. The 3 msec. difference between non-intrusive (1556 msec.) and no aiding(None) (1553 msec.) was not significant in a single factor repeated-measures ANOVA for aiding ($F(1)= .002$, $p= .97$), as seen in Figure 30.

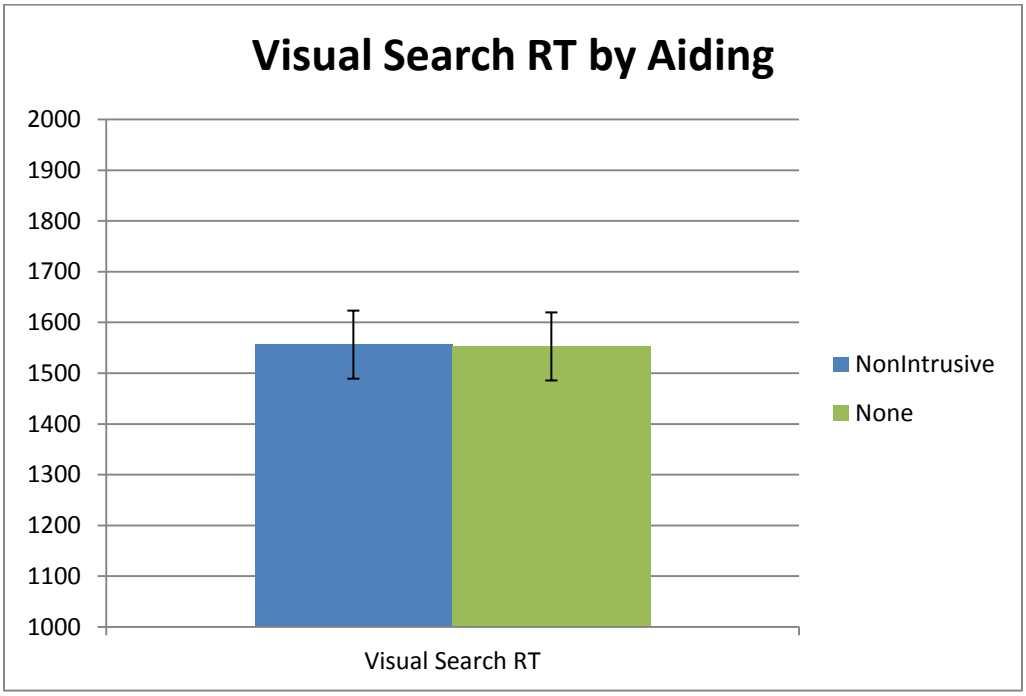


Figure 30: Visual Search RT

PM Error Profile

The PM error types were described previously and counts are summarized within rows by workload and aiding condition in Table 42. Of the 37 total errors, only 4 were committed during aiding trials; all of these errors involved participants performing the wrong action at the right time. Participants committed 27 errors under high workload and 10 errors under low workload. The highest error rate was 43% (14 of 37 errors) for Misses under the high workload no aiding condition.

	High		Low		Grand Total	
Row Labels	**None**	**Non-intrusive**	**None**	**Non-intrusive**		**% of Total Errors (37)**
Too Early, Right Action	9	0	6	0	15	0.405
Right Time, Wrong Action	1	3	1	1	6	0.162
Miss	14	0	2	0	16	0.432
% of Total Errors	**0.649**	**0.081**	**0.243**	**0.027**	**1**	
Error Totals	**24**	**3**	**9**	**1**	**37**	
Correct	24	45	39	47	155	
%Correct	**0.5**	**0.937**	**0.812**	**0.980**	**192**	
Grand Total	**48**	**48**	**48**	**48**	**192**	

Table 43: PM Error Type Counts across Aiding and Workload Levels

As in previous experiments, participants committed a higher percentage of errors on "hard" PM tasks, as illustrated in the following graphs. The pattern is evident for no aiding in Figure 31, while there is a floor effect (very few errors) for non-intrusive trials.

Note: This is represented as error percentage as compared to error counts as in Experiment 1A and 1B since there were more "hard" tasks (8) than "easy" tasks(4) for Experiment 2; whereas there was an equal number of "hard"(6) and "easy"(6) tasks in Experiments 1A and 1B.

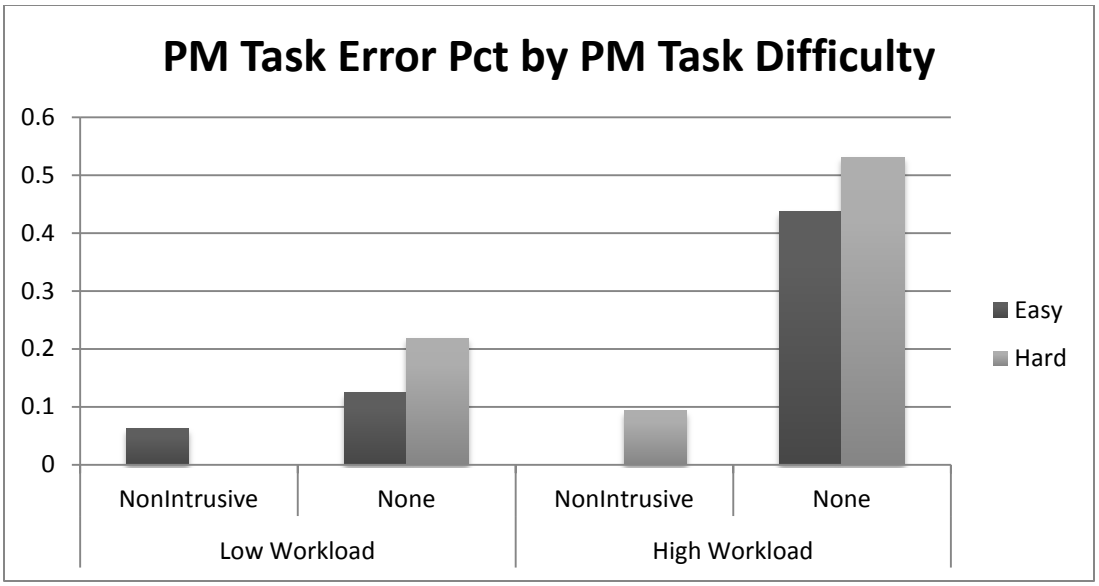


Figure 31: PM Error Pct by PM Task Difficulty across Aiding and Workload Levels

Other NASA TLX Subscales

Graphs of median responses across conditions for NASA TLX subscales Performance, Effort, and Frustration are presented below in Figure 32, Figure 33, and Figure 34 respectively. Please note the response endpoints for Performance are lowest= Good Performance while highest= Poor Performance. All subscales show trends of higher negative impact of high workload compared to low workload. A noteworthy result is non-overlapping standard error bars for low workload condition—higher Frustration for no aiding compared to non-intrusive aiding.

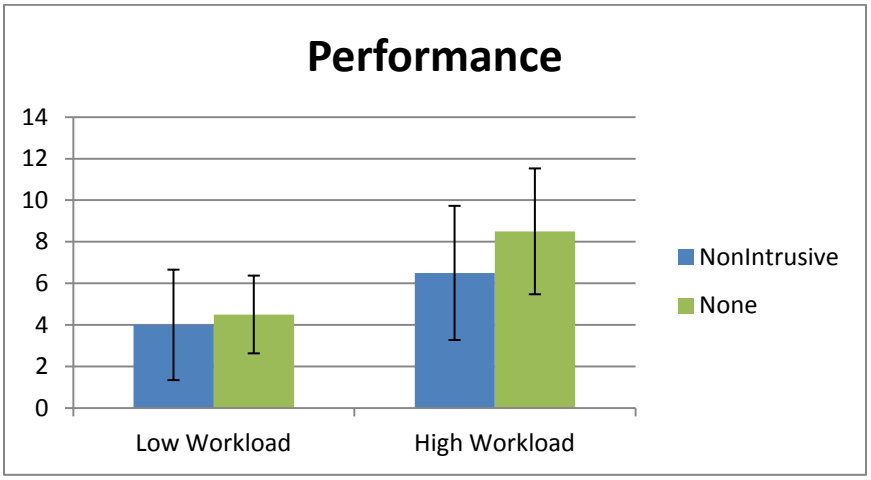


Figure 32: NASA TLX Performance Rating across Aiding and Workload Levels

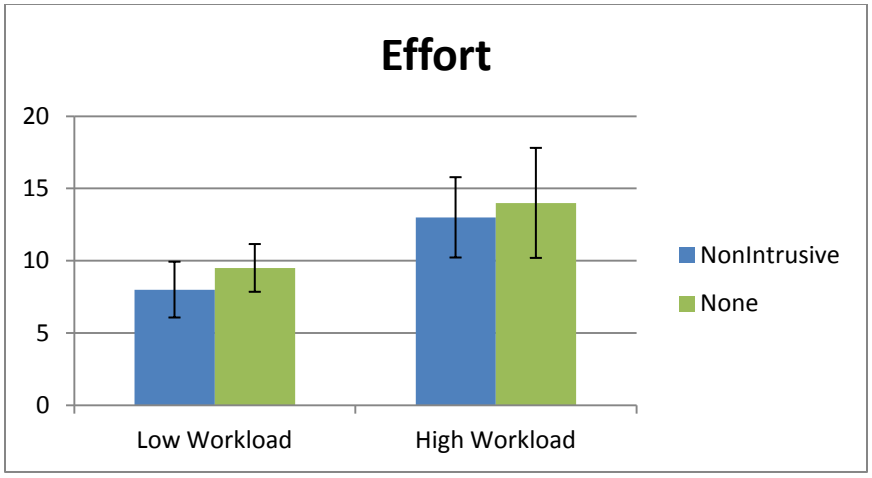


Figure 33: NASA TLX Effort Ratings across Aiding and Workload Levels

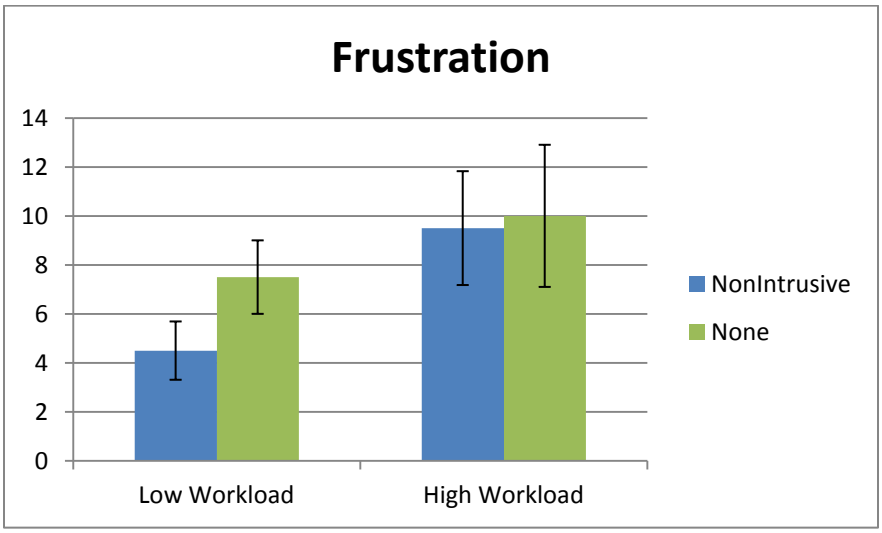


Figure 34: NASA TLX Frustration Rating across Aiding and Workload Levels

PM Aid Subjective Impression

Responses resembled those from Experiment 1B. Participants essentially reported no negative impact of aiding compared to no aiding (Items 1-3). When asked about Primary Task negative impact (items 7-9), the median response was 2.5 (between Disagree and Somewhat disagree). When asked about losing their concentration after attending the PM aid (item 6), the median

response was 2 (Disagree). Participants responded with the following rating scale and median responses are depicted in Figure 35:

- 1 = Strongly disagree
- 2 = Disagree
- 3 = Somewhat disagree
- 4 = Neither agree or disagree
- 5 = Somewhat agree
- 6 = Agree
- 7 = Strongly agree

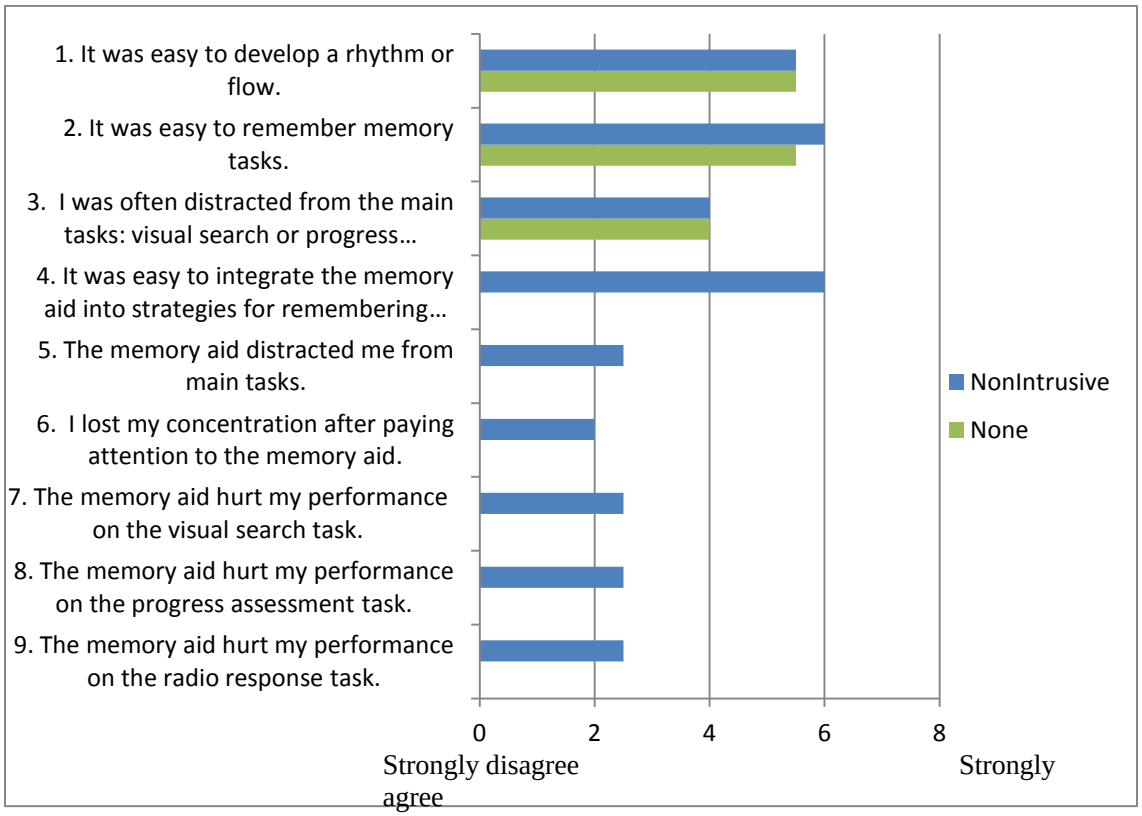


Figure 35: Survey Item Median Results by Aiding Levels

To compare to Experiments 1A and 1B, the median for item 4 was 6.5 and item 5 was 2.0, compared to medians of 6.0 and 2.5 in Experiment 2.

Discussion

Participant subjective workload ratings of Mental Demand, Effort, and Temporal Demand were commensurate with design—higher ratings for high workload as compared to low workload. This affirmatively answered supporting question 1:

- SQ1: Did participants experience different levels of workload between low and high workload Conditions within aiding Condition, as measure by NASA TLX responses?

This result provides confidence that differentiated workload levels were achieved with the primary task manipulation.

With the current aid design, primary task performance was not negatively impacted compared to no aiding condition. Unlike in Experiment 1B, non-intrusive aiding did not produce significantly lower Radio Query performance; in fact, performance was equivalent (73%). When looking at participant RT to Visual Search, aiding and no aiding conditions differed by less than 4 msec. At least for these tasks and measures, there was no cost to the aiding. Accordingly, the second research question was answered negatively; there was no difference, or cost, for primary task performance:

- With the new aid design, was there a difference between Primary Task performance within workload condition (high workload: No aid compared to non-intrusive aid; low workload: No aid compared to non-intrusive aid)?
- This confirmed the hypothesis for **Research Question 7**: There was no differential impact of primary task performance across aiding conditions.

This is important, since in addition to establishing a robust PM performance benefit to the aid, it is also critical that it does not incur a cost to primary task performance.

When evaluated in context of highly differentiated workload levels, the current PM aid realized a substantially greater benefit under high workload compared to low workload (95% more correct compared to 24%). The performance results relevant to **Research Question 3** were as follows The PM aiding benefit, defined as increase in PM performance between no aiding and non-intrusive aiding, was greater (71%) and statistically significant under high workload compared to low workload.

This significant performance benefit supports turning adaptive aiding "on" under high workload; further, the statistically equivalent performance of aiding and no aiding under low workload supports turning adaptive aiding "off" under low workload to achieve an appropriate balance between costs and benefits of PM aiding. This statistically and practically significant benefit supports the contention that adapting based on primary task workload Level could be effective in a complex task environment with complicated PM tasks. Thus, **Research Question 3** was affirmatively answered:

- There was no cost of aiding to primary task performance
- The statistical and practical benefits were sufficient to justify the observed and likely costs.

This will be discussed further in the General Discussion section.

It is also instructive to consider the profile of PM error types without aiding. As expected, 65% (24 of 37) of all errors were committed during high workload no aiding condition. Of the 24 errors, 23 (9- too early; 14- miss) were related to the timing of performing the delayed action; the remaining 1 error was a right time, wrong action type. Based on this profile, over 95% of errors are related to timing and are supported by the current aid design.

The most noteworthy result of Performance, Effort, and Frustration NASA TLX subscales, was the markedly higher Frustration reported for no aiding compared to non-intrusive aiding under low workload. This would suggest that, in addition to PM performance benefits, the non-intrusive aid ameliorated participant Frustration levels in performing complicated PM tasks within complex environment. This is also important since such subjective impressions would be important for fostering trust and acceptance of any aiding solution. While this is relevant to the aid design in general, it is worth noting that this particular benefit would not apply to an adaptive system where aiding was not provided under low workload.

General Discussion

Relationship to Prior Work

Retention Period

While the retention periods, which ranged from ~ 60 seconds to 3 minutes, were commensurate with much of the prior experimental work, this is still much shorter than most real-world intervals. For example, PM tasks from aviation could range from 10s of minutes to several hours. In personal lives, people frequently need to retain PM tasks until the following day which could be up to 24 hours. While such long intervals were not explored in these studies, it is not unreasonable to assume that the current timing-based aid could scale-up to support longer intervals.

PM Difficulty Manipulation

PM tasks were designed by manipulating factors known to impact PM difficulty such as cue salience, duration, and alignment with primary tasks (Einstein et al., 1992, McDaniel & Einstein, 2007). However, these factors had been validated independently in experimental task environments in which there was typically a single, simple primary task and a single, static PM task type; it was unclear if the manipulations would impact PM performance as expected when multiple factors were combined to create unique, complicated PM tasks that were performed concurrently with multiple, dynamic and complex primary tasks. The design manipulation produced consistent results across all three experiments:"hard" tasks were missed significantly more than "easy tasks". This was not just an aggregated effect, since all participants showed the same pattern, more errors on "hard" tasks compared to "easy" tasks, across all three experiments. This contribution should lend confidence to practitioners who want to scale-up PM tasks and primary task to increase ecological validity in more applied research.

Aid Design and Type

The current non-intrusive aid extended the reminder used by McDaniel et al. to include timing information about when an action was due, but also did not have information about the outstanding task (2004). McDaniel et al. effectively used a simple visual reminder, a small blue dot, to remind participants that they had a task to resume, effectively negating the impact of interruptions on participants' PM task execution (2004). Consistent with their findings, the current study found PM benefits despite providing no information about the action to perform. By reducing the cognitive resources required to monitor for the right time to perform delayed action, participants could dedicate more resources to remembering associative links between the aid and the action to be performed. This is supported by the fact that there were very few incorrect actions.

Participants' subjective impression of non-intrusive aid was favorable. The graphical nature of the aid could have reduced processing demands so there was less of a cognitive cost to attending to it. Unlike the intrusive aid, participants could attend to it on their own schedule, thus enabling better integration with ongoing tasks. For example, once participants developed a rhythm for performing the multiple primary tasks, it would be easy to add attending to the aid during lulls in task load or before or after performing a primary task action like scanning for traffic or updating progress assessment.

Primary Task Impact

Prior work has primarily focused on the impact of PM tasks on ongoing task performance (Smith, 2003; Loft & Remington, 2010). The current series of experiments did not consider this directly; instead it explored how different aiding conditions could minimize PM task costs on primary task performance. The primary focus of the reported work is on impact of PM aids, thus all conditions included PM tasks. Presumably there was some cost related to PM tasks, but all primary task performance measures were assessed with accompanying PM tasks. Accordingly, those performance measures all included some impact, likely negative, of PM tasks themselves plus the impact of aiding. Comparing primary task performance between no aiding and aiding conditions revealed an estimate of the impact. The initial non-intrusive aid induced significantly worse performance on the Radio Query task and lower average performance on Visual Search and Progress Assessment tasks in Experiment 1B. After re-designing the aid to be more intuitive and easier to process, there were no significant differences across any of the 3 primary tasks and non-intrusive aiding supported a higher or equivalent average performance score across all task and workload conditions except for Progress Assessment under low workload.

Differential Impact of Aiding Across Workload

Findings from Experiment 1A reiterated the importance of disparate cognitive workload levels when evaluating aiding/automation solutions. Aiding is often predicated on decreasing processing demands such that participants can support performance in spite of high primary task load. More practically speaking, if aiding performance is assessed by reduction in error it is critical to have a high enough error rate to distinguish between conditions; however, if the "high" primary task load condition does not adequately tap their resources, participants could have sufficient resources to complete the aided task, in this case the PM tasks, without the aid. This explanation is consistent with participant self-reports in which they reported frequent "lulls" in

primary task activity during which they could rehearse PM tasks. After the introduction of the radio query task, visual search performance declined and PM error rate increased. Visual search percent correct performance decreased from .95 in Experiment 1A to .80 in Experiment 1B. PM error rate increased, nearly doubling, from .21 in Experiment 1A to .40 in Experiment IB.

The Case for Adaptive Non-intrusive PM Aiding based on Primary Task Workload

Adaptive Aiding based on Workload Instead of PM Difficulty

With any aiding solution the cost/benefit ratio needs to be favorable. In Experiment 1B, the benefit of adaptive aiding based on PM task was assessed by comparing performance improvement when comparing "easy" and "hard" PM tasks across aiding conditions(e.g., relative difference). The lower the relative score, the better the support provided by the aiding since it has equalized the PM difficulty. The relative difference was significantly less for non-intrusive aiding. Despite this statistical difference, the practical impact of this was relatively modest in terms of number of errors due to the very low number of total errors for non-intrusive aiding conditions; in fact, 3 of 6 subjects had the same number of errors between "hard" and "easy" PM tasks.

Adaptive aiding based on workload could be less confusing and a more acceptable option than basing it in PM task difficulty for a number of reasons. First, the aiding function would be more transparent and understandable if it was either "On", under higher primary task load, or "off", under low primary task load. The potential problem with basing it on PM difficulty is that there could be overlapping PM tasks where one is aided and another is not. Second, it would likely be more predictable since operators could more easily identify the conditions that trigger adaptive aiding when it is primary task load since that should be very familiar to them; however, PM task difficulty would likely be determined by multiple factors such as duration, cue salience, and how well it aligns with the primary tasks. This would likely be more difficult to identify. To summarize, primary task workload-based adaptive aiding would be more transparent and predictable which are two factors known to impact operator trust and acceptance of automation (Muir, 1994; Lyons, 2014).

Favorable Subjective Impression and Impact of Non-intrusive Aid

For the critical item ("It was easy to integrate the memory aid into strategies for remembering memory task"), the median score for the updated design in Experiment 2 was 6.0 (6= "Agree"). Further, participants experienced improved PM performance which was immediate and recurring feedback confirming the usefulness of the aid. Given these factors, their positive impression is consistent with what would be expected from the Technology Acceptance Model findings that identified perceived usefulness and perceived ease of use as primary drivers of attitude effects (Davis, 1993). After re-designing the aid to be more intuitive and easier to process, there were no significant differences across any of the 3 primary tasks and non-intrusive aiding supported a higher or equivalent average performance score across all task and workload conditions except for low workload Progress Assessment.

Non-intrusive Aid Supported Superior PM Performance

There was the basic question regarding the effectiveness of a non-intrusive aid compared to an intrusive aid and no-aiding. In Experiment 1A, when there were not disparate primary task workload levels, non-intrusive aids did not support higher PM performance compared to no-aiding. This is consistent with prior work on the differential impact of workload on aiding benefit (Rouse, 1988; Dorneich et al., 2006). After the introduction of a third primary task in Experiments 1B & 2, non-intrusive aiding supported significantly higher PM performance than no-aiding under the higher workload.

Assessing Differential Benefit

In Experiment 2, the workload manipulation simulated a real-time adaptive aiding system by controlling primary task workload within trials; accordingly, the no-aiding low workload would simulate when adaptive aiding is "off" while the high workload non-intrusive aiding condition would simulate when aiding is "on". The benefit of this adaptive approach was established by the pattern of results: significant PM performance benefits under high primary task load but not under low primary task load. This addressed the question: whether the benefits under low workload justify the previously identified costs of an aiding system that is always "on". Costs include complacency, over-reliance, and degradation of native PM skills (Parasuraman et al., 1993; Sitka, 1999; Smith et al., 1997).

To validate the differential benefit, Experiment 2 combined the low workload of Experiment 1A with the high workload of Experiment 1B; NASA TLX scores confirmed that participants perceived disparate levels across the workload conditions. This provided confirmation that

addition of the third primary task induced workload levels commensurate with manipulation. To assess differential benefit, analyses compared no-aiding performance to non-intrusive aiding across workload levels. The premise is that significantly higher performance with aiding under high workload would be a benefit that justifies adaptive aiding to be "on" under high workload; conversely, non-significant difference between aiding conditions under low workload would mean the cost of aiding that is always "on" exceeds the incremental benefit. Experiment 2 analyses revealed a differential benefit such that non-intrusive aid PM performance was significantly higher than no-aiding performance for high workload only. In other terms, the current PM aid supported a substantially higher percent correct compared to no-aiding under high workload (95% more) compared to low workload (24% more).

Conclusions

In summary, there is converging evidence that supports the potential of adaptive PM aiding based on primary task workload. First, participants rated the non-intrusive aid design as easy to integrate into primary task workflow and not distracting to primary tasks. Second, the updated aid design had virtually no negative impact on primary task performance. Third, non-intrusive aiding provided differential benefit under high workload, as indicated by significantly higher PM performance under aiding than no-aiding for high workload only. Given the absence of observed costs and the differential impact across workload levels, we conclude that the cost/benefit analysis is in favor of adaptive aiding for high workload only. Moreover, these findings were investigated within experiments with complex, varied PM tasks embedded within more realistic primary task loading. As is the case when transitioning from laboratory research to the field, the nature of PM aid support should be tailored to the target work domain. The following section details recommendations for how developers of PM aid should apply these research finding.

Recommendations for Developers of PM Aids (re-ordered see Defense Preso)

Based on evidence reported in this work, developers of PM aids are recommended to:

- First determine if they should expect sufficient benefits to operational efficiency and safety to justify the costs. This is critical since there are always costs to aiding solutions.

- Consider PM aids that support/offload long-term monitoring for triggering situations. The non-intrusive aid in the current research likely supported a strategy that allowed participants

to offload long-term monitoring to the graphical aid; real-world retention periods will likely be much longer and operators could benefit from support.

- Consider non-intrusive PM aids that are simple, peripheral, and persistent graphical aids. Under the current work with such an aid, PM performance benefits were established, participants rated them positively, and primary task impact was minimal.

- Consider PM aids that provide the minimal support necessary to realize benefits. Current research established that timing information only could support improved performance. The minimal support would also be less likely to induce complacency and skill degradation.

- One way to minimize support is to provide aiding only when primary task workload is high. Current work established differential benefit under high workload only. This targets aiding when operators would need it most. This simple, predictable adaptation scheme should be easily understood by operators.

- Investigate whether computer-based task environments are amenable to primary task workload estimation based on operator behavior and/or historical workflows.

- Consider both observed and predicted workload since PM performance can be impacted by high workload during retention and retrieval too. In the current work, primary task loading was fairly consistent across the entire experiment such that there was probably equivalent workload, either low or high, at encoding, retention, and retrieval. Estimating primary task load during retention and retrieval periods for use in adaptation will probably be more feasible in domains with very regular workflows.

Bibliography

Adda, C. C., Castro, L. H. ., Além-Mar e Silva, L. C., De Manreza, M. L., & Kashiara, R. (2008). Prospective memory and mesial temporal epilepsy associated with hippocampal sclerosis. *Neuropsychologia*, *46*(7), 1954–1964.

Altgassen, M., Kretschmer, A., & Kliegel, M. (2014). Task dissociation in prospective memory performance in individuals with ADHD. Journal of Attention Disorders, 18(7), 617–624.

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence*, *42*(1), 7–49.

Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. Trends in Cognitive Sciences, 5(12), 533–539.

Burgess, P. W., & Shallice, T. (1997). The relationship between prospective and retrospective memory: Neuropsychological evidence. *Cognitive models of memory*, 247–272.

Ceci, S. J., & Bronfenbrenner, U. (1985). " Don't Forget to Take the Cupcakes out of the Oven": Prospective Memory, Strategic Time-Monitoring, and Context. *Child development*, 152–164.

Cheng, H. D., Wang, K., Xi, C., Niu, C., & Fu, X. M. (2008). Prefrontal cortex involvement in the event-based prospective memory: Evidence from patients with lesions in the prefrontal cortex. *Brain Injury*.

Cohen, A. L., Jaudas, A., & Gollwitzer, P. M. (2008). Number of cues influences the cost of remembering to remember. *Memory & cognition*, *36*(1), 149–156.

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *MEMORY-HOVE-*, *4*, 577–590.

Corbett, A. (1988). An Intelligent Tutoring System. *Journal article by Albert Corbett; THE Journal (Technological Horizons In Education)*, *16*.

Craik, F.I.M. (1986). A functional account of age differences in memory. In F. Klix & H. Hagendorf
(Eds.), Human memory and cognitive capabilities: Mechanisms and performances (pp. 409–422).
Amsterdam: Elsevier North-Holland.

Craik, F. I., & Bialystok, E. (2006). Planning and task management in older adults: Cooking breakfast. Memory & Cognition, 34(6), 1236–1249.

Czerwinski, M., Horvitz, E., & Wilhite, S. (2004). A diary study of task switching and interruptions. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 175–182).

Davis, F. D. (1993). User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. International Journal of Man-Machine Studies, 38(3), 475–487.

Dieckmann, P., Reddersen, S., Wehner, T., & Rall, M. (2006). Prospective memory failures as an unexplored threat to patient safety: results from a pilot study using patient simulators to investigate the missed execution of intentions. *Ergonomics*, *49*(5-6), 526–543.

Dismukes, R. K., Loukopoulos, L. D., & Jobe, K. (2001). The Challenges of Managing Concurrent and Deferred Tasks. In R. S. Jensen (Ed.), Proceedings of the Eleventh International Symposium on Aviation Psychology. Columbus,OH: The Ohio State University.

Dismukes, R. K. (2006). Concurrent task management and prospective memory: pilot error as a model for the vulnerability of experts. *50th annual human factors and ergonomics society conference* (Vol. 50, pp. 909–913).

Dismukes, R. K. (2008). Prospective memory in aviation and everyday settings. *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives*, 411–428.

Dismukes, R. K. (2010). Remembrance of Things Future: Prospective Memory in Laboratory, Workplace, and Everyday Settings. *Reviews of Human Factors and Ergonomics*, *6*(1), 79–122.

Dodhia, R. M., & Dismukes, R. K. (2005). A task interrupted becomes a prospective memory task. *Biennial Meeting of the Society for Applied Research in Memory and Cognition, Wellington, New Zealand*.

Dodhia, R. M., & Dismukes, R. K. (2009). Interruptions create prospective memory tasks. *Applied Cognitive Psychology*, *23*(1), 73–89.

Dorneich, M. C., Ververs, P. M., Mathan, S., & Whitlow, S. D. (2005). A joint human-automation cognitive system to support rapid decision-making in hostile environments. *Systems, Man and Cybernetics, 2005 IEEE International Conference on* (Vol. 3, pp. 2390–2395).

Dorneich, M. C., Ververs, P. M., Whitlow, S. D., Mathan, S., Carciofini, J., & Reusser, T. (2006). Neuro-physiologically-driven adaptive automation to improve decision making under stress.

In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 50, pp. 410–414).

Einstein, G. O, & McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 717.

Einstein, G. O., & McDaniel, M. A. (1996). Retrieval processes in prospective memory: Theoretical approaches and some new empirical findings. In M. A. Brandimonte, G. O. Einstein, & M. A. McDaniel(Eds.), Prospective Memory: Theory and Applications, 115–141. Hillsdale, NJ: Erlbaum.

Einstein, G. O, McDaniel, M. A., Manzi, M., Cochran, B., & Baker, M. (2000). Prospective memory and aging: Forgetting intentions over short delays. *Psychology and Aging*, *15*(4), 671.

Einstein, G. O, McDaniel, M. A., Smith, R. E., & Shaw, P. (1998). Habitual prospective memory and aging: Remembering intentions and forgetting actions. *Psychological Science*, *9*(4), 284.

Einstein, G. O, McDaniel, M. A., Thomas, R., Mayfield, S., Shank, H., Morrisette, N., & Breneiser, J. (2005). Multiple processes in prospective memory retrieval: factors determining monitoring versus spontaneous retrieval. *Journal of Experimental Psychology: General*, *134*(3), 327.

Einstein, G. O, Smith, R. E., McDaniel, M. A., & Shaw, P. (1997). Aging and prospective memory: The influence of increased task demands at encoding and retrieval. *Psychology and Aging*, *12*(3), 479.

Einstein, G.O., McDaniel, M. A., Williford, C. L., Pagan, J. L., & Dismukes, R. (2003). Forgetting of intentions in demanding situations is rapid. *Journal of Experimental Psychology: Applied*, *9*(3), 147.

Einstein, Gilles O., Holland, L. J., McDaniel, M. A., & Guynn, M. J. (1992). Age-related deficits in prospective memory: The influence of task complexity. *Psychology and Aging*, *7*(3), 471–478. doi:10.1037/0882-7974.7.3.471

Eldridge, M., Sellen, A., & Bekerian, D. (1992). *Memory Problems at Work: Their Range, Frequency, and Severity* (No. EPC-1992-129). Rank Xerox Research Centre.

Ellis, J. (1996). Prospective memory or the realization of delayed intentions: A conceptual framework for research. Prospective Memory: Theory and Applications, 1–22.

Ellis, J., Kvavilashvili, L., & Milne, A. (1999). Experimental tests of prospective remembering: The influence of cue-event frequency on performance. *British Journal of Psychology, 90*(1), 9–23.

Fink, N., Pak, R., Bass, B., Johnston, M., & Battisto, D. (2010). A Survey of Nurses Self-Reported Prospective Memory Tasks: What Must they Remember and What do they Forget? Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 54, pp. 1600–1604).

Gawande, A. A., Studdert, D. M., Orav, E. J., Brennan, T. A., & Zinner, M. J. (2003). Risk factors for retained instruments and sponges after surgery. *New England Journal of Medicine*, *348*(3), 229–235.

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American Psychologist*, *54*(7), 493.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology*, *38*, 69–119.

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI magazine*, *22*(4), 39.

Guynn, M. J. (2003). A two-process model of strategic monitoring in event-based prospective memory: Activation/retrieval mode and checking. *International Journal of Psychology*, *38*(4), 245–256.

Guynn, Melissa J., Mark A. McDaniel, and Gilles O. Einstein. "Prospective Memory: When Reminders Fail." Memory & Cognition 26, no. 2 (1998): 287–98.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in Psychology, 52, 139–183.

Helmreich, R. L., Wilhelm, J. A., Klinect, J. R., & Merritt, A. C. (2001). Culture, error and crew resource management. *Improving teamwork in organizations: Applications of resource management training*, 305–331.

Higgins, J. J. (2003). Introduction to modern nonparametric statistics.

Holbrook, J. B., Dismukes, R. K., & Nowinski, J. L. (2005). Identifying sources of variance in everyday prospective memory performance. *Biennial Meeting of the Society for Applied Research in Memory and Cognition, Wellington, New Zealand*.

Kidder, D. P., Park, D. C., Hertzog, C., & Morrell, R. W. (1997). Prospective memory and aging: The effects of working memory and prospective memory task load. *Aging, Neuropsychology, and Cognition*, *4*(2), 93–112.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., Mark, M. A., & others. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education (IJAIED)*, *8*, 30–43.

Kvavilashvili, L., & Fisher, L. (2007). Is time-based prospective remembering mediated by self-initiated rehearsals? Role of incidental cues, ongoing activity, age, and motivation. *Journal of Experimental Psychology: General*, *136*(1), 112.

Kvavilashvili, Lia. (1992). Remembering intentions: A critical review of existing experimental paradigms. *Applied Cognitive Psychology*, *6*(6), 507–524. doi:10.1002/acp.2350060605

Loft, S., & Remington, R. W. (2010). Prospective memory and task interference in a continuous monitoring dynamic display task. Journal of Experimental Psychology: Applied, 16(2), 145.

Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2001). Cockpit interruptions and distractions: A line observation study. *Proceedings of the 11th International Symposium on Aviation Psychology*.

Lyons, Joseph B. "Being Transparent about Transparency: A Model for Human-Robot Interaction." In 2013 AAAI Spring Symposium Series, 2013.

Marsh, R. L., & Hicks, J. L. (1998). Event-based prospective memory and executive control of working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(2), 336.

McDaniel, M. A., & Einstein, G. O. (2007). *Prospective memory: An overview and synthesis of an emerging field*. Sage Publications, Inc.

McDaniel, M. A., Einstein, G. O., Graham, T., & Rall, E. (2004). Delaying execution of intentions: Overcoming the costs of interruptions. *Applied Cognitive Psychology*, *18*(5), 533–547.

McDaniel, M. A., Howard, D. C., & Butler, K. M. (2008). Implementation intentions facilitate prospective memory under high attention demands. Memory & Cognition, 36(4), 716–724.

McDaniel, M. A., Robinson-Riegler, B., & Einstein, G. O. (1998). Prospective remembering: Perceptually driven or conceptually driven processes? *Memory & Cognition*, *26*(1), 121–134.

Mcgann, D., Ellis, J. A., & Milne, A. (2002). Conceptual and perceptual processes in prospective remembering: Differential influence of attentional resources. *Memory & cognition*, *30*(7), 1021–1032.

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

Muir, Bonnie M. "Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems." Ergonomics 37, no. 11 (1994): 1905–22.

Nowinski, J. L., & Dismukes, K. (2005). Effects of ongoing task context and target typicality on prospective memory performance: The importance of associative cueing. *Memory*, *13*(6), 649–657.

Nowinski, J. L., Holbrook, J. B., & Dismukes, R. K. (2003). Human memory and cockpit operations: An ASRS study. *Proceedings of the 12th International Symposium on Aviation psychology* (pp. 888–893).

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced'complacency'. *The International Journal of Aviation Psychology*, *3*(1), 1–23.

Patient Safety Monitor Alert. (2003). *Emergencies, Procedure Changes Contribute to Left-Behind Surgical Instruments*. HCPRO.

Raley, C., Stripling, R., Kruse, A., Schmorrow, D., & Patrey, J. (2004). Augmented Cognition overview: Improving information intake under stress. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, pp. 1150–1154).

Riley, V. (1996). Operator reliance on automation: Theory and data. *Automation and human performance: Theory and applications*, 19–35.

Rouse, W. B. (1988). Adaptive aiding for human/computer control. Human Factors: The Journal of the Human Factors and Ergonomics Society, 30(4), 431–443.

Scerbo, M. (2007). Adaptive automation. *Neuroergonomics: The brain at work*, 239–252.

Schutte, P. C., & Trujillo, A. C. (1996). Flight crew task management in non-normal situations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 40, pp. 244–248).

Scullin, M. K., McDaniel, M. A., & Einstein, G. O. (2010). Control of cost in prospective memory: Evidence for spontaneous retrieval processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 190.

Sellen, A. J., Louie, G., Harris, J. E., & Wilkins, A. J. (1997). What brings intentions to mind? An in situ study of prospective memory. *Memory*, *5*(4), 483–507.

Singh, I. L., Molloy, R., & Parasuraman, R. (1997). Automation-induced monitoring inefficiency: role of display location. *International Journal of Human-Computers Studies*, *46*(1), 17–30.

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

Sleeman, D., & Brown, J. S. (1982). Intelligent tutoring systems. *Intelligent tutoring systems* (Vol. 5091).

Smith, P. J., McCoy, C. E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *27*(3), 360–371.

Smith, R. E. (2003). The cost of remembering to remember in event-based prospective memory: investigating the capacity demands of delayed intention performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 347.

Smith, R. E., & Bayen, U. J. (2005). The effects of working memory resource availability on prospective memory: a formal modeling approach. *Experimental Psychology*, *52*(4), 243.

Stone, M., Dismukes, K., & Remington, R. (2001). Prospective memory in dynamic environments: Effects of load, delay, and phonological rehearsal. *Memory*, *9*(3), 165–176.

Warschauer, M. (n.d.). Mikey's Story. *Mikey's Story*. Retrieved April 20, 2012, from http://www.4rkidssake.org/mikeysstory.htm

West, R., & Craik, F. I. . (1999). Age-related decline in prospective memory: The roles of cue accessibility and cue sensitivity. *Psychology and Aging; Psychology and Aging*, *14*(2), 264.

Whitlow, S.D. (2015, April 16). Personal interview with corporate test pilots.

Wichman, H., & Oyasato, A. (1983). Effects of locus of control and task complexity on prospective remembering. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *25*(5), 583–591.

Wickens, C. D., Gordon, S. E., & Liu, Y. (2004). *An introduction to human factors engineering*. Pearson Prentice Hall.

Appendix A

Participant Experience Survey

Please rate your experience in the most recently completed trial. Indicate your agreement with the following statements according to scale depicted below.

1 = Strongly disagree **2** = Disagree **3** = Somewhat disagree

4 = Neither agree or disagree

5 = Somewhat agree **6** = Agree **7** = Strongly agree

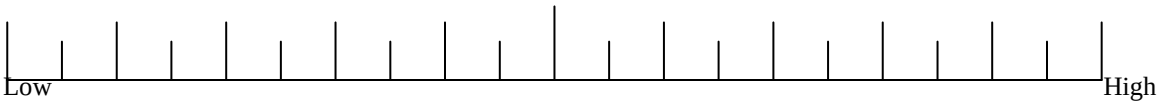
1	It was easy to develop a rhythm or flow.	1 2 3 4 5 6 7 ☐ ☐ ☐ ☐ ☐ ☐ ☐
2	It was easy to remember memory tasks.	1 2 3 4 5 6 7 ☐ ☐ ☐ ☐ ☐ ☐ ☐
3	I was often distracted from the main tasks: visual search or progress assessment.	1 2 3 4 5 6 7 ☐ ☐ ☐ ☐ ☐ ☐ ☐
4	It was easy to integrate the memory aid into strategies for remembering memory tasks.	1 2 3 4 5 6 7 ☐ ☐ ☐ ☐ ☐ ☐ ☐
5	The memory aid distracted me from main tasks.	1 2 3 4 5 6 7 ☐ ☐ ☐ ☐ ☐ ☐ ☐
6	I lost my concentration after paying attention to the memory aid.	1 2 3 4 5 6 7 ☐ ☐ ☐ ☐ ☐ ☐ ☐
7	The PM aid hurt my performance on the visual search task.	1 2 3 4 5 6 7 ☐ ☐ ☐ ☐ ☐ ☐ ☐
8	The PM aid hurt my performance on the progress assessment task.	1 2 3 4 5 6 7 ☐ ☐ ☐ ☐ ☐ ☐ ☐

Trial:___
Condition: ___

NASA Task Load Index (TLX)

Please circle the line which corresponds to the workload experienced during the last trial only.

Mental Demand – How mentally demanding was the task? (e.g., thinking, attending, remembering)

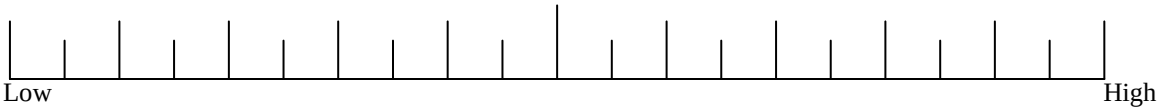


Temporal Demand – How much time pressure did you feel in completing the task (e.g., relaxed pace or fast and furious?)

Performance - How successful were you in performing the task?

Effort - How hard did you have to work to accomplish your level of performance?

Frustration – How irritated and stressed versus content and relaxed did you feel during the task?



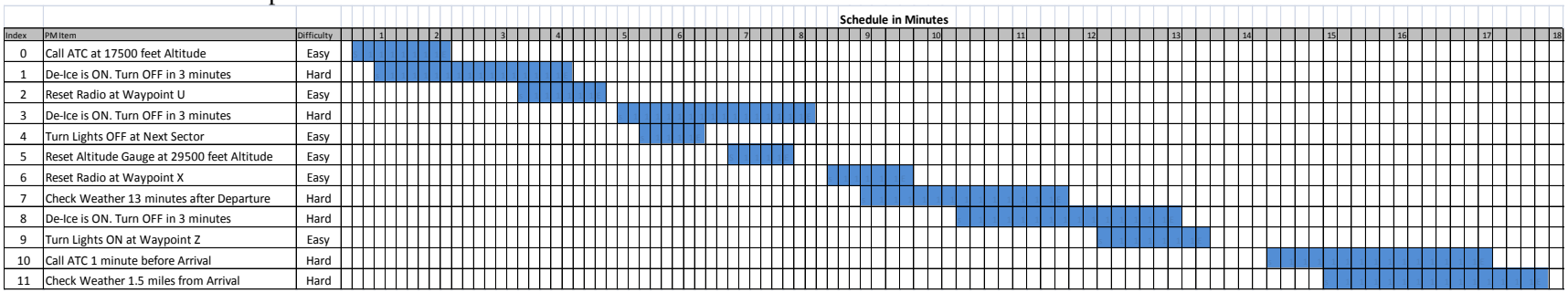
Appendix B

Experimental Scenario Data Example

xpositio n	ypositio n	waypoin t	altitude	speed	ownship timer	distance frompre vious	distance tonext	totaldist ancebet ween	percent agedista nceone xt	totaltim etonext waypoin t	totaltim etonext waypoin tsecond s	elapsedt imeinmi nutesse condstri ng	estimate dtimeHd r	estimate dtimein millisco ndsHdr	percent ageesti matedti me	elapsedt imeinmil liseond s	remaini ngtime	percent agetime	speedch angeper centage	speedch angetre nd	progress assessm entgrou ndtruth	assessm entsche dule	timeSinc eDepart	timeToA rrival	distance SinceDe part	distance ToArriva l	wptTran sition	pmActiv eItems	pmintru siveMiti gationSc hedule	targetSp eedForS egment Schedul e	progress assessm entgrou ndtruth 2
2	60	1	1000	100	1478	0	40	40	0	57108	57.1	0:00	105	65674	0	0	57108	0	0	0	2	0	2:00	18:14	2	65.91	0	0	0	68	9
3	60	1	1191	100	1479	1	39	40	0.025	57108	57.1	0:01	105	65674	0.022	1478	55685	2.578	0	0	3	0	2:00	18:14	2.061	65.85	0	0	0	68	9
4	60	1	1388	101	1468	2	38	40	0.05	57108	57.1	0:02	105	65674	0.044	2946	54162	5.158	0	0	3	0	2:01	18:12	2.122	65.78	0	0	0	68	9
5	60	1	1575	101	1468	3	37	40	0.075	57108	57.1	0:04	105	65674	0.067	4414	52694	7.729	0	0	3	0	2:02	18:11	2.183	65.72	0	0	0	68	9
6	60	1	1767	102	1469	4	36	40	0.1	57108	57.1	0:05	105	65674	0.089	5882	51226	10.29	2	1	3	0	2:04	18:09	2.244	65.66	0	0	0	68	9
7	60	1	1959	102	1469	5	35	40	0.125	57108	57.1	0:07	105	65674	0.111	7345	49768	12.86	1	0	3	0	2:05	18:08	2.306	65.6	0	0	0	68	9
8	60	1	2151	103	1457	6	34	40	0.15	57108	57.1	0:08	105	65674	0.134	8808	48300	15.42	2	1	3	0	2:07	18:07	2.367	65.54	0	0	0	68	9
9	60	1	2348	103	1457	7	33	40	0.175	57108	57.1	0:10	105	65674	0.156	10265	46843	17.97	1	0	3	0	2:08	18:05	2.428	65.48	0	0	0	68	9
10	60	1	2534	104	1452	8	32	40	0.2	57108	57.1	0:11	105	65674	0.178	11722	45386	20.52	2	1	3	0	2:10	18:04	2.489	65.42	0	0	0	68	9
11	60	1	2726	104	1452	9	31	40	0.225	57108	57.1	0:13	105	65674	0.2	13174	43934	23.06	1	0	3	0	2:11	18:02	2.55	65.36	0	0	0	68	9
12	60	1	2918	104	1452	10	30	40	0.25	57108	57.1	0:14	105	65674	0.222	14626	42482	25.61	1	0	3	0	2:13	18:01	2.612	65.3	0	0	0	68	9
13	60	1	3110	105	1447	11	29	40	0.275	57108	57.1	0:16	105	65674	0.244	16078	41030	28.15	1	0	3	0	2:14	17:59	2.673	65.23	0	0	0	68	9
14	60	1	3302	105	1447	12	28	40	0.3	57108	57.1	0:17	105	65674	0.266	17525	39578	30.68	1	0	3	0	2:16	17:58	2.734	65.17	0	0	0	68	9
15	60	1	3494	106	1442	13	27	40	0.325	57108	57.1	0:18	105	65674	0.288	18972	38126	33.22	2	1	3	0	2:17	17:56	2.795	65.11	0	0	0	68	9
16	60	1	3686	106	1442	14	26	40	0.35	57108	57.1	0:20	105	65674	0.31	20414	36674	35.74	1	0	3	0	2:18	17:55	2.856	65.05	0	0	0	68	9
17	60	1	3877	107	1436	15	25	40	0.375	57108	57.1	0:21	105	65674	0.332	21856	35252	38.27	2	1	3	0	2:20	17:53	2.918	64.99	0	0	0	68	9
18	60	1	4069	107	1436	16	24	40	0.4	57108	57.1	0:23	105	65674	0.354	23292	33816	40.78	1	0	3	0	2:21	17:52	2.979	64.93	0	0	0	68	9
19	60	1	4261	108	1431	17	23	40	0.425	57108	57.1	0:24	105	65674	0.376	24728	32380	43.3	2	1	3	0	2:23	17:51	3.04	64.87	0	0	0	68	9
20	60	1	4458	108	1431	18	22	40	0.45	57108	57.1	0:26	105	65674	0.398	26159	30948	45.8	1	0	3	0	2:24	17:49	3.101	64.81	0	0	0	68	9
21	60	1	4645	108	1431	19	21	40	0.475	57108	57.1	0:27	105	65674	0.42	27590	29518	48.31	1	0	3	0	2:26	17:48	3.162	64.74	0	0	0	68	9
22	60	1	4837	109	1426	20	20	40	0.5	57108	57.1	0:29	105	65674	0.441	29021	28087	50.81	1	0	3	0	2:27	17:46	3.224	64.68	0	0	0	68	9
23	60	1	5029	109	1426	21	19	40	0.525	57108	57.1	0:30	105	65674	0.463	30447	26661	53.31	1	0	3	0	2:28	17:45	3.285	64.62	0	0	0	68	9
24	60	1	5220	110	1421	22	18	40	0.55	57108	57.1	0:31	105	65674	0.485	31878	25285	55.81	2	1	3	0	2:30	17:43	3.346	64.56	0	1	0	68	9
25	60	1	5412	110	1421	23	17	40	0.575	57108	57.1	0:33	105	65674	0.506	33294	23914	58.3	1	0	3	0	2:31	17:42	3.407	64.5	0	1	0	68	9
26	60	1	5604	111	1415	24	16	40	0.6	57108	57.1	0:34	105	65674	0.528	34715	22899	60.78	2	1	3	0	2:33	17:41	3.468	64.44	0	1	0	68	9
27	60	1	5796	111	1415	25	15	40	0.625	57108	57.1	0:36	105	65674	0.55	36130	20978	63.26	1	0	3	0	2:34	17:39	3.53	64.38	0	1	0	68	9
28	60	1	5988	112	1410	26	14	40	0.65	57108	57.1	0:37	105	65674	0.571	37545	19569	65.74	2	1	3	0	2:36	17:38	3.591	64.32	0	1	0	68	9
29	60	1	6180	112	1410	27	13	40	0.675	57108	57.1	0:38	105	65674	0.593	38955	18159	68.21	1	0	3	0	2:37	17:36	3.652	64.25	0	1	0	68	9
30	60	1	6372	112	1410	28	12	40	0.7	57108	57.1	0:40	105	65674	0.614	40365	16748	70.68	1	0	3	0	2:38	17:35	3.713	64.19	0	1	0	68	9
31	60	1	6568	113	1405	29	11	40	0.725	57108	57.1	0:41	105	65674	0.636	41775	15338	73.15	1	0	3	0	2:40	17:34	3.774	64.13	0	1	0	68	9
32	60	1	6755	113	1405	30	10	40	0.75	57108	57.1	0:43	105	65674	0.657	43180	13928	75.61	1	0	3	0	2:41	17:32	3.836	64.07	0	1	0	68	9
33	60	1	6947	114	1400	31	9	40	0.775	57108	57.1	0:44	105	65674	0.678	44585	12519	78.07	2	1	3	0	2:43	17:31	3.897	64.01	0	1	0	68	9
34	60	1	7139	114	1400	32	8	40	0.8	57108	57.1	0:45	105	65674	0.7	45995	11128	80.52	1	0	3	0	2:44	17:29	3.958	63.95	0	1	0	68	9
35	60	1	7331	115	1394	33	7	40	0.825	57108	57.1	0:47	105	65674	0.721	47385	9728	82.97	2	1	3	0	2:45	17:28	4.019	63.89	0	1	0	68	9
36	60	1	7528	115	1394	34	6	40	0.85	57108	57.1	0:48	105	65674	0.742	48778	8329	85.41	1	0	3	0	2:47	17:27	4.08	63.83	0	1	0	68	9
37	60	1	7715	116	1389	35	5	40	0.875	57108	57.1	0:50	105	65674	0.763	50179	6985	87.85	2	1	3	0	2:48	17:25	4.141	63.77	0	1	0	68	9
38	60	1	7906	116	1389	36	4	40	0.9	57108	57.1	0:51	105	65674	0.785	51562	5546	90.28	1	0	3	0	2:50	17:24	4.208	63.7	0	1	0	68	9
39	60	1	8098	116	1389	37	3	40	0.925	57108	57.1	0:52	105	65674	0.806	52951	4157	92.72	1	0	3	0	2:51	17:22	4.264	63.64	0	1	0	68	9
40	60	1	8290	117	1384	38	2	40	0.95	57108	57.1	0:54	105	65674	0.827	54340	2768	95.15	1	0	3	0	2:52	17:21	4.325	63.58	0	1	0	68	9
41	60	1	8482	117	1384	39	1	40	0.975	57108	57.1	0:55	105	65674	0.848	55724	1384	97.57	1	0	3	0	2:54	17:20	4.386	63.52	1	1	0	68	9
42	60	2	8674	118	1378	0	58	58	0	76880	76.88	0:00	116	76880	0	0	76880	0	2	1	3	0	2:55	17:18	4.448	63.46	0	1	0	129	1
43	60	2	8866	118	1378	1	57	58	0.017	76880	76.88	0:01	116	76880	0.018	1878	74852	1.805	1	0	1	0	2:57	17:17	4.509	63.4	0	1	0	129	1
44	60	2	9058	119	1378	2	56	58	0.084	76880	76.88	0:02	116	76880	0.086	2756	73574	3.61	2	1	0	0	2:58	17:15	4.57	63.34	0	2	0	129	1

Appendix C

PM Item Schedule Example



Index	PM Item	Difficulty
0	Call ATC at 17500 feet Altitude	Easy
1	De-Ice is ON. Turn OFF in 3 minutes	Hard
2	Reset Radio at Waypoint U	Easy
3	De-Ice is ON. Turn OFF in 3 minutes	Hard
4	Turn Lights OFF at Next Sector	Easy
5	Reset Altitude Gauge at 29500 feet Altitude	Easy
6	Reset Radio at Waypoint X	Easy
7	Check Weather 13 minutes after Departure	Hard
8	De-Ice is ON. Turn OFF in 3 minutes	Hard
9	Turn Lights ON at Waypoint Z	Easy
10	Call ATC 1 minute before Arrival	Hard
11	Check Weather 1.5 miles from Arrival	Hard

Appendix D

Full ANOVA Tables

Experiment 1A PM Task Percent Correct 2 Factor ANOVA Workload x Aid					
Workload = row variable Aid = column variable Subj = subjects					
Source	SS	df	MS	F	P
Subjects	0.1057	5			
Within Subjects					
Workload	0	1	0	0	1.000000
Subj x Workload	0.0255	5	0.0051		
Aid	0.0629	2	0.0314	1.0865	0.374121
Subj x Aid	0.289	10	0.0289		
Workload x Aid	0.0081	2	0.0041	0.1547	0.858683
Subj x Workload x Aid	0.265	10	0.0265		
TOTAL	0.7562	35			

Experiment 1A Visual Search Percent Correct 2 Factor ANOVA Workload x Aid					
Workload = row variable Aid = column variable Subj = subjects					
Source	SS	df	MS	F	P
Subjects	0.0668	5			
Within Subjects					

Workload	0	1	0	0	1.000000
Subj x Workload	0.0027	5	0.0005		
Aid	0.0064	2	0.0032	2.9091	0.100975
Subj x Ai	0.0114	10	0.0011		
Workload x Aid	0.0005	2	0.0002	1	0.401878
Subj x Workload x Aid	0.0022	10	0.0002		
TOTAL	0.09	35			

Experiment 1A Progress Assessment Percent Correct 2 Factor ANOVA Workload x Aid

Workload = row variable
Aid = column variable
Subj = subjects

Source	SS	df	MS	F	P
Subjects	0.1137	5			
Within Subjects					
Workload	0.016	1	0.016	0.6154	0.468281
Subj x Workload	0.1302	5	0.026		
Aid	0.0032	2	0.0016	0.1739	0.842869
Subj x Aid	0.0922	10	0.0092		
Workload x Aid	0.0003	2	0.0001	0.0118	0.988283
Subj x Workload x Aid	0.0848	10	0.0085		
TOTAL	0.4405	35			

ANOVA: Single Factor Aiding for PM Reaction Time					
SUMMARY					
Groups	*Count*	*Sum*	*Average*	*Variance*	
Intrusive	6	39339.13	6556.522	5387470	
None	6	30361.57	5060.262	6029543	
Non-intrusive	6	13549.56	2258.26	777931.5	
ANOVA					
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	57130130	2	28565065	7.027108	0.007025
Within Groups	60974721	15	4064981		
Total	1.18E+08	17			

Experiment 1B PM Task Percent Correct 1 Factor ANOVA Aid					
SUMMARY					
Groups	*Count*	*Sum*	*Average*	*Variance*	
Intrusive	6	39339.13	6556.522	5387470	
None	6	30361.57	5060.262	6029543	
Non-intrusive	6	13549.56	2258.26	777931.5	
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Aid	57130130	2	28565065	7.027108	0.007025
Within Groups	60974721	15	4064981		
Total	1.18E+08	17			

Experiment 1B Visual Search Percent Correct 1 Factor ANOVA Aid					
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	0.002880371	2	0.00144	0.100413	0.905067
Within Groups	0.215139603	15	0.014343		
Total	0.218019975	17			
Experiment 1B Progress Assessment Percent Correct 1 Factor ANOVA Aid					
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	0.01014155	2	0.005071	0.529298	0.599622
Within Groups	0.14370275	15	0.00958		
Total	0.1538443	17			
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	0.04792524	2	0.023963	4.673913	0.026438
Within Groups	0.076903292	15	0.005127		
Total	0.124828532	17			

Experiment 1B Visual Search Reaction Time 1 Factor ANOVA Aid					
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	25203.1	2	12601.55186	0.308926553	0.738796
Within Groups	611871.3	15	40791.41699		
Total	637074.4	17			

Experiment 2 PM Task Percent Correct 2 Factor ANOVA Workload x Aid					
Source	SS	df	MS	F	P
Subjects	0.0673	3			
Within Subjects					
Workload	0.1254	1	0.1254	8.7692	0.059483
Subj x Workload	0.043	3	0.0143		
Aiding	0.3164	1	0.3164	10.7619	0.046407
Subj x Aiding	0.0881	3	0.0294		
Workload x Aiding	0.0734	1	0.0734	8.6353	0.060579
Subj x Workload x Aiding	0.0256	3	0.0085		
TOTAL	0.7391	15			

Experiment 2 PM Task Reaction Time 2 Factor ANOVA Workload x Aid					
Source	SS	df	MS	F	P
Subjects	20029246.0063	3			
Within Subjects					
Workload	438576.7344	1	438576.7344	0.2387	0.658633
Subj x Workload	5511668.8315	3	1837222.9438		
Aid	37288632.9132	1	37288632.9132	7.8217	0.068037
Subj x Aid	14301953.8377	3	4767317.9459		
Workload x Aid	40563.5874	1	40563.5874	0.1	0.772555
Subj x Workload x Aid	1216660.9292	3	405553.6431		
TOTAL	78827302.8396	15			