Remembering to Execute Deferred Tasks in Simulated Air Traffic Control: The Impact of Interruptions

Accepted, Journal of Experimental Psychology: Applied (12/03/2018)

Remembering to Execute Deferred Tasks in Simulated Air Traffic Control: The Impact of Interruptions

Michael David Wilson, Simon Farrell, Troy A. W. Visser, & Shayne Loft

University of Western Australia, Perth.

Author Note

This research was supported by DP12010311 awarded to Loft from the Australian Research Council. Address correspondence to Michael Wilson, School of Psychology, University of Western Australia, Crawley, WA 6009, Australia (michael.wilson@research.uwa.edu.au). The experimental data, all code used to process the data, and all analysis scripts are available at https://github.com/humanfactors/atc-interruptions-impact

Abstract

Air traffic controllers can sometimes forget to complete deferred tasks, with safety implications. In two experiments, we examined how the presence and type of interruptions influenced the probability and speed at which individuals remembered to perform deferred tasks in simulated air traffic control (ATC). Participants were required to accept/handoff aircraft, detect aircraft conflicts, and perform two deferred tasks: a deferred conflict task that required remembering to resolve a conflict in the future; and a deferred handoff task that required substituting an alternative aircraft handoff action in place of routine handoff action. Relative to no interruption, a blank display interruption slowed deferred conflict resumption, but this effect was not augmented by a cognitively demanding *n*-back task or a secondary ATC task interruption. However, the ATC task interruption increased the probability of failing to resume the deferred conflict relative to the blank interruption. An ex-Gaussian model of resumption times revealed that these resumption failures likely reflected true forgetting of the deferred task. Deferred handoff task performance was unaffected by interruptions. These findings suggest that remembering to resume a deferred task in simulated ATC depended on frequent interaction with situational cues on the display and that individuals were particularly susceptible to interference-based forgetting.

Keywords: Interruptions, Prospective Memory, Multitasking, Air Traffic Control, Situation Awareness

Public Significance Statement: The current study found that interruptions negatively impacted the ability of individuals to remember to resume a deferred conflict detection task in simulated air traffic control, particularly when the interrupting task was similar in nature to the primary task. However, remembering to deviate from a routine aircraft handoff procedure was unaffected by interruptions.

In many workplace and everyday situations, we often need to defer a task until a point in the future when it can be carried out. In the psychological science literature, this is referred to as a prospective memory (PM) task. In air traffic control (ATC) settings, situational contexts such as high workload can prevent controllers from immediately completing task actions, requiring them to defer the action to an appropriate point in the future (Shorrock, 2005). During this retention interval, the controller will almost certainly be interrupted by other ongoing task activities. Data from ATC incident reports, controller interviews, and laboratory simulations has found that controllers sometimes forget to complete deferred task actions (Dismukes, Berman, & Loukopoulos, 2007; Shorrock, 2005), and such forgetting can have serious safety implications. For instance, in 2013 two A330s violated minimum separation over Adelaide because, in part, the air traffic controller failed to remember to re-evaluate aircraft separation before issuing instructions (Australian Transport Safety Bureau, 2015). To minimize the risk of such occurrences and maintain system safety, we require a better understanding of the cognitive and situational factors that support the execution of deferred actions after a period interposed activity.

With this goal in mind, recent research has applied theories and methods from basic psychological science to simulations of ATC (for review see, Loft, 2014). In these studies, participants assumed the role of a controller responsible for a region of airspace, called a sector, which varies along both vertical (altitude) and lateral dimensions. Participants monitored a two-dimensional display of their sector, with aircraft flight paths being indicated by lines and aircraft indicated by icons with attached flight information (e.g., call sign, altitude, and airspeed). They were trained to detect and resolve conflicts (i.e., identifying whether any two aircraft will violate minimum separation standards in the future and then changing an aircraft's altitude to resolve that conflict), and to accept and handoff aircraft (i.e., acknowledging responsibility for aircraft entering/exiting their sector by pressing designated response keys). Participants in these studies were also given a deferred (PM) task that required them to remember to deviate from routine operating procedure (by pressing an alternative response key) when accepting 'target' aircraft that met certain conditions (e.g., altitude > 40,000 ft.). Multiple studies have shown that participants often failed to perform this deferred task, and performed ongoing tasks, such as aircraft acceptance and aircraft conflict detection, more slowly than if the deferred task was omitted (Loft, Chapman, & Smith, 2016; Loft, Smith, & Remington, 2013; Loft, Smith, & Bhaskara, 2011; Loft, Finnerty, & Remington, 2011; Loft, Pearcy, & Remington, 2011; Loft & Remington, 2010). This suggests that the cognitive control operations required to monitor the ATC environment for cues associated with the deferred task can impair performance (for further theoretical discussions of the psychological mechanisms underlying "PM cost" effects see, Einstein & McDaniel, 2010; Heathcote, Loft, & Remington, 2015; Smith, 2010; Strickland, Heathcote, Remington, & Loft, 2017).

In the current research we examine the impact of task interruptions on deferred task performance in simulated ATC in order to further understand the cognitive factors that underlie deferred task performance. Instead of manipulating deferred task demands and interpreting performance on other ongoing tasks as evidence for the cognitive processes underlying deferred task retrieval (Loft, 2014); in the current studies, we manipulate the nature of the ongoing task demands and interpret performance on the deferred task as evidence for the cognitive processes underlying deferred task retrieval. We anticipated that this alternative approach will enable us to further understand the cognitive processes underlying storage of deferred task goals and the role of the external environment in supporting their retrieval.

Specifically, we take the approach of past studies from the basic literature and examine the effect of the nature and presence of interruptions on individual's ability to resume a primary task (McDaniel, Einstein, Graham, & Rall, 2004; Monk, Trafton, & Boehm-Davis, 2008; Trafton & Monk, 2007). In these prior interruption studies, participants are faced with a static primary task (e.g. VCR programming) that is interrupted by a secondary task. Typically, disruption is measured by the resumption time; that is the time taken to remember and to perform the first action on the primary task following an interruption (Trafton & Monk, 2007). However, participants can sometimes forget to resume the interrupted primary task (Dodhia & Dismukes, 2009; McDaniel et al., 2004); and if they do remember to resume, they can have difficulty reconstructing the primary task state, resulting in errors on the primary task (Altmann, Trafton, & Hambrick, 2014; Brumby, Cox, & Back, 2013; Monk et al., 2008).

In the current experiments participants completed a series of ATC trials, some of which were interrupted by a secondary task. We examined how the presence and nature of interruptions — presented between encoding a deferred task action and the correct time to perform that action — influenced the probability and speed at which individuals remembered to perform deferred (primary) tasks. The impact of interruptions was examined on two deferred tasks which correspond to operational circumstances that air traffic controllers face in situ: remembering to resume a deferred task and remembering to deviate from a routine procedure.

Resuming Deferred Tasks

Some operational circumstances do not allow controllers to immediately resolve impending conflicts between aircraft (i.e., future violations of minimum aircraft separation; Loft, Bolland, Humphreys, & Neal, 2009). For instance, a controller may identify an aircraft conflict but be unable to immediately issue the conflict resolution because of other air traffic (Loft, Sanderson, Neal, & Mooij, 2007). During this retention interval in which the conflict cannot be resolved, the controller is likely to be interrupted by other ongoing task demands and may have few opportunities for rehearsal of the deferred task (Shorrock, 2005). To what extent does the nature of such interruptions influence a controllers' ability to remember to resume the deferred task?

To simulate the impact of an interruption on deferred task resumption, we included a deferred conflict task that required participants to encode a temporarily unresolvable but impending conflict, and to form the intention to resolve the conflict at an appropriate time in the future. Importantly, while the deferred conflict task was encoded shortly before the interruption, it was only resolvable immediately after the interruption (or equivalent time under conditions in which participants were not interrupted). In this way, our deferred conflict task resumption measure was comparable to the primary task resumption measures typically used in the interruption literature (Trafton & Monk, 2007).

There are several theoretical frameworks that can be leveraged to predict how interruptions might impact deferred task performance in ATC. According to the Memory for Goals theory (MFG; Altmann & Trafton, 2002), deferred task goals are associated with different levels of memory activation. Task goals with higher activation are recalled more quickly following an interruption than goals with lower activation. MFG specifies that task resumption is accomplished by retrieving the primary task goal and task context from memory, and that this process can be facilitated by attending to contextual cues in the post-interruption environment that are linked to the deferred task goal (Altmann & Trafton, 2002). Salvucci and Taatgen, (2011) extend MFG in their cognitive model of human multitasking called *threaded cognition,* which specifies that each task goal is associated with an individual 'thread' which allows for multiple goals to be concurrently activated. Threaded cognition is a general model of human multi-tasking and treats interruptions as a form of "sequential multitasking". According to threaded cognition, individuals would not only maintain a thread related to the deferred task goal, but also maintain a memory representation of the information associated with and required to perform the goal, called the *problem state*. The problem state, like other goals, will decay over time unless it is actively rehearsed (Monk et al., 2008). Rehearsal can be either retrospective (e.g., “What was I doing?”), or prospective (e.g., “what was I about to do”), although evidence suggests individuals prefer prospective rehearsal (Monk et al., 2008). Moreover, rehearsal of the problem state is not limited to repetition of the episodic task goal, but can take advantage of the contextual information associated with the deferred task, thereby increasing the effectiveness of situational retrieval cues under proper conditions (Koriat, Ben-Zur, & Nussbaum, 1990).

Based on MFG and threaded cognition, we might expect that any interruption which prevents visual inspection of the ATC display for a sufficient time should increase the time taken to remember and resolve the deferred conflict, and possibly increase the probability of forgetting the deferred conflict task goal. This is because participants would no longer have access to the display to prime the deferred task goal. In addition, depending on the extent to which a participant’s problem state has decayed during the interruption, the participant will need to take some time to re-develop situation awareness (SA) for the evolved locations and relationships between aircraft in the post-interruption display. Assimilating the mismatch between the pre-interruption and post-interruption displays can take considerable time (Hodgetts, Vachon, & Tremblay, 2013; St. John & Smallman, 2008; St. John, Smallman, & Manes, 2005).

A second central question is whether the nature of the interruption will modulate its effects on deferred task performance. According to MFG, an interruption which is cognitively demanding would be expected to cause greater costs to deferred task performance because it would block mental rehearsal and thus increase decay of the stored task goal (Cades, Werner, Boehm-Davis, Trafton, & Monk, 2008; Hodgetts & Jones, 2006; Monk et al., 2008). Similarly, threaded cognition would predict that a cognitively demanding interruption should be more disruptive because it would introduce competition with the primary task for access to the problem state representation which causes a cognitive bottleneck because only one problem-state can be active at any given time (Salvucci & Taatgen, 2011).

That said, these predictions rely on the assumption that individuals will encode and retrieve detailed internal representations of the task environment (the problem state). In fact, the ATC task may be too complex or feature rich for the creation of such a representation, thus making memory-based retrieval of the problem state an unfeasible interruption recovery strategy. Salvucci and Taatgen (2011) argue that as task complexity increases, individuals are likely to rely more heavily on reconstructing the problem state rather than on storing and retrieving it from memory. Such reconstructive strategies would likely involve the allocation of attentional resources to the contextual features of the ATC display that are known to be associated with task goals (Hunter & Parush, 2010; Ratwani & Trafton, 2008). For instance, a particular location may become associated with the deferred conflict, and following the interruption, visually scanning that location may reinstate the participant's intention to resolve that conflict. A reliance on reconstructive strategies might also be fostered by the dynamic nature of the ATC task which increases the likelihood that the problem state formed from the pre-interruption period does not accurately represent the state of the task at task resumption (St. John & Smallman, 2008). Indeed, several theoretical accounts in the SA literature posit that individuals prefer to store partial representations and rely on frequent interactions with their displays to access "highly selective information on an as-needed basis" (Chiappe et al., 2016; Chiappe, Vu, Rorie, & Morgan, 2012; Gray & Fu, 2004).

If participants rely more heavily on interactions with the ATC display to reconstruct their problem state, as opposed to storing and retrieving the problem state from memory, we would still expect interruptions to negatively impact deferred conflict resumption, because of the time needed to recover SA for the locations of aircraft in the post-interruption scene that are related to deferred task goals. However, we would not expect a more demanding interruption to be more disruptive than a less demanding interruption, because participants would not be storing a problem state in memory that needs to be rehearsed and maintained.

Remembering to Deviate from Routine

Another form of deferred task that controllers perform is to remember to deviate from well-practiced behavioral routines. In these situations, not only must the controller remember a new episodic task, but the intended action must compete for retrieval with task actions strongly associated with primary task goals (Dismukes, 2012; Loft & Remington, 2010). For example, a controller may need to remember to deviate from a routinely performed aircraft handoff procedure and alternatively hold an aircraft when it reaches a specific way-point in the future because of crossing traffic. This type of situation often leads to "habit-capture" (Reason, 1990), where operators fail to perform the intended atypical action, and substitute the routine action instead (Loft & Remington, 2010). To successfully deviate from routine, controllers must inhibit the expectation-driven processing bias cued by automated behavioral routines and notice that the features of the task indicate that the alternative task action should be executed (Norman, 1981; Reason, 1990; Vidulich & Tsang, 2012). Thus, habit-capture is usually presumed to arise because individuals do not perform the required attentional checks of the task environment at the time that deviation from routine is required.

We simulate this type of situation with a deferred handoff (primary) task in which participants were required to remember to handoff a target aircraft with a non-routine response key. There is reason to suspect interruptions might exacerbate habit capture in simulated ATC. For example, in order to link relevant environmental cues to the corresponding requirement to deviate from routine actions, individuals may create a problem state (Salvucci & Taatgen, 2011) by storing information about their intentions to deviate from routine along with information associated with the task goal on the display. Past studies examining deferred tasks in simulated ATC have found that deferred tasks impair performance on the ongoing tasks, indicating that effort is required to maintain the problem state associated with the deferred task goal (for review see, Loft, 2014). It follows that if participants are interrupted and no longer able to monitor the display, their ability to maintain the problem state associated with the deferred task of deviating from routine could be hindered. It is also possible that interruptions may negatively impact habit capture by increasing workload. In an ATC simulation experiment, Stone, Dismukes, & Remington (2001) found that participants were less likely to remember to deviate from a routine procedure when workload was higher (measured by number of aircraft on the screen), and interruptions have been found to increase subjective workload (Adamczyk & Bailey, 2004; Keus van de Poll & Sörqvist, 2016).

On the other hand, interruptions might not necessarily increase habit-capture. Although basic studies have found that interruptions can decrease PM performance (Dodhia & Dismukes, 2009; McDaniel et al., 2004); this is not the case when the task context can be easily reinstated through the provision of contextual cues (Cook, Meeks, Clark-Foos, Merritt, & Marsh, 2014). The ATC display provides rich information at encoding regarding the probable future context of each deferred handoff event, and this may allow participants to retrieve their intention to deviate from routine “only as needed, in a just-in-time manner” (Braver, 2012). Individuals are more likely to remember to perform deferred task actions if they have been associated with specific ongoing task contexts (Bowden, Smith, & Loft, 2017; Cook et al., 2014; Nowinski & Dismukes, 2005), including in simulated ATC (Loft, Finnerty, et al., 2011). To the extent this is also the case in the current experiments, we would not expect an interruption during the retention interval (between encoding intention and the appropriate time to deviate from routine) to impact the probability of habit-capture because the interruption would not influence the extent to which relevant information is available on the display at the time for retrieval of the deferred action.

Experiment 1

We examined the effect of interruptions on two forms of deferred tasks: remembering to resume a deferred conflict task and remembering to deviate from a routine handoff procedure. Participants assumed the role of an air traffic controller responsible for maintaining the safety of aircraft by accepting aircraft entering the sector, detecting and resolving aircraft conflicts, and handing-off aircraft exiting the sector. There were three within-subjects conditions: no-interruption, blank interruption, and *n*-back interruption. All interruptions lasted for 27 s, occluded the display, and were presented without warning. The blank interruption was a blank screen. The *n*-back interruption comprised a visual numerical 2-back task, with 15 random single-digit numbers (2 s duration each; participant pressed the space bar when the stimuli matched the stimuli that appeared two previously). The no-interruption condition served as the baseline comparison. In this condition, participants were not interrupted during the ATC trial.

Method

Participants. Sixty undergraduate students (38 females; median age = 20) from the University of Western Australia participated in the study in exchange for partial course credit or $25 AUD and were tested in groups between one and five.

ATC-Lab$^{\text{Advanced}}$ Simulator. Figure 1 presents a screenshot of the ATC task (Fothergill, Loft, & Neal, 2009). The light grey polygon area is the designated sector, whilst the dark grey area represents flight sectors outside of the participants' control. The black lines denote flight paths that aircraft travel. Aircraft are represented by a circle with a leader-line indicating heading. The aircraft data-blocks specify the aircraft's call sign, speed, type, current altitude, and cleared altitude (the altitude an aircraft is cleared to climb to, descend to, or cruise at). Cleared altitude and current altitude are separated by an arrow that denotes whether the aircraft is climbing (^), descending ($_{\text{v}}$), or cruising (>). Aircraft enter the airspace from the edges of the display, cross sector boundaries, and then exit the display. New aircraft continue to appear throughout the trial, with aircraft positions updated every second (behavioral measures are recorded with millisecond precision). The timer in the lower center of the screen showed how much time had elapsed in the trial and was updated each second.

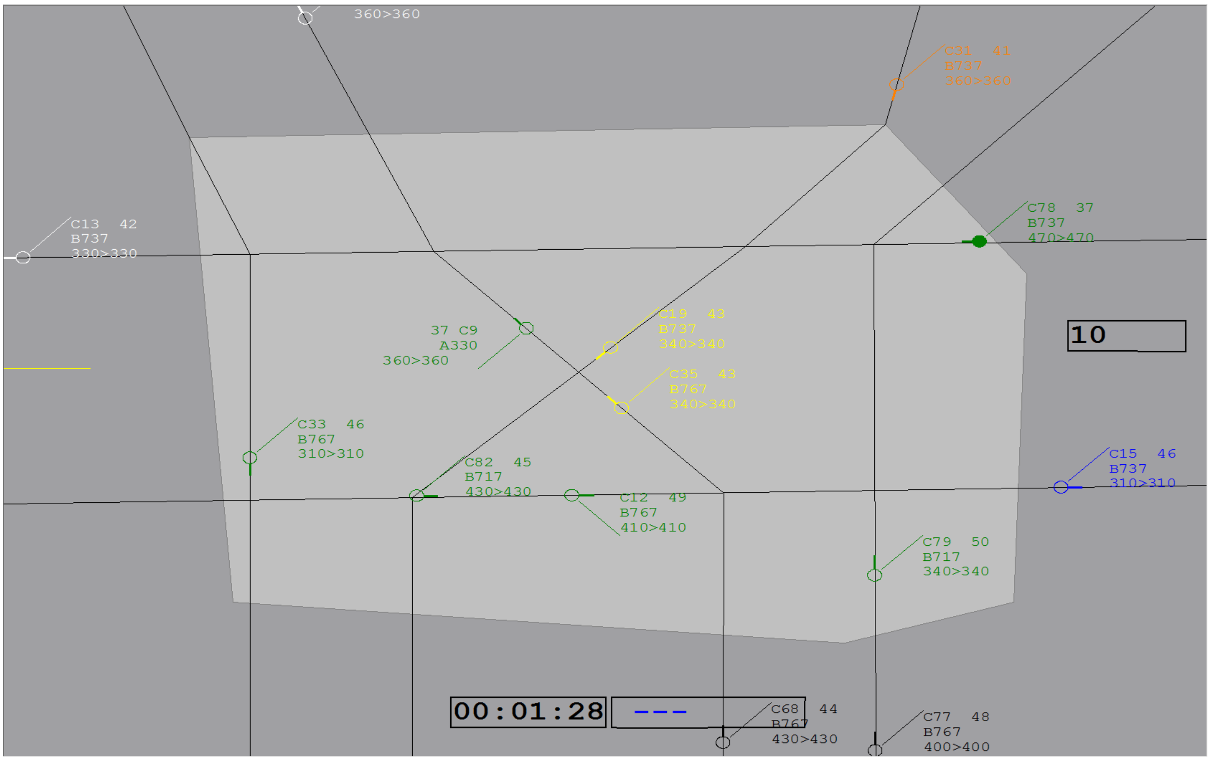

Figure 1. The ATC display. Inbound aircraft are black (e.g., aircraft C77) as they approach the sector, and flash orange for acceptance (e.g. C31) when they reached within 5 miles of the sector boundary. Aircraft turn green (e.g. C37) when accepted. When outbound aircraft cross the sector boundary they flash blue (e.g. C15), and then turn white (e.g. C13) when handed off. Aircraft turn yellow (e.g. C19 & C35) if they violated the minimum vertical and lateral separation. The example in Figure 1 shows that the individual is required to change the altitude of C82 from 430 to an alternative altitude to avoid a conflict with C68. Previously the individual had failed to change the altitude of either C19 or C35 (illuminated in yellow). The running score (10 points) is presented in the middle right hand side of the display.

When aircraft approached the sector, they flashed for acceptance and participants had to accept aircraft by clicking the aircraft icon and pressing the A key within 20 s. As aircraft exited the sector boundary, they flashed for handoff. Participants had to handoff aircraft by clicking the aircraft icon and pressing the H key within 15 s. Aircraft conflicts occurred when an aircraft pair simultaneously violated both lateral (5 nautical miles) and vertical (1,000 feet) separation standards. Participants were required to detect and prevent conflicts from occurring by selecting aircraft in future conflict and changing their cleared altitude to prevent that conflict from occurring.

Participants were awarded points for successfully completing tasks, whilst points were deducted for failure to complete tasks. The current score was continuously updated and displayed in the middle right of the display. Ten points were awarded or deducted for a successful/failed handoff/acceptance. Between 10 and 40 points were awarded for resolving a conflict, depending on how quickly participants resolved the conflict, and 40 points were deducted for failing to resolve a conflict. Forty points were also deducted for unnecessary aircraft interventions (i.e., conflict detection false alarms).

Training Phase. The training phase comprised three tasks: a 10 min audio-visual ATC tutorial; two practice *n*-back trials of 45 s duration; and three 5 min practice ATC trials. Each participant completed one practice trial in each of the three conditions. Results from the training phase (*n*-back and air traffic control practice trials) were screened prior to the test phase to ensure participants understood the task. To demonstrate competence, participants had to perform both the deferred tasks correctly at least once each and achieve 75% accuracy on the *n*-back task. Twenty percent of participants repeated the *n*-back task.

Test Phase. The test phase comprised 15 five-minute trials, with five trials per within-subject condition (no-interruption, blank interruption, and *n*-back interruption). A Latin rectangle scheme was used to counter-balance the 15 trials across the three conditions. To create the scheme, trials were divided (with random assignment) into three equal sized groups, and participants into six equally sized groups. The trial groups were counter-balanced across a 3 (*columns*) × 6 (*rows*) Latin rectangle, where columns assigned a condition to a group of trials; and rows allocated a counter-balancing scheme to a group of participants. This resulted in each of the three conditions being associated equally with each of the 15 ATC trials. The order in which the trials were presented was randomized for each participant.

Each trial comprised a unique set of aircraft and trials differed with respect to the timing and location of events (e.g., conflicts occurred at different times and locations; interruptions began at different times). However, trials had a similar number of ongoing tasks (three conflicts to resolve; between 13 and 19 acceptances; and between 8 and 14 handoffs) and the basic design of the two deferred tasks was consistent amongst trials. The timing of the two deferred tasks was anchored around the programmed ‘interruption start point’ on each trial (90s to 148s). Every trial began with a period of ongoing ATC tasks (acceptances, handoffs, and conflict detection) with no deferred task requirements.

For the deferred handoff task, approximately 60 s into each trial, a message box would appear adjacent to one aircraft instructing participants to handoff that aircraft with an arrow key that corresponded to the aircraft heading (e.g., ↑), instead of the routine ‘H’ key. This message was displayed for 10 s, and participants had to acknowledge it by clicking an "Acknowledge" button. The button only became clickable after 3 s from its initial display to prevent accidental acknowledgement. All messages automatically disappeared if they were not acknowledged within 10 s.

For the deferred conflict task, an aircraft would begin a climb or descent to an altitude that would result in a future conflict with another cruising aircraft. The climb/descent began at approximately the same time the deferred handoff acknowledge button timed out if not previously acknowledged. Participants were instructed that if any aircraft in a conflict pair was changing altitude, then the altitude change functionality would be disabled for both aircraft in the conflict pair until both aircraft were cruising. Twenty seconds prior to the ‘interruption start point’ a message box appeared on the display adjacent to the climbing/descending aircraft instructing the participant to remember to resolve the conflict at the point that both aircraft were cruising in the future. This message also had to be acknowledged within 10 s or it timed out.

At the ‘interruption start point’ of each trial, which varied between trials, participants were either interrupted, or on no-interruption trials, continued ongoing air traffic management. For *n*-back interruption trials, the display would be occluded and participants were presented with a fixation point (+) for 3 s, followed by 24 s of the *n*-back task in which participants were presented with a series of 15 random single-digit numbers (2 s duration each) and were required to press the space bar when the digit matched the digit that appeared two previously (i.e., 2-back). For the blank interruption trials, the display was occluded by a filled black mask for 27 s. Throughout the interruption interval, aircraft continued to move and the simulator automated aircraft handoff and acceptance. No aircraft were programmed to violate separation during the interruption. After 27s (the end of the interruption), both aircraft in the deferred conflict were cruising. Participants then had between 11 and 23 s to resolve the conflict (this time varied from trial to trial). It is critical to highlight that the deferred-conflict could only be resolved immediately after the interruption or at the equivalent time point in the no-interruption condition. This allowed for a direct comparison of performance between the no-interruption and the two interruption conditions. Approximately 60 s later, the deferred-handoff aircraft would flash for handoff and participants would be required to remember to press the correct arrow key in place of the routine ‘H’ key.

After each trial, participants answered two workload questions on a 10-point scale ("How mentally demanding was the task during the last test trial?" and "How hard did you have to work to accomplish your level of performance in the last test trial?"). There was a 30 s break between trials, except after the 7th trial in which there was a 5 min rest break.

Results

Data from the one participant whom did not perform any of the deferred handoffs correctly was not analyzed. Significance was set at an alpha level of .05. Unless otherwise specified, the data was analyzed using two planned contrasts that directly paralleled our hypotheses (Rosenthal & Rosnow, 1985): (1) blank interruption vs no-interruption, and (2) *n*-back interruption vs blank interruption. Effect sizes are estimated using Cohen's *d* (small = 0.3, medium = 0.5, large = 0.8; Cohen, 1988).

All analyses are also reported with associated Bayes Factors (BFs), and BFs are used as the primary framework for inference. Bayesian statistics avoid some pitfalls associated with traditional null hypothesis significance testing (Kass & Raftery, 1995; Wagenmakers, 2007). In particular, BFs quantify the relative evidence favouring the null versus the alternative hypothesis. BFs were computed using the Bayes Factor Package (Morey, Rouder, Love, & Marwick, 2015) for the statistical software, R (R Core Team, 2017). Bayesian analyses were conducted using Baysian paired samples *t*-tests with a medium Jeffreys-Zellner-Siow prior width of $\sqrt{(2)}/2$ (Rouder, Speckman, Sun, Morey, & Iverson, 2009) and were interpreted using Jeffrey's (1998) guidelines with adjustments by Andraszewicz et al. (2015). BFs between 1 and 3 indicate "anecdotal" evidence, Bayes factors greater than 3 indicate moderate evidence, Bayes factors greater than 10 are strong evidence, and Bayes factors greater than 30 are very strong evidence for a given hypothesis relative to an alternative. Bayes factors are represented as BF, and the subscript indicates whether the model comparison is expressed as favoring the alternative hypothesis (BF_{10}) or the null (BF_{01}). To test the extent to which BFs for the main hypotheses varied as a function of prior width, Bayes Factor robustness check plots were inspected for the deferred conflict (resumption time and errors) contrasts and deferred handoff (errors only) contrasts. These revealed that interpretations of our Bayes Factors remained stable over a range of priors (see Appendix for further details).

Deferred Conflict Task. Participants acknowledged all the deferred conflict task instructions. Resumption time was defined as the time taken to change the altitude of one the two aircraft in conflict after the point that both aircraft were cruising. A resumption failure occurred if a participant failed to resolve the conflict before the aircraft violated separation. Mean resumption time and resumption failure proportions are presented in Figure 2.

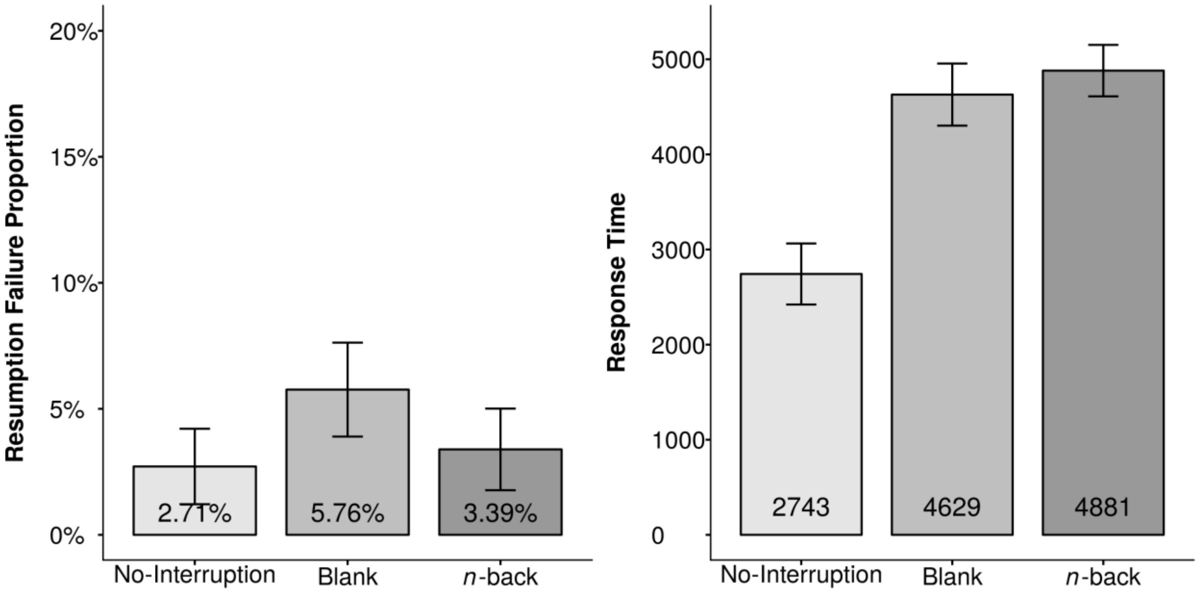

Figure 2. Mean resumption times and failure proportions across the three interruption conditions for the deferred conflict task. Error bars represent 95% within-subjects confidence intervals (Cousineau, 2005).

There was strong evidence that conflict resumption time was slower in the blank condition (*M* = 4629 ms, *SD* = 2095 ms) than in the no-interruption condition (*M* = 2743 ms, *SD* = 1682 ms), *t*(58) = 6.28, *p* < .001, *d* = 0.81, BF_{10} = 287712. However, there was moderate evidence of no difference in conflict resumption time between the *n*-back (*M* = 4881 ms, *SD* = 1526 ms) and blank interruption conditions, *t*(58) = 0.98, *p* = .33, *d* = 0.13, BF_{01} = 4.47. There was minimal evidence that resumption failures were not more likely in the blank condition (*M* = 5.8%, *SD* = 11.2%) than in the no-interruption condition (*M* = 2.7%, *SD*

= 7.8%), *t*(58) = 2.01, *p* = .049, *d* = 0.26, BF$_{01}$ = 1.07. There was anecdotal evidence that resumption failures were not more likely in the *n*-back condition (*M* = 3.4%, *SD* = 7.6%) than in the blank condition *t*(58) = -1.47, *p* = .146, *d* = -0.19, BF$_{01}$ = 2.53.

Deferred Handoff Task. The deferred handoff instruction was not acknowledged on one trial. However, this was not excluded from the analysis as the participant still performed the deferred handoff correctly. We defined a habit-capture as pressing the routine handoff key instead of the instructed alternative arrow key for a target aircraft. Deferred handoff response time (RT) was the time taken to respond to the target aircraft on trials in which the correct PM response was made. No errors of omission were made, that is, all target aircraft were handed off, either correctly (using the correct arrow key: PM response) or incorrectly (using the H key or an incorrect arrow key). False alarms (pressing the arrow key on non-target aircraft) were made to less than 0.5% of aircraft and there was moderate evidence that this did not differ between conditions (smallest *p* = .50, BF$_{01}$ = 6.09). PM response execution errors (remembering to press an arrow key but pressing the incorrect key) were made to 3.06% of target aircraft, and there was anecdotal evidence that this did not differ between conditions (smallest *p* = .14, BF$_{01}$ = 2.71). PM response execution errors were excluded from the analysis, but the pattern of results reported below did not differ when response execution errors were included. Habit-capture rates and correct response times for each condition are presented in Figure 3.

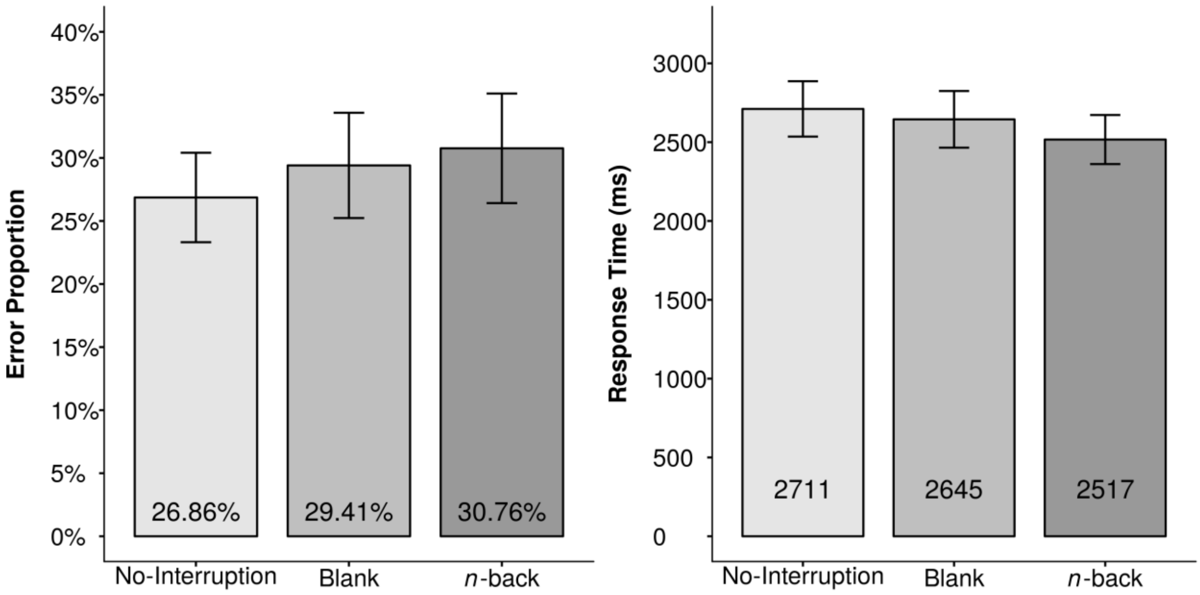

Figure 3. Mean habit-capture rate and RTs across the three interruption conditions for the deferred handoff task. Error bars represent 95% within-subjects confidence intervals (Cousineau, 2005).

There was moderate evidence that the proportion of habit-capture on the deferred handoff task did not differ between the blank (*M* = 29.4%, *SD* = 26.5%) and no-interruption conditions (*M* = 26.9%, *SD* = 25.1%), *t*(58) = 0.78, *p* = .44, *d* = 0.10, BF_{01} = 5.27, or between the *n*-back (*M* = 30.8%, *SD* = 25.7%) and blank conditions, *t*(58) = 0.34, *p* = .73, *d* = 0.04, BF_{01} = 6.64. There was also moderate evidence that deferred handoff RT did not differ between the blank condition (*M* = 2645 ms, *SD* = 1047 ms) and the no-interruption condition (*M* = 2711 ms, *SD* = 1115 ms), *t*(55) = -0.41, *p* = .69, *d* = -0.05, BF_{01} = 6.34, or between the *n*-back (*M* = 2517 ms, *SD* = 883 ms) and blank conditions, *t*(55) = -0.87, *p* = .39, *d* = -0.12, BF_{01} = 4.77.

Subjective Workload. There was strong evidence that the two subjective workload questions were highly correlated, *r*(57) = .96, *p* < .001, BF = 1.81E+29 and they were therefore combined (Wetzels & Wagenmakers, 2012). There was moderate evidence that subjective workload did not differ between the no-interruption (*M* = 5.11, *SD* = 1.74) and blank condition (*M* = 5.05, *SD* = 1.8), *t*(58) = -0.70, *p* = .486, *d* = -0.09, BF_{01} = 5.55; and anecdotal evidence for no difference between the *n*-back (M = 5.2, SD = 1.77) and blank conditions, $t(58) = 1.78$, $p = .081$, $d = 0.23$, $\text{BF}_{01} = 1.61$. Interruptions did not increase subjective workload, but this may be due to a central tendency bias in subjective reporting.

Ongoing task performance. The post-interruption duration varied between two and three minutes across trials. In order to examine whether interruptions had any other impact on ongoing ATC task performance, we assessed six measures of ongoing task performance during the post-interruption period: (1) acceptance RT (mean time taken to accept aircraft), (2) acceptance misses (mean number of failed aircraft acceptances), (3) handoff RT (mean time taken to handoff aircraft), (4) handoff misses (mean number of failed aircraft handoffs), (5) conflict detection time (mean time taken to resolve a conflict), and (6) conflict misses (mean number of aircraft that entered conflict). We did not compare pre-interruption performance to post-interruption performance as systematic scenario design differences between these time periods confound such comparison. For example, the pre-interruption period comprises acknowledgement messages and a period of SA acquisition as the participants have to familiarize themselves with the ATC scenario. Comparing performance in the post-interruption period across conditions was not confounded as our counter-balancing scheme ensured that any differences between scenarios during the post-interruption are controlled for. As can be seen in Table 1, there was no evidence of differences between the three conditions on any of the ongoing performance measures. This indicates that interruptions did not have a chronic impact on task performance, and participants were able to recover from the interruptions relatively quickly. The descriptive statistics associated with each condition can be obtained from the online repository.

Table 1

Grand means for all ongoing task performance measures during the post-interruption period, and associated contrast test results (df = 58).

Variable	Contrast	Mean	*SD*	*t*	*p*	BF_{01}
Acceptance RT (ms)	1	3360	801	1.43	.16	2.7
	2	-	-	1.35	.18	2.96
Acceptance Misses (%)	1	0.01	0.02	0.99	.33	4.42
	2	-	-	0.45	.65	6.37
Handoff RT (ms)	1	2901	810	0.12	.90	6.97
	2	-	-	0.64	.53	5.78
Handoff Misses (%)	1	0.04	0.04	0.51	.61	6.2
	2	-	-	0.05	.96	7.02
Conflict Detection Time (s)	1	56.62	17.30	0.08	.94	7
	2	-	-	0.8	.43	5.18
Conflict Misses (%)	1	0.13	0.18	0.36	.72	6.59
	2	-	-	0.16	.87	6.94

Note: Contrast 1 = between n-back and blank; contrast 2 = between blank and no-interruption.

Discussion

Interruptions slowed deferred conflict resumption time. This suggests that after an interruption, it either took time for participants to retrieve the problem state required to reinstate the deferred task goal, or it took time to develop the sufficient SA after the interruption required to locate and resolve the deferred conflict. However, the moderate Bayesian evidence for no difference between the *n*-back and blank conditions on conflict resumption time suggests that opportunity for rehearsing the problem state during the interruption retention interval was not the primary factor underlying deferred conflict resumption. This provides some evidence that participants relied on interactions with the ATC display following interruptions as opposed to storing problem states prior to the interruption.

The 30% habit-capture error rate is consistent with previous work demonstrating participant vulnerability to habit-capture in simulated ATC (Loft, 2014). It is likely that habit captures occurred because individuals failed to perform the required attentional checks of the ATC task environment when deviation from routine was required (Norman, 1981; Reason, 1990). However, moderate Bayesian evidence in favor of the null indicated that interruptions did not further increase habit-capture. This suggests that either individuals did not actively maintain the intent to deviate from routine over the deferred task retention interval, or that the interruption conditions used in Experiment 1 failed to impede the maintenance of the intent.

Experiment 2

Taken together, the empirical outcomes of Experiment 1 suggest that rather than storing and rehearsing the task problem state and pre-interruption ATC display scene, individuals relied on interactions with the display during the post-interruption period to reconstruct the problem state and information required to perform deferred task goals. However, it may be the case that individuals did store, rehearse, and retrieve problem states but that the *n*-back task failed to impede that process because it was too dissimilar to the ATC task. Interference accounts of working memory capacity posit that our ability to concurrently hold several memory representations is limited by the mutual interference between these representations (Hurlstone, Hitch, & Baddeley, 2014; May, Hasher, & Kane, 1999; Oberauer, Farrell, Jarrold, & Lewandowsky, 2016). As the similarity between information cues increase, interference between associated memory items occurs within working memory (Bunting, 2006; Norman, 1981). Several studies in the interruptions literature have found that when an interrupting and primary task share high visual similarity or have similar goals or required task actions, resumption time and post-completion errors increase (Borst, Taatgen, & van Rijn, 2010; Edwards & Gronlund, 1998; Gillie & Broadbent, 1989; Ratwani & Trafton, 2008). Indeed, both MFG and Threaded Cognition specify an interference mechanism, in addition to a decay mechanism, to account for the disruptive effects of interruptions (Altmann & Trafton, 2002; Borst et al., 2010; Salvucci & Taatgen, 2011). On this logic then, an interruption that is more similar to the primary ATC task may be more likely to interfere with stored memory associations.

To test this possibility, in Experiment 2 we replaced the *n*-back task with a separate ATC sector that participants had to monitor and control during the interruption period. We expected this interrupting ATC task might interfere with participants' memories for pre-interruption locations and associated episodic task goals (problem state) due to the overlap in the task demands and visuo-spatial memory representations. For example, the interrupting ATC task included conflicts that needed to be resolved, and aircraft that need to be handed-off, in similar locations as the deferred conflict and handoff tasks in the primary ATC task.

We expected to replicate the effect of interruption on conflict resumption time when comparing the blank condition to the no-interruption condition. Furthermore, to the extent that problem states are stored and retrieved, we expected slower resumption time in the ATC interruption condition compared to the blank condition. In contrast, if individuals rely primarily on reconstruction strategies, the ATC interruption condition should not differ from the blank condition, replicating the results from the *n*-back condition in Experiment 1.

For the deferred handoff task, Experiment 1 provided reasonably strong evidence that interruptions do not impact habit-capture. If participants rely on post-interruption contextual cues to prime deferred task goals, then it is unlikely that even an ATC task interruption that causes interference would increase habit-capture. This point notwithstanding, the ATC interruption task provides a stronger test of the alternative hypothesis that some form of internal cognitive control, along with information about the information on the ATC display associated with that task goal, is required to maintain the intention to deviate from routine.

Method

Participants. Sixty undergraduate students (33 female; median age = 20) from the University of Western Australia participated in the study in exchange for partial course credit.

Materials and Procedure. Experiment 2 differed from Experiment 1 in three ways. Firstly, instead of an *n*-back interruption condition, participants were interrupted by a separate

ATC scenario (ATC-interruption condition). Secondly, to ensure that the onset of the ATC-interruption was distinctive from the primary scenario and to avoid participant confusion about which sector they were currently controlling, a textbox was added to the top-left corner of the display indicating either ’primary scenario’ or ’interrupting scenario’. Thirdly, at the end of the test phase, participants completed a brief open-ended short-answer questionnaire that asked: “Describe how you felt when you were interrupted” and “Describe what (*if any*) strategies you used when you were interrupted”, and were asked “How did this differ between forms of interruptions you were exposed to (i.e., blank vs filled)?”

ATC-Interruption. The interrupting ATC scenario required participants to monitor an ATC sector which was displayed in place of the primary scenario sector (see Figure 4). The task objectives were identical to those of the ongoing ATC task. Each interrupting ATC scenario comprised two or three acceptances, two or three handoffs, and two conflicts requiring resolution. Twenty percent of conflicts violated minimum separation (i.e., turned yellow) during the interrupting period. Conflicts did not always violate minimum separation during the interrupting period because of the limited trial duration of 27 s. Participants were instructed that performance on the interrupting ATC scenario was of equal importance to the primary ATC task but that no deferred task demands would be presented during the interrupting ATC task. The timer normally on the primary ATC display was removed in the interrupting scenario display to ensure it was comparable to the blank and *n*-back conditions in which participants did not receive feedback regarding the duration of the interruption. There were five unique interrupting scenarios and the order in which they were presented was randomized for each participant.

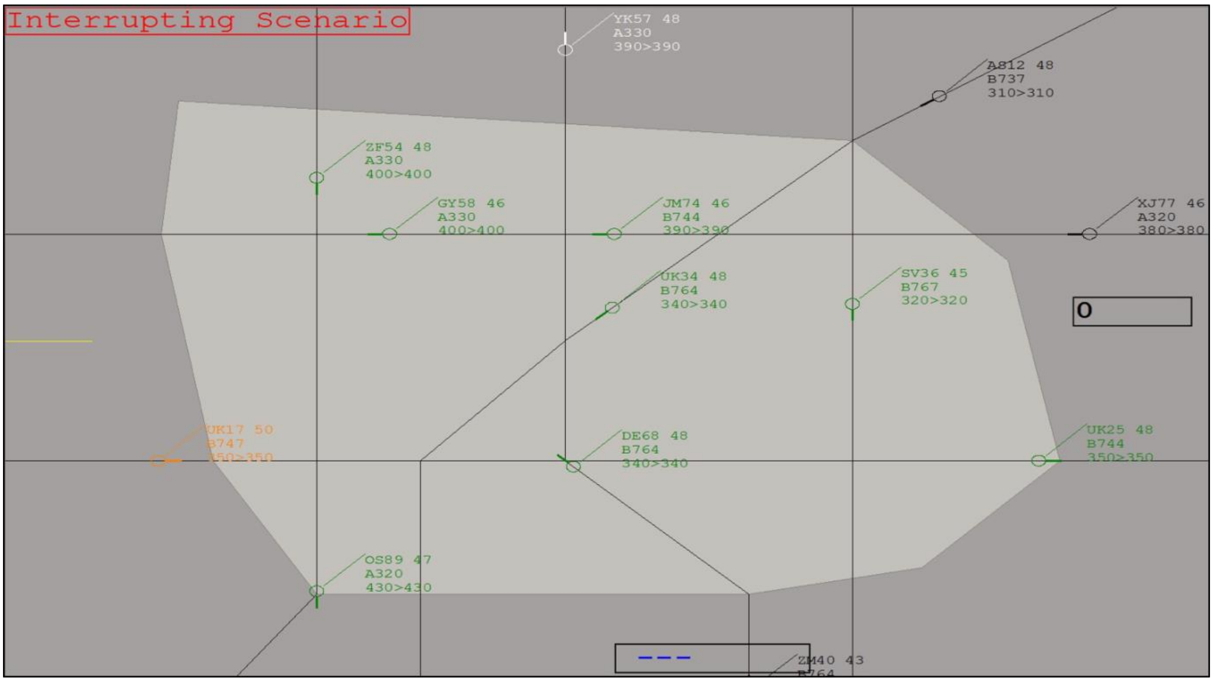

Figure 4. An example of an interrupting ATC scenario. The differences from the primary scenario to note are: the different sector boundaries (i.e., the shape), different flight paths, the "interrupting scenario" textbox, and the absence of a timer. The two conflicts are between ZF54 and GY58; and DE68 and UK34.

Consistent with the interruption conditions in Experiment 1, the ATC-interruption duration was 27 s. The ATC-interruption comprised three temporal parts: firstly, a crosshair was presented for approximately[1] 2500 ms; next participants completed 24 s of ongoing ATC management with no special instructions; finally, participants were presented with a blank screen for approximately 500 ms to provide a brief visual buffer between the interrupting scenario and the primary scenario.

Results

Data from one participant was excluded from all analyses as this participant failed to perform any of the deferred handoff tasks correctly.

[1] This is an approximation due to minor random variability in the time it took to load the interrupting scenario in the order of up to 100-200 ms. This did not affect interruption end time.

Deferred Conflict Task. The deferred conflict instruction was not acknowledged on one trial. However, this was not excluded from the analysis as the participant correctly resolved the deferred conflict. The mean resumption time and resumption failure proportions for each condition are presented in Figure 5.

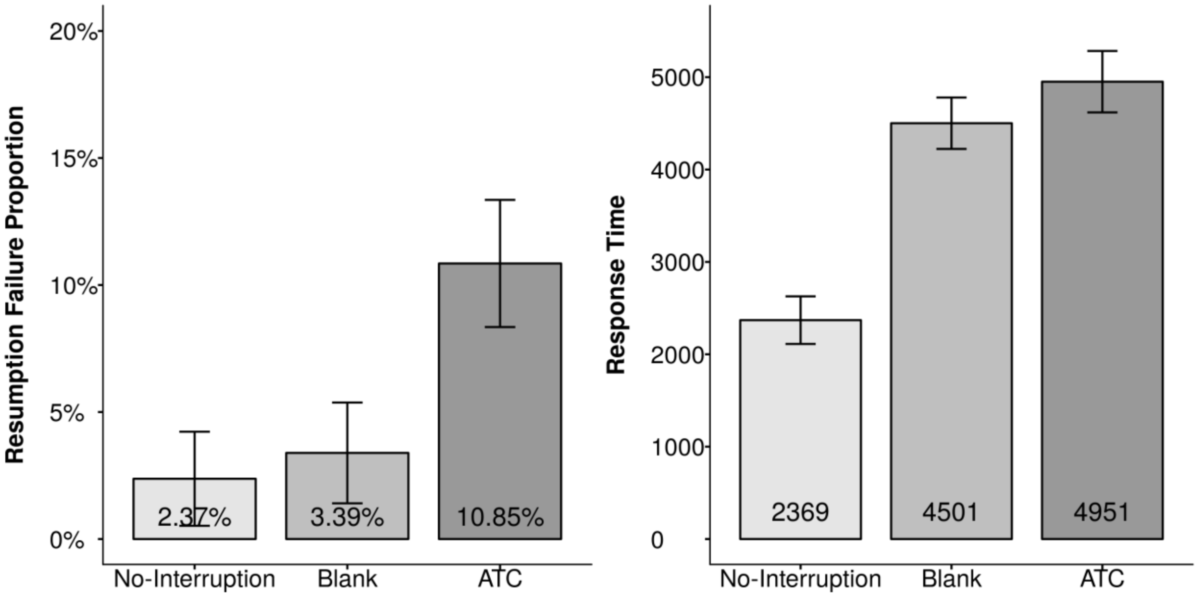

Figure 5. Mean resumption times and failure proportions across the three interruption conditions for the deferred conflict task. Error bars represent 95% within-subjects confidence intervals (Cousineau, 2005).

There was strong evidence that conflict resumption time was slower in the blank condition (*M* = 4501 ms, *SD* = 1789 ms) than in the no-interruption condition (*M* = 2369 ms, *SD* = 1214 ms), *t*(58) = 9.93, *p* < .001, *d* = 1.28, BF_{10} = 1.84E+11. There was anecdotal evidence of no difference in conflict resumption time between the ATC (*M* = 4951 ms, *SD* = 1937 ms) and blank interruption conditions, *t*(58) = 1.58, *p* = .12, *d* = 0.20, BF_{01} = 2.17. There was moderate evidence that resumption failures were not more likely in the blank condition (*M* = 3.4%, *SD* = 8.4%) than in the no-interruption condition (*M* = 2.4%, *SD* = 7.5%), *t*(58) = 0.69, *p* = .49, *d* = 0.09, BF_{01} = 5.62; but there was strong evidence that resumption failures were more likely in the ATC condition (*M* = 10.8%, *SD* = 14.5%) than in the blank condition *t*(58) = 3.55, *p* < .001, *d* = 0.46, BF_{10} = 33.84.

Deferred Handoff Task. Participants failed to acknowledge the deferred handoff instruction on 0.88% of trials. However, such trials were not excluded from the analysis as in all cases the deferred handoff was still performed correctly. Overall, 0.66% of responses were errors of omission in which the aircraft was not handed off at all, but there was moderate evidence that this did not differ between conditions (smallest *p* = .79, BF_{01} = 6.81). Participants made false alarms to less than 0.5% of non-target aircraft and there was anecdotal evidence that this did not differ between conditions, (smallest *p* = .09, BF_{01} = 1.86). PM response execution errors (remembering to press an arrow key, but pressing the incorrect key) were made to 3.06% of target aircraft, and there was moderate evidence that this did not differ between conditions (smallest *p* = .82, BF_{01} = 5.02). The habit-capture rates and correct RT means are presented in Figure 6.

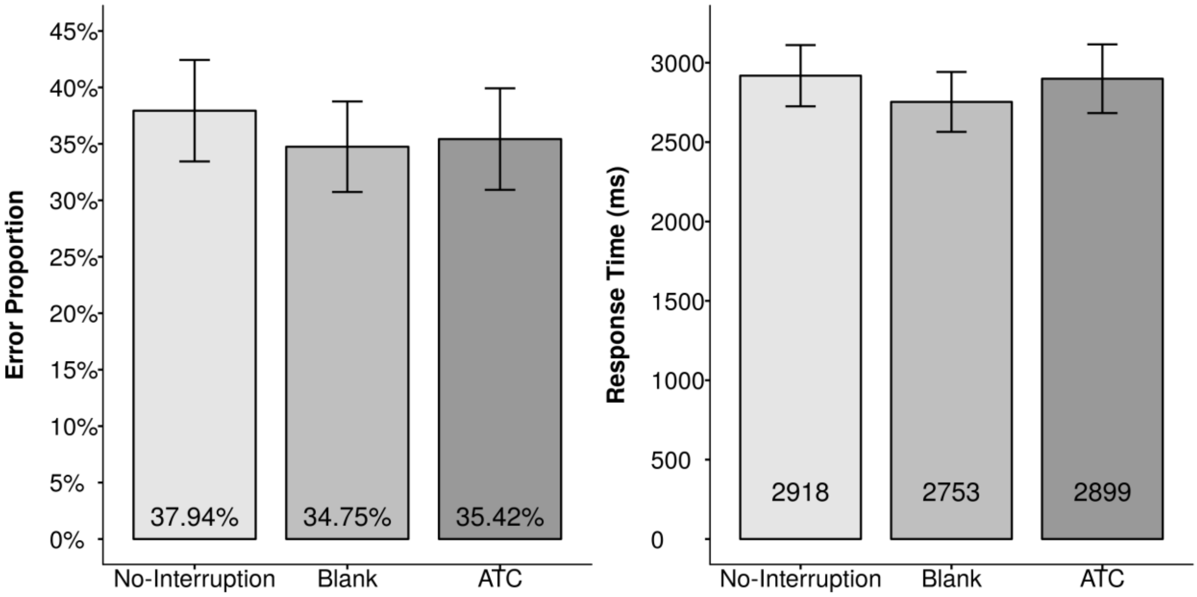

Figure 6. Mean habit capture error rate and response times across the three interruption conditions for the deferred handoff task. Error bars represent 95% within-subjects confidence intervals (Cousineau, 2005).

There was moderate evidence that the proportion of habit-capture on the deferred handoff task did not differ between the blank (*M* = 34.7%, *SD* = 29.6%) and no-interruption conditions (*M* = 37.9%, *SD* = 30.6%), *t*(58) = -0.86, *p* = .39, *d* = -0.11, BF_{01} = 4.92, or between the ATC (*M* = 35.4%, *SD* = 32%) and blank conditions, *t*(58) = 0.18, *p* = .86, *d* = 0.02, BF_{01} = 6.91. There was also moderate evidence that deferred handoff RT did not differ between the blank (*M* = 2753 ms, *SD* = 951 ms) and the no-interruption conditions (*M* = 2918 ms, *SD* = 1374 ms), *t*(49) = -1.03, *p* = .31, *d* = -0.14, BF_{01} = 3.94, or between the ATC (*M* = 2899 ms, *SD* = 1268 ms) and blank conditions, *t*(49) = 0.80, *p* = .43, *d* = 0.11, BF_{01} = 4.8.

Subjective Workload. There was strong evidence that the two subjective workload questions were highly correlated, *r*(57) = .96, *p* < .001, BF_{10} = 7.89E+17, and thus they were combined. There was moderate evidence that subjective workload did not differ between the no-interruption (*M* = 5.31, *SD* = 1.68) and blank condition (*M* = 5.37, *SD* = 1.67), *t*(58) = 0.75, *p* = .454, *d* = 0.10, BF_{01} = 5.36, or between the ATC (*M* = 5.45, *SD* = 1.71) and blank conditions, *t*(58) = 0.83, *p* = .410, *d* = 0.11, BF_{01} = 5.06.

Ongoing Task Performance. Of all the ongoing task measures for the post-interruption period, only post-interruption aircraft acceptance response time was significantly faster following a blank interruption compared to no-interruption (mean difference = 124.5 ms), however there was only anecdotal Bayesian evidence favoring this effect. The results of all other ongoing task performance measures are reported in Table 2. The descriptive statistics associated with each condition can be obtained from the online repository.

Table 2

Grand means for all ongoing task performance measures during the post-interruption period, and associated contrast test results (df = 58).

Variable	Contrast	Mean	*SD*	t	p	BF_{01}
Acceptance RT (ms)	1	3262.17	772.89	0.98	.33	4.45
	2	-	-	-2.45	.018	0.46
Acceptance Misses (%)	1	1	2	1.04	.30	4.18
	2	-	-	0.08	.94	7
Handoff RT (ms)	1	3313.54	842.79	-1.65	.10	1.96
	2	-	-	0.33	.74	6.67
Handoff Misses (%)	1	4	5	0.24	.81	6.84
	2	-	-	-0.72	.47	5.48
CDT (s)	1	54.12	16.39	-0.49	.63	6.26
	2	-	-	0.28	.78	6.77
Conflict Misses (%)	1	9	15	0.49	.63	6.26
	2	-	-	-0.8	.43	5.18

Note: Contrast 1 = between ATC and blank; contrast 2 = between blank and no-interruption.

Interrupting ATC task performance. Performance across the five ATC-interruption trials was aggregated for each participant. Three aircraft which flashed for acceptance in the last 2 s of two trials were excluded from analysis. Table 4 presents the means and standard deviations of ongoing task measures. While the proportion of missed conflicts, accepts and handoffs was considerably higher during the ATC-interruption trials than on the primary ATC scenarios, it is important to note that due to the abrupt onset and short duration of the interrupting scenario, this difference is not likely indicative of reduced task effort. In primary scenarios participants had considerably more time to detect conflicts before they violated minimum separation (*M* = 82.28 s) and had more time to develop their SA of the sector. Additionally, the conflict performance measures reported in Table 3 include conflicts that did not violate minimum separation during the interruption (i.e., did not turn yellow), therefore participants may have detected such conflicts if given more time.

Table 3

Means and standard deviations for ATC-Interruption performance on handoffs, acceptances, and conflict detection.

Variable	M	*SD*
Handoff Miss Proportion	6.39%	7.72%
Handoff Response Time	2691 ms	736 ms
Accept Miss Proportion	8.57%	6.55%
Accept Response Time	4124 ms	720 ms
Conflict Miss Proportion	45.67%	19.43%
Conflict Detection Time	14.57 s	2639 ms
Conflict False Alarms	0.17	0.59

Discussion

Experiment 2 replicated the finding of Experiment 1 that interruptions slowed deferred conflict resumption time. In addition, we found strong Bayesian evidence that individuals failed to detect more deferred conflicts following an ATC task interruption compared to a blank interruption. We further explore the nature of this effect below using Ex-Gaussian distribution modeling, but at a minimum, the increased resumption errors indicate that storing and retrieving the problem state plays a role in conflict detection task resumption.

In Experiment 1, there was a significant difference in conflict detection resumption failures between the blank and the no-interruption conditions, but the Bayesian evidence revealed anecdotal evidence for the null hypothesis. In Experiment 2, there was no significant difference between these two conditions for conflict detection resumption failures, and there was moderate Bayesian evidence in favor of the null hypothesis. Given that the blank and no-interruption conditions were identical across the two experiments (other than the presence of the "primary" scenario textbox), the resumption failure data for the blank and no-interruption conditions were combined for meta-analysis. There was no significant difference in resumption failures between the combined blank (*M* = 4.58%, *SD* = 9.93%) and combined no-interruption (*M* = 2.54%, *SD* = 7.64%) conditions, *t*(117) = 1.92, *p* = .057, *d* = 0.18, BF_{01} = 1.66, and the Bayesian evidence favored the null hypothesis with anecdotal support. Thus, there is no clear evidence one way or the other as to whether the blank interruption increased resumption failures over and above the no-interruption condition.

In Experiment 2, we replicated the moderate Bayesian evidence from Experiment 1 that interruptions do not increase habit-capture. Either participants were not actively maintaining the intent to deviate from routine over the deferred task retention interval, or both the *n*-back and ATC interrupting tasks failed to impede the maintenance of that intent. We return to these possibilities in the General Discussion.

Ex-Gaussian Distribution Modeling

Resumption failures and resumption time are not independent measures of performance. A resumption failure occurs when a participant fails to ‘resume’ the deferred conflict task (i.e., fails to redetect and/or resolve the conflict) before violation of minimum separation standards (herein, referred to as the cutoff time, which varied between 11–23 s post interruption). It is possible that some participants would have eventually remembered to resolve the deferred conflict, given more time. If so, the increased resumption failures in the ATC condition may reflect a delayed retrieval of the problem state, rather than complete forgetting of the deferred conflict task goal. We can adjudicate between these possibilities by inspecting and modelling the entire deferred conflict RT distribution (i.e., resumption time distribution) and assessing whether the probability of correctly resolving the deferred conflict has plateaued prior to the cutoff time (consistent with forgetting), or whether the cutoff has effectively censored some slower responses that would otherwise have been made if there was further time available to resolve the conflict. In addition, consideration of the entire distribution of RTs allows us to more specifically identify the nature of any differences between all the conditions (Ratcliff, 1979; Wixted & Rohrer, 1994).

Figure 7 plots the empirical distribution function for deferred conflict RTs in the four unique conditions in Experiment 1 and 2. For each time, the plot shows the probability that the deferred conflict has been resolved. Given the data were right-censored (RTs are not observed for responses that did not occur prior to the cutoff), the functions were estimated using the Kaplan-Meier estimator (Kaplan & Meier, 1958), using the 'survival' package in R (Therneau, 2015). The crosses denote individual trials that were censored; that is, they indicate the cutoff times for cases where the conflict was not resolved prior to the cutoff. The crosses indicate that in some situations the cutoff may have prevented a slower response from being executed as the distribution functions continue to gradually rise to the right of the earlier (left-most) cutoffs. However, that rise is gradual, and it is not clear whether it can fully explain the differences in failure probabilities between the conditions.

We modelled the data with an exponentially modified Gaussian distribution. The ex-Gaussian distribution is frequently used to model RT distributions (Balota & Yap, 2011; Luce, 1986; Ratcliff, 1979) and recall times (Rohrer & Wixted, 1994; Wixted & Rohrer, 1994) in psychology. Here, we use the ex-Gaussian model in a manner similar to its application to free recall (Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). We assume that participants constantly monitor the display, and in doing so have a constant probability of resolving the deferred conflict. If it is assumed that locations in the display are sampled with replacement, then an exponential distribution of resumption times will be observed. In addition, if the time for non-decisional processes outside of this sampling process (e.g., responding to the deferred conflict when detected) are normally distributed, the resulting distribution is ex-Gaussian in nature. We model complete resumption failure as the asymptote of the cumulative ex-Gaussian distribution function, which is estimated as a separate parameter in the model.

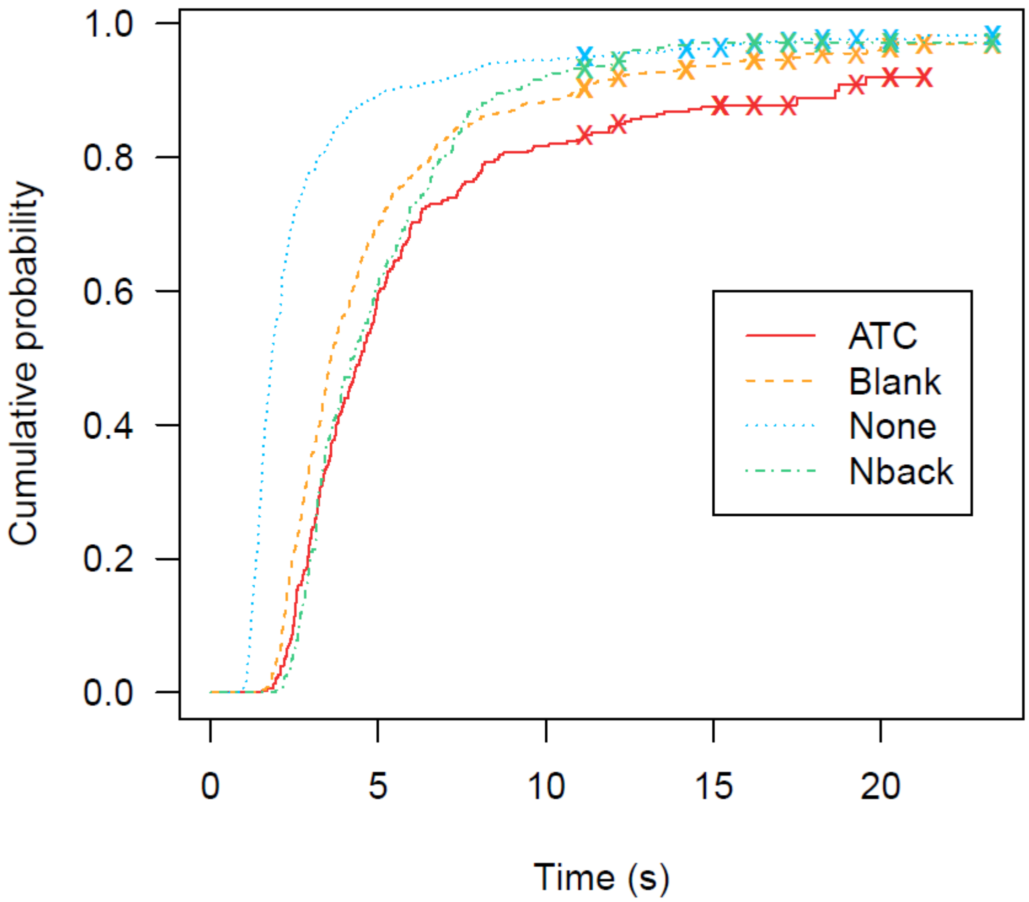

Figure 7. Cumulative RT function for the four unique conditions in Experiment 1 and 2. The lines plot out the Kaplan-Meier estimator of the cumulative density for right-censored data and the crosses indicate the censored data (deferred conflict resumption failures).

The ex-Gaussian model was applied to latencies from all four unique conditions in both experiments, and parameters were estimated using ‘Stan’ (Carpenter et al., 2017), which uses the no-U-turn sampler to obtain Bayesian posterior distributions on parameters. Critically, the truncated responses were fit using a censored distribution: the known cutoff for each trial was used to estimate the probability, under the model, that a resumption time would exceed that cutoff for those trials where the conflict was not resolved, and the resulting probability entered in the likelihood calculation. Specifically, the log-likelihood was given by

$$lnL = \Sigma_{i=1}^{N_c} \ln(1-p_f)+logf(y_i|\mu,\sigma,\lambda)+\Sigma_{j=1}^{N_e} \ln(p_f+(1-p_f)(1-F(cutoff_j,\mu,\sigma,\lambda)), (1)$$

where N_c is the number of correct responses on the deferred conflict task, N_e is the number of resumption failures, f is the ex-Gaussian density function, and F is the cumulative distribution function for the ex-Gaussian. The parameter μ is the shift of the ex-Gaussian function, and captures how long it takes for the cumulative functions in Figure 7 to kick up from 0. The σ parameter is the standard deviation of the Gaussian component of the ex-Gaussian, and is treated here as a nuisance parameter not of theoretical interest. The λ parameter is the rate parameter of the exponential component of the ex-Gaussian, and captures the rate of increase towards the asymptote in Figure 7. Finally, p_f is a parameter capturing the probability of complete resumption failure, represented by the asymptote of the empirical cumulative distribution functions shown in Figure 7. As explained earlier, for each of the observed resumption failures, it is unclear whether the intention to resume the deferred conflict task was forgotten, or whether a correct conflict resolution that would have occurred was prevented from occurring by the imposition of a cutoff for that trial. The right term of Equation 1 models and weights each of these possibilities: either a resumption failure was due to complete forgetting (with probability p_f), or was due to early termination, obtained by calculating the area under the tail of the ex-Gaussian lying above the cutoff for that trial.

As there was a limited number of observations per participant, all observations in a condition were fit using a common set of parameters (Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). The data were fit using the rstan package (Stan Development Team, 2016) in R (R Core Team, 2017). Weakly informative priors were specified for μ ($N(0, 10)$), λ ($U(0, 10)$), and p_f ($U(0, 1)$). A stronger prior was placed on σ ($N(0, 0.4)$) to prevent inflated estimates σ. The parameters σ and λ were bounded at 0, and p_f was restricted to lie between 0 and 1.

Table 4 gives the posteriors for each of the parameters by condition. The top part of the table shows that the estimates of p_f correspond to the empirical frequencies of resumption failures plotted in Figure 7. In other words, the model fits suggest that the empirical distributions plotted in Figure 7 have reached asymptote, and that resumption failures in our experiments were almost solely due to participants forgetting to return to the deferred conflict task (rather than not having time to resolve the deferred conflict). The μ estimates suggest that participants were slower to initiate responding to the deferred conflict in the three interruption conditions compared to the no-interruption condition, and that this slowing was greater for the ATC and *n*-back interruption conditions compared to the blank interruption condition. Finally, the λ estimates at the bottom of Table 4 show a higher sampling rate in the no-interruption condition, with little evidence of difference between the three interruption conditions. In other words, at each time step, if the deferred conflict task had not yet been resolved on that trial, the conflict was more likely to be resolved in the no-interruption condition, producing lower (faster) and less variable resumption times.

Table 4.

Summary statistics of Bayesian posteriors of the ex-Gaussian parameters μ, σ, and λ and the probability of failure p_f.

Parameter	Condition	Mean	.025 quantile	.975 quantile
p_f	None	0.97	0.95	0.98
	Blank	0.95	0.93	0.97
	n-back	0.96	0.94	0.98
	ATC	0.90	0.86	0.93
μ	None	1.04	1.01	1.08
	Blank	1.87	1.80	1.96
	n-back	2.38	2.24	2.52
	ATC	2.09	1.93	2.25
σ	None	0.08	0.06	0.12
	Blank	0.23	0.17	0.30
	n-back	0.28	0.17	0.41
	ATC	0.31	0.20	0.46
λ	None	0.66	0.60	0.72
	Blank	0.38	0.35	0.42
	n-back	0.40	0.03	0.45
	ATC	0.33	0.02	0.38

General Discussion

We conducted two experiments investigating how interruptions impact deferred primary task performance in simulated ATC. We examined two forms of deferred tasks which represented tasks that controllers perform in-situ: remembering to resume a task at a later point in time and remembering to deviate from routine. The empirical findings of the present study can be summarized as follows. The blank interruption slowed the time taken to resume and resolve the deferred conflict. In Experiment 1 we did not find the more demanding *n*-back interruption increased resumption time over and above the blank interruption. In Experiment 2, we also did not find evidence favoring a difference in resumption time between the ATC-interruption and the blank interruption; however, the ATC-interruption was associated with an increased likelihood of resumption failures — that is, forgetting to resolve the deferred conflict altogether — compared to the blank condition. On the deferred handoff task, we replicated findings from previous work that individuals are vulnerable to habit-capture in simulated ATC (Loft, 2014). However, habit-capture was not affected by interruptions. Interruptions also did not affect post-interruption ongoing task performance or ratings of subjective workload.

Theoretical Implications

Resuming Interrupted Tasks. Controllers must sometimes defer task actions and remember to resume them at the appropriate point in the future (Shorrock, 2005). We simulated this situation with a deferred conflict task that required participants to acknowledge a temporarily unresolvable conflict, and then to remember to resolve it at the appropriate time. We expected that an interruption which occluded the display would increase the time taken to remember and resolve the deferred conflict and possibly increase the probability of resumption failure – that is forgetting the task goal altogether. We reasoned that this would occur because individuals would no longer have access to the ATC display to prime the deferred task goal and because individuals would require some time to redevelop SA for the post-interruption display. In line with this, the data from both experiments revealed strong evidence that the time to resolve the deferred conflict was slower under blank interruption conditions than when not interrupted. This suggests that participants used the information in the display to perform the deferred task goal, and individuals required some time to re-develop SA for the post-interruption display after an interruption. This finding supports Salvucci's (2010) argument that in complex tasks, an important factor underlying interruption recovery is the time required for deliberate and strategic reconstruction of the task environment (i.e. problem state).

However, the meta-analysis revealed no clear evidence one way or the other as to whether the blank interruption increased conflict resumption failures. One possible reason for this may be because there was substantial variability in how individuals utilized the blank interruption time interval. We had anticipated *a priori* that the blank interruption would provide individuals the opportunity to rehearse their deferred conflict task goal; however, the results from our post-experiment open-ended survey revealed there was considerable variability in participants' strategic response to the blank interruption. For instance, 37.5% of participants reported finding the blank interruption to be restful, enjoyable, or quite manageable (compared to only 4.36% for the ATC-interruption); whilst 58% of participants reported feelings of anxiety, stress, or frustration (compared to 62.50% for the ATC-interruption).

We also expected that an interruption that placed significant cognitive demands on participants would further slow conflict task resumption, because it would impede rehearsal of the problem state and the deferred task goal (Altmann & Trafton, 2002; Borst et al., 2010). However, in Experiment 1, resumption time did not further increase following a cognitively demanding *n*-back interruption, relative to the blank interruption where individuals were presumably free to rehearse deferred task goals. We reasoned this null effect might have occurred because individuals were not storing or rehearsing the task problem state and pre-interruption ATC display scene. However, the fact that the ATC-interruption increased resumption failures in Experiment 2 suggested instead that individuals did store and actively maintain representations of the ATC task problem state, but that the *n*-back task failed to impede this process. This contrasts with previous research that found an *n*-back task interruption increased resumption time and errors relative to unfilled or undemanding interruptions on a VCR programming task (Monk et al., 2008; Cades et al., 2008).

One possible explanation for this discrepancy is that more basic static tasks may be more sensitive to the effects of rehearsal inhibition because the problem-state can be more effectively stored in memory and task resumption can be achieved with fewer contaminant cognitive operations (Salvucci & Taatgen, 2011). In comparison, resumption in the more complex and dynamic ATC task involves prioritizing attention across several competing goals whilst also reacquiring SA of the updated display scene. This increase in the number of cognitive operations and problem-state complexity in ATC likely limits the effectiveness of memory-based rehearsal strategies, and this increased range of possible resumption strategies could mean that isolating the influence of rehearsal on a given primary task resumption process is more difficult.

In Experiment 2, while there was no clear evidence as to whether the ATC-interruption increased resumption time compared to the blank condition or not, there was strong evidence that it increased resumption failures (by 7.46%). Modelling the entire distribution of RTs provided evidence that resumption failures were not simply due to a slowed resumption process, but were due to complete forgetting of the deferred intention. This finding extends basic laboratory studies that have found that interruptions can sometimes cause forgetting to return to a primary task (Dodhia & Dismukes, 2009; McDaniel et al., 2004).

Resumption failures constitute a substantial error, as they indicate that the contextual cues present in the environment (e.g., the presence of the conflict on the display) failed to prompt retrieval of the deferred task goal through associative memory cueing (Cook et al., 2014). Consistent with our motivation for Experiment 2, a likely explanation for increased resumption failures in the ATC-interruption (but not in the *n*-back condition) is that the visual-spatial memory representations and task goals were similar to those of the deferred conflict task (Borst et al., 2010; Bunting, 2006; Norman, 1981). This argument is based on interference accounts of working memory which posit that our capacity to concurrently hold several memory representations is limited by the mutual interference between representations (May et al., 1999; Oberauer et al., 2016). Alternatively, resource-based theories posit that the efficiency of a cognitive function is monotonically related to the amount of cognitive resources allocated to it (Ma, Husain, & Bayes, 2014; Tombu & Jolicœur, 2003). As a result, tasks will interfere with each other to the extent that they require the same processing resource at the same time (Logie, Zucco, & Baddeley, 1990; Navon & Gopher, 1979; Ratwani & Trafton, 2008). Based on these accounts, it is possible that the increased resumption failures occurred not due to high similarity, but because the ATC-interruption required access to the same resources that are required to maintain the primary task goals or problem state (i.e., overlapping visual-spatial memory demands). Future research might attempt to tease apart the independent contributions of similarity and resource demands by examining the effect of a dissimilar interrupting task which placed significant demands on visual-spatial working memory, for instance, a visual-spatial *n*-back task.

Remembering to Deviate from Routine. Operators sometimes must remember to deviate from firmly reinforced behavioral routines. We simulated this situation with a deferred handoff task in which participants had to remember to handoff a target aircraft with a non-routine response key. Consistent with previous research examining deferred handoff tasks in simulated ATC (Loft, 2014), participants failed to remember to deviate from routine on a significant proportion of occasions (31% and 37.9% habit capture rate in Experiments 1 & 2, respectively). Habit-capture likely occurred because participants failed to adequately attend to the relevant features of the task at the time that deviation from routine was required (Norman, 1981; Reason, 1990). Findings of PM costs to ongoing ATC tasks in prior research suggests that individuals may maintain some form of cognitive control over the deferred task retention interval in order to remember to attend to the ATC task features associated with the intention to deviate from routine (Loft, 2014). Thus, we reasoned interruptions could interfere with individual's ability to maintain the problem state associated with the deferred task goal, particularly when the interruption was more cognitively demanding.

In contrast to this prediction, none of our interruptions, whether cognitively demanding or not, affected habit-capture. One possible explanation is that participants engaged in a cue driven, "just-in-time" reactive strategy, allocating their attention to the deferred handoff tasks only at the time that deviation was required (Braver, 2012). This strategy would have been particularly effective for the deferred handoff task because not only was the task associated with a specific future context, but the target aircraft flashed blue when the handoff was required which provided a clear cue regarding the exact context of the deferred task handoff event. This explanation is consistent with basic and applied PM research that has found that PM tasks are more likely to be completed if they are reinstated through the provision of contextual cues (Bowden et al., 2017; Cook et al., 2014; Loft, Finnerty, et al., 2011). If the ability to deviate from routine is indeed impacted more by the attentional demands at retrieval compared to those over the PM retention interval, it will be important for future research to examine whether distractions or increased workload at the time the deviation from routine is required increases habit-capture (e.g., Stone et al., 2001).

An alternative possibility is that participants did proactively maintain the problem state associated with their intention to deviate from routine, but interruptions simply had no effect on their ability to do so. This could have well been the case because the deferred handoff task was performed a reasonably long time after the interruption had ended, thereby allowing participant's sufficient time to re-engage with the deferred task goal. An important avenue for future research will be to examine the effects of manipulating the time that the act of deviating from routine is required, relative to the end of the interruption. Another possibility is that the intention to deviate from routine was forgotten during the interruption, but was also forgotten during the no-interruption condition at a similar rate. In any case, the implications of our findings are clear: overcoming a habitual response poses a significant cognitive challenge, and the cognitive processes required to do this are not influenced by interruptions in this version of the ATC task.

One factor that might help to explain the higher habit-capture rate in this study compared to previous simulated ATC studies was the fact that the deferred conflict task was nested within the retention interval of the deferred handoff task, whilst in previous studies participants only completed one deferred task per trial (Loft, 2014; cf. Stone et al., 2001). The extent to which this would have triggered performance costs to one or both of the deferred tasks is unclear. However, in an electronic order-entry task, Sasangohar, Donmez, Easty, & Trbovich (2017) examined the effects of nested interruptions where individuals had to resume multiple tasks and reported that this nesting further increased the impact of interruptions. In field operations, it is likely individuals would be faced with remembering multiple future intentions. An important next step will be to extend current cognitive theory to account for the costs associated with holding multiple deferred intentions, and for human factors practitioners to determine whether these costs translate to increased risk of human error and under what situational contexts.

Practical Implications

ATC is a complex, safety-critical task characterized by time pressure, and the concurrent monitoring and execution of multiple tasks. Interruptions are frequent, and examples include communication requests from aircraft or other controllers, unexpected events requiring immediate attention, or the performance of other routine tasks (Kontogiannis & Malakis, 2009). Observational and ethnographic studies examining interruptions in ATC have suggested that interruptions may increase the likelihood of controllers forgetting deferred tasks and failing to detect or monitor data (Jones & Endsley, 1996; Shorrock, 2005). For instance, a common form of radar monitor error is when controllers form an “intention to monitor” a situation, but then fail to remember to do so due to other concurrent task demands or interruptions (Shorrock, 2005). The findings from Experiment 1 suggest the negative effects of interruption on memory may be a result of attention being shifted away from the situational cues in the ATC display that could prompt retrieval of intended actions. The fact that the *n*-back interruption was not associated with higher levels of disruption than the blank interruption suggests that increased opportunity for problem state rehearsal may not be sufficient to decrease the negative impact of interruption.

The increased resumption failures in the ATC-interruption condition suggests that under some circumstances, monitoring multiple displays that share information processing requirements could have safety implications. On July 1st, 2002, a Tupolev-Tu-154M and a Boeing 757 collided in mid-air over Überlingen in Germany. Subsequent investigation revealed that a major contributor to the accident was the fact that the controller was monitoring a secondary ATC display at the time and had failed to subsequently divert attention back to the impending conflict on the primary display in sufficient time (Shorrock, 2007). The present work suggests that one element in the Überlingen accident may have been the effect that switching between multiple similar ATC displays has on memory for deferred tasks. Additionally, the evidence from the ex-Gaussian model suggested that resumption failures were likely due to complete forgetting, rather than a slowed resumption process. This suggests that if a controller fails to resume an interrupted task at the correct reentry point due to memory interference, they may not eventually recover the intention from the perceptual elements of the display, no matter how much time they have. Indeed, several of Shorrock's (2005) case reports specifically note that in most "intention to monitor" memory errors, controllers only recovered the information with explicit prompting from either automated ATC systems or other controllers on duty.

The findings from Experiment 2 also may be particularly relevant to research on interruption recovery aids which explore whether 'event reviews' and 'change logs' on secondary displays can improve recovery from interruption. These tools are designed to assist operators restore SA after an interruption by visually representing important changes that occurred in the environment during the interruption (St. John & Smallman, 2008; Sasangohar, Scott, & Cummings, 2014; Scott, Mercier, Cummings, & Wang, 2006). While our findings support the idea that helping operators locate and attend to important items after an interruption could be useful; it is possible that these tools could cause interference-based forgetting if operators are switching between a primary task and recovery tool which share visually similar interfaces. Indeed, this may explain why several studies have found paradoxically negative effects of interruption recovery aids on interruption recovery (Hodgetts, Tremblay, Vallières, & Vachon, 2015; St. John et al., 2005; Scott et al., 2006).

Considerations and Conclusions

Several studies using dynamic control tasks have reported that interruptions can negatively impact ongoing task performance (Hodgetts et al., 2015), and that it can take considerable time to return to baseline performance (Loft, Sadler, Braithwaite, & Huf, 2015). However, we failed to detect any effect of interruptions to ongoing task performance in the post-interruption period. One possible reason for this finding is that our analysis of ongoing task performance was limited in resolution to the entire post-interruption period, but the effects of the interruption may have only occurred for a shorter duration, more immediately after the interruption. Therefore, it is possible that there was an effect of interruptions to post-interruption ongoing task performance, but due to the low frequency of events in the ATC task, we were unable to detect it as we did not have enough ongoing task events to analyze post-interruption performance as a function of time after interruption.

The use of a student sample with limited training does limit our ability to generalize the results to expert controllers with lengthy training histories. Experienced controllers receive intensive training with well-defined performance standards and as such are far less likely to forget to perform deferred tasks as often as we observed in our experiments. Nonetheless, given the high frequency of aircraft and concomitant controller actions required each day, even small changes in error probabilities associated with deferred intentions could translate into large differences in incidents (Dismukes, 2012; Loft et al., 2013). There is a critical need to determine whether current display technology/aids are equipped to prevent the source of human error identified in the current study (Loft et al., 2016).

Additionally, the interruption manipulations we examined in the present study do not represent the range of interruptions ‘in the wild’. For instance, controllers may face interruptions in the order of several minutes. We selected the interruption duration of 27 s to ensure that items on the display were not entirely different from the pre-interruption state. If this was the case, participants would have to fully reconstruct (rather than recover) SA of the ATC display scene. Additionally, controllers in operational settings would often receive warnings about impending interruptions, allowing some strategic control over when to engage the pending interruption (McFarlane, 2002). In our studies, interruptions were presented with no explicit warning which may have inhibited participants from using interruption preparation strategies (Trafton, Altmann, Brock, & Mintz, 2003). Nevertheless, several participants reported in follow-up interviews that they strategically memorized the location of important aircraft when the deferred conflict encoding message appeared (note however, an interruption only followed this message 2/3 of the time, and participants did not know in advance whether they would be interrupted). Future research should examine whether preparation strategies during the 'interruption lag' can minimize deferred task performance decrements and whether strategic preparation is associated with performance costs to other ongoing primary tasks.

It is pertinent that both researchers and practitioners understand the factors underlying deferred task performance in safety-critical complex work environments. The present study has used a "use-inspired" basic research paradigm (Stokes, 1997) that allowed the systematic investigation of the effects of interruptions on multiple forms of deferred task. The results of this paper suggest that while interruptions can negatively impact deferred tasks under some circumstances, the presence and extent of the negative effects will depend heavily on the characteristics of both the interrupting task and deferred task at hand.

References

Adamczyk, P. D., & Bailey, B. P. (2004). If not now when?: the effects of interruption at different moments within task execution. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *6*(1), 271–278. https://doi.org/10.1145/985692.985727

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*, *26*(1), 39–83. https://doi.org/10.1016/S0364-0213(01)00058-1

Altmann, E. M., Trafton, J. G., & Hambrick, D. Z. (2014). Momentary interruptions can derail the train of thought. *Journal of Experimental Psychology: General*, *143*(1), 215–226. https://doi.org/10.1037/a0030986

Australian Transport Safety Bureau. (2015). *Loss of separation between Airbus A330 VH-EBO and Airbus A330 VH-EBS*. Retrieved from https://skybrary.aero/bookshelf/books/3045.pdf

Balota, D. A., & Yap, M. J. (2011). Moving Beyond the Mean in Studies of Mental Chronometry. *Current Directions in Psychological Science*, *20*(3), 160–166. https://doi.org/10.1177/0963721411408885

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2010). The problem state: A cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 363–382. https://doi.org/10.1037/a0018106

Bowden, V. K., Smith, R. E., & Loft, S. (2017). Eye movements provide insights into the conscious use of context in prospective memory. *Consciousness and Cognition*, *52*, 68–74. https://doi.org/10.1016/j.concog.2017.04.003

Braver, T. S. (2012). The Variable Nature of Cognitive Control: a Dual Mechanisms Framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Brumby, D. P., Cox, A. L., & Back, J. (2013). Recovering from an interruption: Investigating speed-accuracy tradeoffs in task resumption strategy. *Journal of Experimental Psychology: Applied*, *19*(2), 95–107. https://doi.org/10.1037/a0032696

Bunting, M. (2006). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(2), 183–196. https://doi.org/10.1037/0278-7393.32.2.183

Cades, D. M., Werner, N., Boehm-Davis, D. A., Trafton, J. G., & Monk, C. A. (2008). Dealing with Interruptions can be Complex, but does Interruption Complexity Matter: A Mental Resources Approach to Quantifying Disruptions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(4), 398–402. https://doi.org/10.1177/154193120805200442

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chiappe, D., Morgan, C. A., Kraut, J., Ziccardi, J., Sturre, L., Strybel, T. Z., & Vu, K.-P. L. (2016). Evaluating probe techniques and a situated theory of situation awareness. *Journal of Experimental Psychology: Applied*, *22*(4), 436–454. https://doi.org/10.1037/xap0000097

Chiappe, D., Vu, K.-P. L., Rorie, C., & Morgan, C. (2012). A Situated Approach To Shared Situation Awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 748–752. https://doi.org/10.1177/1071181312561156

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 edition). Hillsdale, N.J: Routledge.

Cook, G. I., Meeks, J. T., Clark-Foos, A., Merritt, P. S., & Marsh, R. L. (2014). The Role of Interruptions and Contextual Associations in Delayed-Execute Prospective Memory. *Applied Cognitive Psychology*, *28*(1), 91–103. https://doi.org/10.1002/acp.2960

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Dismukes, R. K. (2012). Prospective Memory in Workplace and Everyday Situations. *Current Directions in Psychological Science*, *21*(4), 215–220. https://doi.org/10.1177/0963721412447621

Dismukes, R. K., Berman, B. A., & Loukopoulos, L. (2007). *The Limits of Expertise: Rethinking Pilot Error and the Causes of Airline Accidents*. Burlington, VT: Routledge.

Dodhia, R. M., & Dismukes, R. K. (2009). Interruptions create prospective memory tasks. *Applied Cognitive Psychology*, *23*(1), 73–89. https://doi.org/10.1002/acp.1441

Edwards, M. B., & Gronlund, S. D. (1998). Task Interruption and its Effects on Memory. *Memory*, *6*(6), 665–687. https://doi.org/10.1080/741943375

Einstein, G. O., & McDaniel, M. A. (2010). Prospective memory and what costs do not reveal about retrieval processes: A commentary on Smith, Hunt, McVay, and McConnell (2007). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1082–1088. https://doi.org/10.1037/a0019184

Fothergill, S., Loft, S., & Neal, A. (2009). ATC-labAdvanced: An air traffic control simulator with realism and control. *Behavior Research Methods*, *41*(1), 118–127. https://doi.org/10.3758/brm.41.1.118

Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research*, *50*(4), 243–250. https://doi.org/10.1007/bf00309260

Gray, W. D., & Fu, W. T. (2004). Soft constraints in interactive behavior: the case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science*, *28*(3), 359–382. https://doi.org/10.1207/s15516709cog2803_3

Heathcote, A., Loft, S., & Remington, R. W. (2015). Slow down and remember to remember! A delay theory of prospective memory costs. *Psychological Review*, *122*(2), 376–410. https://doi.org/10.1037/a0038952

Hodgetts, H. M., & Jones, D. M. (2006). Interruption of the Tower of London task: Support for a goal-activation approach. *Journal of Experimental Psychology: General*, *135*(1), 103–115. https://doi.org/10.1037/0096-3445.135.1.103

Hodgetts, H. M., Tremblay, S., Vallières, B. R., & Vachon, F. (2015). Decision support and vulnerability to interruption in a dynamic multitasking environment. *International Journal of Human-Computer Studies*, *79*, 106–117. https://doi.org/10.1016/j.ijhcs.2015.01.009

Hodgetts, H. M., Vachon, F., & Tremblay, S. (2013). Background Sound Impairs Interruption Recovery in Dynamic Task Situations: Procedural Conflict? *Applied Cognitive Psychology*, *28*(1), 10–21. https://doi.org/10.1002/acp.2952

Hunter, A. C., & Parush, A. (2010). Where Did they Go? Recovering Dynamic Objects after Interruptions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(23), 1990–1994. https://doi.org/10.1177/154193121005402318

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin*, *140*(2), 339—373. https://doi.org/10.1037/a0034221

Jeffreys, H. (1998). *The Theory of Probability*. OUP Oxford.

St. John, M., & Smallman, H. S. (2008). Staying Up to Speed: Four Design Principles for Maintaining and Recovering Situation Awareness. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 118–139. https://doi.org/10.1518/155534308x284408

St. John, M., Smallman, H. S., & Manes, D. I. (2005). Recovery from Interruptions to a Dynamic Monitoring Task: The Beguiling Utility of Instant Replay. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(3), 473–477. https://doi.org/10.1177/154193120504900355

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, and Environmental Medicine*, *67*(6), 507–512.

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457. https://doi.org/10.2307/2281868

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773. https://doi.org/10.2307/2291091

Keus van de Poll, M., & Sörqvist, P. (2016). Effects of Task Interruption and Background Speech on Word Processed Writing. *Applied Cognitive Psychology*, *30*(3), 430–439. https://doi.org/10.1002/acp.3221

Kontogiannis, T., & Malakis, S. (2009). A proactive approach to human error detection and identification in aviation and air traffic control. *Safety Science*, *47*(5), 693–706. https://doi.org/https://doi.org/10.1016/j.ssci.2008.09.007

Koriat, A., Ben-Zur, H., & Nussbaum, A. (1990). Encoding information for future action: Memory for to-be-performed tasks versus memory for to-be-recalled tasks. *Memory & Cognition*, *18*(6), 568–578. https://doi.org/10.3758/bf03197099

Loft, S. (2014). Applying Psychological Science to Examine Prospective Memory in Simulated Air Traffic Control. *Current Directions in Psychological Science*, *23*(5), 326–331. https://doi.org/10.1177/0963721414545214

Loft, S., Bolland, S., Humphreys, M. S., & Neal, A. (2009). A theory and model of conflict detection in air traffic control: Incorporating environmental constraints. *Journal of Experimental Psychology: Applied*, *15*(2), 106–124. https://doi.org/10.1037/a0016118

Loft, S., Chapman, M., & Smith, R. E. (2016). Reducing prospective memory error and costs in simulated air traffic control: External aids, extending practice, and removing perceived memory requirements. *Journal of Experimental Psychology: Applied*, *22*(3), 272–284. https://doi.org/10.1037/xap0000088

Loft, S., Finnerty, D., & Remington, R. W. (2011). Using Spatial Context to Support Prospective Memory in Simulated Air Traffic Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(6), 662–671. https://doi.org/10.1177/0018720811421783

Loft, S., Pearcy, B., & Remington, R. W. (2011). Varying the Complexity of the Prospective Memory Decision Process in an Air Traffic Control Simulation. *Zeitschrift Für Psychologie*, *219*(2), 77–84. https://doi.org/10.1027/2151-2604/a000051

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), 145–157. https://doi.org/10.1037/a0018900

Loft, S., Sadler, A., Braithwaite, J., & Huf, S. (2015). The Chronic Detrimental Impact of Interruptions in a Simulated Submarine Track Management Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(8), 1417–1426. https://doi.org/10.1177/0018720815599518

Loft, S., Sanderson, P., Neal, A., & Mooij, M. (2007). Modeling and Predicting Mental Workload in En Route Air Traffic Control: Critical Review and Broader Implications. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(3), 376–399. https://doi.org/10.1518/001872007x197017

Loft, S., Smith, R. E., & Bhaskara, A. (2011). Prospective memory in an air traffic control simulation: External aids that signal when to act. *Journal of Experimental Psychology: Applied*, *17*(1), 60–70. https://doi.org/10.1037/a0022845

Loft, S., Smith, R. E., & Remington, R. W. (2013). Minimizing the disruptive effects of prospective memory in simulated air traffic control. *Journal of Experimental Psychology: Applied*, *19*(3), 254–265. https://doi.org/10.1037/a0034141

Logie, R. H., Zucco, G. M., & Baddeley, A. D. (1990). Interference with visual short-term memory. *Acta Psychologica*, *75*(1), 55–74. https://doi.org/10.1016/0001-6918(90)90066-o

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press on Demand.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition*, *27*(5), 759–767. https://doi.org/10.3758/bf03198529

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. https://doi.org/10.1002/acp.1002

McFarlane, D. (2002). Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction. *Human-Computer Interaction*, *17*(1), 63–139. https://doi.org/10.1207/s15327051hci1701_2

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, *14*(4), 299–313. https://doi.org/10.1037/a0014402

Morey, R., Rouder, J., Love, J., & Marwick, B. (2015). *BayesFactor: 0.9.12-2 CRAN*. R package version 0.9.12-2. https://doi.org/10.5281/zenodo.31202

Navon, D., & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review*, *86*(3), 214–255. https://doi.org/10.1037/0033-295X.86.3.214

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1–15. https://doi.org/10.1037/0033-295x.88.1.1

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. https://doi.org/10.1080/09658210444000313

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446–461. https://doi.org/10.1037/0033-2909.86.3.446

Ratwani, R. M., & Trafton, J. G. (2008). Spatial memory guides task resumption. *Visual Cognition*, *16*(8), 1001–1010. https://doi.org/10.1080/13506280802025791

Reason, J. (1990). *Human error*. Cambridge University Press. https://doi.org/10.1017/CBO9781139062367

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*(5), 511–524. https://doi.org/10.3758/bf03198390

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast Analysis: Focused Comparisons in the Analysis of Variance*. Cambridge, New York: Cambridge University Press.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/pbr.16.2.225

Salvucci, D. D. (2010). On reconstruction of task context after interruption. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*. https://doi.org/10.1145/1753326.1753341

Salvucci, D. D., & Taatgen, N. A. (2011). *The Multitasking Mind*. New York: Oxford University Press.

Sasangohar, F., Donmez, B., Easty, A. C., & Trbovich, P. L. (2017). Effects of Nested Interruptions on Task Resumption: A Laboratory Study With Intensive Care Nurses. *Human Factors*, *59*(4), 628–639. https://doi.org/10.1177/0018720816689513

Sasangohar, F., Scott, S. D., & Cummings, M. L. (2014). Supervisory-level interruption recovery in time-critical control tasks. *Applied Ergonomics*, *45*(4), 1148–1156. https://doi.org/10.1016/j.apergo.2014.02.005

Scott, S. D., Mercier, S., Cummings, M. L., & Wang, E. (2006). Assisting Interruption Recovery in Supervisory Control of Multiple Uavs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(5), 699–703. https://doi.org/10.1177/154193120605000518

Shorrock, S. T. (2005). Errors of memory in air traffic control. *Safety Science*, *43*(8), 571–588. https://doi.org/10.1016/j.ssci.2005.04.001

Shorrock, S. T. (2007). Errors of perception in air traffic control. *Safety Science*, *45*(8), 890–904. https://doi.org/10.1016/j.ssci.2006.08.018

Smith, R. E. (2010). What costs do reveal and moving beyond the cost debate: Reply to Einstein and McDaniel (2010). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1089–1095. https://doi.org/10.1037/a0019183

Stan Development Team. (2016). *RStan: the R interface to Stan. R package version 2.14.1.* Retrieved from http://mc-stan.org/

Stokes, D. E. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution Press.

Stone, M., Dismukes, K., & Remington, R. (2001). Prospective memory in dynamic environments: Effects of load, delay, and phonological rehearsal. *Memory*, *9*(3), 165–176. https://doi.org/10.1080/09658210143000100

Strickland, L., Heathcote, A., Remington, R. W., & Loft, S. (2017). Accumulating evidence about what prospective memory costs actually reveal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(10), 1616–1629. https://doi.org/10.1037/xlm0000400

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. R package version 2.41. Retrieved from https://CRAN.R-project.org/package=survival

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance, 29*(1), 3-18. http://dx.doi.org/10.1037/0096-1523.29.1.3

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, *58*(5), 583–603. https://doi.org/10.1016/s1071-5819(03)00023-5

Trafton, J. G., & Monk, C. A. (2007). Task Interruptions. *Reviews of Human Factors and Ergonomics*, *3*(1), 111–126. https://doi.org/10.1518/155723408x299852

Vidulich, M. A., & Tsang, P. S. (2012). Mental workload and situation awareness. *Handbook of Human Factors and Ergonomics*, 243. https://doi.org/10.4135/9781483328768.n8

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*(1), 89–106. https://doi.org/10.3758/bf03200763

Appendix

The Bayesian *t*-tests reported in Experiment 1 and 2 were conducted using a default prior distribution on the effect size for the alternative hypothesis. Specifically, we used a Cauchy distribution with a medium Jeffreys-Zellner-Siow prior that specifies a scale parameter of, $r = \sqrt{(2)}/\ 2$ (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Increasing the scale parameter *r* widens the prior distribution, reflecting a prior expectation of larger effect sizes; whilst *r* values closer to zero make the distribution narrower, reflecting prior expectations of smaller effect sizes.

An important question is whether the findings related to the deferred task performance measures were robust to the choice of prior selected. To answer this question, we conducted a Bayes Factor robustness check by recalculating and plotting BFs under a range of different priors. Specifically, each panel in Figures 1A and 2A contains a robustness check for one Bayesian *t*-test, and shows a range of scale parameters on the *x* axis and the corresponding recalculated Bayes Factors on the *y* axis. Across all panels, it can be seen that our interpretations of the BFs based on the user selected (default) prior remained consistent across a wide range of possible priors, indicating that our findings are robust to the choice of prior we selected.

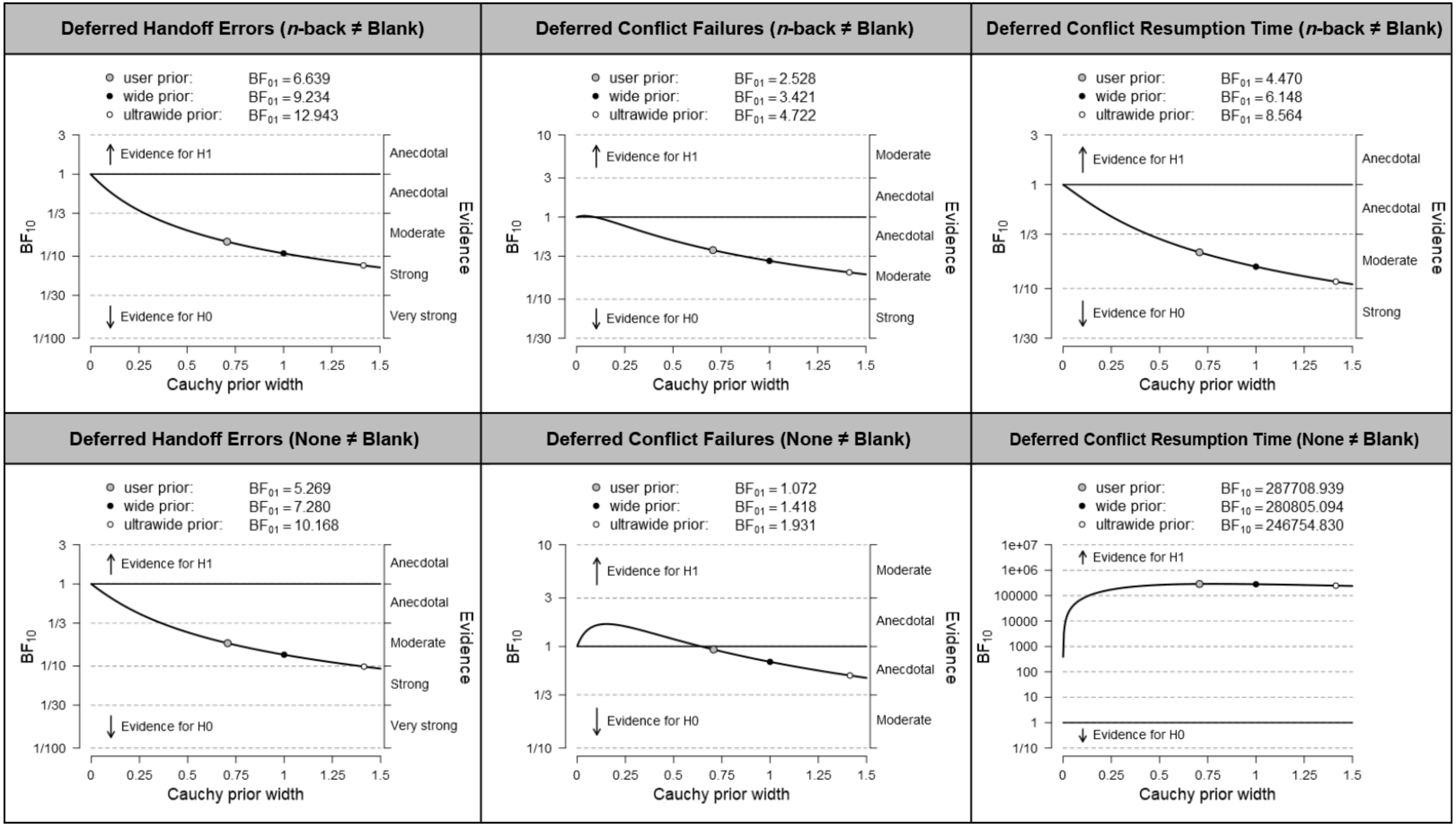

Figure 1A. Bayes Factor’s robustness plots for deferred handoff errors, deferred conflict failures (misses), and deferred conflict resumption time in Experiment 1.

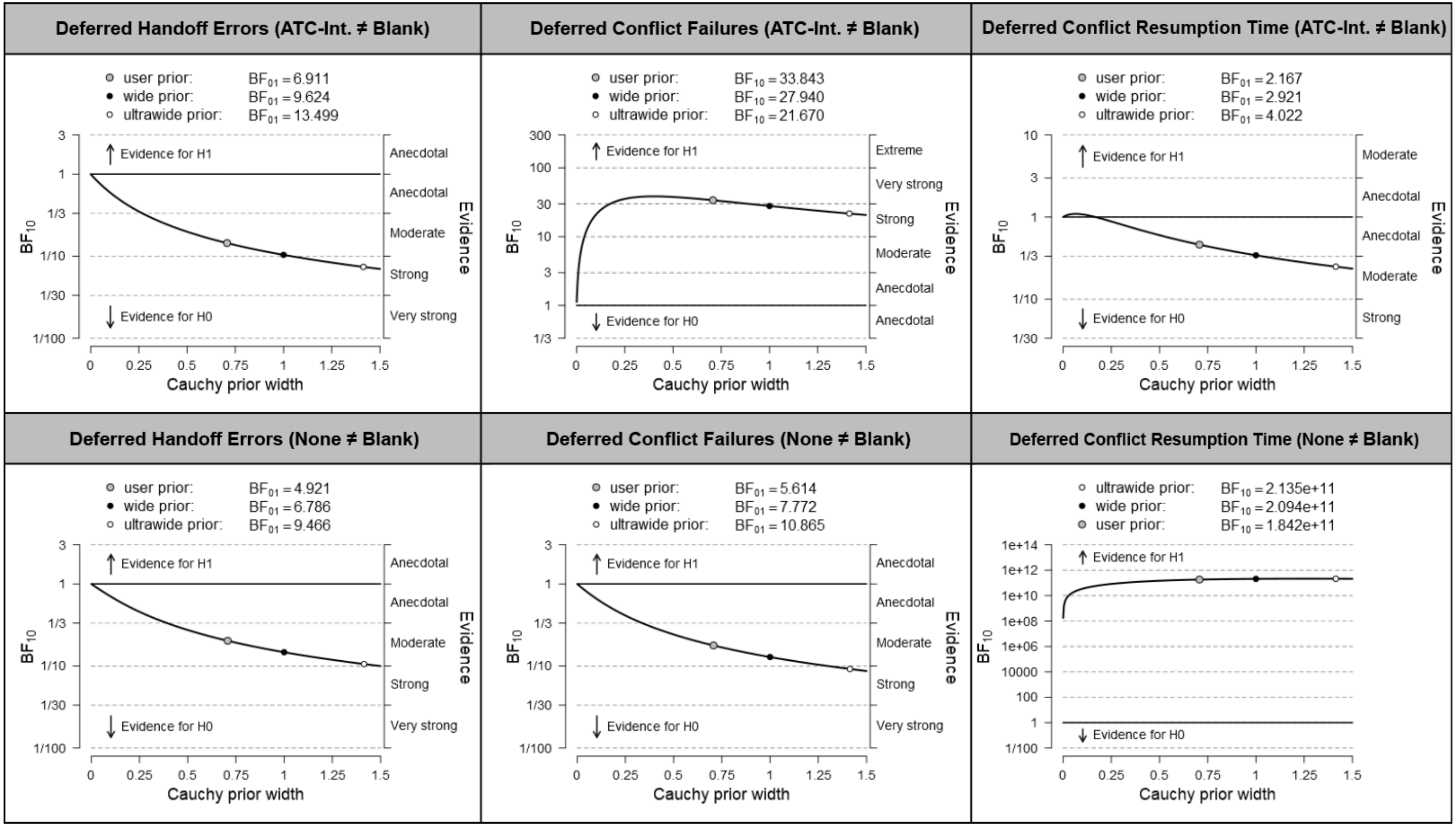

Figure 2A. Bayes Factor's robustness plots for deferred handoff errors, deferred conflict failures (misses) and deferred conflict resumption time in Experiment 2.