What Makes an Interruption Disruptive? Understanding the Effects of Interruption Relevance and Timing on Performance

CANDIDATE

Alexander J J Gould

SUPERVISORS

Dr Anna L Cox

Dr Duncan P Brumby

EXAMINERS

Dr Phillip L Morgan, *University of South Wales*
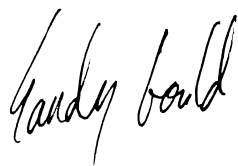
Prof Bradley C Love, *University College London*

A dissertation submitted in partial fulfilment of the requirements for the degree of:

Doctor of Philosophy of **University College London**

UCL Interaction Centre,
Department of Computer Science,
University College London

UCL

DECLARATION

I, Alexander Gould, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Interruptions disrupt activity, hindering performance and provoking errors. They present an obvious challenge in safety-critical environments where momentary slips can have fatal consequences. Interruptions are also a problem in more workaday settings, like offices, where they can reduce productivity and increase stress levels. To be able to systematically manage the negative effects of interruptions, we first need to understand the factors that influence their disruptiveness.

This thesis explores how the disruptiveness of interruptions is influenced by their relevance and timing. Seven experimental studies investigate these properties in the context of a routine data-entry task. The first three experiments explore how relevance and timing interact. They demonstrate that the relevance of interruptions depends on the contents of working memory at the moment of interruption. Next, a pair of experiments distinguish the oft-conflated concepts of interruption relevance and relatedness. They show that interruptions with similar content to the task at hand can negatively affect performance if they do not contribute toward the rehearsal of goals in working memory. By causing active interference, seemingly useful interruptions that are related to the task at hand have the potential to be more disruptive than entirely unrelated, irrelevant interruptions. The final two experiments in this thesis test the reliability of the effects observed in the first five experiments through alternative experimental paradigms. They show that relevance and timing effects are consistent even when participants are given control over interruptions and that these effects are robust even in an online setting where experimental control is compromised.

The work presented in this thesis enhances our understanding of the factors influencing the disruptiveness of interruptions. Its primary contribution is to show that when we talk about interruptions, 'relevance', 'irrelevance' and 'relatedness' must be considered in the context of the contents of working memory at the moment of interruption. This finding has implications for experimental investigations of interrupted performance, efforts to understand the effects of interruptions in the workplace, and the development of systems that help users manage interruptions.

Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Interruptions disrupt activity, often hindering performance and provoking errors. Understanding why interruptions have the effects that they do is a problem of critical importance in environments where errors can have serious, irreversible consequences. Whilst less pressing, interruptions are also a challenge in less safety-critical environments where interruptions reduce productivity (Sykes, 2011) and increase stress levels (Mark, Gudith, & Klocke, 2008).

However, research has also demonstrated that interruptions are not universally deleterious. In some situations, people switch to other tasks to improve their productivity (Jin & Dabbish, 2009; Duggan, Johnson, & Sørli, 2013). This implies that the effects of interruptions depend on contextual factors at the moment of interruption. For instance, attending to emails whilst writing a conference paper is likely to be disruptive, but attending to an email from the organizers with some important news about a change in deadline might be sufficiently beneficial so as to outweigh the disruption caused by switching away from the paper. Conversely, an email from the organizers might seem worth looking at, but could turn out to be an unhelpful, distracting message about accommodation arrangements.

The seven empirical invesitigations presented in this thesis explores the extent to which two factors — timing and relevance — influence the disruptiveness of interruptions during routine data-entry tasks. This work draws on a variety of research that has been conducted in the area of interruptions and multitasking. An important theory covered by the review in Chapter 2 is Memory for Goals (Altmann & Trafton, 2002), which provides a theoretical memory-centric account of the process of task suspension and resumption before, during and after an interruption. Previous work has made some progress toward understanding the role of relevance and timing in determining the disruptiveness of interruptions

(e.g., Czerwinski, Cutrell, & Horvitz, 2000; Iqbal & Bailey, 2005) and researchers have used these findings to implement systems that control the flow of interruptions to users (e.g., Arroyo & Selker, 2011; Iqbal & Bailey, 2010). However, previous investigations have failed to describe the properties that make interruptions relevant, or the role that interruption timing can play in determining relevance. Instead, studies and systems have classified interruptions as a relevant or irrelevant on the basis of assumptions about the relevance of study materials, limiting the generalizability of findings.

This thesis refines and extends our understanding of the effects of interruption relevance and timing on performance in a routine data-entry task. The task, a toy pharmacy-inspired task, requires participants to copy values from a central order sheet into the fields in each of the five subtasks that make up the task. The properties of the task and a justification of its suitability for the research questions are provided in Chapter 3.

Chapter 4 describes a series of three experiments that investigate the effects of timing and relevance on interruption disruptiveness. Disruptiveness was measured by how quickly participants resume the Pharmacy Task after the end of an interruption. This is known as the resumption lag, and is widely used as a measure of disruptiveness (e.g., Andrews, Ratwani, & Trafton, 2009; Cades, Boehm-Davis, Trafton, & Monk, 2011). Taken together, these experiments, which use the same design, contribute the first memory-centric account of interruption relevance. By framing relevance as a working memory problem, it is possible for interruptions to be dynamically categorised on the basis of their effects on the ephemeral contents of working memory, rather than through an *a priori* classification scheme. This chapter delivers a simple two-state model of interruption relevance; interruptions are categorised as relevant or irrelevant based on their effects on representations in working memory.

Chapter 5 recognises that a simple relevant/irrelevant dichotomy does not fully describe the variety of interruptions that are encountered; a stronger distinction between relevant and related interruptions is required. This chapter comprises two experiments of similar design that distinguish interruptions relevant to the task at hand from those that are merely related or have similar content, but are irrelevant. Experiment 4 investigates the effect of an interruption that has nothing to do with the task compared to interruptions that either encourage participants to rehearse their primary task progress or that interfere with their memory for primary task progress. The results show that a completely irrelevant interruption is less disruptive than an interruption that is related to the task at hand, and can interfere with recall of progress on resumption. Experiment 5 has the same design as

Experiment 4, but segments the process of resumption in order to examine the process in more detail. The results provide further evidence that interruption relevance stems from memory processes, and that caution is required when interpreting relevance as the relatedness of an interruption, or the degree to which content in a task and an interruption are similar.

Chapter 6 reports two experiments that investigate the robustness of the effect of relevance found in previous experiments. Experiment 6 explores the reliability of the relevance effect when participants are given control over when they are interrupted, while Experiment 6 takes the Pharmacy Task online to determine whether the relevance effect stands up to a loss of experimental control. Both experiments reproduce the previous finding: relevant interruptions that reinforce goals in working memory are less disruptive to performance than interruptions that interfere with the contents of working memory. Replicating the effect in paradigms different to the previous experiments gives greater confidence to the findings and strengthens the conclusions of the thesis with respect to interruption management systems.

The primary contribution of these experiments is to demonstrate that a simple relevant–irrelevant dichotomy is insufficient for describing the reinforcing and interfering effects that act on people's memory of where they are and what they are doing in a task. To determine the relevance or irrelevance of an interruption, one first has to understand which goals reside in working memory, and how these could be reinforced or disrupted by an incoming interruption. Later experiments suggest that how we measure the disruptiveness of interruptions may need to be re-evaluated in some situations. In particular, researchers should make more judicious use of resumption lag as a primary measure of disruptiveness; resumption lag should be augmented with other measures of disruptiveness such as error rates and subjective experience. Finally, the work demonstrates the reliability and generalizability of these working memory effects in less constrained environments by replicating the findings of the first five experiments: first by handing control over when interruptions appear to participants; and then by showing resilience to loss of experimental control in an online study.

This work has implications for researchers investigating the cognition of interruptions, for those trying to understand the effects of interruptions in the workplace, and for the development of systems that help users to manage interruptions. For experimental psychologists, this work shows that memory effects need to be considered when considering how interruptions affect task performance, in particular how goals change over task execution,

and consequently how the relevance of an interruption changes from moment-to-moment. For situated researchers and those developing interruption management systems, the experiments presented in this thesis show that the criteria for judging relevance need to be fundamentally re-evaluated. Rather than assume that an interruption is relevant because it has content in common with the task-at-hand, relevance also needs to include an understanding of the interfering and reinforcing effects of an interruption with respect to what users are keeping track of in working memory.

Chapter 2

Background and motivation

Interruptions stemming from devices, ourselves and other people are a ubiquitous feature of our personal and professional lives. The perception that we spend our lives bombarded by constant demands for our attention is prevalent, receiving regular coverage in the popular media (e.g., Thompson, 2005; Tugend, 2008; Carey, 2009; Wearden, 2010). It is not surprising, then, that researchers from a number of disciplines have picked up on interruptions and multitasking as an area where research has the potential to solve — or at least better understand — some of the problems that arise as a result of distractions and disruptions.

Multitasking behaviour and interruptions are of particular concern in safety-critical environments such as healthcare (e.g., Chisholm, Collison, Nelson, & Cordell, 2000; Westbrook, Coiera, et al., 2010), aviation (e.g., Loukopoulos, Dismukes, & Barshi, 2001; Latorella, 1998) and driving (e.g., Brumby, Salvucci, & Howes, 2009; Horrey & Lesch, 2009) where the consequences of errors are severe. Although the cost of errors is lower in less safety-critical environments like offices, the negative effects interruptions often have on performance and mood have motivated various situated explorations of interruptions (e.g., Mark, Gonzalez, & Harris, 2005; González & Mark, 2004).

To combat the proliferation of interruptions and mitigate their negative effects, researchers have developed systems that control the flow of interruptions to users (e.g., Iqbal & Bailey, 2010; Arroyo & Selker, 2011). These systems often assess the content of an interruption before delivering it at the moment it is most likely to be useful and least likely to be disruptive. Although such systems have been largely successful, some of the assumptions they make about the properties of interruptions do not have empirical support. This is particularly so when they consider the relevance of an interruption to the task at hand.

This thesis focuses on understanding the properties of interruption timing and relevance and the effects they have on how quickly and accurately people can resume after interruption. This chapter reviews existing work in the area of interruptions in multitasking to identify how prior work has conceptualised interruption relevance and timing. To begin, the problem of interruptions is broadly explored. The effects of interruptions on performance are used to contextualise efforts to develop interruption management systems. The assumptions of these systems are outlined with respect to psychological theories of interruption management. Questions that form the basis of the experiments in this thesis are developed on the basis of discord between the assumptions of management systems and these theories.

The problem with interruptions

Interruptions occur in almost every domain in our working and personal lives. The propensity of interruptions to negatively affect performance in a number of domains has made them a target for research. Although this thesis is focused on the cognitive processes involved in dealing with interruptions, many of the problems investigated by experimental work are motivated by real-world problems.

It is important to understand the motivations for experimental studies of interruption. Often work of this kind is motivated by the real-world problems that interruptions are thought to cause. Given this motivation, it seems prudent to first consider wether observational work supports the idea that interruptions are indeed problematic.

Situated work is not the primary concern of this thesis though, so rather than enumerate the totality of situated work conducted over the last few decades, this section focuses on interruptions in the healthcare domain. This is an area where interruptions have been well-studied and the routine procedural tasks that characterise much of the work in healthcare settings also provide the inspiration for the experimental tasks used in this thesis.

2.1   Interruptions in safety-critical settings

The potential for interruptions to negatively affect performance in safety-critical environments has been motivation for investigations of interrupted performance and multitasking behaviour. This section focuses specifically on understanding interruptions research in the healthcare domain. The idea that interruptions in this setting are prevalent and can lead

to errors is critically assessed on the basis of the evidence available. Attempts to design practice and artefacts in ways that mitigate their prevalence are also briefly evaluated.

Healthcare settings have been fertile territory for researchers investigating interruptions in safety-critical environments. This is perhaps because the dynamic and often busy nature of healthcare work precipitates interruptions, but it mostly likely a consequence of the potential for very high cost errors in healthcare settings. Healthcare is an ideal lens for reviewing interruptions research in a safe-critical environment because interruptions have been extensively investigated in the area. Due to the challenges of conducting research on wards, healthcare-based work is a good example of both the advantages and limitations of observational interruptions research.

Many on-ward investigations of interruptions in healthcare have focused on characterising the prevalence of interruptions on wards. In one of the early studies of healthcare interruptions, Chisholm et al. (2000) describe medical professionals as being 'interrupt-driven', with physicians experiencing interruptions every six minutes, on average. Other in situ studies have followed-up on this work with broad agreement that clinicians are frequently interrupted; on average, between six and twelve times every hour (e.g., Brixey et al., 2008; Kalisch & Aebersold, 2010; Westbrook, Coiera, et al., 2010).

Moving beyond raw measures of frequency, some descriptive accounts of interruptions on wards have attempted to characterise the sources of interruptions in addition to their frequency. Interruptions on the ward have a number of sources: medical professionals might be interrupted by other members of staff, patients or themselves (Biron, Lavoie-Tremblay, & Loiselle, 2009). Other researchers have noted that telephones, missing medication, emergencies or visitors can all prompt interruptions (Relihan, O'Brien, O'Hara, & Silke, 2010). Hillel and Vicente (2003) found that other clinicians were responsible for the majority of interruptions on a post-anaesthetic recovery ward. Evidence from other investigations has also shown that staff members are responsible for the majority of interruptions (McGillis Hall, Pedersen, & Fairley, 2010). Unfortunately, reducing the number of interruptions clinicians endure is difficult; Tucker and Spear (2006) show that 95% of all interruptions on a ward are directly related to patient care.

With a large literature strongly suggesting that interruptions are a frequent and persistent feature of healthcare settings, researchers have moved on to trying to understand the consequences of interruptions on performance. In particular, there has been a focus on the relationship between interruptions and errors. Errors are common during medical

procedures, with 18 errors made per 100 patients across all stages of treatment (Fordyce et al., 2003). Although only a small proportion of these errors (2%) have a significant effect on patient outcomes, given the number of procedures performed in healthcare systems on a daily basis, the absolute number of serious errors is still high. Even relatively old figures (Committee on Quality of Health Care in America & Institute of Medicine, 2000) suggest that almost 100,000 people in the United States die each year as a direct consequence of medical errors.

While the relationship between interruptions and errors is stated as a motivating factor in medically-situated interruptions research (e.g., Chisholm et al., 2000; Brixey et al., 2008), the link between the two has been hard to demonstrate empirically, largely due to method-ological difficulties. Mapping interruptions at a particular moment to errors which might only manifest minutes or hours later is difficult to do with any certainty. Nevertheless, some efforts have been made to investigate whether interruptions cause errors in medical settings.

Westbrook, Coiera, et al. (2010) investigated the relationship between interruptions and clinical task completion rates and found that 11% of all tasks were interrupted and that after 18% of interruptions, clinicians never returned to the task at hand. On some of these occasions, the task at hand would have been made redundant by the interruption. However, in other cases, important tasks may have been left uncompleted, compromising patient care. The same data revealed rushing behaviour as a result of interruptions, which itself could lead to an increase in error rates. In another study, Westbrook, Woods, Rob, Dunsmuir, and Day (2010) looked directly at error rates in interrupted tasks, finding that each interruption increased the chance of making a clinical error by 13%.

Being able to observe systematic errors has been one of the major challenges in human error research, as experts are generally accurate in resuming tasks. Although comprehen-sive on the surface, observational studies will need more inconspicuous methods of study — combined with orders of magnitude more data — to be able to satisfactorily describe the relationship between interruptions and errors on wards.

In a review, Grundgeiger and Sanderson (2009) also concluded that systematic investi-gation of the link between interruptions and error is difficult to achieve, adding that the complexity of the environment makes attribution of error through post hoc investigations difficult. This difficulty is in many ways unsurprising: given how few serious errors are

made as a proportion of all procedures (Fordyce et al., 2003), attributing causation to interruptions is a practical impossibility for ward-based research.

Despite the difficulties in matching interruptions to errors and other negative effects, a few studies have examined the effect of introducing interruption-mitigating practices to wards. Relihan et al. (2010) introduced training, checklists, special clothing and signage to a ward and saw a 58% decrease in the number of interruptions that occurred. The number of interruptions from other nurses was most affected by the intervention, indicating that the training had been successful. An evaluation period took place six months after the initial study and indicated that the effect of the interventions persisted into the long term. Pape (2003) also found a decrease in the number of interruptions that occurred on a ward after training had taken place. There is, however, no panacea for interruptions in healthcare settings; wards host a diverse array of actors, so solutions that effectively minimise one problem might create others. One hospital had to withdraw its red 'Do Not Disturb' tabards after patients complained that it made nursing staff unapproachable (Beckford, 2011).

Overall, situated investigations of interruption in healthcare have been valuable. Most importantly, they have demonstrated that interruptions are a real problem and have the potential to cause serious deviations in performance. The work also shows that interventions, whether through teaching or the introduction of supporting artefacts, can reduce the prevalence of interruptions. On the other hand, the healthcare-based research that has been done demonstrates the challenges of trying to understand the effects of interruptions in open, dynamic environments; tracing causal relationships between interruptions of various types and from a number of sources to deviations in performance is extremely challenging.

2.2   Experimental investigations of interruptions

Observational studies of interruptions have shown that interruptions are frequent and that their prevalence can be reduced with interventions. Although there is some evidence that increasing interruption frequency has a deleterious effect on performance, the nature of this relationship in real world environments is murky. This is because interruptions are mediated by a number of factors, not all of which are well understood. As Grundgeiger, Sanderson, MacDougall, and Venkatesh (2010) showed, mapping cause and effect are very difficult in complex environments like hospitals.

Barriers to the systematic *in situ* study of interruptions limits what can be said about the effects of interruptions on performance. More significantly, it is almost impossible to understand *why* interruptions have the effects that they do with situated approaches as they rarely contribute to or test theory. This is where experimental psychology, which has a relatively long history of interruptions research (see, e.g., Gillie & Broadbent, 1989), can make a contribution. Laboratory-based experiments afford the degree of control that is required to investigate the cognitive processes that determine how people deal with interruptions.

In this section, experimental investigations of interruptions are explored. There is a particular focus on how the timing and relevance of interruptions affects their disruptiveness. Important theoretical work is described alongside corroborating and disconfirming evidence. Finally, hypotheses about interruption timing and relevance are developed based on extrapolation from existing theory.

2.2.1 Timecourse of an interruption

When interrupted, people must suspend activity on the task that they are working on in order to attend to an interrupting task. After dealing with an interruption it is necessary to recall progress made before the onset of the interruption. The characteristics of interruptions and tasks modulate effects on performance over the timecourse of an interruption.

Trafton, Altmann, Brock, and Mintz (2003) identify two critical periods in the timecourse of an interruption, the *interruption lag* and the *resumption lag* (see Figure 2.1). The interruption lag is the period between an interruption being forewarned and it being attended to. For instance, an alert of notification might appear indicating that an interruption is imminent some time before the interruption occurs. After an interruption has been dealt with, it is then necessary to resume the original task. The period of time between completing an interruption and resuming a primary task is the resumption lag.

This theory of interruptions makes several assumptions about how people deal with interruptions. Firstly, an interruption lag presupposes an alert or notification before the onset of an interruption. Secondly, this theory assumes a linear path from primary task to interrupting task and back to primary task. In reality, people may not return to the task they were working on before they were interrupted because the task is forgotten or is made redundant by new information (Westbrook, Coiera, et al., 2010). Previous attempts to use this theory to understand interruptions in real workplaces have found a number of difficulties in matching messy events in reality to the structure of the model (Grundgeiger et al., 2010). As such, this model is most relevant and useful in environments where researchers
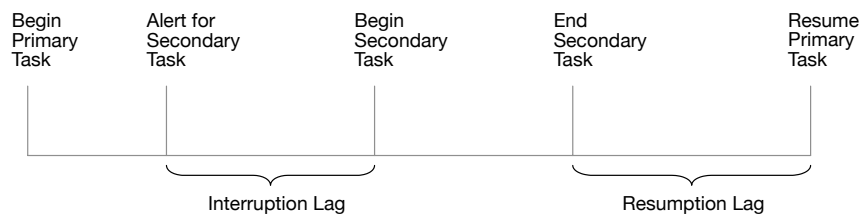
Figure 2.1: The timecourse of an interruption, redrawn from (Trafton et al., 2003). Note that interruption lag only exists when interruptions are forewarned through an alert or notification. The resumption lag period only exists if the primary task is resumed after the completion of an interruption.

have systematic control over alerts and can ensure participants return to their task after working on an interrupting task.

This thesis focuses specifically on how the time it takes to resume after an interruption — the *resumption lag* — is affected by the attributes of interruptions. Although the effects of pre-interruption events on interrupted performance are of research interest (e.g., Morgan, Patrick, & Tiley, 2013; Trafton et al., 2003), research has generally focused on the resumption component of interruptions. This is most likely because errors and deviations in performance that can be measured manifest at the point of resumption. This thesis also focuses on the process of recovering after interruption. This component of the timecourse of an interruption is thus of particular interest.

2.2.2 Recovering after an interruption

Interruptions cause degraded performance, both in terms of the time it takes to resume after an interruption and the number of errors made during a task (e.g., Altmann & Trafton, 2004; Monk, Boehm-Davis, Mason, & Trafton, 2004; Brumby, Cox, Back, & Gould, 2013). How people resume a task after interruption motivated much of the early interruption research. This was partly because it is at the point of resumption that an error is most likely to occur, but also because resumption after interruption provided an excellent way for researchers to test their theories of activation-based memory.

Of particular interest has been how people resume their primary task, because it is the point at which errors manifest. Bailey and Konstan (2006) found that interruptions doubled the incidence of errors on a routine task. They also found that interruptions increased participants self-reports of stress (see also Adamczyk & Bailey, 2004). Monk et al. (2004) studied interruptions during a driving task and found that after interruptions, as well as being slower to resume a task after interruption, participants resumed their primary task in the wrong place and did not realise their error. This is commonly called a *sequence error*

in the interruption literature. Ratwani and Trafton (2006) also showed that interruptions result in a resumption time that is longer than simply moving between subtasks, indicating some additional cost associated with resuming after interruptions irrespective of the normal motor costs of moving the mouse and eyes to the next subtask (see also Ratwani & Trafton, 2010). Overall the resumption literature paints a picture of slower, more error-prone performance as a result of interruption.

2.2.2.1   The role of memory in post-interruption resumption

Altmann and Trafton's (2002) Memory for Goals theory is an activation-based model of memory that describes how goals representing progress through a task are managed before, during and after interruption. This model of interrupted performance holds that when interrupted during a routine task, representations in working memory ('control codes') allow people to keep track of the step they are working on (Trafton, Altmann, & Ratwani, 2011). At the moment of interruption occurring, the goal representing the last complete subtask should have the highest activation.

In providing an account of interrupted performance Memory for Goals leverages a number of concepts that have been developed over several decades in the human memory literature. Terms and concepts used in Memory for Goals theory are frequently referenced in the experiments that comprise this thesis. To ensure that these terms are unambiguous, they are individually described here in the style of an extended glossary.

Control codes   The most component of Memory for Goals theory are control codes. Control codes are goals in working memory that store place-keeping information about task progress (Altmann & Trafton, 2002). These goals can be retrieved after a change of context – an interruption, for instance – so that a task can be resumed from the point at which it was left. In Memory for Goals theory, control codes represented the last completed step (Trafton et al., 2011). On post-interruption resumption the next step is inferred from the last completed step and knowledge of the task.

Control codes represent place-keeping information and are created during the course of task execution. They do not represent task-specific information that has been encoded in working memory. Memory for Goals makes no claims about the encoding of task-specific information, although other theories like Threaded Cognition (Salvucci & Taatgen, 2010) do provide an account of how these content-related goals are managed.

Goal decay   Memory for Goals is based on an activation model of working memory. Control code goals in working memory have some level of activation associated with them. Over time the activation level of a particular goal falls. The result is that goals can be forgotten.

The activation level of a goal is significant because it determines which goal is most likely to be retrieved from working memory. The higher the activation of a goal the more likely it is to be retrieved. If the wrong goal has the highest activation the wrong step in a task is likely to be executed (Trafton et al., 2011). Furthermore, when goals are marked as completed they lose a significant amount of activation. This means that any task steps required after a goal has been completed are likely to be omitted (Byrne & Bovair, 1997; Li, Blandford, Cairns, & Young, 2008).

Goal reinforcement   Although most of the time the activation level of a particular goal is falling, goal activation can be boosted such that mean activation remains high. Goals can be boosted directly through rehearsal of a goal. Rehearsal occurs every time that a goal is accessed. Goals might be accessed because of a conscious effort to retain a particular goal or because they are necessarily accessed as part of the execution of a task.

Goals also receive activation boosts because of spreading activation from other goals (Altmann, Trafton, & Hambrick, 2013). This means that when a related goal receives a boost in activation from rehearsal any related goals also receive a smaller activation boost. In this way goals that are related to the current goal in focus can be primed.

Interference threshold   Goals in the form of control codes have a particular level of activation in memory. This level of activation decays over time. After a goal has decayed sufficiently, retrieval will be unreliable. The point at which this occurs is known as the interference threshold.

As a task is executed, goals are created. Ones that are not reinforced have their activation drop. The process of retrieving goals from memory is stochastic and goals with the highest activation are the most likely to be retrieved. Eventually the activation of a goal will drop sufficiently that the probability of it being retrieved is effectively zero: the goal has been forgotten.

The stochastic nature of the process means that forgetting is not the only possibility. The interference threshold is determined by the activation of a target goal compared to other goals in working memory. As the activation of a target goal falls, the relative difference

between its level of activation and that of old goals in working memory shrinks. This means that the likelihood of the target goal being retrieved shrinks relative to old goals. The result of this is that errors can occur. For instance, perseveration errors occur when the activation level of a goal that has previously been completed is sufficiently high that it interferes with the retrieval of the 'correct' goal. The previously completed goal is thus retrieved instead of the target goal and the step gets inappropriately repeated (Trafton et al., 2011).

The combination of goal decay, goal reinforcement and the interference threshold determine how disruptive a given interruption will be. Some interruptions will allow reinforcement, boosting the activation of the target goal. Other interruptions will preclude rehearsal, resulting in goal decay and – near the interference threshold – unreliable retrieval. These concepts are applied throughout this thesis to explain the observed phenomena.

Memory for Goals suggests that place-keeping control codes must be maintained in working memory during interruptions. As with any representation in human memory they are susceptible to a variety of memory effects (see, e.g., Ratwani, Andrews, Sousk, & Trafton, 2008; Morgan, Patrick, Waldron, King, & Patrick, 2009). Reinforcement of place-keeping goals for the duration of an interruption broadly reduces the likelihood of slow or inaccurate resumptions (Monk, Trafton, & Boehm-Davis, 2008; Monk et al., 2004). Interference has the opposite effect on the speed and accuracy of recall (Altmann et al., 2013). Other empirical results corroborate the theoretical assumptions of Memory for Goals theory (see Monk et al., 2008; Trafton et al., 2003), even if in reality the process of resumption might be moderated by the requirements of the primary task and strategic decision-making (e.g., Brumby et al., 2013; Morgan & Patrick, 2013; Salvucci, 2010).

Following the assumptions of the Memory for Goals theory, one can predict that goals representing task progress should be susceptible to both interference and reinforcement. This means that using content similarity as a proxy for relevance has the potential to cause unanticipated increases to the disruptiveness of interruption. For example, an incoming interruption could have similar content to that of the current activity; an interruption management system might see this as relevant, and prioritise it. However, the interruption might relate to a step in the task that has yet to be started or has been completed already. This could cause proactive interferences with memory of task progress, increasing interruption disruptiveness.

To illustrate this point consider a user working on a particular step of a routine data-entry task. When they are interrupted, it is important that they remember where they were in the primary task. In one scenario, the incoming interruption calls for direct rehearsal of progress by asking participants explicitly about their progress through the primary task. In another scenario, participants are asked random questions about the task environment. While from a content similarity perspective both interruptions are relevant to the primary task, they are likely to have different effects on memory: the former forces rehearsal of the point where the user needs to resume, whereas the latter might encourage proactive interference by forcing participants to divert their attention to other parts of the task which may have already been completed. Indeed, an interruption with no content similarity might prove to be less disruptive by virtue of it merely preventing rehearsal task progress, rather than causing proactive interference.

Of course, the capacity of an interruption to elicit reinforcement or interference effects depends on the representation of progress that is maintained for the duration of the interruption. These representations change moment-to-moment during task execution, suggesting that relevance should also be seen as dynamic property, the magnitude of which depends on the timing of an interruption.

2.2.3 The role of artefacts in resumption

People are usually not completely reliant on their memory for quick and accurate resumption. Research has examined the extent to which cues that indicate where to resume help increase resumption speed and reduce the occurrence of errors. Prior work shows that introducing explicit cues that indicate what needs to be done next improves speed and accuracy of resumptions (Jones, Gould, & Cox, 2012; Hodgetts & Jones, 2006a; Altmann & Trafton, 2004; Chung & Byrne, 2008).

In all but the most linear of routine tasks, programatically determining which step comes next is difficult (Jones et al., 2012). Researchers have therefore investigated whether showing participants what they were doing before they were interrupted can improve performance. While less effective than cues indicating which step comes next (Trafton, Altmann, & Brock, 2005; Jones et al., 2012), such cues are still more effective than having no cues at all (Kern, Marshall, & Schmidt, 2010; Parnin & DeLine, 2010).

The effectiveness of cues in improving resumption performance is not a focus of this thesis. However, the role of artefacts in a broader sense is investigated. In the context of resumption, artefacts are any non-memory resource that can be leveraged to aid resump-

tion. Artefacts do not necessarily represent progress; instead they provide resources that people can leverage to make the process of resumption easier. This might include the visual layout of the task or indirectly useful information presented through the interface.

While there have been a number of investigations into the role of cues, there have been fewer studies of how the general characteristics of a task influence resumption performance. Ratwani and Trafton (2008) investigated how the layout of a task influenced how it was represented in memory. They had participants perform a moderately taxing transcription exercise that required them to work downward through a list of digits, typing only even numbers. Ratwani and Trafton then interrupted participants either with a simple arithmetic task or a mental rotation task. They showed that performing mental rotations was more disruptive than doing arithmetic. This, the authors suggested, showed that place-keeping goals in the task were in visual memory. This finding is interesting because it suggests that the design of a task can induce people to encode working memory representations in a specific way. This encoding then determines the potential for an interruption to be disruptive. However, their study does not provide unequivocal evidence of the role of primary task artefacts. It did not investigate whether systematic manipulation of the primary task led to corresponding changes in working memory representations and, in turn, interruption disruptiveness.

Other work has attempted to systematically manipulate primary task artefacts and measure interruption disruptiveness. Borst, Buwalda, van Rijn, and Taatgen (2013) investigated how the external environment can support maintenance of representations in working memory during multitasking. They compared two interfaces for a mental arithmetic task, one that kept track of units carried over and another that did not. Their results showed that an external representation of carried units reduced load and improved performance. While this study is a step towards a systematic evaluation of interfaces, this investigation was in the same vein as prior work on the effect of explicit place-keeping cues on interrupted performance. To date no work has focused on how the layout of interfaces or the representation of information in primary tasks affects interruption disruptiveness.

2.3 Interruption relevance and timing

2.3.1 Interruption timing

The contributions of working memory and task structure to interrupted performance manifest in a number of ways. One of these is how the timing of an interruption affects its disruptiveness. Interruptions are usually disruptive at any time, but one might intuitively

feel that there are better and worse moments to be interrupted. Indeed, investigating what makes particular moments better or worse has been an active area of research in the community. For instance, previous work has shown that interruptions arriving at subtask boundaries are less disruptive than those arriving in the middle of completing a subtask (Cutrell, Czerwinski, & Horvitz, 2000; Monk et al., 2004). Iqbal and Bailey (2005) suggest that this is because subtask boundaries represent moments of low workload; it is easier to be interrupted at the end of a sentence than in the middle of writing one.

This is reflected in the design of systems that manage interruptions. The *Oasis* system developed by Iqbal and Bailey (2010) holds notifications until a user reaches a natural breakpoint in their activity. They define three degrees of granularity for breakpoints: coarse, medium and fine. At one end of the spectrum, fine-grained breakpoints involve switching between writing and reviewing a document. At the other end of the spectrum, coarse-grained breakpoints occur when switching to an entirely independent task. Interruptions are better left until coarse-grained breakpoints when there is no ongoing context that needs to be maintained.

Further support for the notion of better and worse moments for interruption comes from the way that people themselves manage interruptions. Salvucci and Bogunovich (2010) investigated how people managed interruptions in a simple task. Participants were asked to manage an email inbox of incoming requests. These requests required that participants fetch some product information from another another window and hold it in memory until they wrote it down. They also had to attend to instant messages. Salvucci and Bogunovich (2010) found that people tended to wait until they were not storing product information in their head before they switched to the instant messaging task. They suggested that maintaining representations of product information in working memory (they refer to it as the problem-state) required significant cognitive resources. Once a request was completed these resources were no longer required. Without the costs associated with forgetting the product information people were more inclined to switch to the instant messaging task. Their experiment demonstrates that variations in workload are reflected in the decisions that people make about whether or not to deal with a discretionary interruption: given the choice, people will wait until subtask boundaries before interrupting themselves.

Using task boundaries as a proxy for workload is likely to work well as long as moments of low workload align with task boundaries. However, it is unclear whether subtask boundaries are still a useful heuristic for interruption management systems if workload does not

align with subtask boundaries. For example, in some data-entry tasks, planning where information needs to be entered might be more taxing than the act of entering the values itself. If reduced disruptiveness at subtask boundaries were simply a corollary of workload, one would expect that tasks with consistently low workload would confer no advantage on between-subtask interruptions. Returning to theory might help to make sense of this conundrum.

Previous work has developed a theory of interruption timing for routine tasks. Botvinick and Bylsma (2005) investigated interrupted performance after their network-inspired modelling work (Botvinick & Plaut, 2004) suggested that interruptions were likely to be most disruptive in the middle of the subtask, rather than at subtask boundaries. In their experiment they interrupted participants at various points while they were performing the routine task of making coffee. They found that when participants were interrupted in the middle of a subtask, they were more likely to make errors than they were at subtask boundaries.

Botvinick and Bylsma (2005) attributed the increase in errors during subtasks to a weakening of temporal context: information that was needed to make a decision about what had been done and what was left to do were most strongly represented at decision points. These decision points occur between subtasks when people need to decide what needs to be done next. Representations of what has been done and what is left to do are a necessary component of the decision-making process, so these elements are most strongly represented between subtasks. Within a subtask, information associated with decision points is less strongly represented because it is not required. It is therefore more easily disrupted.

Memory for Goals theory (Altmann & Trafton, 2002) describes interrupted performance in a similar manner, but makes no explicit predictions about the effect of interruption timing. Rather than talking about the strength of a representation in a network it explains resumption through place-keeping goals which decay over time. The extent to which these goals have been accessed or rehearsed determines their activation. At any given point in time the activation of a goal exhibits stochastic variation. The degree of activation at the point of resumption determines how likely it is that people will be able to recall where they are in a task. The relative activation of a goal depends on the time since it was last accessed. It follows that activation of a particular goal changes over the course of a task.

Memory for Goals does not attempt to predict how within-subtask interruptions are processed (Trafton et al., 2011). Rather, it assumes that place-keeping information is stored

discretely and represents the last subtask completed, rather than progress through a sub-task. No clear mechanism for keeping track of progress in the middle of a task is eluci-dated. When resuming after an interruption, the last completed task is recalled and within subtask progress is, presumably, deduced from visual cues. How then do people resume accurately in the middle of subtasks in the absence of such cues?

Despite being the dominant theoretical account of the cognition of interruptions, Memory for Goals theory does not seem to account for resumption behaviour after mid-subtask interruptions. However, it does provide a well conceptualised theoretical framework such that it can be extended to describe within-subtask interruptions. The final section of this chapter develops a Memory for Goals-based theory of the effect of interruption timing. Before this theory can be elucidated however, it is first necessary to consider the role that interruption relevance plays in interruption disruptiveness. As will be been, this thesis holds that these concepts are intimately related.

2.3.2 Interruption relevance

Intuitively, one might expect that interruptions that are relevant to the task at hand will be less disruptive than irrelevant ones because they might contain useful information and maintain context. Support for this intuition has been found in number of studies (e.g., Czerwinski et al., 2000; Adamczyk & Bailey, 2004; Iqbal & Bailey, 2008). Assessing the relevance of an interruption has the advantage of allowing for the blocking of irrelevant interruptions and the prioritisation of relevant interruptions. This makes it possible for the beneficial aspects of interruptions — timely delivery of information, for instance — to be exploited to maximum effect. Of course, the effectiveness of relevance as a tool for distinguishing more useful interruption from less useful ones is dependent on the opera-tionalisation of relevance.

Arroyo and Selker (2011) developed an interruption management system that used rel-evance to determine whether incoming notifications should be released to users or held back. Their system computed relevance by comparing the content of the current activ-ity and an incoming notification by examining files. If there was a good correspondence, the interruption is deemed to be relevant. For example, if one was working on a report on Cornish viaducts, interruptions containing the words 'Cornish', 'viaduct', 'Brunel' and 'rail-way' might be flagged as relevant. However, such a system cannot distinguish between an interruption that is relevant to the task at hand and one that has similar content, but is nonetheless irrelevant. In other words, the same system might also label interruptions about Cornish holidays and Roman viaducts as relevant.

Other investigations of interruption relevance have also treated 'content similarity' and 'relevance' as equivalent. For instance, Czerwinski et al. (2000) defined relevant interruptions as those that gave participants the solution to a question they were trying to answer in the primary task. Irrelevant interruptions gave participants some fact about the environment in which they were working. Iqbal and Bailey (2008) make a similar distinction: relevant interruptions were related to the current task, irrelevant interruptions were not. Although these studies have shown that relevant interruptions are less disruptive than irrelevant interruptions, defining relevance on the basis of content similarity ignores the memory processes that play a significant role in resuming after interruption.

How, then, can the concept of relevance best be represented under existing memory-centric theories of interrupted performance? Existing theories do not explicitly discuss the effects of interruption relevance in the context. Whether they can support a memory-based definition is therefore uncertain.

In the context of Memory for Goals theory, it might be possible to define relevance through the rehearsal and interference effects that constitute the basis of the theory. An interruption might be relevant if it encourages rehearsal and irrelevant if it causes interference. Such a definition would have the advantage of fitting neatly alongside existing findings without requiring additional theory to explain the relationship between relevance and disruptiveness. More fundamentally, it would provide a theory-backed definition of relevance. Previous work has investigated the effects of relevance on disruptiveness (e.g., Iqbal & Bailey, 2008; Czerwinski et al., 2000) but has failed to provide a convincing explanation of what it means for an interruption to be relevant or irrelevant in the first place.

An alternative theory of relevance is posited by Salvucci and Taatgen (2010). They hold that the processing of relevant interruptions described in a study by Cutrell et al. (2000) shows that "differences between conditions are not due to the retrieval of problem state[...]; rather, they arise from the need to create a new problem state for irrelevant interruptions." (pp. 168) Which is to say that performance is better with relevant interruptions because the cost of creating a new problem-state is avoided.

In the problem-state view, an interruption can be seen as relevant if the currently loaded problem-state could be used to address the interruption. An irrelevant interruption would require the deconstruction of the existing problem space, construction of a new problem-state for the interruption and the reconstruction of the primary-task problem-state once

the interruption was finished. This account suggests that any interruption that makes use of the currently loaded problem-state must be relevant.

Although the role of relevance in interruption disruptiveness has been investigated, there are a number of gaps in the motivation and results of prior work. This thesis develops a theoretically-supported definition of relevance that is empirically tested.

This section has explored the possibility that interruption relevance is the product of the contents of working memory at the moment of interruption and the working memory requirements of interrupting tasks. A corollary of this formulation is that relevance must vary over the course of task execution as the content of working memory is changed to meet task demands.

This temporal component to relevance suggests that the timing of interruptions is a critical factor in determining the relevance of a given interruption. An interruption might be relevant at one point during task execution, but, as time goes on and working memory representations change, the same interrupting task has the potential to interfere with the contents of working memory, reducing performance. The next section explores the connection between interruption relevance and timing from a theoretical perspective. It builds on Memory for Goals and explains the relationship between working memory and interruption timing and relevance. A number of assumptions are made; these are empirically tested in the experimental chapters of this thesis.

2.3.2.1 A Memory for Goals-inspired theory of interruption timing and relevance

This section proposes an Memory for Goals-inspired theory of the effects of interruption timing and relevance on resumption performance. The purpose of this theory is twofold. First, it provides a theory of resumption after within-subtask interruptions that is compatible with Memory for Goals. Secondly, it ties together the concepts of interruption timing and relevance, giving them a consistent theoretical underpinning.

To develop a theory of within-subtask interruption is first assumed, in keeping with Memory for Goals theory, that place-keeping representations ('control codes') are changed at subtask boundaries and that goal decay is continuous over the period of a subtask. Based on this premise it is possible to make predictions about the effects of within-subtask interruptions. Let us work up to these predictions from the principles of Memory for Goals. First, consider the execution of a routine task made up of three subtasks, *A*, *B* and *C*. These are represented visually in Figure 2.2.
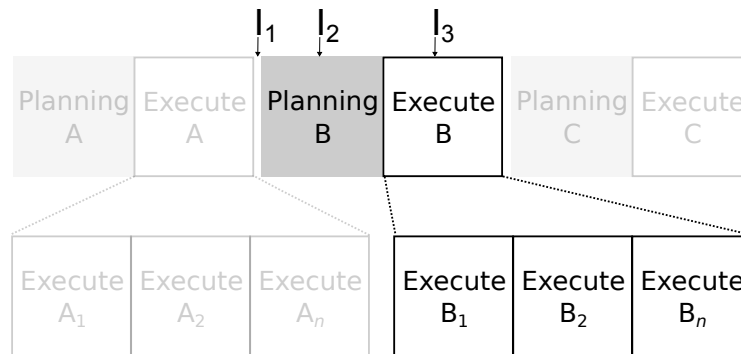
Figure 2.2: Illustration of sequence of subtask planning and execution with notional points of interruption marked.

Consider a scenario through the lens of standard (i.e., Altmann & Trafton, 2002; Trafton et al., 2011) theory. Figure 2.2 represents three subtasks of a notional task and the components of each subtask that require execution. At the moment of an interruption, represented by I_1, work on subtask *A* has been completed, but planning has not taken place for the execution of subtask *B*. This is the between-subtask situation that the Memory for Goals model was designed to describe. When resuming after I_1, the model assumes that the place-keeping goal (i.e., control code) currently held in memory is for the last completed subtask (*A*). It then infers that the next step that needs to be worked on is subtask *B*. At this point the model might make a *perseveration error*, resulting in subtask *A* being repeated. This occurs because of the noisy nature of memory in Memory for Goals theory: old place-keeping goals from previous subtasks can intrude when activation for the correct place-keeping goal drops sufficiently for old place-keeping goals to be picked.

It is also possible for the model to skip forward, missing a subtask. These *anticipation errors* are most likely to result from interruptions at I_2. In this situation work has been completed on subtask *A*, and planning has begun on subtask *B*. When resuming after interruption at this point, the place-keeping goal is now set for the next *intended* step rather than a *completed* step. The model is insensitive to this, and as a result begins work on subtask *C* because the place-keeping goal is interpreted as being a signal that subtask *B* is complete.

The model does not make a prediction about the process of resumption after interruption at point I_3. This is because for each subtasks is made up of a number of steps. The interruption has occurred within a subtask. Memory for Goals explicitly does not account for this scenario (Trafton et al., 2011). However, extending the logic of the model means it is possible to make some predictions about resumption after a within-subtask interruption. The predictions depend largely on the representational model of subtasks that is used;

specifically whether new place-keeping goals are introduced during the course of a sub-task (i.e., subtasks of subtasks are atomic) or whether subtasks are completed under a single place-keeping goal.

The Memory for Goals model does not provide for the mid-subtask introduction of place-keeping goals, so here it is assumed that a place-keeping goal is maintained with high activation as participants work on a subtask. Assuming this assumption is correct, one can predict that anticipation errors are more likely to occur than they would be at subtask boundaries. As the place-keeping goal represents the last completed subtask, on re-sumption after interruption the model predicts that participants should assume they have completed the subtask during which they were interrupted and will attempt to begin work on the next subtask. The likelihood of a perseveration error depends on the point of inter-ruption in the subtask.

Activation of the current place-keeping goal is maintained over the course of subtask execution and only begins to fall at the onset of an interruption. Therefore if interruption is earlier in the subtask, it is more likely that an old place-keeping goal will be picked up on resumption than it would be if the interruption were later in the subtask, due to the larger difference in activation between the correct and old place-keeping goals. This is intuitively plausible: the further one gets from an old subtask, the less likely one is to accidentally go back to it. Figure 2.3 is a diagrammatic representation of how interruption timing can affect the likelihood of successful resumption after interruption. The figures illustrate that the closer an interruption comes to the completion of a *previous* subtask, the more likely it is that a resumption error will be made. This is because old goals are not rehearsed and as a consequence their activation falls; the longer the activation falls, the less likely it is that the old goal will be accidentally retrieved when participants resume after interruption. As well as increasing the likelihood of error, lower activation levels also makes goal retrieval more effortful and thus time consuming (although see Brumby et al., 2013, for a contrasting explanation).

This theory of subtask execution does not account for accurate resumption within a sub-task. For accurate within-subtask interruption participants must create place-keeping goals: subtasks of this task are themselves made up for further subtasks and that these sub-subtasks are represented in the same manner as their parents. Figure 2.4 repre-sents progress through two subtasks where each subtask is itself constructed of discrete subtasks. Rather than representing exclusive descriptions of within-subtask resumption it

(a) Decay when interruption is early in subtask

(b) Decay when interruption late in subtask
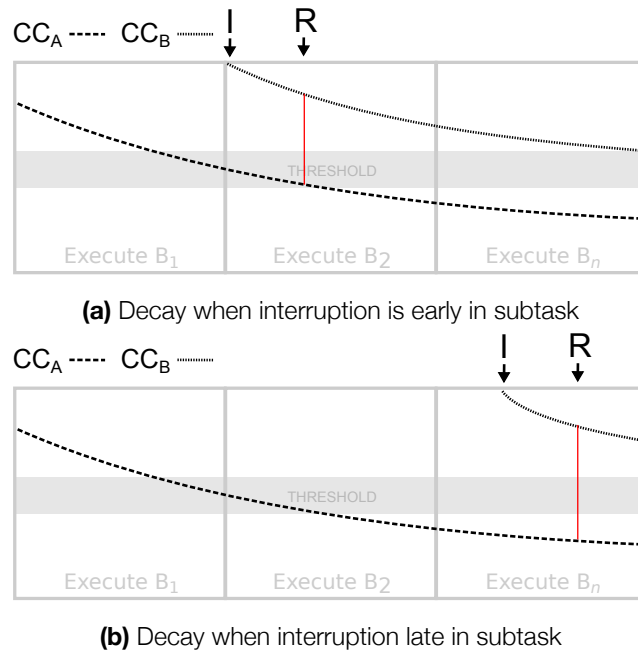
Figure 2.3: Illustrative diagrams showing notional decay of place-keeping goals for subtask A (CC_A) and subtask B (CC_B) when interruption arrives early (a) and late (b) in subtask B. Labelled points I and R indicate interruption and resumption points respectively. Vertical line at point R indicates notional differences in activation levels between old (CC_A) and new (CC_B) goals at the point of resumption.

might be that they merely represent two potential strategies for subtask completion: one strategy is picked based on the demands of the environment.

If the environment provides sufficient cues for resumption such that in-memory place-keeping is not required or that the cost of keeping and manipulating place-keeping goals exceeds the benefit of doing so, then it might be that a place-keeping goal-free strategy is adopted. If there are few or no cues for resumption and the code of maintaining place-keeping goals is exceeded by the cost of forgetting one's position in the task, then one might expect that participants would utilise place-keeping goals at a within-subtask level. How participants respond to interruptions will depend heavily on the strategy participants use to keep track of where they are in the task. Given that the interruptions used in this experiment are designed to help rehearse or interfere with place-keeping goals at the subtask level, if within-subtask place-keeping goals are stored as a different representation then one might expect these place-keeping goals to be less affected by the interruptions.

As previously discussed, Memory for Goals does not provide an account of goal management between subtasks, except to say that goals decay over time. Prior investigations of interruption timing have characterised it as the effect of variations in workload on interruptability. The extent to which 'good' interruption timing simply represents the coincidence
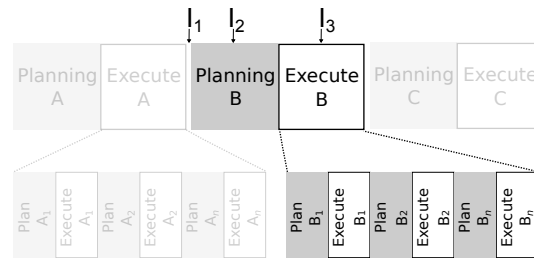
Figure 2.4: Planning and execution of subtasks with potential points of interruption marked. Note that unlike Figure 2.2, sub-subtasks also have place-keeping goals and planning associated with them.

of a particular set of operations with a particular set of working memory representations has not been adequately explored.

Interruption timing has been characterised here as the moment of interruption relative to the changes in working memory representations during task execution. This thesis proposes that the concept of relevance can be viewed in the same terms. It is hypothesised that interruption relevance can be defined as the effect of an interrupting task on representations in working memory of the course of an interruption. As there is currently no theory of interruption relevance, this represents the first attempt to produce a theoretically informed concept of relevance.

One of the assumptions of Memory for Goals theory is that representations in working memory decay. Place-keeping goals' activation levels fall over time, but this activation can be boosted when goals are rehearsed. During an interruption, rehearsal is inhibited. This is why longer interruptions are more disruptive (Monk et al., 2008; Li, Cox, Blandford, Cairns, & Abeles, 2006; Hodgetts & Jones, 2006b). Memory for Goals makes no requirement that interruptions block rehearsal however. This often assumed to be the case because interruptions used in experiments usually require participants to work on interrupting tasks that have little to do with the task at hand. This means that place-keeping goals in memory decay and are forgotten. If interruptions make use of working memory representations created for the interrupted task however, there is no reason that activation levels need fall.

The hypothesis tested by this thesis is that relevance simply represents the degree of correspondence between the representations stored in working memory and the demands of an interruption. Relevant interruptions encourage rehearsal of place-keeping goals stored in working memory. This boosts their activation and as a result reduces disruptiveness. Irrelevant interruptions block rehearsal or introduce proactive interference. The conceptualisation ties interruption timing and relevance together. Timing relates to the changing

contents of working memory over task execution. Relevance relates to the effects of interruptions on the contents of working memory at the moment of interruption. Whether this relationship holds true is an empirical question that is tested by the experiments presented in this thesis.

2.4 Summary

This chapter explores work that has previously been done to understand interrupted performance. In keeping with the methodological approach of this thesis, there is a focus on laboratory-based experiments and the role of memory in resuming after interruption. However, situated and non-experimental work has added significantly to the body of knowledge about interruptions. To understand the context of the experimental work, interruptions-focused work from other paradigms has also been considered, with a focus on hospital-based observational studies.

Gaps in existing work emerge as prior work is reviewed. Although the role of working memory in interruption management is well explored (e.g., Altmann & Trafton, 2002; Iqbal & Bailey, 2005; Trafton et al., 2003), some of the implications of models of working memory have not been empirically tested. In particular, there has been little work to understand how a particular interruption affects working memory representations over the course of task execution.

The correspondence or disjunction of interrupting tasks to working memory representations might provide a definition of interruption relevance that is grounded in psychological principles, rather than dictionary definitions or rough heuristics. Part of understanding relevance will require understanding how relevance varies over the course of task execution; exploring interruption timing is also necessary. This thesis presents an empirical exploration of these issues. As the answers to these questions are necessarily couched in the tasks that are performed, the next chapter describes the experimental paradigm through which these questions are explored.

Chapter 3

The Pharmacy Task

3.1 Outline

This thesis makes use of a single task paradigm, the Pharmacy Task. The task is a re-designed version of the Doughnut Machine (Li et al., 2006), a well developed routine procedural task that has been used in various forms in a number of studies of routine procedural tasks (see Li et al., 2008; Ratwani, McCurry, & Trafton, 2008; Back, Brumby, & Cox, 2010; Ament, Cox, Blandford, & Brumby, 2010; Hiltz, Back, & Blandford, 2010; Jones et al., 2012; Brumby et al., 2013; Gould, Brumby, & Cox, 2013, for a sample). While the original task requires users to make doughnuts, the Pharmacy Task used in this thesis replaces the task of creating virtual doughnuts with the task of entering data to order sets of medication prescriptions.

The Doughnut Machine was designed for the study of post-completion errors, a particular type of cognitive slip. In the original version of the task, interruptions were used as a mechanism for eliciting errors. While discussions of error are a component of this thesis, the focus is on interruption processing; it is not an aim of this thesis to develop or test a theory of human error. As such, the major changes to the task come from the removal of device-oriented steps in the task. Device-oriented steps are not fundamental to the completion of a goal, but rather are features of a particular instantiation of a system that allows for the completion of a goal. They take longer to complete and cause more errors than task-oriented steps (Ament, Cox, Blandford, & Brumby, 2013). For this reason, the Pharmacy Task contained as few device-oriented steps as possible.

This chapter provides a general description of the task and its constituent parts. Each experiment in this thesis uses a variation of the Pharmacy Task. Differences between the

model presented in this chapter and the particular instantiation used for each experiment are detailed in the respective method sections.

3.2 Features of the Pharmacy Task

3.2.1 The Pharmacy Task as a data-entry task

The Pharmacy Task is fundamentally a routine data-entry task. Target quantities are presented and participants are asked to copy these quantities into the corresponding parts of each subtask. The target quantities are given as a list of three 'prescriptions' in each trial.

Each prescription is composed of one quantity value and four attributes. For each subtask, participants must identify the attributes that are present in the *Prescription Sheet* and copy the quantity associated with the attribute into the appropriate subtask element. Using the scenario in Figure 3.1b as an example, if a participant was working on the *Shape* subtask, they would need to enter: *30 Triangle*, *10 Rectangle* and *40 Round*.

While the order of attributes in subtasks is fixed, the order of attributes in the *Prescription Sheet* is randomly generated and thus do not usually appear in the order that they do in the subtasks. It is therefore necessary for participants to remember quantities and attributes as pairs. Remembering *40* is of little use unless the associated attribute, *Tablet*, is also remembered with it. For example, the first prescription shown in Figure 3.1b is *30 Gum Triangle Brown Tub*. This means that the quantity *30* needs to be entered in the *Gum*, *Triangle*, *Brown* and *Tub* elements of the respective subtasks.

Of course, entering one order at a time would require participants to make three trips through the task. Therefore, participants are told to enter multiple quantities into each subtask. For the prescriptions shown in Figure 3.1b, this would mean entering *30 Gum*, *10 Capsule* and *40 Lozenge* into the first subtask, *Type*.

Once participants had entered the quantities next to the correct subtask attributes, they clicked *OK*. If the quantities had been entered correctly, they then started working on the next subtask. (The processing of errors is reported below.) Once the fifth subtask had been completed, participants clicked *Process* to complete the trial. At the start of each trial the *Prescription Sheet* was hidden. Trials started and the prescriptions were revealed by clicking the *Next Prescription* button, which took the place of the *Prescription Sheet* until it was clicked.

Figure 3.1: The Pharmacy Task has five subtasks (shown in Figure 3.1a) and these sub-tasks need to be completed in a specific order, *Type* (1), *Shape* (2), *Colour* (3), *Packaging* (4) and *Label* (5). For each of these subtasks, three quantities need to be entered. In the example shown in (b), *30* needs to be entered into the *Triangle* box in the second sub-task along with *10 Rectangle* and *40 Round*. Note that subtasks 1 and 5 have the same elements, so the same quantities are entered in the same order for these subtasks. The mapping of information from the *Prescription Sheet* to the subtasks is shown in Figure 3.1b.

In the majority of the experiments in this thesis there are no explicit or implicit cues for place-keeping. Participants must remember which subtasks they have completed and which subtasks they have left to complete. Once participants click the *OK* button at the end of a subtask, the values entered are reset to zero. If participants forget where they are in the task there is no way to reconstruct the information from the environment. However, this does not mean that the task is without features that could help participants to remember where they need to resume. For instance, looking over the contents of the *Prescription Sheet* might aid participants in recalling where to resume. The parts of the task that might be utilised in this manner are described later in this chapter.

3.2.2 Subtasks and subtask elements in the Pharmacy Task

The Pharmacy Task is made up of five subtasks: *Type*, *Shape*, *Colour*, *Packaging*, *Label*. Each subtask comprises five subtask elements. Subtask elements are made up of a fixed attribute (e.g., *Diamond*) and a text field for a quantity. Of the five subtask elements in each subtask, only three, selected at random, need to be filled in during each trial. Subtask elements are completed with values from the *Prescription Sheet*. Each subtask has unique attributes, except for the first and last subtasks, which share the same attributes.

Step #	Action
1	**Click** *Next Prescription*
2	**Begin subtask 1,** *Type*
3	Retrieve and enter first quantity-attribute pair from *Prescription Sheet*
4	Retrieve and enter second quantity-attribute pair from *Prescription Sheet*
5	Retrieve and enter third quantity-attribute pair from *Prescription Sheet*
6	**Click** *OK* **to complete subtask 1**
7	**Begin subtask 2,** *Shape*
8	Retrieve and enter first quantity-attribute pair from *Prescription Sheet*
9	Retrieve and enter second quantity-attribute pair from *Prescription Sheet*
10	Retrieve and enter third quantity-attribute pair from *Prescription Sheet*
11	**Click** *OK* **to complete subtask 2**
12	**Begin subtask 3,** *Colour*
13	Retrieve and enter first quantity-attribute pair from *Prescription Sheet*
14	Retrieve and enter second quantity-attribute pair from *Prescription Sheet*
15	Retrieve and enter third quantity-attribute pair from *Prescription Sheet*
16	**Click** *OK* **to complete subtask 3**
17	**Begin subtask 4,** *Packaging*
18	Retrieve and enter first quantity-attribute pair from *Prescription Sheet*
19	Retrieve and enter second quantity-attribute pair from *Prescription Sheet*
20	Retrieve and enter third quantity-attribute pair from *Prescription Sheet*
21	**Click** *OK* **to complete subtask 4**
22	**Begin subtask 5,** *Label*
23	Retrieve and enter first quantity-attribute pair from *Prescription Sheet*
24	Retrieve and enter second quantity-attribute pair from *Prescription Sheet*
25	Retrieve and enter third quantity-attribute pair from *Prescription Sheet*
26	**Click** *OK* **to complete subtask 5**
27	**Click** *Process*

Table 3.1: An example action sequence for an uninterrupted, correct trial with the Pharmacy Task. Subtasks actions (in bold) have to be completed in order. Participants have some discretion over the order in which they enter values into subtasks. They could, for example, enter value-attribute pairs in the order that they appear on the *Prescription Sheet*, or, alternatively, they could enter values in the order that attributes appear in subtasks. Items presented on the *Prescription Sheet* are not necessarily in the same order as subtask attributes.

Figure 3.2: Each of the five subtasks (3.2a) that comprise the Pharmacy Task itself consists of five sub-subtasks, or *subtask elements*. Subtask elements (3.2b) are the slots into which quantity attribute-pairs from the *Prescription Sheet* are copied. Participants only need enter three quantity-attribute pairs in each subtask.

3.2.3 Making errors in the Pharmacy Task

The Pharmacy Task requires participants to take a largely linear path through each trial. Subtasks must be completed in the order indicated in Figure 3.1a. If participants attempt to work on anything other than the correct subtask they will have made a sequence error. Participants are alerted to errors with a statement informing them of the task they were expected to work on. For example, if a participant finishes the *Type* (first) subtask and then tries to work on the *Colour* (third) subtask, participants are given a warning message in the form of 'Error: expected Shape' (the second subtask). As the Pharmacy Task does not provide progress cues, this ensures that participants do not have to resort to random clicking in the event that they forget where they are in the task. To discourage guessing behaviour, any error in which the participant selects the wrong subtask results in a lockout of eight seconds. During the lockout participants cannot perform any activity and must wait for the lockout period to finish before they can continue working. Lockouts were used because they have been shown to reduce error rates in routine procedural tasks like the Pharmacy Task (e.g., Back et al., 2010; Brumby et al., 2013).

Progress through a subtask is visible as long as participants are still working it: if a participant has entered two quantities, they can see the two values they have already entered on the screen. These values would provide a strong resumption cue to participants resuming after an interruption that appeared as they finished entering the second value. Therefore, when participants are interrupted within a subtask the values are stored in the internals

40

of the program but reset to zero on the interface visible to participants. This destroys the resumption cue. If participants are able to remember what they were about to work on when they were interrupted, their previous progress on the subtask is restored. If participants try to enter a quantity they have already entered for the subtask they are required, via a red flash in the text field, to enter all three quantities for the subtask all over again. This deters guessing behaviour.

The order in which quantities are entered into subtask elements is one of the few parts of the procedure that participants have control over; values can be entered into subtask attributes in any order. However, the correct quantities must be entered next to the correct attributes before participants can proceed to the next subtask. Failure to do so results in a warning message requesting that participants check their submission. There is no lockout period for these errors, but participants cannot continue until the correct quantity-attribute pairs have been entered. This behaviour differs to that of the Doughnut Machine, where entering the wrong values into a subtask did not cause an error message to appear.
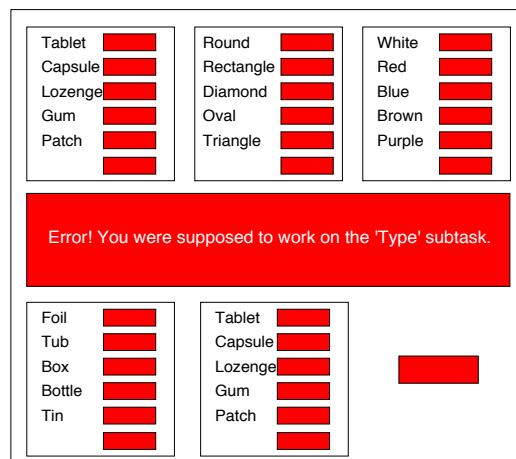
Figure 3.3: When participants make an error during execution of the task, the screen is locked and all fields and buttons turn red. A notice telling participants what they should have been working on replaces the *Prescription Sheet*.

3.3   Interruptions during the Pharmacy Task

Large parts of this thesis focus on how interruption timing affects post-interruption resumption performance. In the Pharmacy Task, there are a variety of points at which participants could be interrupted. Participants could be interrupted between trials, in the middle of typing numbers or when their mouse cursor moved over particular parts of the screen.

Rather than distribute interruptions throughout the tasks at random, this thesis investigates the effects of interruptions during subtasks and at subtask boundaries, as this has been an area of particular research interest (e.g., Iqbal & Bailey, 2005; Bailey & Konstan, 2006). Subtask boundaries are the period of time between the completion of one subtask and start of the next. In this thesis, interruptions that appear at this point are referred to as *between-subtask* interruptions. Between-subtask interruptions can be contrasted with interruptions that occur during the completion of a subtask. In this thesis these are called *within-subtask* interruptions.

A within-subtask interruption can occur at any point within a given subtask. For the Pharmacy Task, this would include the times at which participants were entering quantities into subtask elements. While there are interesting questions about what constitutes a subtask (and therefore when something is between or within a subtask), to improve control and ease implementation, within-subtask interruptions in the Pharmacy Task always arrive after entering a quantity into one element but before entering anything into the next element. Within-subtask interruptions can therefore arrive at the moment a participant is about to work on the second or third subtask elements in a subtask, but not before the first has been started or after the third has been completed. This gives more certainty that a participant is truly in the middle of working on the subtask, rather than, for example, planning their work on the next subtask. Given the constraints on where between- and within-subtask interruptions could appear, there were a limited number of interface elements that an interruption could be 'attached' to. These are illustrated in Figure 3.4.

The interrupting tasks varied in each experiment. Rather than provide a detailed exploration of the various interrupting tasks here they are instead described in the method section of each experiment. Covering interrupting tasks on a per-experiment basis provides better context and motivation for the design decisions taken.

3.3.1   Resources for resumption

Section 3.2.1 described how values entered into subtask elements are hidden once a subtask is completed. A consequence of this behaviour is that there are no explicit cues that would allow participants to reconstruct their location from the task environment. However, the task does have resources that could aid the development of strategies for place-keeping.

One of these resources is the spatial layout of the task. Subtasks and the attributes that comprise them are always in the same location in the task. Previous work has suggested

42
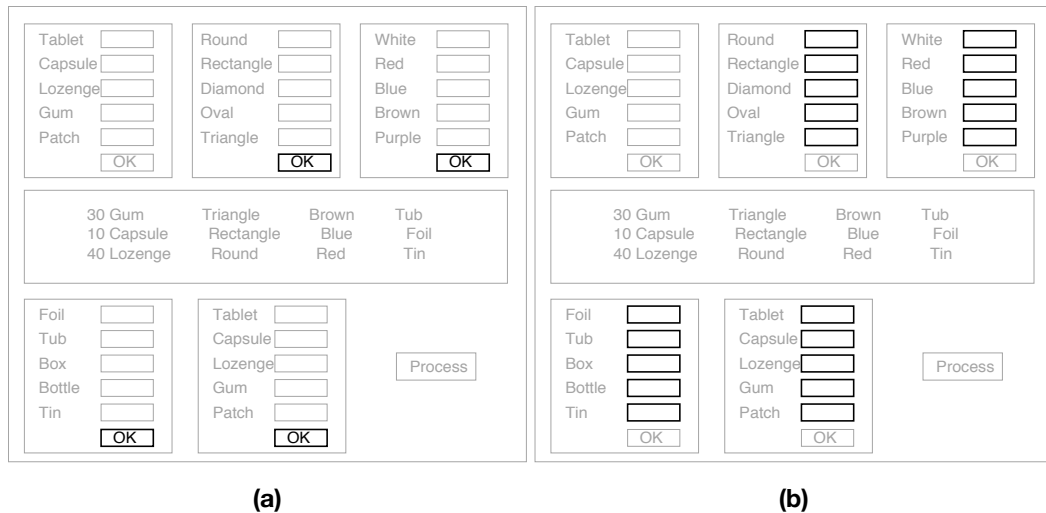
Figure 3.4: These figures show the locations in which interaction with the interface might result in an interruption. Figure 3.4a highlights the *OK* buttons for subtasks 2 to 5. These are the elements which could result in a between-subtask interruption when clicked. Figure 3.4b shows the text-entry boxes for the elements of subtasks 2 to 5. Within-subtask interruptions could appear when any of these are clicked.

that spatial memory of layout guides post-interruption resumption in tasks like the one described here (Ratwani & Trafton, 2008; Werner et al., 2011). If spatially-oriented strategies are utilised at the point of resumption, it suggests that spatial layout plays a role in the execution of the task by participants in uninterrupted scenarios. It is possible to imagine a scenario in which making the spatial layout of the task unreliable would result in a change to the strategies participants used to deal with interruptions.

Another resource is the *Prescription Sheet*. This contains the three prescriptions that participants must enter into the subtasks. The *Prescription Sheet* is different for each trial, but remains unchanged over the course of a particular trial. Therefore, the *Prescription Sheet* has the potential to act as a strategic resource during the course of a trial; while the information in the *Prescription Sheet* is essential to the task, there are multiple ways of employing the *Prescription Sheet* in the course of completing the task.

The order in which participants retrieve information from the *Prescription Sheet* and the amount they retrieve each time they look at the sheet could affect the way they completed the task. Of particular interest is the relationship between the *Prescription Sheet* and resumption performance. In the event that participants encode and retain *Prescription Sheet* information during an interruption then the *Prescription Sheet* would likely play no role in resumption after interruption. The most likely scenario would be that participants forget some of the details needed to complete the task and have to re-encode information from

the *Prescription Sheet* when resuming after interruption. The strategies that participants adopt will depend on the relationship between interruption and task. The relationship between resumption performance and fixed resources that might act as cues is explored in depth in Experiment 5.

3.4   Summary

The seven experiments reported in this thesis use the Pharmacy Task. This is a routine procedural data-entry task that is modelled on the Doughnut Machine. Versions of the Doughnut Machine have been used extensively in laboratory-based investigations of interruption. The task requires participants to copy values from a central location, the *Prescription Sheet*, to the correct fields in each of the five subtasks. Subtasks have to be completed in a set order, but data can be entered into the subtask elements in any order. To discourage satisficing behaviour, errors made during task execution are subject to a variety of penalties. To achieve the variety of experimental manipulations required by this thesis, each experiment uses a slightly different version of the Pharmacy Task. These changes are covered in detail in the method section for each experiment.

Chapter 4

The effect of interruption timing and relevance on post-interruption resumption performance

4.1 Outline

This chapter comprises a set of three experiments that investigate the effects of timing and relevance on interruption disruptiveness. Disruptiveness was measured by how quickly participants resume the Pharmacy Task after the end of an interruption. This is known as the resumption lag, and is widely used as a measure of disruptiveness (e.g., Andrews et al., 2009; Cades et al., 2011). Taken together, these experiments, which use the same design, contribute the first memory-centric account of interruption relevance. By framing relevance as a working memory problem, it is possible for interruptions to be dynamically categorised on the basis of their effects on the ephemeral contents of working memory, rather than through an *a priori* classification scheme. This chapter delivers a simple bifurcated model of interruption relevance with two states, relevant and irrelevant.

The experiments described in this chapter are variations on a theme: all three focus on the relationship between interruption relevance and timing, and resumption performance. Experiment 1 investigates how relevance and timing affect between-subtask place-keeping goals. These goals are what allow participants to resume on the right subtask after interruption. Experiments 2 and 3 look more closely at within-subtask place-keeping goals, which allow participants to accurately resume their activities at a particular location within a subtask.

4.2 Experiment 1: Understanding the effect of relevance and timing on interruption disruptiveness

Part of this study and a subset of the results described here were published in Gould, Brumby, and Cox (2013) and presented at the annual meeting of the Human Factors and Ergonomics Society.

4.2.1 Motivation

To develop effective interruption management systems it is necessary to understand how the properties of interruptions affect their disruptiveness. Previous studies have focused in particular on the effects of timing and interruption. Czerwinski et al. (2000) investigated the effects of relevance and timing of interruptions during an instant-messaging task. Interruptions were 'relevant' when they gave participants the answer to a question they were trying to answer. 'Irrelevant' interruptions gave participants a random fact about what they were doing. Iqbal and Bailey (2008) compared 'relevant' interruptions, which "provided examples [and] useful tips", with 'general interest' interruptions that presented news headlines from Google News. Both of these studies find that 'relevant' interruptions are less disruptive to the task being interrupted than 'irrelevant' interruptions, but it is not clear why this should be the case.

It may seem trivial that relevant interruptions are less disruptive than irrelevant ones, but our limited understanding of what makes interruptions relevant means it is not possible to proceduralize relevance in a way that allows the concept to be deployed systematically. Part of the problem is that researchers have not given sufficient care to defining their use of the term *relevance*. Relevance might describe the utility of information carried by an interruption or it might encapsulate broader congruity of topics for interruptions and primary tasks. Such definitions are still too nebulous to be useful, as they themselves comprise an aggregation of a number of more basic measures.

This study develops a measure of relevance based on the effects of interruptions on working memory. During the execution of routine tasks, goals in working memory represent progress. The representations in memory vary moment-to-moment as the task is executed. When a particular interruption requires access to a particular goal in working memory, that goal is rehearsed and strengthened, reducing disruptiveness. Therefore, interruptions that reinforce the contents of working memory at the moment of interruption are necessarily relevant to the task at hand.

The goal of this study is to explore the relationship between working memory and interruption timing and relevance. It aims to determine whether interruption relevance stems from the effect of a particular interruption on goals held in working memory at the point of interruption. It is hypothesised that the shifting contents of working memory will make a given interruption more or less relevant over the course of task execution. Experiment 1 tests this hypothesis by introducing both relevant and irrelevant interruptions at different points during execution of the task. Relevant interruptions require participants to answer questions about their progress through the Pharmacy Task. Irrelevant interruptions test knowledge about the features of the Pharmacy Task. As the Pharmacy Task hides place-keeping cues from participants, participants must maintain representations of progress in working memory during the course of an interruption (Borst et al., 2013). Therefore, it is hypothesised that relevant interruptions that encourage rehearsal of progress-representing place-keeping goals will be less disruptive than irrelevant interruptions that either do not encourage such rehearsal, or actively interfere with place-keeping goals.

4.2.2 Method

4.2.2.1 Participants

Twenty-four participants (14 male) with a mean age of 23 years (*SD*=5 years) took part in the study. Participants were drawn from the UCL participant pool and were paid £7 for approximately one hour of their time.

4.2.2.2 Design

The experiment used a 2×2 within-subjects design. There were two independent variables: interruption relevance, which had two levels, relevant and irrelevant; and interruption timing, which also had two levels, within-subtask and between-subtask. The primary measure was resumption lag, which is the time between a participant being returned to the primary task after interruption and interacting with the interface. This gives the earliest indication of whether participants have successfully remembered where to resume the task. Task, subtask and interruption timings were also recorded for analysis along with error rates. Conditions were randomly allocated in three blocks of four (within-subtask/relevant; within-subtask/irrelevant; between-subtask/relevant; between-subtask/irrelevant) for a total of twelve trials.

4.2.2.3 Materials

Primary task   This study used the Pharmacy Task as the primary task. This is a routine procedural task with five subtasks, themselves comprising five elements. This task
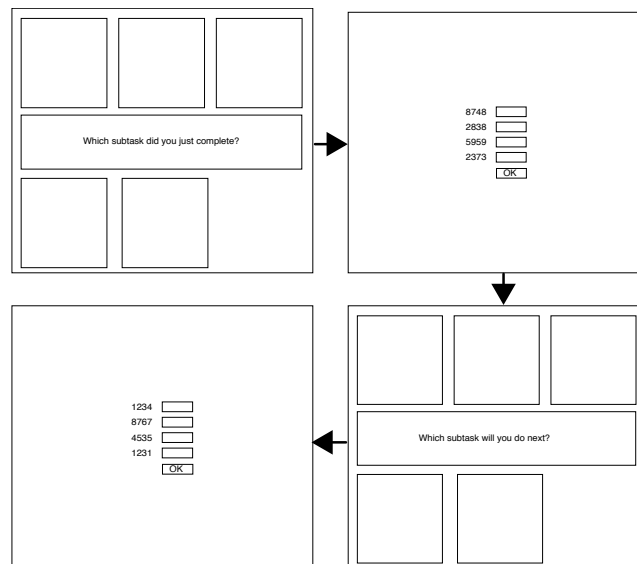
Figure 4.1: Each interruption consists of four tasks, two audit tasks and two transcription tasks. The order in which these subtasks were completed is shown clockwise from top-left.

was described in detail in Chapter 3. This section describes variations on the description previously given as well as the interruptions used in Experiment 1.

Interruption timing   Interruptions arrived either between a subtask or within a subtask. Between-subtask interruptions occurred immediately after a participant clicked on the 'OK' button of a subtask (i.e., when one subtask had been finished but the next one hadn't been started). Within-subtask interruptions occurred when participants had entered either one or two pieces of data in a subtask. There were two interruptions per trial and both were of the same type. The timing of interruptions was randomised. The order of trials was pseudo-random: blocks of four trials (one of each combination of relevance and timing) were internally randomised.

Interrupting tasks   Interruptions comprised two blocks of tasks. Each block consisting of one filler task and one target task, presented sequentially. The order of these tasks during an interruption is illustrated in Figure 4.1. The filler task was a transcription task that was same across conditions. The content of the target tasks, which were called *audit tasks*, varied depending on whether participants were working on a relevant or irrelevant interruption. These tasks and their relationship to the manipulations of timing and relevance are detailed below.

Audit task   The audit task asked participants to report on some aspect of the primary task. In relevant interruptions (*progress audits*), this was related to a participant's progress

48

through the task (e.g., 'What subtask did you just complete?'). In irrelevant interruptions (*knowledge audits*), it related to participants' knowledge of the subtask names (e.g., 'Which is the Label subtask?'). Participants responded by clicking the (now blanked-out) subtask that they though correctly answered the question. To preclude the possibility of the audit task explicitly cueing resumption in relevant interruptions, there was no feedback on this task for either kind of interruption.

Both relevant and irrelevant interruptions had the same task sequence. The manipulation of relevance occurred in the 'audit' phases of the interruption, which differed in the question posed, but not in complexity or duration. Note that audit questions for both relevant and irrelevant interruptions were at the subtask level: participants were never asked questions about subtask element features or progress. Audit tasks which formed part of within-subtask interruptions were worded slightly differently to those in between-subtask interruptions. For example, instead of asking participants 'Which subtask will you work on next?' in a between-subtask interruption, participants would be asked 'After completing the current subtask, which subtask will you work on next?' Note that despite the change to the wording, the question is still about subtask-level place-keeping.

Relevant audit tasks   When designing the relevant interruptions for this task, it was important that relevant interruptions did not have an explicit or obvious advantage over irrelevant interruptions. Therefore, the two progress audit sections of the relevant interruptions all asked questions (see Figure 4.2 for an example) that related to the participant's progress through the primary task (i.e., its current state). The intention was that the relevant tasks should cue, or strengthen the activation of the place-keeping goal. Previous studies (Ratwani & Trafton, 2008; Werner et al., 2011), have suggested that for the type of routine task used in this study resumption goals are stored spatially rather than declaratively. Therefore, the relevant interruption tasks are represented spatially in this study.

Irrelevant audit tasks   Irrelevant interruptions differed only in the question asked — the layout and means of response were identical to the relevant audit. Participants were asked questions about their knowledge of the names of the subtasks (e.g., 'Which is the Type subtask?'), which they answered by clicking on the appropriate area of the screen.

Transcription task   One of the tasks participants were required to complete as part of the interruption was a transcription task. This was a simple filler task; it required participants to transcribe four four-digit numbers from one part of the screen to a text entry panel on another part of the screen. Participants were required to correctly complete all tran-
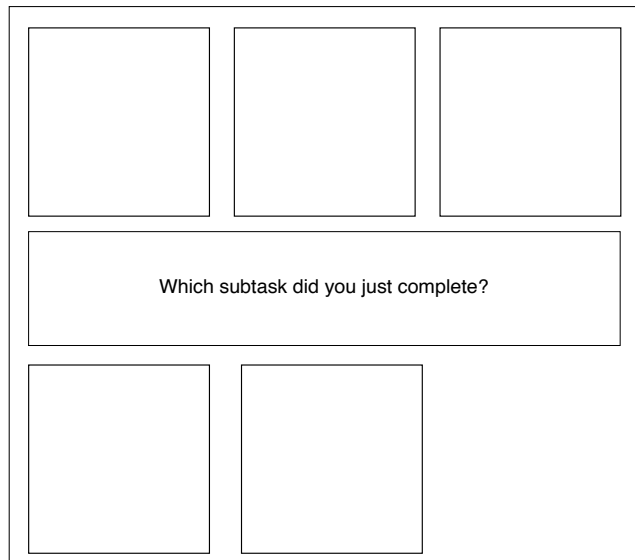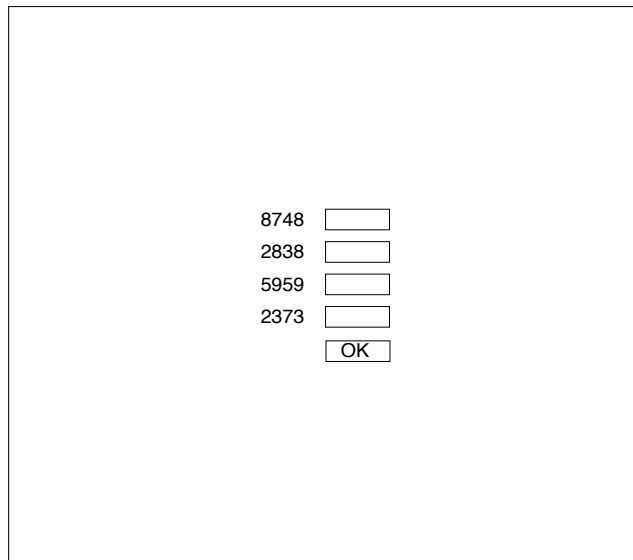
Figure 4.2: The audit tasks made use of the Pharmacy Task interface. All task features were removed and the *Prescription Sheet* was replaced with a question, the content of which could be relevant or irrelevant. Participants answered the question by hovering their mouse over the appropriate subtask.

scriptions and were prevented from continuing until they had done so. The transcription task was the only time during interruptions when participants were not immersed in the Pharmacy Task, so acted as an opportunity to distract participants from place-keeping rehearsal. The layout of this task is shown in Figure 4.3.

Post-experiment questionnaire   A six-item questionnaire was prepared for the purpose of gauging participants' perceptions of the interruptions they encountered and the task as a whole. It was issued on paper after the experiment had concluded. Additionally, four free-response questions were posed to participants. They gave verbal responses which were recorded in note form by the experimenter. The contents of the questionnaire are presented in Figure A.2 in Appendix A.

4.2.2.4   Procedure

Participants were given a consent form to read and sign followed by an information sheet which outlined the purpose of the study and what they were expected to do during the course of the experiment. A video demonstration of the task was shown to participants. The video demonstration provided a consistent introduction for all participants and showed a trial being completed, with interruptions occurring as they would in the experimental trials.

Figure 4.3: The transcription task is a simple filler task. Participants must correctly copy four values from one part of the screen to another. Any mistakes or missing values are flagged and participants are required to correct them before continuing.

After resolving any uncertainties arising from the training video, participants proceeded to the training trials. All participants completed one training trial with no interruptions, followed by three that had interruptions (one with a between-subtask interruption followed by two trials with within-subtask interruptions). Where participants were interrupted within a subtask, they were told that they should try to resume at the point they were interrupted within the subtask. If a participant had completed two of the three parts of a subtask, they should have tried to resume on the third one. If they attempted to resume in a textbox for which values had already been entered, the textbox flashed red and participants had to enter all values for the subtask again.

After a brief pause for questions, participants proceeded to the experimental trials. Participants completed 12 experimental trials, with an optional two minute break after completing six trials. During the experimental trials there were no cues to guide participants on resumption.

In order to discourage guessing, a sequence error at any point in the primary task (including on resumption after an interruption) resulted in an eight-second lockout during which participants were not able to continue with the task. Participants were informed of the correct action when the errors message appeared on the screen. Participants were required to correct any errors they made in entering values into the subtasks but there was no lockout associated with errors of this kind.

4.2.3 Results

4.2.3.1 Data processing — error and outlier identification

Each participant was interrupted 24 times during the course of the experimental trials, yielding a total of 576 interruptions across all participants. Data were processed, with errors and outliers dropped from the analysis.

First, interruptions that resulted in incorrect resumptions were excluded from the analysis of resumption lag. These included sequence errors at the subtask level, for example if a participant was supposed to be working on *Packaging* on resumption but clicked on something in the *Label* subtask instead. There were 86 instances of this kind of resumption error and all were excluded from the analysis.

Secondly, within-subtask resumptions where the participant selected a subtask element which they had already completed before the interruption arrived were dropped (e.g., a participant selected *Box* in the *Packaging* subtask when they had already completed the *Box* element before being interrupted). Although the benefits of remembering within-subtask location were explained to participants, there was no mechanism for enforcement or feedback. It was not known whether participants would continue to attempt to resume within-subtask resumptions accurately. However, there were 57 incorrect within-subtask resumptions out of a total of 288 within-subtask resumptions (i.e., 20%). These incorrect resumptions were assumed to be indications of erroneous resumptions rather than the result of a strategy to ignore within-subtask location on interruption. Thus they were discarded from the analysis of resumption lag data.

After removing these erroneous resumptions from the dataset, resumptions that were ±1.96 standard deviations (i.e., 95%) from a participant's individual mean resumption time were also removed as outliers. Data from 23 resumptions were classified as outliers and were excluded from the analysis. This left a total of 410 trials from the original 576 (71%) for the analysis of resumption lag. The trials with incorrect resumptions are analysed separately below.

4.2.3.2 Resumption lag

Resumption lag was the primary measure of performance in Experiment 1. Figure 4.4 and Table 4.1 shows the mean resumption times for each combination of interruptions. Participants resumed more quickly after relevant interruptions that arrived between-subtasks

than those arriving within a subtask. For irrelevant interruptions, resumptions were faster after within-subtask interruptions than after between-subtask interruptions.

	Within	Between
Relevant	4532 (SD=1454)	3777 (SD=1613)
Irrelevant	4510 (SD=1377)	4883 (SD=1324)
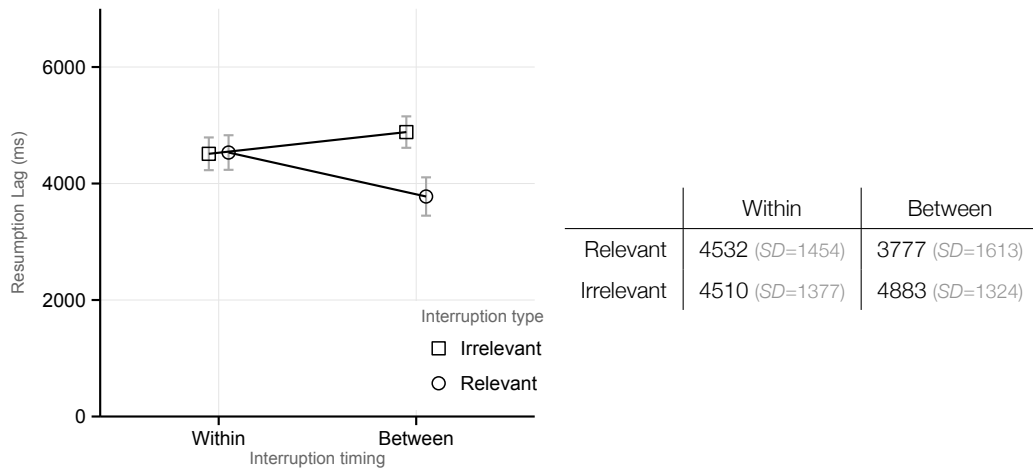
Figure 4.4 and Table 4.1 show aggregate resumption lags for Experiment 1. Note the interaction between interruption timing and relevance is significant. All times in milliseconds.

A factorial repeated measures ANOVA was used to analyse the resumption lags. It revealed a significant main effect of interruption relevance (relevant vs irrelevant) on resumption lag ($F(1,23)$=12.94, p<.01, η_P^2=.36). However, there was no significant main effect of interruption timing (within-subtask vs between-subtask) on resumption lag ($F(1,23)$=0.49, p=.491, η_P^2=.02). The main effects were superseded by a significant interaction between the type and timing of interruptions ($F(1,23)$=5.04, p<.05, η_P^2=.18). Achieved power for the interaction effect was good (.87), even with the highly conservative assumption that there was no correlation among repeated measures.

A pair of simple effects analyses were conducted to explore the nature of the interaction effect. These analyses focused on the simple effect of interruption relevance on between- and within-subtask interruptions. First, the effect of relevance on between-subtask interruptions was examined. An analysis of the simple effect of relevance on resumption after between-subtask interruptions showed that participants resumed significantly more quickly after relevant between-subtask interruptions than after irrelevant between-subtask interruptions ($F(1,23)$=11.35, p<.01, η_P^2=.33). A corresponding analysis of the simple effect of relevance on within-subtask interruptions revealed no significant effect ($F(1,23)$=0.007, p=.932, η_P^2=0). Together these results show that resumptions after between-subtask interruptions were affected by the manipulation of relevance but that within-subtask interruptions were not.

	Within		Between
	Element	Subtask	Subtask
Relevant	10% (28)	3% (5)	21% (30)
Irrelevant	10% (29)	13% (18)	23% (33)

Table 4.2: Rates of resumption errors by condition and error type. Figures in grey are frequency.

4.2.3.3 Resumption error frequency and distribution

During the course of the experiment, participants were interrupted 576 times across all conditions. A total of 143 resumption errors were made, meaning one in four resumptions after an interruption was inaccurate in some way.

Sequence errors made up the majority of the errors. This kind of error occurs when the wrong step in a procedure is selected (e.g., *Type* was selected instead of *Shape*). In the Pharmacy Task, these manifested as the selection of the wrong subtask on resumption after interruption. A total of 86 of these subtask-level resumption errors were made (15%). Of these errors, 63 occurred after a between-subtask interruption (30 after relevant interruptions, 33 after irrelevant interruptions). The remaining 23 occurred after within-subtask interruptions (5 after relevant interruptions, 18 after irrelevant interruptions). Participants were more accurate when choosing which subtask to resume on if they were in the middle of a subtask than when they finished one subtask but had not started the next. These rates are summarised in the 'Subtask' columns of Table 4.2.

In addition to choosing the wrong subtask, participants could also choose the wrong subtask element; they could resume at the wrong element, but in the correct subtask. Errors made in the selection of a subtask element could only be made after a within-subtask interruption. After a between-subtask interruption, participants could resume on whichever subtask element they liked, as long as the subtask was correct. Participants were told to continue with what they were about to do before they were interrupted. Due to this requirement, an error occurred whenever a participant selected a subtask element (e.g., *Diamond* instead of *Triangle* in the *Shape* subtask) other than the one they were about to work on when they were interrupted. There were a total of 57 incorrect within-subtask resumptions out of a total of 288 within-subtask resumptions, giving an error rate of 20%. Twenty-eight of these incorrect within-subtask resumptions came after relevant interruptions and 29 came after irrelevant interruptions.

4.2.3.4 Testing methodological assumptions

The majority of laboratory studies of interruption have used interruptions that have a fixed duration of time, rather than asking participants to complete a fixed quantity of work (e.g., Li et al., 2008; Back et al., 2010, although see Altmann et al., 2013 for an example of self-paced interruptions being used). Studies that have investigated the relationship between the length of an interruption and subsequent resumption lag have shown that resumption lag is proportional to interruption length (Hodgetts & Jones, 2007; Monk et al., 2008). While the progress and knowledge audits were designed to be as similar as possible, it was necessary to determine whether interruption relevance affected total interruption time. After removing interruptions that were followed by errors and outliers in the manner previously described, both interruption types averaged around 25 seconds in duration (relevant M=25009ms, SD=6101ms; irrelevant M=24965ms, SD=5653ms). A dependent measures t-test was used to determine whether interruption relevance had an effect on total interruption time. No effect was detected (t(23)=-0.20, p=.84).

Between-subtask interruption always appeared in the same position: at the moment immediately after participants clicked the *OK* button to complete the task. Within-subtask interruptions, on the other hand, could arrive at one of two points during a subtask. To ensure that the within-subtask interruptions could be treated as homogenous, t-tests were used to determine whether there was any difference in performance between the interruptions that arrived within a subtasks at position one as opposed to position two. Within-subtask interruption position had no effect on resumption lag (t(23)=1.05, p=.31) or on total interruption time (t(23)=-0.15, p=.88). All within-subtask resumptions could therefore be considered simultaneously.

4.2.3.5 General performance on interrupting tasks

When participants were interrupted, the first item in the interruption they had to work on was an audit task. Depending on the condition, this audit task was relevant (*progress audit*) or irrelevant (*knowledge audit*). It is possible that participants were strategically rehearsing place-keeping information before acting in the first audit tasks. In this way, participants could have been strategically creating their own period of *interruption lag*, during which they held off working, choosing instead to rehearse place-keeping goals. Previous work (e.g., Trafton et al., 2003) has focused on an interruption lag period before the interruption *appears* but the term fits just as well to the period before a participant *attends* to an interruption.

To determine whether interruption relevance and timing had an effect on how quickly participants answered the first audit of an interruption a factorial repeated measures ANOVA was used. The test revealed a significant ($F(1,23)=9.76$, $p<.01$, $\eta_P^2=.30$) effect of interruption relevance on the time between the interruption appearing on the screen and the answer to the question being clicked, with relevant interruptions yielding in faster responses (M=3103ms, SD=1274ms) than irrelevant interruptions (M=3306ms, SD=1376ms). There was no significant effect of interruption timing ($F(1,23)=0.47$, $p=.50$, $\eta_P^2=.02$) and there was no significant interaction ($F(1,23)=0.65$, $p=.43$, $\eta_P^2=.30$).

The effect of relevance on response time at the start of interruptions might reflect the easy availability of place-keeping goals in working memory. Alternatively, it could be the result of intrinsic differences in the complexity or difficulty of the audit tasks themselves. To get a more reliable estimate of the intrinsic duration of audit tasks, the second audits during an interruption were examined. These second audits would not have been subject to any effects arising from the transition from primary task to interruption and therefore are a better representation of how long each audit took participants. Running tests on this subset of the data revealed no significant effects of relevance ($F(1,23)=2.80$, $p=.11$, $\eta_P^2=.11$) or timing ($F(1,23)=1.02$, $p=.32$, $\eta_P^2=.04$) on time spent on second audits. There was no interaction ($F(1,23)=0.06$, $p=.814$, $\eta_P^2=.003$). This suggests that participants were not faster at completing relevant audits at the start of interruptions because these audits were intrinsically less effortful. Instead, it seems that either retrieval times from memory, or strategic decision-making resulted in faster completion times for relevant audits.

4.2.3.6 Uninterrupted baseline performance

In addition to post-interruption processes, the resumption lags measured in Experiment 1 included time spent performing activities that would have had to be undertaken even in the absence of interruptions. These activities include finding the next subtask on the interface, looking up an attribute and looking up the value associated with an attribute. To understand the relative contribution of these activities to the task

The first baseline examines the time taken between uninterrupted subtasks. Subtasks finish when participants click the 'OK' button in a subtask with the correct information filled out in the subtask elements. The inter-action period for subtasks is the time between a participant finishing one subtask and the first action on the next subtask. This period covers the time it takes to look at the Prescription Sheet, decide an order for entry, to encode a number and then select the appropriate location on the subtask. It mirrors the metrics used to determine resumption lag after between-subtask interruptions.

The inter-action period between subtasks was computed by taking the completion time of an subtask that was not interrupted during execution (within-subtask interruption) or immediately after completion (between-subtask interruption). This completion time was taken from the time of the first action on the next subtask. It was not necessary that the next subtask was uninterrupted because interruptions of all types occur after the first action in a subtask. This inter-action interval thus gives an accurate idea of the time taken to transition from one subtask to the next in the absence of interruptions.

Across all participants and trials a total of 1440 subtasks were completed. Due to the pseudo-random allocation of interruptions to subtasks in any particular trial, the total number of inter-action intervals between subtasks that could be extracted was 689. As with resumption lags, outliers were computed based on individual participants' interaction interval and were removed using a 95% criterion. This left a total of 652 between subtask inter-action intervals to analyse. After outlier removal, participant means were computed so that the measure was not biased by the uneven numbers of samples for each participant that was caused by the pseudo-random allocation of interruptions. The mean time between finishing one subtask and starting the next was 3234-ms (*SD*=1233-ms) with a range of 1703-ms to 6805-ms. This was faster than resumptions after relevant (3777-ms) and irrelevant (4883-ms) between-subtask interruptions.

The second baseline measure examines the time taken between a subtask element being selected in the first click made in the selected subtask. The purpose of this baseline is to give an indication of marginal time cost involved in selecting a subtask element and beginning to enter values in it. This baseline measure mirrors the way that resumption lags after within-subtask interruptions are computed. To maintain equivalence with these post-interruption resumption lags only the time taken to start working on the second subtask element in a task is used.

There were a total of 1440 subtask completed across all participants and trials. If any within-subtask interruptions occurred within the subtask it could not be reliably used for baseline. This criterion resulted in the removal of two of five subtasks for all trials with within-subtask interruptions. This left 1152 subtasks for analysis. The within subtask inter-action period was computed by calculating the gap between the last digit being typed in the first subtask element and the second subtask element being clicked.

As with the previous analyses, participant means were calculated and outliers were re-moved based on this mean using the 95% criterion. This left 1097 within subtask inter-

action intervals for analysis. Participants took a mean of 1838-ms (SD=652-ms) to move from one subtask element to the next. Participant means ranged from 884-ms to 3488-ms. The inter-action period was less than the resumption lag after relevant (4532-ms) and irrelevant (4510-ms) within-subtask interruptions. The relative difference between resumption lag and inter-action interval was greater for within-subtask interruptions than it was for between-subtask interruptions.

4.2.4 Discussion

The goal of Experiment 1 was to understand whether interruption relevance could be defined as the product of the interaction of an interruption and the contents of working memory at the moment of interruption. To do this, it investigated how interruption relevance and timing affected resumption performance. Participants were interrupted either by a task that required them to rehearse their progress through the task (relevant) or tested their knowledge of the features of the task (irrelevant).

Although there was a significant main effect of interruption relevance on resumption performance, with relevant interruptions faster in aggregate, this effect was subordinate to a significant interaction of interruption timing and relevance. There was no effect of interruption relevance on within-subtask interruptions, but relevant between-subtask interruptions were significantly less disruptive than irrelevant between-subtask interruptions. As such, the effect of relevance was not reliable for the whole duration of task execution. These results are consistent with the idea that interruption relevance is the product of reinforcing and interfering effects on working memory. As the contents of working memory fluctuate over the course of task execution, so the reinforcing and interfering powers of a particular interruption task fluctuate.

There was no evidence that the manipulation of relevance affected resumption lag after within-subtask interruptions; rehearsal and interference effects on place-keeping goals only manifested when interruptions appeared between subtasks. This suggests that the contents of working memory (i.e., place-keeping goals) within a subtask were encoded in such a way that the progress and knowledge audit tasks did not have an effect. This is perhaps not surprising, given that relevant audits only asked participants about their progress through the task and not their progress through each of the subtask elements. Likewise, irrelevant interruptions only tested knowledge of subtasks, and not their constituent elements.

Participants resumed more quickly after relevant interruptions, but only when the interruption arrived between subtasks. While it is possible that these results are anomalous, there is a good explanation for them based on existing theory. The differences in resumption lags across conditions can be explained through existing models of memory utilization during interrupted work, such as Memory for Goals (Altmann & Trafton, 2002). In Memory for Goals theory, goals are created in working memory during the process of task execution. If these goals are not rehearsed, either because they are complete, or because other goals are receiving activation, they decay and are forgotten. The more that a goal decays, the longer it takes to retrieve; this is why resumption lag is used as a measure of interruption disruptiveness (Trafton et al., 2011). In Experiment 1 relevant interruptions that arrived between subtasks encouraged participants to rehearse the current goal in memory, which represented the next action that they needed to perform. This rehearsal strengthened the goal, making it easier to retrieve after the interruption. Conversely, irrelevant interruptions encouraged participants to rehearse goals that had already been or were yet to be completed. While these goals were rehearsed, the current goal experienced time-based decay. This decay, combined with interference from the goals that where rehearsed, made it more difficult and therefore slower to retrieve the correct goal on resumption. Relevance manipulations do not seem to have had the same effect on within-subtask interruptions. A possible explanation for this is the encoding strategies that participants used.

The implication of these results is that determining the relevance of an interruption is difficult without modelling working memory. There is no way of determining, *a priori*, whether an interruption will be relevant to a given task. Interruption relevance varies with changes in working memory over the course of task execution. Interruption management systems that are insensitive to these changes are unlikely to be effective. Interruption management systems that compare the content of an interruption and a task (e.g., Arroyo & Selker, 2011) are particularly likely to be caught out because they are more likely to produce a consistent scoring of similarity over the course of task execution. As Experiment 1 shows, relevance is a transient property; interruptions may only be relevant for a few short steps.

The results obtained from Experiment 1 also suggest that interruption management systems that introduce interruptions at natural breakpoints (e.g., Iqbal & Bailey, 2010)) may be unnecessarily restrictive. The within-subtask interruptions during Experiment 1 did not occur at a subtask boundary, which is generally considered a breakpoint (e.g., Adamczyk & Bailey, 2004; Fischer, Greenhalgh, & Benford, 2011; Bailey & Iqbal, 2008; Bogunovich & Salvucci, 2011). Nevertheless, all interruptions that arrived in the middle of a subtask

were less disruptive than an irrelevant interruption arriving a subtask boundary; it seems that unless the contents of working memory can be discarded at the end of subtask — and they cannot in this task — then boundary effects are not significant compared to reinforcing and interfering effects. This fits with previous work that demonstrates that conflicts over working memory resources are responsible for many of the costs incurred during multitasking (Borst, Taatgen, & van Rijn, 2010; Salvucci & Taatgen, 2010).

An open question that the next experiment addresses is whether the interaction observed in Experiment 1 can be mirrored by changing the demands of the interrupting task. In Experiment 1, between subtask interruptions produced relevance-contingent swings in performance because they were focused on goals that were most likely to be active between subtasks. Just as interruptions focusing on knowledge of subtask progress were most effective when delivered between subtasks, interruptions that focus on progress though subtask *elements* should be most effective when delivered within subtasks. Whether this is indeed the case is explored in the next study, Experiment 2.

4.3 Experiment 2: Understanding hierarchical effects on interruption timing and relevance

Part of this study and a subset of the results described here were published in Gould, Cox, Brumby, and Wiseman (2013) and presented at the AAAI Meeting on Human Computation & Crowdsourcing.

4.3.1 Motivation

Experiment 1 showed that interruption relevance can be thought of as the extent to which place-keeping goals in working memory are subject to interference or reinforcement by an interruption. Interruptions relevant to goals in memory at the moment of interruption resulted in goal rehearsal, causing less disruption to working memory and resulting in correspondingly shorter resumption lag. Conversely, when interruptions were not relevant to the current goal in memory, they caused active interference and longer resumption lags. However, in Experiment 1 interference and reinforcement effects were only observed when participants were interrupted between subtasks. No such effects were observed when participants were interrupted during the execution of a subtask.

During the discussion of the results of Experiment 1 it was hypothesised that the interaction of interruption relevance and timing could be explained by variation in encoding strategies for place-keeping. As well as encoding place-keeping goals at the subtask level, it is possible that place-keeping goals were also being encoded separately to keep track of progress at the within-subtask level. The fact that the relevance of within-subtask interruptions had no effect on performance suggested that the audit tasks used in Experiment 1 acted on representations of between-subtask but not within-subtask task progress. The question, then, is whether a different kind of audit task would result in different schemes of reinforcement and interference.

Experiment 2 is designed to reveal whether the relevance of an interruption is contingent on the scope of place-keeping goals at the moment of interruption. Does it make a difference what people are actually keeping place of? Experiment 1 used knowledge audits (irrelevant) and progress audits (relevant) as the interrupting tasks. These were designed to either interfere with, or encourage rehearsal of, place-keeping goals at the between-subtask level. They did this by asking participants to think about their progress through subtasks or to demonstrate their knowledge about randomly-selected features of the Pharmacy Task.

Here, the interruptions used in Experiment 1 are replaced with interruptions designed to encourage either interference or rehearsal of within-subtask place-keeping goals. To do this, questions in knowledge audits focus on knowledge about subtask elements, rather than subtasks as a whole. Progress audits focus on progress through subtask elements, rather than through subtasks. In doing this, it was expected that the effects on between-subtask interruptions observed in Experiment 1 would be transferred to within-subtask interruptions.

Previous work has suggested that different interrupting tasks can have a wide variety of effects on performance. Ratwani and Trafton (2008) have shown that when task progress is spatially represented, interruptions that require the manipulation of visual representations in memory are significantly more disruptive than those that do not. Other work, using a Pharmacy Task-like setup demonstrated that the disruptiveness of an interruption in a visually-oriented task could be reduced by delivering interruptions aurally (Ratwani, Andrews, et al., 2008). In a non-routine problem solving task, Hess and Detweiler (1994) demonstrated that when interruptions were similar to the main task, interruptions were less disruptive. However, it is not clear how similar or dissimilar interruptions need to be before effects on performance are measurable. The effect of the small changes made to the interrupting task in Experiment 2 are therefore of significant interest.

4.3.2 Method

4.3.2.1 Participants

24 participants (15 female) with a mean age of 24 years (*SD*=6 years) took part in the study. Participants were drawn from the UCL participant pool and were paid £7 for approximately one hour of their time. None of the participants were involved in Experiment 1.

4.3.2.2 Design

Experiment 2 used the same 2×2 within-subjects design as Experiment 1. There were two independent variables: interruption relevance, which had two levels, relevant and irrelevant; and interruption timing, which also had two levels, within-subtask and between-subtask.

4.3.2.3 Materials

To facilitate more complex and thorough analyses of participant performance, the Pharmacy Task was reprogrammed for Experiment 2. The Python-based task used in Experiment 1 was replaced with one developed in HTML and Javascript. Implementing the study as a browser-based task made it easier to deploy the experiment, improved the

aesthetics of the task environment and allowed for the easy storage of richer metrics that were not constrained by tabular data formats.

Despite the reimplementation, few changes were made to the design of the task. The main interface change was to introduce unique elements for the last subtask ('Label'). In Experiment 1, the elements in the first and last subtasks were the same (e.g., *Tablet*, *Capsule*, *Lozenge*...). The duplication of elements from the first (*Type*) to last (*Label*) subtask had caused confusion in the previous experiment, and there was no experimental rationale for retaining it. Other than this small change to the primary task, the majority of the changes made were to the interrupting tasks.

Changes to the scheduling of interruption subtasks   As in Experiment 1, interruptions were made up of a mix of audit tasks, which tested knowledge of the task or progress through it, and transcription tasks, which provided a filler transcription task. In Experiment 1, there were two audit tasks and two transcription tasks in each interruption in the order *Audit $\rightarrow$ Transcription $\rightarrow$ Audit $\rightarrow$ Transcription*. This meant that an audit task was always immediately adjacent to the switch from the primary task. This may have distorted responses made to audit questions, so the schedule was changed.

For Experiment 2, each interruption comprised three transcription tasks with two audit tasks interposed between them. In this way, no audit tasks followed or were followed by the primary task. Both types of interruptions (relevant and irrelevant) had identical schedules, except for the 'audit' phases of the interruption, which differed by the question posed, but not by complexity or duration.

Changes to the transcription task   The number of transcription tasks per interruption was increased from two to three for this study. To mitigate the increase in trial time that would have resulted from this addition, the number of problems in each transcription task was reduced from three to two. This maintained a total of six transcriptions over the course of an interruption. Otherwise, the transcription task remained functionally identical — participants had to copy values from one column of values to an adjacent column of text fields.

Changes to the audit task   This study used audit tasks that were similar in structure to those used in Experiment 1, except that they had different content. Some design changes were required to accommodate the changes to the content. For relevant, within-subtask audit tasks, participants were questioned about the subtask they were working on at the moment of interruption. For example, participants might be asked 'What was the second

element you worked on?' Irrelevant audits appearing during a within-subtask interruptions tested participants' knowledge of subtask elements for the current subtask. For example, participants might be asked 'Which is the *Brown* element in the *Colour* subtask?' Once again participants indicated their choice by clicking on a blanked-out version of the task.

Some modifications were necessary in order to accommodate these interruptions that arrived between subtasks. This was because between-subtask interruptions arrive when participants are not currently working on a subtask. This means that questions cannot be asked about the current subtask. For relevant audits that arrived between subtasks, participants were questioned about the subtask elements of the task that they had just completed in the previous subtask.

4.3.2.4 Procedure

The procedure for Experiment 2 was broadly the same as the one used for Experiment 1. The biggest change was to the questionnaire and introduction. After completing the experiment, participants completed a twelve point questionnaire on the computer, followed by five free response questions asked by the experimenter. This differed from the Experiment 1, where the questionnaire was administered entirely in paper form. Introductory materials presented entirely in the browser rather than through a combination of paper and computer-based videos. Consent forms were still issued on paper. Otherwise there were no procedural changes; the number and scheduling of training and experimental trials were identical and rest periods had the same duration and schedule.

4.3.3 Results

4.3.3.1 Data processing — error and outlier identification

Before analysing resumption lags, incorrect and outlying resumptions were removed from the data. Incorrect resumptions occurred either when participants tried to work on the wrong subtask or when participants chose the wrong subtask element (for within-subtask resumptions). Of a total of 576 resumptions (24 participants, 12 trials, 2 interruptions per trial), 83 were incorrect and discarded from the resumption lag analysis. These incorrect resumptions and their relative proportions across conditions are considered and analysed separately.

After removing the incorrect resumptions from the dataset, outliers were removed from the remaining data. Experiment 2 used the same criterion as previous experiments: resumptions that were ±1.96 standard deviations (i.e., 95%) from a participant's mean resumption time for a particular condition were removed. This accommodation of individual differ-

ences and variation across conditions mean that outlier removal was not unnecessarily conservative: a total of 14 trials were removed from the analysis of resumption lags.

Therefore, of a total of 576 resumptions (24 participants, 12 trials, 2 interruptions per trial), 83 were incorrect and 14 were outliers. This left 470 resumptions for analysis of resumption lags. This meant there were 80 more resumptions for analysis than there were in Experiment 1.

4.3.3.2 Resumption lag

As in Experiment 1, resumption lag was the primary measure of performance. Resumption lag is the period between work finishing on an interruption an restarting on the Pharmacy Task. Lags were computed from the resumptions that remained after incorrect and outlying interruptions were removed. Incorrect resumptions are considered separately.

As can be seen in Table 4.3, participants tended to resume more quickly after relevant interruptions than they did after irrelevant interruptions. Participants also tended to resume more quickly after within-subtask interruptions than they did after between-subtask resumptions. Fastest resumptions occurred after within-subtask relevant interruptions.

	Within	Between
Relevant	3456 (*SD*=1652)	4062 (*SD*=1911)
Irrelevant	3815 (*SD*=1455)	4680 (*SD*=1938)
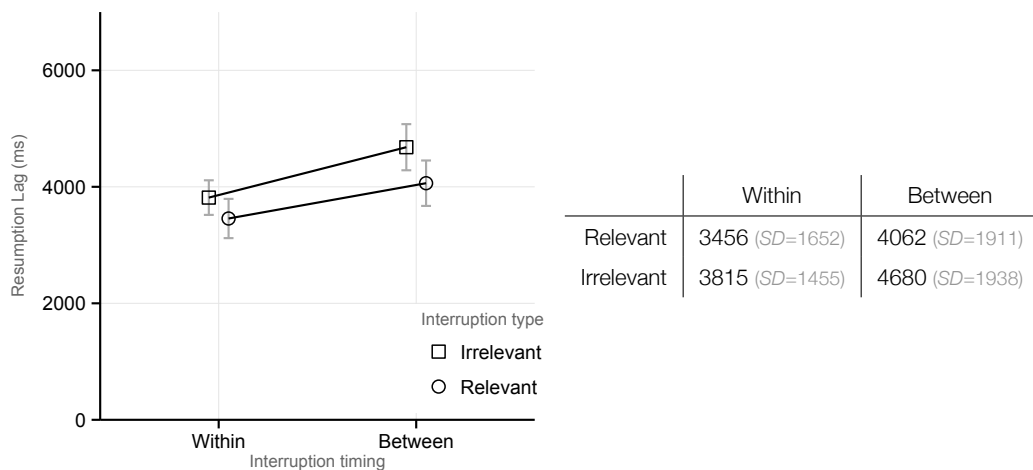
Figure 4.5 and Table 4.3 show aggregate resumption lags for Experiment 2. All times in milliseconds.

A factorial repeated measures ANOVA was used to determine whether there were any significant differences in performance arising from the experimental manipulations. There was a significant effect of interruption relevance (relevant vs irrelevant) on resumption performance (F(1,23)=5.04, p<.05, η_P^2=.18). There was no significant effect of interruption timing (within vs between) (F(1,23)=4.07, p=.055, η_P^2=.15). Even with conservative assumptions about correlation between measures (no correlation assumed), a good level of

power was achieved for the effects of relevance (.87) and timing (.80), suggesting the non-significant result can be trusted. There was no interaction between the type and timing of interruptions ($F(1,23)=0.58$, $p=.45$, $\eta_P^2=.02$).

These results suggest that disruptiveness was affected by the content of an interruption (relevant vs irrelevant). However, there was no statistically robust evidence of an effect of timing (within-subtask vs between-subtask). This was contrary to the hypothesis that within-subtask interruptions would be subject to exaggerated reinforcement and interference effects but that between-subtask interruptions would not.

4.3.3.3 Resumption error frequency and distribution

Due to the hierarchical nature of the task, the kinds of errors participants could make when resuming were constrained by the interruption timing. When participants were interrupted within subtasks, on resumption they could make two kinds of errors; either they could pick the incorrect subtask, or they could choose the correct subtask but the incorrect element (e.g., choosing *Foil* instead of *Bottle*). After within-subtask interruptions, participants chose the wrong subtask eight times, and the wrong element 31 times. When resuming after between-subtask interruptions, choosing the incorrect subtask was the only kind of error participants could make; they made 44 of these. When aggregated by interruption relevance, participants made more errors after irrelevant interruptions (48) than after relevant interruptions (35). As in the previous experiments, participants were far less likely to choose the wrong subtask on resumption after within-subtask interruptions (8) than they were after between-subtask interruptions (44). A complete breakdown of resumption error rates is shown in Table 4.4.

	Within		Between
	Element	Subtask	Subtask
Relevant	8% (11)	3% (5)	13% (19)
Irrelevant	14% (20)	2% (3)	17% (25)

Table 4.4: Rates of resumption errors by condition and error type. Values in grey are frequency.

4.3.3.4 Interrupting task performance

As in the previous experiments, interruptions were not of a fixed duration, but rather required participants to perform a set unit of work before they could return to the primary task. Interruption duration can affect the disruptiveness of interruptions, so it was important that there were no significant differences in total interruption duration (from onset to completion, but not resumption).

	Within	Between
Relevant	24665 (SD=7193)	24923 (SD=7343)
Irrelevant	24229 (SD=6267)	23915 (SD=6055)

Table 4.5: Interruption duration for each combination of relevance and timing. All times in milliseconds.

A factorial repeated measures ANOVA showed no significant effect of timing ($F(1,23)=0.01$, $p=.92$, $\eta_P^2=0$), no significant effect of relevance ($F(1,23)=4.16$, $p=0.053$, $\eta_P^2=.15$) and no interaction ($F(1,23)=1.24$, $p=0.28$, $\eta_P^2=.05$). Table 4.5 details mean interruption duration by condition.

Participants were not given feedback on their responses to audit questions. As such, there was no time-based disincentive to guessing. On any given audit, participants could choose from one of five options, so a 20% correctness rate would be indicative of guessing behaviour. Table 4.6 details audit accuracy for each condition. Participants were most accurate when answering relevant questions within subtasks (61%) and accuracy in all conditions precludes the possibility of guessing.

	Within	Between
Relevant	61%	47%
Irrelevant	51%	53%

Table 4.6: Audit response accuracy by condition. The expected rate from simple guessing behaviour is 20%.

4.3.4 Discussion

There was a significant effect of interruption relevance on resumption lag; participants were significantly faster resuming after relevant interruptions than after irrelevant ones. However, there was no effect of interruption timing on resumption lag and there was no interaction effect. This is consistent with previous studies that show relevant interruptions are less disruptive to performance across categories (e.g., Czerwinski et al., 2000; Adamczyk & Bailey, 2004; Iqbal & Bailey, 2008), but it is not consistent with the effects observed in Experiment 1. How can this conflict be explained?

In Experiment 1, an interaction effect appeared which, when broken down, showed that within-subtask interruptions did not reflect the manipulation of interruption relevance. Between-subtask interruptions, on the other hand, showed a significant effect of relevance; relevant interruptions were less disruptive than irrelevant interruptions. This was

attributed to the audit tasks causing goals representing subtask progress to receive interference or reinforcement. Representations of within-subtask progress received neither rehearsal nor additional inference; the relevance manipulation had no effect on them.

In contrast, Experiment 2 used audit tasks that were designed to act on within-subtask place-keeping. It was therefore expected that the effect would be mirrored and that this time, between-subtask interruptions would not reflect an effect of relevance. Although the results of Experiment 2 do not share the symmetry of those in Experiment 1, the explanation for the results of that experiments also applies to the results here.

When participants had to rehearse their progress within subtasks, they necessarily also rehearsed their progress at the between subtask level. This can be explained by the concept of spreading activation. When a goal's activation is strengthened, it is not the only goal that receives activation. Other goals that are related to the goals receiving activation are also strengthened (Altmann et al., 2013). Activation is also spread from the current focus of attention (Ratwani, Andrews, et al., 2008), such that environmental cues and result in activation spreading to a number of goals (Trafton et al., 2011).

In this case, focusing on within-subtask goals and cues resulted in spreading activation to parent goals. Due to the hierarchical structure of the task, in this case progress through each subtask was subordinate to progress across the subtasks. Therefore, when within-subtask place-keeping goals were rehearsed, goals representing between-subtask progress were also rehearsed. To put this another way; one cannot know which subtask element one is working on, without also knowing which subtask one is also working on.

Further support for this hypothesis comes from the variation in rates of resumption errors. In both Experiment 2 and Experiment 1, participants were least likely to choose the wrong subtask on resumption when they were interrupted within a subtask. This can be attributed to the fact that rehearsal of within-subtask place-keeping goals necessarily strengthens between-subtask place-keeping goals; within-subtask progress goals also represent between-subtask progress and any superordinate representations of between-subtask progress also receive spreading activation.

This fundamental asymmetry was difficult to circumvent in the experimental design. In Experiment 1, for example, participants were only ever asked questions about their progress at the subtask level. Participants were not asked about their progress within a subtask. Conversely, in Experiment 2, participants were only asked about within-subtask progress. This meant that when interrupted between subtasks, participants had to answer questions

about the subtask they had just finished. This meant that the questions asked were not always completely congruous with the place-keeping requirements of the task at the moment of interruption. Nevertheless, the results of Experiment 2 suggest that even with this mismatch between audit questions and interruption timing, relevant interruptions, which required participants to reflect on their progress through the task, still produced faster resumption performance than irrelevant interruptions.

An alternative explanation for the trend toward faster within-subtask resumptions might be that participants were making strategic speed-accuracy tradeoffs. When resuming after a between-subtask interruptions, participants had to pick the correct subtask or they incurred an eight second lockout. When resuming after within-subtask interruptions, the cost of choosing the incorrect subtask element was re-entering one or two values into the subtask.

If participants perceived the cost of re-entering values to be lower than a lockout, they may have resumed faster but made more errors. The data suggest that even if participants perceived a difference in the costs of the two kinds of error, it was not a factor in their decision making; participants made fewer errors when resuming after within-subtask interruptions, even though the potential for making errors was greater (because they could choose the wrong subtask or the wrong element; there were two ways to make an error).

The results of Experiment 2 contribute further evidence to the idea that interruption relevance can be viewed as the extent to which interruptions interfere with or reinforce the contents of working memory. Over the course of task execution, the contents of working memory changes; this is why the moment of interruption is critical to its disruptiveness.

Although this account is clear about the links between the requirements of an interruption and the rehearsal of goals, it is not clear how this link might be mediated by the presentation of tasks, rather than their functional consequences. As Experiments 1 and 2 make use of similar interrupting tasks, it could be that the behaviour observed is a product not of underlying reinforcement and interference processes, but is rather an artefact of a particular interrupting task. If this were the case, it would add significantly to the design of interruption management systems; the disruptiveness of a particular interruption would stem from its presentation and not just the steps involved in its execution.

The thesis that is developed here posits that relevant interruptions encourage rehearsal of place-keeping goals while irrelevant interruptions cause active interference. While the fact that the interrupting tasks used are designed to interact with the place-keeping goals, the

extent to which this interaction is responsible for the result observed is unclear; the design of interrupting tasks might determine not only which goals are affected, but how they are affected.

The next experiment, Experiment 3, aims to understand the extent to which the design of an interrupting task can affect disruptiveness, even when the goals in working memory that are affected by an interruption are the same. It does this by changing the audit tasks such that they operate differently but act upon the same goals.

4.4 Experiment 3: Investigating the effect of interrupting task on disruptiveness

4.4.1 Motivation

Experiments 1 and 2 show that the relevance of an interruption depends on the interaction of interrupting tasks with the contents of working memory at the moment of interruption. The previous experiments investigate how interruptions that target different place-keeping goals are affected at various moments.

In Experiment 1 participants were given questions to answer about the task or their progress in the task. In Experiment 2 this was changed to focus on subtask elements: rather than answer questions about subtask progress (e.g., *Type*, *Shape* etc), participants had to answer questions about subtask element progress (i.e., within subtask progress, *Diamond*, *Triangle* etc). This change was introduced to investigate the hypothesis that place-keeping goals were encoded differently at various points in the task, in turn determining the disruptiveness of a particular interruption.

Although the results of these experiments show that there is a direct relationship between interruptions and rehearsal and interference in working memory, it is not clear what, if anything, mediates the interaction between interrupting tasks and goals in working memory. The interrupting tasks in Experiments 1 and 2 were very similar; both asked participants questions about their progress through the task or subtasks (progress audits) or tested their knowledge of features of the task overall (knowledge audits). It is possible that the effects observed were due, in part, to the similarity of the interrupting tasks that were used. To what extent would a different interrupting task have produced different results?

Experiment 3 seeks to understand how the design of the interrupting tasks mediates the effects of interruptions on working memory. To do this, the visual representation of the interrupting task is changed, but the goal of the task is left unchanged. Whereas interruptions in Experiments 1 and 2 asked questions, the interruptions in Experiment 3 require participants to fill-in missing information. As part of filling these blanks in the relevant condition, participants have to rehearse place-keeping goals. Irrelevant interruptions require participants fill in different information that requires that they retrieve and rehearse goals that might interfere with place-keeping goals. The differences between audit tasks from previous experiments and the current experiment are illustrated in Figure 4.6.

Of course, radically different interrupting tasks that result in different reinforcement and rehearsal schedules in working memory are likely to elicit large differences in disruptiveness. Cades, Davis, Trafton, and Monk (2007) showed that more taxing interrupting tasks reduced participants' opportunities for rehearsal, increasing the disruptiveness of interruptions. Changing the modality of interruptions can change their disruptiveness (Ratwani, Andrews, et al., 2008) and longer interruptions are more disruptive (Monk et al., 2008). These results are not the focus here; rather, the focus is on interrupting tasks that have the same effect on working memory but different presentation.

In the same way that different representations of the same information can result in different effects on performance (Larkin & Simon, 1987; Peebles & Cheng, 2001; Gould, Cox, & Brumby, 2013b), the question is whether the presentation of interrupting tasks in Experiments 1 and 2 influenced their disruptiveness. There has been work that attempts to dissociate the task-oriented operations required to reach a goal from the steps involved in manipulating a device or interface along the way.
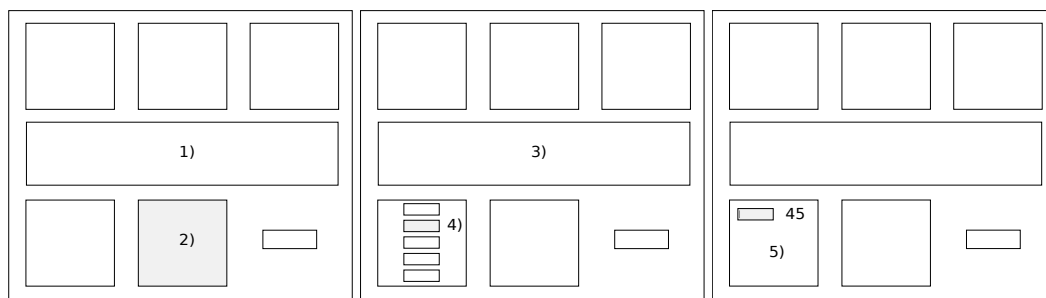
Figure 4.6: In Experiments 1 and 2, interruptions featured a question area (1), (3) and an answer area, which could be any of the five subtasks (2) or subtask elements (4). Experiment 3 takes a different approach; rather than being asked a question, participants had to complete missing details (5)

4.4.2   Method

4.4.2.1   Participants

26 participants (16 female) with a mean age of 26 years (SD=8 years) took part in the study. Participants were drawn from the UCL participant pool and were paid £7 for approximately one hour of their time. None of the participants in the experiment had participated in Experiments 1 or 2.

4.4.2.2   Design, Materials, Procedure

Experiment 3 used the same 2×2 repeated measures design as Experiments 1 and 2. The computer program, primary task, interruption timing, interruption composition and transcription task were the same in this study as they were in Experiment 1. The main
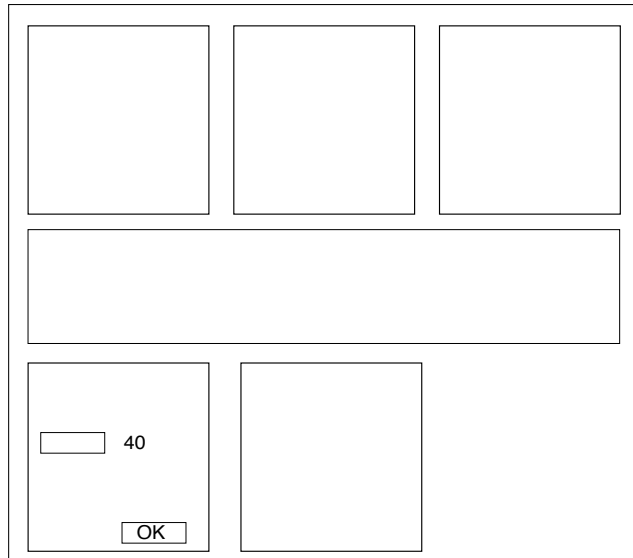
Figure 4.7: The audit task was designed to probe knowledge at the level of subtask elements. In contrast, Experiment 1 was designed to probe knowledge only of the subtask level. In this example, a participant is required to type in the element name that matches the number 40. In this example, the correct response would be 'Box'.

difference between Experiment 3 and Experiments 1 and 2 was the design of the audit task. The post-experiment questionnaire was also modified to explicitly solicit responses from participants about their strategy for place-keeping (see Figure A.10 in Appendix A).

Changes to the audit task   In Experiment 1, the audit task required participants to answer a question about their progress through the task (in the relevant condition) or about their knowledge of the task (irrelevant condition). The questions in Experiment 1 always asked questions at the subtask level; there were no questions about within-subtask progress or subtask elements (sub-subtask). It was hypothesised that the use of these subtask-level questions for both between- and within-subtask interruptions were the cause of the interaction between interruption relevance and timing observed in Experiment 1.

The design of audit tasks in Experiment 1 was based on the findings of previous literature, which has suggested that in Pharmacy Task-like tasks, place-keeping at the level of subtasks is encoded spatially (Ratwani & Trafton, 2008). The results from Experiment 1 generally support this claim, but the behaviour observed at the within-subtask level was different to that observed at the between-subtask level. The results suggest that as well as encoding separate place-keeping goals for within-subtask activities, these within-subtask place-keeping goals are encoded in a non-spatial manner. On the basis of participant performance and post-experiment questioning, the audit tasks in this study attempt to act on phonologically-encoded place-keeping goals.

In Experiment 3, the audit tasks focused on participants' knowledge of subtask elements and within-subtask progress. Interruptions asked participants to 'fill-in the gaps' for within-subtask elements. Participants were shown an interface with all information redacted, except for either a quantity or an attribute associated with the subtask they were working on at the moment of interruption (for within-subtask interruptions) or for the subtask they had just completed (for between-subtask interruptions). The 'gap' that participants were asked to fill could either be one of the subtask attributes (e.g., for colour it would be *White*, *Red*, *Brown* or *Blue*) or it could be a quantity associated with one of the subtask attributes (e.g., *5*, *15*, *20*, *35*). Participants had to fill in the missing piece of information from memory.

During relevant interruptions participants were asked to fill in information about the within-subtask element they were working on at the moment of interruption. As within-subtask interruptions appear when a text-box is clicked, it is assumed that participants click on the text-box that corresponds to the within-subtask element they are working on at the moment of the click. Irrelevant interruptions asked participants to complete the missing quantity or attribute for a within-subtask element they had either already, or had yet to, complete, but not the element they were about to work on at the point of interruption.

When participants were interrupted between subtasks, they completed the same exercise, except in this instance they were directed to fill in gaps for a within-subtask element from the subtask they had just completed. Thus, for between-subtask interruptions, irrelevant and relevant interruptions were identical. This was a consequence of the scoping of activities and place-keeping between- and within-subtask. Place-keeping at the subtask level is still necessary when resuming within a subtask — one has to go to the correct subtask before one can go to the correct subtask element. However, due to the hierarchical nature of the task, the reverse is not true: there is no reason for place-keeping to occur at the level of subtask elements because between finishing one subtask and starting the next, there are no subtask elements.

4.4.3 Results

4.4.3.1 Data processing — error and outlier identification

Raw output was processed using the same data processor used in Experiment 1. This meant that within- or between-subtask interruptions that resulted in incorrect resumptions were dropped. There were 180 instances of resumption error and all were removed from the data. There were 57 incorrect within-subtask resumptions out of a total of 312 within-subtask resumptions (i.e., 18%), all of which were excluded from analysis of resumption lag. This rate was similar to the rate of within-subtask resumption errors in Experiment 1. These incorrect resumptions and their relative proportions across conditions are considered and analysed separately.

After removing erroneous resumptions from the dataset, resumptions that were more than 1.96 standard deviations from a participant's individual mean resumption time were also removed as outliers. Eighteen resumptions were classified as outliers and were excluded from the analysis. This left a total of 370 trials included in the analysis of resumption lag.

Therefore, of a total of 624 resumptions (26 participants, 12 trials, 2 interruptions per trial), 237 were incorrect and 18 were outliers. This left 369 resumptions for analysis of resumption lags. This meant there were 41 fewer accurate resumptions for analysis than there were in Experiment 1.

4.4.3.2 Resumption lag

	Within	Between
Relevant	4232 (*SD*=1326)	5046 (*SD*=2443)
Irrelevant	4515 (*SD*=1669)	5174 (*SD*=2377)
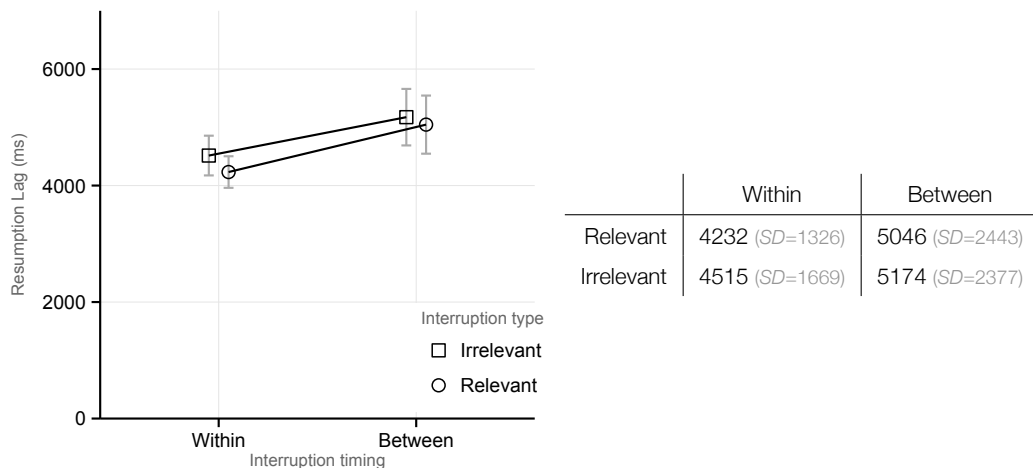
Figure 4.8 and Table 4.7 show aggregate resumption lags for Experiment 3. All times in milliseconds.

As in the previous experiment, resumption lag was the primary measure of performance. The resumption lag was the period between a participant completing an interrupting task

and restarting work on the Pharmacy Task. Restarting work was measured by the first click on an interface element on resumption. Participants resumed more quickly after relevant interruptions than after irrelevant interruptions. Participants resumed more quickly after within-subtask interruptions than after between-subtask interruptions. Figure 4.8 and Table 4.7 show the mean resumption times for each combination of interruptions.

A factorial repeated measures ANOVA was used to determine whether there were any significant differences in performance arising from the experimental manipulations. There was a significant effect of interruption relevance on resumption performance ($F(1,25)=5.54$, $p<.05$, $\eta_P^2=.18$). There was no effect of interruption timing ($F(1,25)=0.38$, $p=.54$, $\eta_P^2=.01$) and there was no interaction between the relevance and timing of interruptions ($F=0.59$, $p=.46$, $\eta_P^2=.02$).

These results suggest that disruptiveness was affected by the content of an interruption (relevant, irrelevant), but not its timing (within-subtask, between-subtask). This study replicates the main effect of relevance that appeared in Experiment 2. There was no interaction effect, meaning the effect of relevance was consistent with respect to the timing of interruptions.

4.4.3.3 Resumption error frequency and distribution

On resuming the primary task after interruption, there were two kinds of error that participants could make. The primary task required participants to resume on the subtask they are currently working on or about to work on. This might be a subtask they were in the middle of working on (within-subtask interruption) or it could be one that they had yet to start (between-subtask interruption). If participants attempted to resume on a subtask other than the one they were about to work on before the were interrupted, they will have made a sequence error.

When resuming after interruption participants made 180 sequence errors out of 624 resumptions. Of these errors, 135 were made when resuming after a between-subtask interruption; 66 were made after irrelevant interruptions and 69 after relevant interruptions. The remainder occurred after within-subtask interruptions; 28 errors occurred resuming after relevant interruptions while 17 errors occurred after irrelevant interruptions. These errors are summarised in the 'Subtask' columns of Table 4.8.

As well as selecting the wrong subtask, when participants resumed after a within-subtask interruption they could also select the incorrect subtask element. There were a total of 57

occasions when participants selected the wrong subtask element on resumption. Nineteen of the 57 errors came after within-subtask relevant interruptions, with the remaining 38 coming after within-subtask irrelevant interruptions.

Overall the error rates were higher in this study than in the previous experiments. Error rates across conditions in Experiment 3 were approximately double those of Experiments 1 and 2, although the pattern of results is broadly maintained.

	Within		Between
	Element	Subtask	Subtask
Relevant	12% (19)	18% (28)	44% (69)
Irrelevant	24% (38)	11% (17)	42% (66)

Table 4.8: Rates of resumption errors by condition and error type. Figures in grey are frequency.

4.4.4 Discussion

The results show that participants resumed more quickly after relevant interruptions than they did after irrelevant interruptions. This effect was consistent with the results of Experiment 2 and was expected because the audit tasks were focused on subtask elements. There was no main effect of interruption timing, although there was a trend toward faster resumption after within-subtask interruptions, as there was in Experiment 2.

This result supports a memory-centric account of interruptions relevance, and demonstrates that when the goals affected by an interruption are the same, small variation in on-screen appearance has no discernible effect on performance. Experiment 3 used an interrupting task that required participants to perform a gap filling question. The results were the same as previous experiments', which used an interrupting task involving questions and answers.

The consistency in performance between the experiments is to be expected given the structure of the primary task. Although the audit tasks differed in Experiment 3, the primary task was the same and activity type was the same across the audit tasks, even though its presentation changed. There is some evidence to suggest that participants use spatial encodings to keep track of their progress through tasks like the Pharmacy Task (Ratwani & Trafton, 2008; Werner et al., 2011). Thus, the results of Experiment 3 provide evidence for the idea that as long as a set of interruptions have the same reinforcing and interfering effects on place-keeping memory in a given task, their disruptiveness at a given moment will be predictable.

This has implications for the design of interruption management systems; it suggests that interruption management systems need to pay special attention to the cognitive factors that contribute to the effects of interruptions on users. While some efforts have been made to understand how the 'cognitive context' contributes to interruption management (Grandhi & Jones, 2010), there has been no real acknowledgement of the role that working memory plays in determining interruption relevance.

The results of Experiment 3 showed no main effect of interruption timing on resumption performance. This suggests that previous work may have over-stated the disruptiveness of within-subtask interruptions. Previous work has shown that participants often defer switches until the breakpoints between subtasks (e.g., Janssen, Brumby, & Garnett, 2012; Bogunovich & Salvucci, 2011) and developed interruption management systems around the concept (Iqbal & Bailey, 2010).

The findings of this study do not call these results into question, but they do suggest that other factors may be a bigger influence on disruptiveness. The data support the idea that interruption disruptiveness arises from the interaction between place-keeping goals at the point of interruption and the contents of the interruptions themselves. If the interaction is minimal or positive, interruptions within a subtask may not be all that disruptive.

This effect is likely to be contingent on the type of task being executed, however. In the Pharmacy Task, place-keeping cues are removed on resumption. Relying on memory to guide resumption improves resumption performance (Morgan et al., 2013; Morgan & Patrick, 2010), probably because there are few of the re-encoding costs associated with memory-based resumption that come when state is offloaded to the task (Salvucci, 2010). In tasks where state is not or cannot be represented in memory, within-subtask interruption may incur significantly higher resumption costs.

In addition to resumption lags, resumption errors were also recorded. These showed that participants made proportionally more errors resuming after between-subtask interruptions than after within-subtask interruptions. The previous experiments also reflect this trend.

The distribution of resumption errors in Experiments 1 (Table 4.2) and 2 (Table 4.4) also show that participants were more likely to resume at the correct subtask when they were interrupted within-subtask than when they were when interrupted between-subtask. Given that Experiment 3 differed substantially in the presentation of audit tasks when compared to Experiments 1 and 2, the evidence suggests that the skewed proportions are not

attributable to the design changes made to the task for Experiment 3. It is difficult to account for this discrepancy with Memory for Goals theory.

The Memory for Goals account of resumption (see Trafton et al., 2011) states that when resuming after interruption the memory system retrieves a place-keeping goal ('control-code'). Usually this place-keeping goal represents the last *completed* task. Due to noise in the memory system, sometimes the wrong place-keeping goal is retrieved from memory; this is the source of sequence errors on resumption. The theory implies that there should be no difference in resumption performance as a function of interruption timing. Goals receive spreading activation from the focus of attention (Trafton et al., 2011) while working through a task, which is to say that activation of goals only begins to fall when participants move on to the next goal or they are interrupted. All things being equal, decay over the period of an interruption should be the same for between- and within-subtask interruptions. This means the probability of retrieving the correct goal on resumption should also be the same.

An intuitively plausible suggestion for the effect is that place-keeping goals are created at the level of both subtasks and subtask elements. Having multiple place-keeping goals representing different levels of progress through a task means that when resuming, multiple goals are retrieved, which introduces a degree of redundancy. In this account, the hierarchical nature of the task would clearly have an effect on the strategies available for place-keeping. Memory for Goals does not explicitly account for hierarchical effects, although it could be extended to account for them; previous work has focused heavily on the hierarchical effects on performance in routine tasks (e.g., Gray, 2000).

The results of this study once again demonstrate that the relevance of an interruption varies with the contents of working memory at the moment of interruption. Experiment 3, along with Experiments 1 and 2 show that even though irrelevant interruptions were still related to the primary task they were still disruptive. This suggests that previous work that has used relatedness (e.g., Iqbal & Bailey, 2008) or content similarity (e.g., Arroyo & Selker, 2011) as proxies for relevance may not have made use of a reliable conceptualisation of relevance.

Although these experiments show that the content of working memory is a critical factor in the relevance of an interruption, they do not explore a variety of interruptions; both relevant and irrelevant interruptions were still concerned with the primary task. What happens if an interruption has nothing to do with the primary task? This is explored in the next chapter.

4.5 Summary

This chapter examined the effect of interruption relevance and timing on in-memory place-keeping. It demonstrated that interruptions can interfere with or reinforce place-keeping goals such that small changes in an interrupting task can have significant effects on resumption performance. It also demonstrated that these effects are highly sensitive to context. Place-keeping goals change over the course of task execution and as a consequence, the disruptiveness of an interruption will vary both with the timing of an interruption and the contents of working memory at the moment of interruption.

Experiment 1 used interrupting tasks that focused on progress across subtasks, whereas Experiments 2 and 3 focused on progress within subtasks. The results of all three experiments in this chapter demonstrated consistent main effects of interruption relevance on resumption lag. Experiment 1 compared the effects of relevant and irrelevant interruptions at different moments in task execution. It showed that relevance had a significant effect on post-interruption resumption after between-subtask interruptions. No effect of relevance was detected for within-subtask interruptions. To investigate whether this asymmetry was a result of hierarchical place-keeping effects, Experiment 2 focused on place-keeping behaviour within a subtask, changing the interrupting task to probe participants' knowledge of subtask elements. Statistical tests revealed a significant main effect of relevance on resumption lag. This result was indicative of people using different strategies for encoding place-keeping goals over the course of task execution. Finally, Experiment 3 examined whether the representation of interrupting tasks affects how they interact with place-keeping goals in working memory. The results replicated those of Experiment 2, implying that as long as interruptions cue rehearsal of the same place-keeping goals, their precise representation on the screen is not an important factor.

Although the experiments in this chapter compare relevant and irrelevant interruptions, both types of interruption were still related to the primary task. In reality, interruptions can come from a variety of sources - they may be entirely unrelated to the task at hand. The next chapter investigates how interruptions that are unrelated to the task at hand affect performance.

Chapter 5

Can related and relevant interruptions be distinguished?

5.1 Outline

Chapter 4 presented three experiments that investigated the effects of interruption timing and relevance on resumption after interruption. The results of these studies showed that the relevance of interruptions is partly the product of the interaction of interrupting tasks and the contents of working memory at the moment of interruption.

These studies provide strong evidence, but the conclusions that can be drawn are limited by the experimental materials. In Experiments 1, 2 and 3, interruptions could be relevant or irrelevant. However, both relevant and irrelevant interruptions were still focused on the Pharmacy Task; relevant interruptions asked participants to recall their progress through the task whilst irrelevant interruptions asked participants to recall information about the Pharmacy Task itself.

That significant differences in performance were observed between these two kinds of task emphasises how strongly rehearsing and interfering effects act on working memory and affect resumption performance. However, in most settings it would be unusual to only receive interruptions that are related to the task at hand. Workplace studies have shown that people receive interruptions from a wide variety of sources (e.g., Mark et al., 2005; McGillis Hall et al., 2010; Rouncefield, Hughes, Rodden, & Viller, 1994; González & Mark, 2004).

The experiments presented in this chapter look to understand how the relatedness — and not just the relevance — of interruptions affects their disruptiveness. In addition to the irrelevant and relevant interruptions related to the primary task from the previous chapter, this chapter investigates how an irrelevant interruption that is unrelated to the primary task affects performance.

Experiment 4 tests whether unrelated irrelevant interruptions are more or less disruptive than related irrelevant interruptions. It shows that in certain circumstances, interruptions that have no relation to the primary task can be less disruptive than irrelevant interruptions that are related to the task at hand. Experiment 5 builds on the results of Experiment 4 by making changes to the process of resumption. These changes allow for a closer inspection of the process of post-interruption resumption.

5.2 Experiment 4: Distinguishing related, relevant and irrelevant interruptions

5.2.1 Motivation

The previous experiments in this thesis all compared the effects of relevant and irrelevant interruptions on resumption performance. Relevant interruptions asked participants to bring their progress through the task to mind. Irrelevant interruptions asked participants to recall features of the Pharmacy Task. Both kinds of interruption were related to the task at hand; there was no change of context. How disruptive would an interruption be that was entirely unrelated to the primary task?

Experiment 4 builds directly on the findings of the experiments described in Chapter 4 to answer this question. The previous experiments in this thesis compared relevant and irrelevant interruptions that arrived between or within subtasks. As Figure 5.1 demonstrates, when interruptions focused on knowledge of progress within subtasks in Experiments 2 and 3, the results differed from those in Experiment 1. The present study combines the interruptions that focus on within and across subtask progress and adds an additional type of interruption: irrelevant and unrelated. Adding an interruption that is not related to the primary task means that Experiment 4 is able to test whether irrelevant but related interruptions are more or less disruptive than irrelevant and unrelated interruptions.

Being able to distinguishing different types of interruption is of significant practical and theoretical importance. It is of practical importance because of the nature of real work. Rather than solely being interrupted by things related to the task at hand, previous situated work has shown that people are interrupted from a variety of sources (e.g., Mark et al., 2005; McGillis Hall et al., 2010; Rouncefield et al., 1994; González & Mark, 2004). Interventions designed to reduce the impact of interruptions in safety-critical environments have focused on minimising distractions that are not related to the current activity (e.g., Anthony, Wiencek, Bauer, Daly, & Anthony, 2010; Colligan, Guerlain, Steck, & Hoke, 2012).

As well as its importance for practical applications, the relative disruptiveness of interruptions that are unrelated to the task at hand is also a theoretically interesting problem. Changes in working context may precipitate degradation of existing representations of context (Botvinick & Bylsma, 2005), resulting in errors. One might thus expect interrup-

tions that are unrelated to the task at hand to force a change in context and reduce resumption performance.

In complex tasks that require significant re-encoding of information from the environment on resumption (see Salvucci, 2010; Iqbal & Bailey, 2005), the costs of these context switches are very high (although they can be mitigated with cues: see, e.g., Parnin & DeLine, 2010; Jones et al., 2012); re-encoding the environment is time consuming. However, in tasks where there is little re-encoding required, it is less clear what the effects of unrelated interruptions might be.

Previous work (e.g., Cades et al., 2007) and the experiments previously reported in this thesis suggest that for routine tasks, the relationship between working memory and interruption is critical. Some interruptions afford rehearsal of goals, others cause interference. The question that Experiment 4 aims to answer is whether irrelevant interruptions that are unrelated to the task at hand cause more or less interference than irrelevant interruptions that are related to the task at hand.
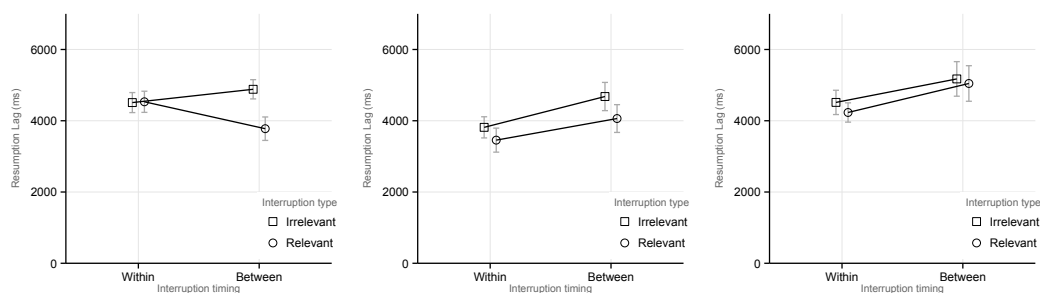
Figure 5.1: A review of resumption lags from experiments (left-to-right) 1, 2 and 3. Experiment 1 used interrupting audits focused on between-subtask place-keeping. Experiments 2 and 3 used interrupting audits focused on within-subtask place-keeping. All times in milliseconds.

5.2.2 Method

5.2.2.1 Participants

Thirty-five participants (14 male) with a mean age of 23 years (*SD*=4 years) took part in the study. Participants were drawn from the UCL participant pool and were paid £10 for approximately one and a half hours of their time.

5.2.2.2 Design

This study used more conditions than previous studies. It used a 5×2 within-subjects design. As in previous experiments, one independent variable was interruption timing. It had the same levels as the previous experiments, within-subtask and between-subtask. Inter-

ruption relevance was the other independent variable. Unlike previous studies, this had five levels: element-reinforcing, element-interfering, subtask-reinforcing, subtask-interfering and irrelevant. The differences between these conditions are explained further below.

5.2.2.3 Materials

Experiment 4 was deployed using an expanded version of the reimplemented task developed for Experiment 2. There were several changes made to accommodate the increased number of conditions in this study. Other changes were made to eliminate potential confounds. These changes are detailed below.

Interrupting task changes   Experiment 4 used the same schedule of tasking within each interruption as Experiment 2. All interruptions involved three audit tasks interposed between three transcription tasks. As in Experiment 2, there were two transcriptions to perform in each transcription task for a total of six transcriptions per interruption. The transcription task was the same as in previous experiments.

Changes to the audit task   Element interruptions asked questions about the elements that make up a subtask (e.g., *Red*, *Gum*, *Triangle*). Subtask questions asked questions about the subtasks that make up the subtask (e.g., *Type*, *Shape*, *Colour*). Irrelevant interruptions had no relationship to the task whatsoever, and therefore did not require element and subtask varieties. Relevant (i.e., element and subtask interruptions) questions always appeared in the *Prescription Sheet* area and the answers occupied the same location as the corresponding subtasks and elements. Irrelevant interruptions situated questions and answers within the prescription task to prevent the position confounding place-keeping. Element, subtask and irrelevant interruptions all required participants to pick one of five possible responses. This allowed for easier and fairer comparison of accuracy percentages. Below are some example questions with potential answers.

5.2.2.4 Procedure

There were a number of changes to the procedure, partly to accommodate the extra conditions, partly to remove confounds which may have existed in the previous experiment. These changes are enumerated below.

Changes to number and variety of trials   In previous experiments, participants experienced three trials for each combination of conditions (e.g., three relevant-within, three, irrelevant-between etc). In Experiment 4 there are ten combinations of conditions instead of four. It would not have been possible to collect data from 30 trials during the time

Timing	Relevance	Example question
Within	Element-reinforcing	"Which element were you about to work on before this audit task?"
	Element-interfering	"Which is the 'Gum' element?"
	Subtask-reinforcing	"Which subtask were you about to work on before this audit task?"
	Subtask-interfering	"Which is the 'Colour' subtask?"
Between	Element-reinforcing	"In the subtask you'll return to after this audit, which element will you work on first?"
	Element-interfering	"Which is the 'Gum' element?"
	Subtask-reinforcing	"Which subtask were you about to work on before this audit task?"
	Subtask-interfering	"Which is the 'Colour' subtask?"
Both	Irrelevant	"Which of these numbers is biggest?"

Table 5.1: A sample of the questions delivered as part of audit tasks in Experiment 4

available, so instead the number of trials was set at 20, giving room for two of each combination. To make up for the reduced number of trials per condition for a given participant, the number of participants was increased proportionally.

Changes to the cost of making errors   In the preceding experiments, the way that errors were penalised varied depending on the kind of error that participants made. In previous studies, sequence errors made at any point (during normal activity or on resumption) led to an eight second lockout. Sequence errors only occurred when participants selected the wrong subtask. This remains unchanged in this study.

Participants can enter values into the subtask elements in any order they like. In previous studies, when resuming after within-subtask interruptions, participants had to select the element that they were about to work on when they were interrupted. If they chose the correct subtask element, then their progress on previously completed subtask elements was restored and they continued working. If they selected a subtask element that they'd already worked on, or one that did not correspond to one of the orders on the *Prescription Sheet*, then they lost any progress they'd made on the subtask and had to re-enter the values. The cost of making an error therefore depended on the progress participants had made before they were interrupted and their typing speed. (The faster they could type, the less having to restart the subtask cost them.)

86

These fluctuations in error cost could have led to participants making strategic adaptations — speeding up resumptions when the cost of error was low, slowing down when the cost of error was high. However, data from Experiment 2 showed that participants resumed with the same level of caution after interruptions regardless of interruption timing; the error rates were effectively equivalent in both conditions (and even absolutely lower after within-subtask interruptions).

Nevertheless, it was felt that the cost of making errors needed to be normalised across the study. Therefore, in Experiment 4 the cost of making a resumption error after interruption were the same, regardless of whether participants had to choose the correct subtask or subtask element on resumption. Selecting the wrong subtask, or selecting a subtask element other than the one which initiated in the interruption resulted in an eight second lockout.

This study also introduced a new cost for errors made on the audit tasks during interruptions. In previous experiments, participants were not given any feedback on their answers to audit questions. This was to ensure that participants were not given extra place-keeping information during interruptions. However, not giving feedback on responses precluded the possibility of disincentivising guessing behaviour.

Although data from Experiment 3 shows that response accuracy was well above chance, this does not mean that participants performed to their full ability. Given that the experimental manipulations in this study are contingent on participants actively engaging with audit tasks, it was felt necessary to introduce some cost to making errors on audit tasks.

Of course, these costs could not be realised during the interruptions (because they would give participants place-keeping information and increase the duration of the interruptions) or on resumption (because it would also increase the duration of the interruption). Instead, participants accumulated the costs of audit errors, and 'repaid' them between trials. Each inaccurate response to an audit question added a two seconds to a lockout on trial completion. As there were two interruptions per trial, participants could be locked-out out for a maximum of 10 seconds.

Changes to the interruption schedule   In previous experiments, the questions asked during audits were randomly selected from a selection appropriate for the condition. Sometimes participants had the same question twice, although they usually had two different audit questions during the interruption. This presumed that questions for a given condition had precisely equivalent effects.

It is conceivable that questions from the same condition could have subtly different effects on performance, particularly in conditions where participants are asked about their progress. With this potential confound in mind, for this study participants were asked the same question for each audit during an interruption. This opens the possibility of investigating whether there is any variation in performance contributed by questions from the same condition.

Changes to the rest schedule   Experiment 4 required 10 conditions, so the way that trials were blocked did not allow for two rests. To compensate for the fact that there was only one rest among 20 trials, participants were given up to four minutes to rest. To ensure all participants had some degree of rest, the first 60 seconds of the rest period could not be skipped. This differed to Experiment 3 where participants were offered rests of 45 seconds after four trials and again after eight trials.

5.2.3 Results

Of the 35 participants to take part in the experiment, two had to be excluded from the analysis because they made so many resumption errors that no usable data were produced for one or more conditions. These participants were excluded from all analyses.

5.2.3.1 Data processing — error and outlier identification

Before the data were analysed, they were filtered. First, all erroneous resumptions were removed from the resumption lag data; a total of 168 cases were removed. The frequency and distribution of these errors across conditions is considered below.

As in previous experiments, outliers were removed from the sample at the 95% criterion (i.e., ±1.96 SDs). Whereas in previous experiments the 95% criterion was based on the standard deviation of a particular condition for the participant in question, Experiment 4 simply used all resumptions by a participant to compute a mean and standard deviation. This was because the 5×2 design of this study mean that each condition had a maximum for four data points (20 trials, 2 trials per condition, 2 interruptions per trial), increasingly the likelihood of individual values distorting the mean and standard deviation for a particular condition. Fifty-nine resumptions were removed from the sample on this criterion.

The thirty-three participants included in the analysis made a total of 1320 resumptions. Of these resumptions a total 227 (17%) were removed from the sample for being errors or outliers.

5.2.3.2 Resumption lag

Resumption lag — the time it took for participants to resume work on the primary task after interruption — was again the primary measure of performance. Resumption lags were computed from the 1093 resumptions that remained after removing outliers and errors.

	Within	Between
Irrelevant	3245ms (SD=1389)	3763ms (SD=1270)
Element-reinforcing	2748ms (SD=1199)	3831ms (SD=1415)
Element-interfering	3293ms (SD=1522)	4305ms (SD=1581)
Subtask-reinforcing	2804ms (SD=878)	3288ms (SD=1100)
Subtask-interfering	4003ms (SD=2383)	3921ms (SD=1281)

Table 5.2: This table shows aggregate resumption lags for Experiment 4. All times in milliseconds.

As in previous studies, the resumption lags were analysed with a factorial repeated measures ANOVA. The results of the ANOVA revealed a significant interaction of interruption timing and relevance ($F(3.09, 98.77) = 3.891$, $p < .05$, $\eta_P^2 = .11$). This test was Greenhouse-Geisser corrected due to a violation of sphericity ($\varepsilon = .48$, $p < .05$). Achieved power for the interaction was high ($> .99$). There were subordinate main effects of timing ($F(1,32) = 13.59$, $p < .01$, $\eta_P^2 = .30$) and relevance ($F(4,128) = 7.63$, $p < .001$, $\eta_P^2 = .19$, uncorrected; $\varepsilon = .69$, $p = .25$).

To better understand the nature of the interaction effect, an analysis of simple effects was carried out. These tests examine the simple effect of relevance on the two timing conditions and the simple effect of timing on the five relevance conditions.

First, the simple effect of relevance on timing is considered. Interruptions arrived either with a subtask or between a subtask. There was a significant simple effect of interruption relevance on how quickly participants resumed after within-subtask interruptions ($F(4,29) = 5.33$, $p < .01$, $\eta_P^2 = .42$). There was also a significant simple effect of interruption relevance on how quickly participants resumed after between-subtask interruptions ($F(4,29) = 5.13$, $p < .01$, $\eta_P^2 = .41$). These results show that the relevance of an interruption affected resumption performance for both within- and between-subtask interruptions.

Next, the simple effect of timing on relevance is considered. There were five levels of interruption relevance; subtask-reinforcing, subtask-interfering, element-reinforcing, element-interfering and irrelevant. There was a significant simple effect of timing on element-reinforcing ($F(1,32) = 20.57$, $p < .001$, $\eta_P^2 = .39$) and element-interfering ($F(1,32) = 13.10$, $p < .01$,
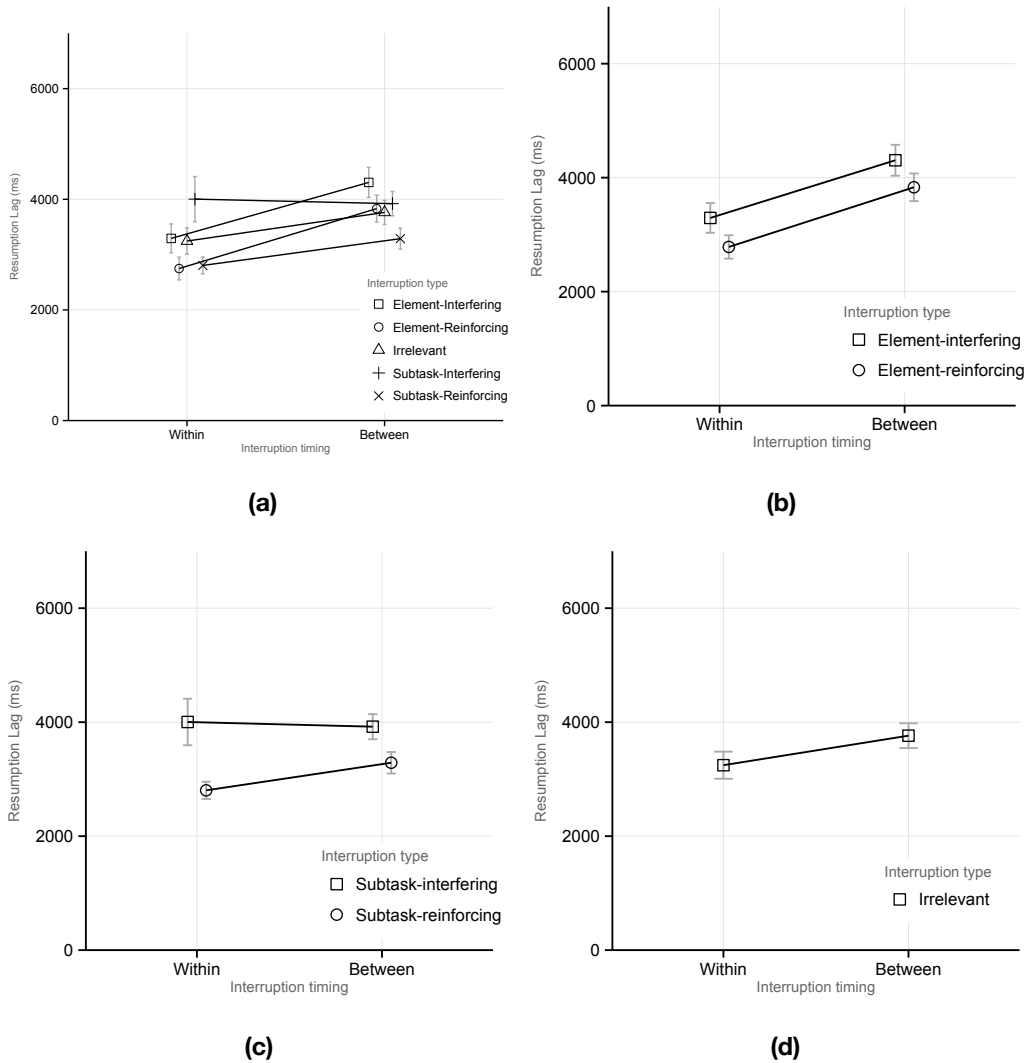
Figure 5.2: Figure 5.2a shows the complete interaction plot of resumption lags in Experiment 4. The other three figures show different features of this interaction plot for clarity. Figure 5.2b shows resumptions after element-level interruptions, which were similar to those used in Experiments 2 and 3. Figure 5.2c shows the effect of subtask-level interruptions, similar to those used in Experiment 1. Finally, Figure 5.2d shows the effect of irrelevant interruptions — these were a new type of interruption added in Experiment 4. All times in milliseconds.

η_P^2=.29) interruptions. There was no significant simple effect of timing on subtask-reinforcing (F(1,32)=4.13, p=.051, η_P^2=.41), subtask-interfering (F(1,32)=0.06, p=.80, η_P^2=.002) or irrelevant (F(1,32)=4.10, p=.051, η_P^2=.11) interruptions.

These results show that the overall interaction effect is driven by the non-significant simple effect of interruption timing on resumption after subtask-reinforcing, subtask-interfering and irrelevant interruptions. The significant effects for the other levels of relevance and for timing fit with the main effects observed. Timing affected the disruptiveness of particular levels of relevance on some occasions, but not on others.

Of particular interest was the differences between the interfering and irrelevant interruptions. Participants resumed significantly faster after within-subtask irrelevant interruptions than subtask-interfering (p<.05) interruptions but were no quicker than after element-interfering subtasks (p=.86). This suggests that when interrupted within a subtask, irrelevant interruptions were no more disruptive than element-interfering interruptions and were less disruptive than subtask-interfering.

When interrupted between subtasks, participants resumed significantly faster after irrelevant interruptions than element-interfering (p<.05) but were no quicker than after subtask-interfering subtasks (p=.46). This suggests that when interrupted between subtasks, irrelevant interruptions were no more disruptive than subtask-interfering interruptions and were less disruptive than element-interfering interruptions.

Taken together, these results show two things. First, they show that interruptions that are unrelated to the primary task are at worst no more disruptive than interruptions that are related to the primary task. In some scenarios, unrelated irrelevant interruptions are significantly less disruptive than some related by interfering interruptions. Secondly, the results again demonstrate hierarchical differences in the disruptiveness of interruptions. In this study, interruptions that focus on knowledge of subtask elements were significantly affected by interruption timing. Interruptions that focused on knowledge of between-subtask placekeeping were not affected by interruption timing.

5.2.3.3 Resumption error frequency and distribution

Participants each completed twenty trials. Each trial was interrupted twice yielding 1320 interruptions in total. Participants made an error when resuming on 168 occasions, giving an error rate of 13%. These 168 errors could be further broken down by experimental

	Within		Between
	Element	Subtask	Subtask
Irrelevant	9% (12)	1% (1)	11% (14)
Element-reinforcing	11% (15)	1% (1)	14% (18)
Element-interfering	9% (12)	2% (3)	13% (17)
Subtask-reinforcing	8% (11)	2% (3)	19% (25)
Subtask-interfering	12% (16)	2% (2)	14% (18)

Table 5.3: Rates of resumption errors by condition and error type. Figures in grey are frequency.

condition. Half of the interruptions that participants experienced came between subtasks. Of the 660 between-subtask interruptions, 92 were resumed incorrectly (i.e., 14%).

This means that there were 76 erroneous resumptions after within-subtask interruptions. When resuming after within-subtask interruptions, participants had to chose the correct subtask and the correct subtask element. Participants were extremely accurate in choosing the correct subtask when resuming after within-subtask interruptions, making only 10 errors (1.5%).

In terms of picking the correct subtask element after resumption, participants made errors on 66 occasions (10%). The distribution shown in Table 5.3 suggests they were fairly evenly distributed.

5.2.3.4   Primary task performance

Performance on the primary task outside of the period immediately following an interruption was not the main focus of Experiment 4. However, these data can provide some insight into the extent to which task execution was routine and the effect of interruptions on the strategies employed during execution of the primary task.

The number of errors made outside of the immediate post-interruption period provides some insight into the extent to which the task was routine for participants. Over a total of 700 trials, participants made 89 non-resumption errors. These errors were distributed over 78 trials, meaning that participants made non-resumption errors in 11% of trials. This error rate is quite high, but analysis of the errors suggests that they were fairly straightforward slips, and were not indicative of a lack of understanding.

	Within	Between
Irrelevant	7%	8%
Element-reinforcing	13%	9%
Element-interfering	8%	8%
Subtask-reinforcing	15%	13%
Subtask-interfering	10%	9%

Table 5.4: This table gives describes the distribution of all non-resumption errors across the ten conditions. A total of 89 non-resumption errors were made over a total of 700 trials.

There were two types of error that participants could make during the course of task execution. The first was element entry errors. These occurred when participants did not enter the correct values into the subtask elements. This could occur because the values were mistyped, were typed into the wrong box or were missed out.

Of the 89 errors made, 42 were element entry errors. The other 47 errors were sequence errors: these occurred when participants tried to work on an incorrect subtask. This often happened when participants tried to work on the next subtask having forgotten to click the 'OK' button to confirm their entries on the current subtask. Participants skipping ahead accounted for 21 of the 47 resumption errors.

Alternatively participants might not have realised they had pressed the 'OK' and tried to work on a subtask they had already finished. These errors could be due to a motor slip (accidentally double clicking) or because of a failure to notice that the button had been pressed (the values on the current subtask were cleared when participants clicked the 'OK' button) and accounted for a further 23 of the 47 errors.

Participants made fewer sequence errors as the experiment went on: 20 errors were made in the first five trials, 17 in the next five, 6 in the next and 4 in the final five trials. Table 5.4 provides a breakdown of all errors made by condition.

The 'downstream' effects of interruptions were also of interest. Participants were interrupted twice in each of the 20 experimental trials and this may have driven the strategies they chose for executing the primary task, which may have had some measurable effects on primary task performance. Mean trial time, shown in Table 5.5 shows that trials take approximately two minutes to complete — this is fairly consistent across conditions, with a high degree of variability, reflecting individual differences in speed of task execution.

	Within	Between
Irrelevant	117s (SD=20)	110s (SD=20)
Element-reinforcing	118s (SD=21)	115s (SD=21)
Element-interfering	120s (SD=22)	118s (SD=21)
Subtask-reinforcing	116s (SD=20)	109s (SD=18)
Subtask-interfering	121s (SD=22)	115s (SD=21)

Table 5.5: Mean trial duration for each combination of relevance and timing. All times in seconds.

	Before	Inter-interruption	After
Subtask, Within	10534ms (SD=2847)	10634ms (SD=2619)	10394ms (SD=2326)
Subtask, Between	10243ms (SD=2730)	13349ms (SD=4445)	11982ms (SD=3596)
Element, Within	1397ms (SD=698)	1419ms (SD=714)	1358ms (SD=611)
Element, Between	1360ms (SD=620)	1469ms (SD=743)	1379ms (SD=638)

Table 5.6: This table shows the mean of the aggregate time it took participants to complete subtasks (top two rows) and elements (bottom two rows) before, between and after interruptions. Note that times from interrupted tasks are not included in this figures. All times in milliseconds.

To get a better idea of the effects of interruptions, subtask and element execution durations were binned relative to their execution with respect to interruptions during a trial. Either execution of a subtask or element occurred during a subtask that came before any interruptions, or they came after one interruption but before the next, or they came after both interruptions had been completed. To keep calculations simple, the two tasks in a trial which had interruptions associated with them were ignored.

Mean completion times for elements and subtasks, split by trial interruption timing, are given in Table 5.6. Generally, the results show little difference in execution durations as a function of whether they came before, between or after interruptions. Subtask execution between one interruption and the next (i.e., the inter-interruption period) is slightly slower and more variable for between subtask interruptions, but this can be accounted for by the fact that at least one of the trials in the inter-interruption period includes the resumption period after the first interruption, which would add some time to the subtask completion duration.

	Within	Between
Irrelevant	23s $(SD=4)$	22s $(SD=4)$
Element-reinforcing	24s $(SD=5)$	25s $(SD=5)$
Element-interfering	24s $(SD=5)$	25s $(SD=5)$
Subtask-reinforcing	22s $(SD=4)$	23s $(SD=4)$
Subtask-interfering	24s $(SD=4)$	22s $(SD=4)$

Table 5.7: Interruption duration for each combination of relevance and timing. All times in seconds.

	Within	Between
Irrelevant	95%	92%
Element-reinforcing	70%	54%
Element-interfering	69%	67%
Subtask-reinforcing	75%	89%
Subtask-interfering	88%	93%

Table 5.8: Audit response accuracy by condition. Expected rate produced by guessing would be 20%.

5.2.3.5 Interrupting task performance

In previous experiments, there was no penalty for guessing the answers to audit tasks. In Experiment 4 accuracy on audit tasks was recorded and participants were given a time penalty of two seconds for each error made on audit tasks during a trial. The penalties were aggregated and served at the end of the trial. As in previous experiments, there were five possible answers to each question. Guessing would therefore yield a 20% accuracy rate. Aggregated audit accuracies are shown in Table 5.8, and show a high level of accuracy compared to chance and results from previous experiments.

The *element-reinforcing* and *element-interfering* audit tasks were analogous to the *relevant* and *irrelevant* audit tasks in Experiment 2. The results suggested that participants were more accurate on audit tasks in Experiment 4 than they were in Experiment 2.

Also of interest was the time it took participants to start working on interruptions after they had appeared. This time period is in some ways analogous to *interruption lag* (Trafton et al., 2003), which is the time period between being alerted to the imminent arrival of an interruption and the appearance of an interruption.

In Experiment 4 participants were not warned about the onset of an interruption, therefore any costs of switching are borne in the first moments after being switched to the interrupting task. In this case, the first moments after being switched occurred in the first of three transcription tasks. Examining how quickly participants started typing in each of the three transcription task might therefore reveal these switching costs, if they are indeed manifested in the first moments after interruption.

Figure 5.3 illustrates the time it took participants to start typing on the transcription tasks for first, second and third transcription tasks. As can be seen, participants are slower to start typing on the first transcription task than they are in the second and third transcription tasks.

To investigate these differences in more detail, a series of factorial repeated measures ANOVAs were conducted. The results of these tests show that there was a significant effect of interruption timing on the length of time it took participants to start typing on the first transcription task ($F(1,32)=52.12$, $p<.001$, $\eta_P^2=.62$). No such effect was observed for the second ($F(1,32)=1.31$, $p=0.26$, $\eta_P^2=.04$) or third ($F(1,32)=0.25$, $p=0.61$, $\eta_P^2=.01$) transcription tasks. These results show that in the moments immediately after interruption onset, participants took significantly longer to start working when interrupted within subtasks than they did when interrupted between subtasks.
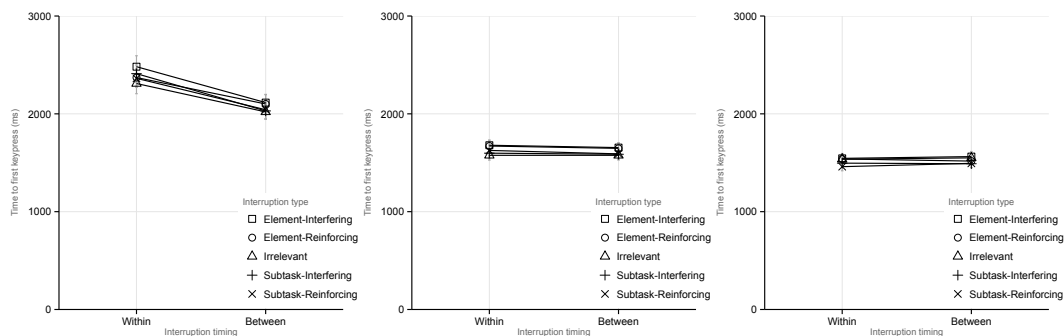
Figure 5.3: These figures show the time taken for participants to begin typing on the (left to right) first, second and third transcription tasks in each interruption. Note that the first transcription task is presented immediately at the moment of interruption. All times in milliseconds.

5.2.3.6 Post-experiment questionnaire responses

After completing the experiment, participants completed a 17 item questionnaire. The purpose of the questionnaire was to provide descriptive information about participants' subjective experiences of different parts of the experiment.

No statistical tests were run, instead the median and modal responses for each question were calculated. Two questions related to perceptions of interruption timing were of particular interest. When participants were asked the extent to which they agreed with the statement "The audits were harder when they came in the middle of a subtask" (i.e., within-subtask interruptions). The median and modal response was "Agree". When asked "The audits were harder when they came after finishing one subtask but before starting the next" (i.e., between-subtask), the median and modal response was "Disagree".

Discussions with participants after the completion of the post-experiment questionnaire confirmed that participants found it more difficult to deal with within-subtask interruptions than they did between-subtask interruptions. A complete list of questions with median and modal responses is given in the Appendix A.

5.2.4 Discussion

The results of Experiment 4, Experiment 4, demonstrate that an interruption being related to a task does not guarantee it will be less disruptive than an unrelated interruption. In this study, there were instances where the unrelated, irrelevant interruption was less disruptive than an irrelevant but still unrelated interruption. This is an important finding because it contradicts the common assumption (e.g., Iqbal & Bailey, 2008; Arroyo & Selker, 2011) that interruptions that are related to the task at hand are fundamentally less disruptive.

The five different levels of interruption relevance in Experiment 4 can be thought of as asking participants three different kinds of questions. In Experiment 4, irrelevant interruptions asked participants to choose the largest or smallest number from a list of five numbers. These questions had no relationship at all with the primary task. The other two categories, subtask and element, were both related to the primary task. *Subtask* interruptions asked participants about their knowledge of subtasks or their progress through them. *Element* interruptions asked participants questions about their knowledge of subtask elements or their progress through them. Both subtask and element interruption types came in reinforcing and interfering varieties.

Although the interaction effect observed in Experiment 4 had different characteristics to the one in Experiment 1 by virtue of representing more conditions, the same conclusion can be made from these results, and with more confidence. Experiment 1 had relevant and irrelevant interruptions, but both types of interruption were still related to the primary task — they mapped to the subtask-reinforcing and subtask-interfering conditions in Ex-

periment 4. The interaction effect in Experiment 1 seemed to be driven by timing effects — that relevance of the interruptions varied over the course of task execution.

The goal of Experiment 4 was to investigate the effect of interruptions that were completely unrelated to the task at hand alongside interruptions designed to have reinforcing and interfering effects. Although the significant interaction in Experiment 4 indicates that timing moderated relevance in Experiment 4 also, more important was the relationship between the irrelevant interruptions compared to the reinforcing and interfering interruptions that were related to the primary task.

As predicted, completely irrelevant interruptions tended to be more disruptive than reinforcing interruptions, but less disruptive than certain interfering interruptions. This finding lends support to the idea that similarity of content is not a sufficient criterion for judging interruption disruptiveness: interruptions with high content similarity have the potential to interfere with place-keeping goals in memory through proactive interference to a greater extent than interruptions that have nothing to do with the task at hand.

5.2.4.1 Hierarchical effects

As in previous experiments, hierarchical effects were evident in the distribution of resumption errors. When participants resumed after between-subtask interruptions, their chance of selecting the wrong subtask was approximately 14%. However, when resuming after within-subtask interruptions participants chose the wrong subtask only 2% of the time. When resuming after within-subtask interruptions participants also had to remember which subtask element they were about to work on at the moment they were interrupted; they chose the wrong subtask element on approximately 10% of resumptions. Despite having more opportunity for making an error on resuming after a within-subtask interruption (the wrong subtask or subtask element could be chosen), participants made fewer errors when resuming after within-subtask interruptions.

These results suggest that participants found it easier to remember which subtask they were working on when interrupted within a subtask than they did when interrupted between subtasks. This likely stems from the hierarchical arrangement of the task; if participants are able to successfully remember which element they are working on, they get information about the subtask they are working on for 'free'. The low probability of selecting the wrong subtask when trying to choose the correct element suggests that even when recall for the current element fails, there is still a strong memory for the subtask that needs to be worked on.

This fundamental asymmetry might explain why the simple effects of timing were consistent across relevance conditions; within- and between-subtask interruptions are not discrete in terms of their place-keeping requirements. Within-subtask place-keeping requires participants also keep track of which subtask they are working on, but between-subtask place-keeping is ignorant of which element needs to be worked on next.

The hierarchical nature of many routine tasks has been mentioned previously in the interruptions literature (e.g., Ratwani & Trafton, 2008; Monk et al., 2008; Altmann & Trafton, 2007) and more broadly as part of discussions of routine action (e.g., Gray, 2000; Botvinick & Plaut, 2004). However, no hierarchical effects have previously been documented in the context of recovering after interruption. This is a novel finding.

Alternatively, the slow resumptions after between-subtask, element-level interruptions might be a simple result of the constraints imposed by the task on resumption. As (Salvucci, 2010) notes, the time it take to resume a task depends on the amount of re-encoding that is required. This is also the case with the Pharmacy Task. In particular, the results of this study may simply be an artefact of the steps involved in resuming, rather than the strength of goals in working memory. For example, when resuming after an interruption, participants have the choice; they can recall what they need to do next and select the appropriate subtask or element, or they can recall what they need to do next, get the information they need from the *Prescription Sheet* to complete the next step, and then select the appropriate subtask or element.

There are several effects that these strategies could have on resumption lag. Some participants will select a subtask or element much more quickly than others who perform other operations before indicating where they intend to resume. As well as variation across individuals, there is also the possibility that this behaviour varies depending on the moment of interruption. Participants might demonstrate where they intend to resume more quickly after within-subtask interruptions than after between-subtask interruptions. This would explain why participants generally resume faster after within-subtask interruptions. The next experiment in this chapter explores how the process of resumption might contribute to the performance observed in the studies so far.

5.2.4.2   Strategic adaptation to perceived disruptiveness

The results of the post-experiment questionnaire and verbal free-response questions suggested that, by and large, participants found that within-subtask interruptions were more difficult to deal with than between-subtask interruptions. If interruptions at a certain time

are more difficult to deal with, then one would expect resumption performance to be worse after such interruptions.

However, the results of Experiment 4 show participants resumed more quickly after within-subtask interruptions than between-subtask interruptions. There are two ways that this can be interpreted. Either there is no direct relationship between perceived and actual disruptiveness of interruptions, or participants made strategic adaptations to accommodate the increased disruptiveness of within-subtask interruptions.

Timing had no significant effect on resumption lag after irrelevant interruptions. This is informative because irrelevant interruptions give an insight into the disruptiveness of different interruption timings without the interactions between timing and relevance that occur when the interruptions are related to the primary task (as they were for the other 4 levels of interruption relevance). This could indicate that there was no effect of interruption timing when the interaction effects are discounted, or could simply reflect that strategic adaptations prevented resumption performance after within-subtask interruptions being worse than it might otherwise have been.

Stronger evidence for strategic adaptation comes from timing data. Examining the total time spent on interruptions and the total trial time, there is nothing to suggest that participants were taking extra time to strategically rehearse. However, at this level the data are quite noisy. If participants were strategically rehearsing their progress when they were interrupted, it is likely that this rehearsal would occur at the moment of interruption.

The first task that participants were faced with when interrupted was a transcription task. Participants were generally slower to start typing on the first transcription, but more importantly and less obviously, they were significantly slower to start working on the first transcription when interrupted within a subtask than they were when interrupted between subtasks. No such effect was apparent for the second and third transcriptions.

This result provides strong evidence in support of the theory that participants spent longer rehearsing place-keeping goals at the moment of interruption when they were interrupted within a subtask. This demonstrates that even in a heavily constrained task, people make small strategic adaptations to their place-keeping behaviour which in turn can have significant effects on performance. Such adaptations have been noted in the past; Altmann and Trafton (2004) showed that the longer participants had to prepare for an interruption, the easier they found it to resume. Morgan et al. (2013) showed that making encoding of progress before an interruption more effortful led to improved resumption performance.

Other work has demonstrated strategic adaptation of task switching behaviour to meet the demands of task environments whether effectively (e.g., Morgan et al., 2013; Bogunovich & Salvucci, 2011) or not (e.g., Katidioti & Taatgen, 2014).

Memory for Goals (Altmann & Trafton, 2002) suggests that strategic rehearsal can increase the activation of place-keeping goals. When place-keeping goals are rehearsed, subsequent retrieval is quicker and more accurate. This is reflected in shorter resumption lags. It seems that in Experiment 4 participants were rehearsing their progress at the moment of interruption, and that the extent of rehearsal was determined by participants' subjective experiences of the disruptiveness of interruptions.

5.2.4.3 Strategic adaptation to frequent interruption

In Experiment 4, each trial was interrupted twice by interruptions from the same condition. The points at which interruptions appeared in a trial were randomly allocated to one of the possible interruption positions in the primary task. The order of trials was also randomised for each participant. Despite this randomisation, there was a good deal of information about the interruptions that participants could infer; participants knew that they would be interrupted at some point in each trial, and that the second interruption would have the same timing and relevance as the first. Knowing this information would allow participants to adapt their behaviour to exploit the knowledge.

In post-experiment questioning, some participants suggested that they were careful to remember what they were about to click on while they were working on the subtask. Other participants suggested that once they had completed the two interruptions in a trial, they could speed up because they knew they wouldn't be interrupted again. The data collected suggest that if participants did use these strategies, they were not universally adopted; participants largely worked at a steady rate throughout the trials irrespective of how many times they had been interrupted.

Although participants did not seem to adapt their behaviour depending on where they were in a trial, it is still likely that participants adapted their place-keeping behaviour to the regularity of the interruptions over the course of the experiment. If interruptions were far less frequent, it is unlikely that participants would have strong, ongoing place-keeping strategies: the benefit would not exceed the cost of enacting these strategies.

Being interrupted frequently favours a strong, well proceduralised strategy. Indeed, Monk (2004) found that participants interrupted every 10 seconds resumed more quickly than

those interrupted every 30 seconds. Monk attributed the apparent reduction in disruptiveness to the adoption of more aggressive goal maintenance strategies. Future experiments should investigate the extent to which interruption frequency affects resumption performance in this task.

5.3 Experiment 5: Deconstructing resumption lag, understanding interruption disruptiveness

5.3.1 Motivation

Experiment 4 showed that, in some situations, interruptions that are unrelated and irrelevant to the task at hand can be less disruptive than interruptions that are related but irrelevant to the task at hand. This suggests that content similarity and relatedness are not reliable proxies for interruption disruptiveness. However, close scrutiny of the results raised another question; to what extent does the structure of the Pharmacy Task — rather than the content of interruptions — determine resumption speed?

In the experiments previously reported, resumption lag was the period between an interruption finishing and a participant resuming activity on the primary task. Resumption usually takes around four seconds in the Pharmacy Task, but as Salvucci (2010) points out, resumption lags of this order are far longer than it should take to simply retrieve a goal from memory. Instead, periods of resumption are made up of a combination of recall from memory, re-encoding of information from the task and sensorimotor pointing tasks.

Previous work to investigate how the complexity of interruptions affects resumption performance backs up this argument. Iqbal and Bailey (2005) recorded resumptions of over ten seconds in a document-editing task. Other work has shown that interruptions have a disproportionately negative effect on complex tasks compared to simple tasks (Speier, Vessey, & Valacich, 2003). This supports the idea that accessing the environment makes up a significant proportion of resumption lags; as more complex tasks incur higher encoding and re-encoding costs. Oulasvirta and Saariluoma (2006) demonstrated the costs of re-encoding by showing that encoding representations in longer-term memory avoided these costs.

Further evidence for the contribution of re-encoding and sensorimotor costs to resumption lag comes from efforts to strengthen memory before interruptions. Morgan et al. (2013) showed that by forcing participants to make more effort to commit goal-state to memory resumption time was reduced; a second experiment showed this was because less re-encoding was required on resumption.

Resumption in the Pharmacy Task is reliant on memory because the interface removes place-keeping cues at the moment of resumption. Nevertheless, there are features of the task that remain visible and that could aid memory, or augment a memory-based

resumption strategy. Re-encoding these aids so that they can facilitate resumption takes time. These extra operations will manifest in increased resumption lags. In particular, resumption lags might be extended by participants looking at the *Prescription Sheet* to work out what the next input value will be before they perform any actions that would demonstrate they knew where to resume.

Experiment 5 seeks to determine the contribution of task features and post-interruption operations to resumption lag by forcing participants to make 'naked resumptions', where they are required to indicate where they intended to resume working before returning to the primary task. In this way, participants cannot utilise features of the primary task or perform pre-resumption operations such as collecting new information. This involves making changes to the resumption process. Whereas in all previous experiments the interruption finished and participants were put straight back into the primary task, in Experiment 5 participants are shown the task with the *Prescription Sheet* missing and asked to indicate where they intend to resume the task.

It is expected that participants will be faster to indicate their resumption destination than they would be to resume work on the primary task because there will be less chance of determining resumption location from task features. Additionally, there will be no time added by participants encoding target information for the next task before performing any actions. However, without some of these implicit cues to support resumption, it is expected that participants will make more errors when indicating their resumption intentions than in previous studies where such cues were present.

5.3.2   Method

5.3.2.1   Participants

23 participants (14 male) with a mean age of 23 years (SD=4 years) took part in the study. Participants were drawn from the UCL participant pool and were paid £10 for approximately one and a half hours of their time.

5.3.2.2   Design

Experiment 5 had the same 5×2 within-subjects design as Experiment 4. There were two independent variables, relevance and timing. Relevance had five levels; element-reinforcing, element-interfering, irrelevant, subtask-reinforcing, element-interfering. Timing had two levels; within-subtask and between-subtask. There were two measures of interest in this study. The first was choice lag. This was the time it took participants to make a decision about where they intended to resume the primary task. The second was

resumption lag. This was the period of time it took participants to start working on the primary task after they'd chosen where they'd resume. Other measures such as error rate and trial duration were also collected.

5.3.2.3 Materials

The materials used in Experiment 5 were the same as those used in Experiment 4 with a significant change to the way that participants resumed after interruption. In previous experiments participants completed the interruption and were sent back to the Pharmacy Task, where they had to try and resume accurately. Any progress that participants had made was hidden on resumption and during completion of the task. Experiment 5 re-organised the process of resumption such that resumption aids and sources for pre-resumption operations were hidden.

First, rather than being put straight back into the Pharmacy Task on resumption, participants are first given a 'resumption selection' screen. On this screen, subtask progress and the *Prescription Sheet* are hidden. Participants select where they intended to resume by selecting the appropriate subtask or subtask element.

Secondly, because resumption progress is hidden on the selection screen, progress is no longer hidden once participants have indicated where they intend to resume. In a significant change to previous experiments, values are left in place as participants work through subtasks. Once participants have dealt with the resumption selection screen, progress becomes visible once again. To make progress even clearer the 'OK' buttons are removed once a subtask is complete.

If participants resumed in the wrong place on the resumption selection screen, they were given an eight second lockout penalty just as they were in previous experiments. This penalty was served immediately after the resumption selector, before participants returned to working on the primary task. In all other respects the experiment was identical to Experiment 4.

5.3.2.4 Procedure

The procedure differed in some ways from previous experiments. Data in previous studies was collected participant-by-participant. In this study participants were run simultaneously in blocks of up to ten participants.

Participants arrived ten minutes before the start of the session and were taken to a computer lab. To start, all participants were shown the introductory video on a data projector,
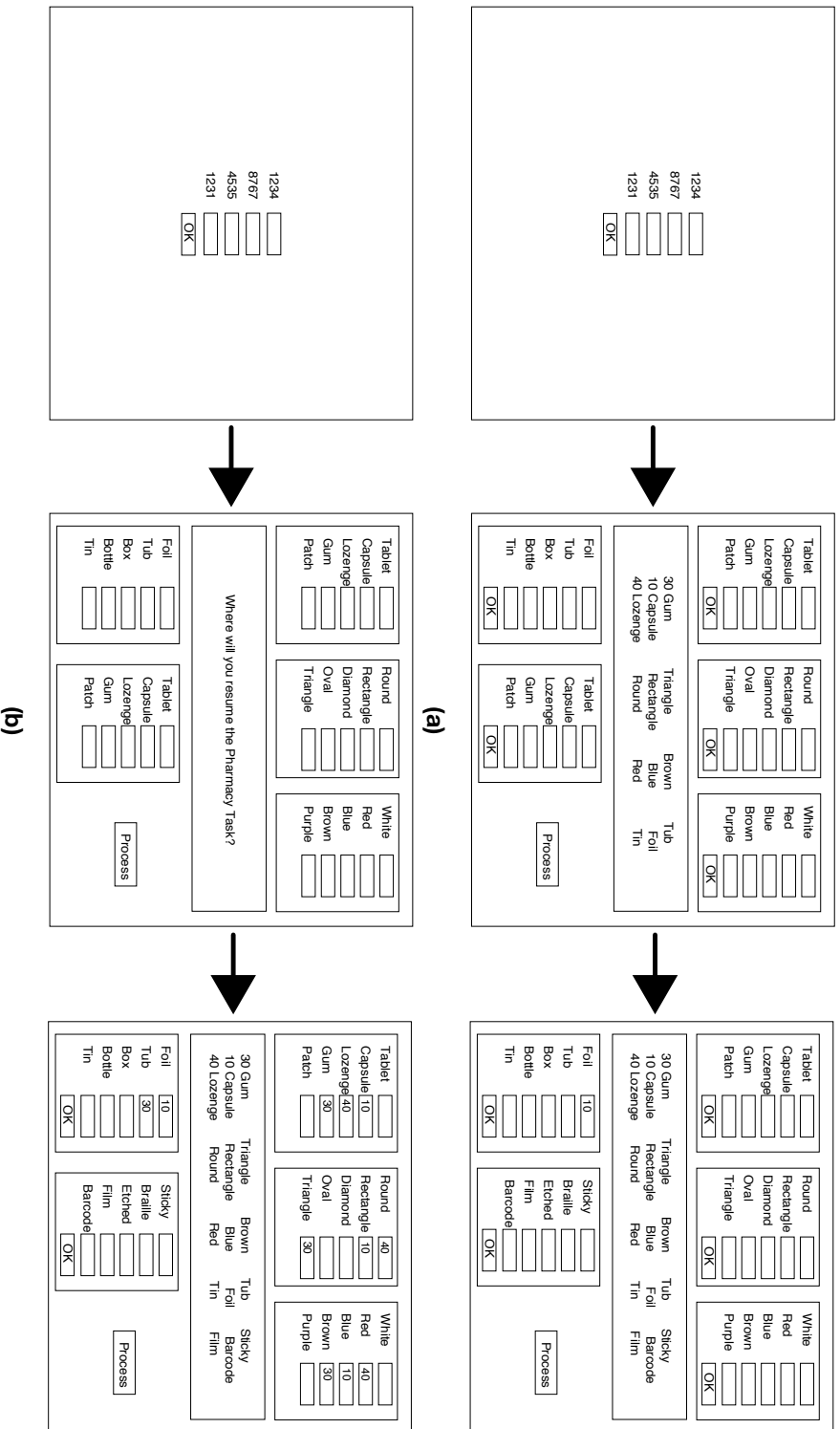
Figure 5.4: This figure illustrates the modifications made to the resumption procedure for Experiment 5. Figure 5.4a shows the process of resumption for Experiment 4. In Experiment 5, participants complete an interruption and return to the primary task to resume on whatever they were about to work on when they were interrupted. Progress through task is not displayed, but the *Prescription Sheet* is available on resumption. Figure 5.4b shows the process for Experiment 5. After an interruption, participants are asked to indicate where they will resume. They are then returned to the Pharmacy Task. No *Prescription Sheet* is visible when choosing where to resume. Cues display task progress once participants have returned to the Pharmacy Task.

106

after which they were directed to the information page. Participants then had an opportunity to ask questions, which were answered to the group as a whole.

After this, participants gave consent and proceeded as they would have in Experiment 4. While participants run individually were given verbal debriefings, this was not possible for groups. Participants who took part in a group were given a written debriefing.

Additionally, free response questions, which were answered verbally and recorded by the experimenter in individual sessions, were instead asked alongside the post-experiment questionnaire. Participants responded in free-response text fields. Participants could leave once they had finished and were paid as they left the lab.

5.3.3 Results

One participant did not complete the experiment. Their data are excluded from all analyses. This left data from 22 participants for use in the following analyses.

5.3.3.1 Choice lag

Experiment 5 measured a new construct, choice lag. Choice lag was the time it took participants to decide where they intended to resume. Choice lag was measured from the moment participants completed the final transcription task of an interruption to the moment that they made a selection on the resumption indication screen. Resumption lag came after choice lag and measured the period of time between a participant indicating where they intended to resume and actually resuming the task.

It had been planned to use the same data processing technique for choice lags as had been used previously for resumption lags. Incorrect choices would be discarded from the analysis and outliers would be removed. However, the planned approach presented significant difficulties. As described in the next section, participants had great difficulty with the choice task. Ten of 22 participants failed to correctly indicate where they were going to resume in at least one condition. Six of these participants failed to correctly indicate where they were going to resume in two or more conditions.

The distribution of the errors that participants made is considered below. For the analysis of choice lags, both correct and incorrect errors were included. Outliers were still removed against the 95% criterion (i.e., ±1.96 SDs) in the same way that they were for resumption lags. Ninety-four outliers were discarded from the analysis of choice lags. This left a total of 786 choice lags for analysis.
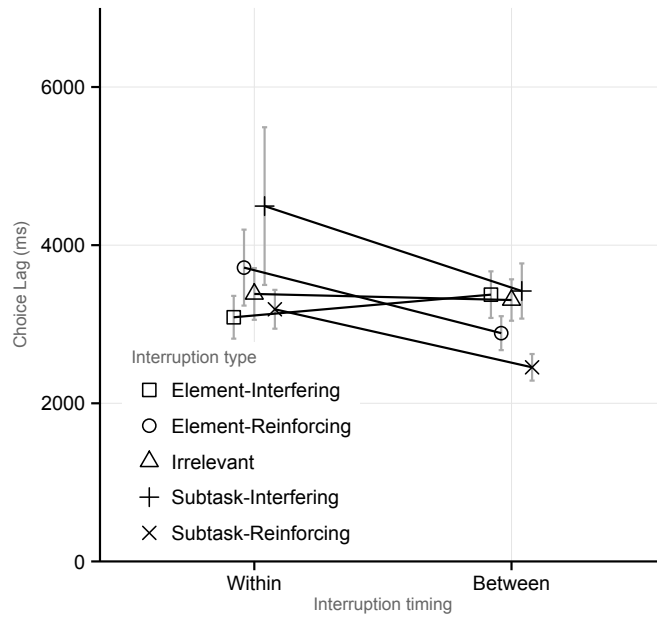
	Within	Between
Irrelevant	3383ms (*SD*=1534)	3306ms (*SD*=1229)
Element-reinforcing	3716ms (*SD*=2253)	2887ms (*SD*=1007)
Element-interfering	3088ms (*SD*=1269)	3374ms (*SD*=1382)
Subtask-reinforcing	3189ms (*SD*=1152)	2455ms (*SD*=789)
Subtask-interfering	4494ms (*SD*=4677)	3420ms (*SD*=1637)

Figure 5.5 and Table 5.9 show aggregate choice lags for Experiment 5. There were no significant main effects or interactions. All times in milliseconds.

In all but one condition (element-interfering) participants tended to make their choices more quickly when before resuming after between-subtask interruptions. Choice lags were analysed with a factorial repeated measures ANOVA. The tests revealed no significant interaction or main effects. There was no effect of timing ($F(1,21)=2.20$, $p=.15$, $\eta_P^2=.10$). There was no effect of relevance ($F(1.6,33.2)=2.56$, $p=.10$, $\eta_P^2=.11$) once a Greenhouse-Geisser correction had been applied to compensate for a violation of sphericity ($\varepsilon=.05$, $p<.001$). Achieved power was high for the effects of timing (.95), relevance (.98). The interaction term was subject to a significant violation of the assumption of sphericity ($\varepsilon=.02$, $p<.001$). With a Greenhouse-Geisser correction applied, there was no significant interaction ($F(1.3,28.2)=1.36$, $p=.26$, $\eta_P^2=.06$). Although there was a trend toward faster choices after between-subtask interruptions, the results of the ANOVA suggest that none of the manipulations had an effect on how quickly participants chose where they intended to resume.

5.3.3.2 Choice error frequency and typology

It is possible that by including both correct and incorrect choices in the analysis of choice lags the power of the test was diminished. However, the inclusion of errors in the analysis of choice lags was made necessary by the high error rates elicited by the choice task; some conditions did not have an analysable quantity of correct choices.

When presented with the choice task, participants had to indicate where they intended to resume on the Pharmacy Task. Participants were told that they should select the subtask or subtask element that they were about to work on before they were interrupted. From a total of 880 choice tasks, participants selected the wrong subtask or subtask element on 319 occasions (36%). Of these errors, 134 were made when resuming after between-subtask interruptions. The remainder came after within-subtask interruptions. Of the 185 incorrect choices made after within-subtask interruption, 144 involved the selection of the correct subtask but incorrect subtask element. The other 40 errors occurred because participants selected the wrong subtask.

There was significant variation in the number of errors participants made. Over the course of the experiment, participants made between 4 and 37 choice errors. The mean number of errors participants made over all conditions was 15 (*SD*=8).

The high error rates, and the variation in error rates across participants suggests that some participants found the choice task difficult. Although error rates were generally high, the pattern of choice errors was reflected the pattern of resumption errors observed in

previous experiments; participants were less likely to choose the wrong subtask after a within-subtask interruption than they were to choose the wrong subtask element after a within-subtask interruption or the wrong subtask after a between-subtask interruption.

	Within		Between
	Element	Subtask	Subtask
Irrelevant	20% (18)	16% (14)	35% (31)
Element-reinforcing	28% (25)	9% (8)	25% (22)
Element-interfering	43% (38)	2% (2)	32% (28)
Subtask-reinforcing	28% (25)	9% (8)	27% (24)
Subtask-interfering	43% (38)	10% (9)	33% (29)

Table 5.10: Rates of choice errors by condition and error type. Figures in grey are frequency.

5.3.3.3 Resumption lag

Before resumption lags were analysed, instances of participants trying to work on the wrong subtask or subtask element — resumption errors — were discarded. A total of 16 errors were made when participants were resuming the primary task from a total of 880 resumptions (2%). Given the introduction of place-keeping cues in the task, this low error rate is not surprising. The breakdown of these errors is covered later.

After discarding errors, outliers were removed from the sample at the 95% criterion (i.e., ±1.96 SDs). As in Experiment 4, outliers for a given participant were computed against all of their resumptions over the experiment. Outliers were computed in this way because the 5×2 design meant there were four resumptions recorded per condition; this made the outlier computation for resumption lags in a particular condition too sensitive. From the 864 resumptions that remained after discarding errors, a further 98 resumptions were discarded as outliers.

After filtering outliers and errors, 766 resumption lags were left for resumption lag analysis. In Experiment 5, resumption lag was the period of time that it took participants to start working on the Pharmacy Task once they had returned from an interruption. As place-keeping cues were visible in Experiment 5, this was the time it took for participants to work out where they were, encode any information require for the next task, and to select the appropriate interface widget.

As can be seen in Figure 5.6, there is a clear trend across all conditions. Participants resume more quickly after within-subtask interruptions than they do after between-subtask
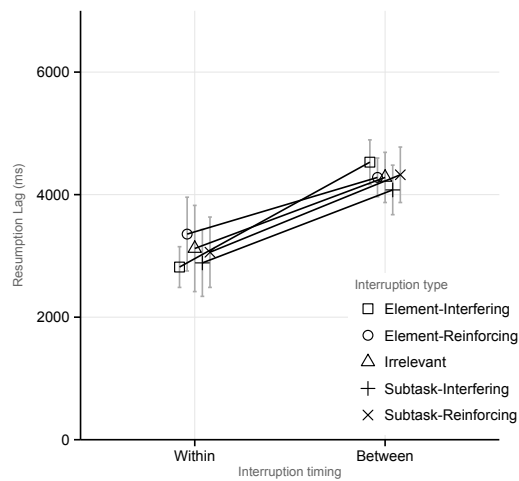
	Within	Between
Irrelevant	3121ms (*SD*=3299)	4281ms (*SD*=1920)
Element-reinforcing	3356ms (*SD*=2823)	4207ms (*SD*=1485)
Element-interfering	2818ms (*SD*=1559)	4540ms (*SD*=1706)
Subtask-reinforcing	3060ms (*SD*=2693)	4324ms (*SD*=2123)
Subtask-interfering	2886ms (*SD*=2563)	4077ms (*SD*=1896)

Figure 5.6 and Table 5.11 show aggregate resumption lags for Experiment 5. There is a significant main effect of interruption timing on resumption lag; this interaction plot suffices for illustrating this effect. All times in milliseconds.

interruptions. The resumption lag data were analysed with a factorial repeated measures ANOVA. Participants generally resumed more quickly after within-subtask interruptions than after between-subtask interruptions; the test revealed a significant main effect of interruption timing on resumption lag ($F(1,21)=11.94$, $p<.01$, $\eta_P^2=.36$). After correcting for a violation of sphericity ($\varepsilon=.25$, $p<.01$) there was no significant main effect of relevance ($F(2.3,48.3)=0.58$, $p=.59$, $\eta_P^2=.03$). After correcting for a violation of sphericity ($\varepsilon=.23$, $p<.01$) there was no significant interaction ($F(2.2,45.9)=0.73$, $p=.50$, $\eta_P^2=.03$). These results show that after having indicated where they would resume, participants were faster to identify progress, encode target information and interact with the task after within-subtask interruptions than after between-subtask interruptions.

5.3.3.4 Resumption error frequency and distribution

Each participant completed twenty trials, two for each condition. As each trial was interrupted twice, this gave a total of 880 interruptions. Participants made resumption errors on 16 occasions (2%). Interruptions arriving between subtasks resulted in resumption errors on seven occasions (2%). For the other 440 interruptions that arrived in the middle of a subtask, a total of nine resulted in resumption errors (2%).

Within-subtask resumption errors could be further broken down. When resuming after a within-subtask interruption, on no occasion did a participant select the wrong subtask. The wrong subtask element was chosen after a within-subtask interruption after nine resumptions. In only one condition, between-subtask interruptions with element-interfering audits did the error rate reach the 5% level required for errors to be systematic (Byrne & Bovair, 1997). This suggests that errors were the result of lapses in concentration rather than systematic failures of resumption strategy. It was previously predicted that introducing resumption cues to the task would result in a reduced error rate.

	Within		Between
	Element	Subtask	Subtask
Irrelevant	2% (2)	0% (0)	0% (0)
Element-reinforcing	0% (0)	0% (0)	0% (0)
Element-interfering	2% (2)	0% (0)	5% (4)
Subtask-reinforcing	2% (2)	0% (0)	1% (1)
Subtask-interfering	3% (3)	0% (0)	2% (2)

Table 5.12: Rates of resumption errors by condition and error type. Figures in grey are frequency.

5.3.4 Discussion

The results of Experiment 5 show that participants were able to start working more quickly after within-subtask interruptions than they were after between-subtask resumptions. As there was no effect of memory-affecting relevance on resumption performance, this result is indicative of participants making use of the environment to resume the task.

Given that previous research has suggested that subtask boundaries might be the best place to interrupt people (Iqbal & Bailey, 2010) and that people will defer switching tasks until subtask boundaries (Janssen et al., 2012; Bogunovich & Salvucci, 2011; Salvucci & Bogunovich, 2010), why was resumption faster after within-subtask interruptions in Experiment 5? The simple answer is that people make tradeoffs when dealing with interruptions.

When working through a task, goals are created in working memory to represent information about a task and progress through it. The complexity of the representations stored in this 'problem-state' (Borst et al., 2010; Salvucci & Taatgen, 2010) directly contributes to the disruptiveness of an interruption and the strategies that people employ to mitigate the negative effects of interruptions.

After switching to an interrupting task, representations in memory decay and may be forgotten (Altmann & Trafton, 2002). As Salvucci (2010) points out, the more complex the representations stored in the problem-state, the more time consuming the process of re-encoding any forgotten information on resumption. Indeed, when a primary task is more complex it takes longer to resume after interruption (Iqbal & Bailey, 2005; Magrabi, Li, Day, & Coiera, 2010). People are sensitive to these re-encoding costs and modify their behaviour accordingly; thus, there is a tendency to avoid interruptions when there are currently representations in working memory (Salvucci & Bogunovich, 2010). When a task or subtask is completed, representations associated with that task are no longer of use and can be discarded (Bailey & Iqbal, 2008). Therefore, it is best to interrupt people between-subtasks when their problem-state is empty.

Of course, such an account presupposes that the complexity of information stored in the problem-state varies over time. While problem solving tasks (e.g., Hodgetts & Jones, 2006b) that might require complex representations are likely to cause large differences in the disruptiveness of within- and between-subtask interruptions, this does not necessarily follow for routine procedural tasks.

In the Pharmacy Task, the only thing that needed to be consistently represented throughout task execution was progress. In this version of the task, progress cues were added; this made even representations of progress redundant. The process of resumption becomes an identification task; participants must determine what they have done, what comes next and what information they need before continuing. In scenarios where information access costs are high, participants might commit this information to memory (Morgan et al., 2013). However, the cost of obtaining this information in the Pharmacy Task was minimal because it was displayed continuously in the middle of the screen.

The main effect of timing thus represents how long it took participants to work out what had been done, what needed to be done next and what information was required. After resuming following a within-subtask interruption, progress was highly visible; values that had already been entered into subtask elements were displayed. A value that had not

already been entered had to be fetched from the *Prescription Sheet* and entered. Given participants had already consulted the section of the *Prescription Sheet* for the target subtask, the process of identifying which values matched which labels would have been hastened. After between subtask interruptions, participants had to work out what the last completed subtask was, identify the next subtask and the relevant part of the *Prescription Sheet*, fetch the first value and enter it. As the results demonstrate, this process of starting a new subtask was more time consuming than picking up one that had already been started.

This novel finding has implications for the design of interruption management systems. While previous systems have focused on subtask boundaries as a good place to interrupt people (Iqbal & Bailey, 2010), Experiment 5 shows that it is important to account for the costs of re-encoding primary task information on resumption. In tasks where representations are complex and re-encoding costs are high, then subtask boundaries make an ideal moment for interruption because representations are no longer required. Conversely, when representations are trivial or redundant, re-encoding costs are small and the effect of subtask familiarity may make within-subtask interruptions less disruptive compared to between-subtask interruptions.

The reduction or elimination of resumption errors with the introduction of place-keeping cues was expected. Previous work has shown that the introduction of cues to routine tasks can eliminate (Chung & Byrne, 2008) or reduce (Jones et al., 2012) errors. Although a resumption strategy that is reliant on memory can improve performance (Morgan et al., 2013, 2009), accurate and up-to-date place-keeping artefacts in the environment are more reliable. This finding echoes previous work that has shown that designing tasks to increase the physical representation of information boosts multitasking performance (Borst et al., 2013).

The difficulty of resuming without cues was illustrated by the poor performance on the choice task. The error rate was too low to imply participants were guessing and the eight-second lockout for incorrect choices would have discouraged fast guesses anyway. Rather, the results are indicative of the difficulties participants had in showing where they intended to resume without the aid of cues or the *Prescription Sheet*.

Experiment 5 has provided further confirmation of the importance of working memory effects on the disruptiveness of interruptions. It shows that when cues are available for resumption, and re-encoding costs are small, within-subtask interruptions can be less

disruptive than between-subtask interruptions. As well as modelling the complexity of interrupting tasks, interruption management systems should also model the complexity of the interrupted task. The complexity of the tasks will determine which moments are appropriate for interruption and which are not.

5.4 Summary

This chapter comprised two experiments. The first experiment, Experiment 4, investigated the extent to which relatedness and relevance represent different concepts. The experiment compared interrupting tasks that were related to the primary task with interruptions that had no relationship at all to the primary task. The results showed that interruptions that have nothing to do with the task at hand can be less disruptive than interruptions that are related to the task at hand but cause interference with place-keeping behaviour.

The results of this experiment suggest that the definition of 'relevant' used by previous investigations is insufficiently rigorous. If an interruption has content that is similar to that of the task at hand but is actually irrelevant, it has the potential to interfere with place-keeping. Entirely unrelated interruptions to do with a different task require a change in context, but are unlikely to actively interfere with representations related to the task at hand. This suggests that interruption management systems should not use the content of interruptions as a predictor of disruptiveness to primary tasks unless they can also determine the potential for an interruption to interfere with place-keeping.

The second experiment, Experiment 5 used the same design as Experiment 4 to investigate the process of resumption. It focused on how environment artefacts aid resumption after interruptions. The results of the experiment demonstrate that when resumption is cued and re-encoding costs are small, the timing of an interruption is less important than the operations that need to be performed before resumption.

Previous work has suggested that interruptions between subtasks are less disruptive. Experiment 5 shows that this is only the case when there are significant re-encoding costs associated with resumption. When re-encoding costs are low, within-subtask interruptions may be less disruptive because pre-existing familiarity with subtasks might ease resumption. Resuming between-subtasks instead makes it necessary to re-orient oneself to determine what needs to be done next.

The five experiments reported in this chapter and Chapter 4 demonstrate that memory is a critical factor in the resumption process. Of course, these experiments have been rigorously controlled; participants were interrupted in a lab setting without any choice. As Experiment 5 demonstrates, other factors, like the composition of a primary task, can influence resumption performance. It might be that in less controlled settings, the memory effects observed in the previous experiments are overwhelmed by other influences. There-

fore, the purpose of the next chapter is to assess how robust and reliable the previous findings are. To do this, Chapter 6 the effects of interruption relevance in timing in experimental setups where other factors have the potential to overwhelm the effects observed in this and previous experiments.

Chapter 6

How reliable are the effects of interruption relevance and timing on post-interruption resumption?

6.1   Outline

The previous five experiments presented in this thesis consistently show that when interruptions encourage rehearsal of place-keeping goals they are less disruptive to routine tasks. This suggests that in order to effectively aid work, interruption management systems must be cognisant of the effects of interruptions on memory. This chapter is concerned with the reliability of these findings.

The studies that have been conducted so far have been constrained in terms of the environments they were run in and the range of behaviour elicited from participants. They were run in controlled environments and participants had no control over when they were interrupted. Although these experiments have produced reliable results, they have been limited to using forced interruptions. In this interruption paradigm interruptions appear without interaction from a participant, and participants must complete interruptions before they can return to the main task. Forced interruptions are useful because they increase control over both participants and the concepts under scrutiny. They allow for small changes in, for example, interruption relevance or timing. Having a high level of control allows for the inference of causality between manipulations and performance.

In reality, interruptions make up just one part of a messy working environment. In complex settings, interruptions come from a number of sources (McGillis Hall et al., 2010) and

118

the process of resumption is not always as easy to map in the way that it is in the lab; people often do not resume their original task after an interruption (Westbrook, Coiera, et al., 2010; Grundgeiger et al., 2010). In the kinds of knowledge-working environments for which interruption management systems are designed, workers move between different working domains and have discretion over when they deal with interruptions (González & Mark, 2004; Mark et al., 2005).

With so many potential confounding factors for an interruption management system to concern itself with, is the relationship between working memory and interruption disruptiveness worth considering? It is possible that the effects observed previously in this thesis might appear as noise in the wider scheme of managing interruptions and tasks. The goal of this chapter is therefore to test the reliability of the previous results in this thesis in less controlled settings.

This chapter comprises two experiments. Experiment 6 investigates whether giving participants control over when they switch to interruptions alters the effects of relevance on performance. It does this by testing a version of the Pharmacy Task with discretionary rather than forced interruptions. Experiment 7 replicates Experiment 2 in an online setting. The aim of the study is to determine whether the effects previously observed are robust in environments that are uncontrolled. The question is whether the effects of interruption relevance that were previously observed still hold once other factors come into play. The extent to which the proposed changes to interruption management systems are practicable depends on how reliable these memory effects are when there are fewer constraints on the properties of interruptions.

6.2 Experiment 6: The effects of relevance on the management of deferrable interruptions

6.2.1 Motivation

The experiments that have so far been reported have all made use of a paradigm where participants are forced to stop working on the Pharmacy Task and switch their attention to an interruption. Participants have had no control over the timing of these interruptions. Of course, in real world situations people usually have some control over when they deal with an interruption — they can attend to them at a more convenient moment, depending on the nature of the alert and the properties of the interruption. Interruptions are often prompted by device alerts (McGillis Hall et al., 2010; Chisholm, Weaver, Whenmouth, & Giles, 2011) and notifications (Fischer et al., 2011) or because people interrupt themselves (Jin & Dabbish, 2009; Dabbish, Mark, & González, 2011).

How people manage interruptions when given the opportunity to defer them to a later time has been an area of active research. Salvucci and Bogunovich (2010) demonstrated that participants will defer interruptions that arrive during moments of high load until periods of low load. However, this behaviour is sensitive to the demands of tasks; participants will attend to interruptions more quickly if there is only a small window of time in which to deal with it (Bogunovich & Salvucci, 2011). Other research has shown that participants tend to switch tasks when they feel negatively about the task that they are working on (Adler & Benbunan-Fich, 2012) or because they are not making enough progress (Payne, Duggan, & Neth, 2007).

Despite these investigations, there have been no attempts to empirically examine whether and how people manage effects of interruption relevance when they have control over when they attend to interruptions. This thesis and previous studies show that people tend to find it easier to resume after relevant interruptions (Adamczyk & Bailey, 2004; Iqbal & Bailey, 2008).

Czerwinski et al. (2000) examined the effect of interruption relevance on post-interruption performance in a sub-task, but not the effect of interruption relevance on the decision to defer an interruption. Their experimental design required users to attend to the interruption before they could determine its relevance; consequently participants only realised if an interruption was relevant after they had switched to it.

Experiment 6 seeks to determine whether the effect of interruption relevance on disruptiveness are persistent even if participants have control over when they deal with interruptions. Are relevance effects minimised once participants have strategic control over when they deal with interruptions?

The study described here employs a novel experimental design in which the relevance of an interruption is revealed in the primary task through a cue. The decision to attend to the interruption is therefore an informed one. This is particularly important because people have the propensity to investigate 'mysterious' cues that are underspecified (Wainer, Dabbish, & Kraut, 2011).

The study compares how likely participants are to defer an interruption based on whether it is relevant or irrelevant. It also investigates how relevance affects resumption performance when participants have control over when they are interrupted. It is expected that participants will adapt their behaviour to fit the interruptions; they will be more likely to attend to relevant interruptions but that this behaviour with mitigate the effect of relevance on resumption lag; participants will defer irrelevant interruptions to moments when they are less disruptive.

6.2.2 Method

6.2.2.1 Participants

13 participants (9 female) with a mean age of 23 years (*SD*=4) took part in the study. Participants were drawn from the UCL participant pool and were paid £7 for approximately one hour of their time. None of the participants had previously taken part in any of the experiments described in this thesis.

6.2.2.2 Design

Experiment 6 used a one-way within-subjects design with counterbalancing. The independent variable was the relevance of an interruption and it had two levels; relevant and irrelevant. The primary dependent variable was resumption lag. Interruption lag, error rate and deferral rate were also measured.

6.2.2.3 Materials

Experiment 6 used a significantly modified version of the Pharmacy Task. Modifications were made to both the primary task and the interrupting tasks. These changes are outlined below.

Primary task   The version of the Pharmacy Task used in Experiment 6 was a significant variation on the task described in Chapter 3 and deployed in Chapters 4 and 5. Whilst the goal of the task — to copy information from a central area and into one of five subtasks — was the same, some of the steps required to achieve this goal were different.

One of the major changes in this version of the Pharmacy Task was the introduction of a selector panel that was a feature of the Doughnut Machine (Li et al., 2006, 2008) that provided the inspiration for the Pharmacy Task. The selector step required participants to indicate which subtask they were going to work on before they started working on it. After completing a subtask, participants were instructed to use a selector panel to 'unlock' the subtask they intend to work on next. If participants tried to work on a subtask without first having unlocked it with the selector, the action was deemed to be an error. Buttons on the selector panel were re-ordered in line with Li et al. (2006). This meant the buttons were not spatially congruent with the location of the subtasks on the screen.

As well as the introduction of the selector panel, changes were made to the method of input for each subtask. In previous experiments, numbers were typed in the same way for each subtask. Experiment 6 followed Li et al. (2006) in that entering values for subtask necessitated interacting with a variety of interface widgets. In the *Type* and *Label* subtask participants typed as usual. However, for the *Shape* subtask participants had to use incrementing arrows; for the *Colour* subtask participants had to use a slider; and for the *Packaging* subtask participants had to choose from a drop-down menu.

In addition to having a variety of entry interfaces, each subtask has selector steps for each subtask element. To enter values into a particular subtask element, regardless of interaction method participants had to select an adjacent radio button. If the radio button was not selected, the subtask element would not accept input.

Interrupting task   Interruptions used in the studies described previously in this thesis have typically been forced; during or at the end of a subtask an interruption would appear that covered the whole task interface. Participants had no choice over when they attended to an interruption. Experiment 6 used a different interruption paradigm — deferrable interruptions — with a different kind of interrupting task.

The intention of Experiment 6 was to examine whether having the option to defer an interruption would affect when participants attended to interruptions. With this in mind, interruptions to the primary task took the form of instant messages (IMs). IMs were viewed

in the same window as the primary task but were on a separate tab, meaning it was only possible to look at the primary task interface or the IM interface at any one time.

The interruption cue — i.e., the alert signifying the arrival of a new message — was the top section of the interface flashing in a colour corresponding to the sender of the message. The name of the sender was also displayed while a message was pending. Although the tabs occluded one another, participants were able to switch between tabs as many times as they wished. However, there were costs to switching; there was a time cost associated with having to manipulate the task interface to switch tabs and all subtask progress was lost on switching. Subtask progress was removed because it presented a strong cue for resumption.

Messages were cued for the duration of a subtask, and vanished when participants started the next subtask. This was to ensure that relevant interruptions could not receive attention when they were no longer relevant (because the participant had moved to the next subtask). Using a tabbed layout afforded similar switching behaviour to Mark et al.'s (2008) and Salvucci and Bogunovich's (2010) tasks without the technical complexity and potential confounds of having multiple overlapping windows.

The IMs in the relevant condition came from a 'pharmacy technician' who always asked a question relevant to the current subtask, for example: "How many tablets are you ordering?". In the irrelevant trials, IMs came from 'Phil', who asked participants questions about potential flat-shares. Participants were told to provide single word responses to the questions from both sources.
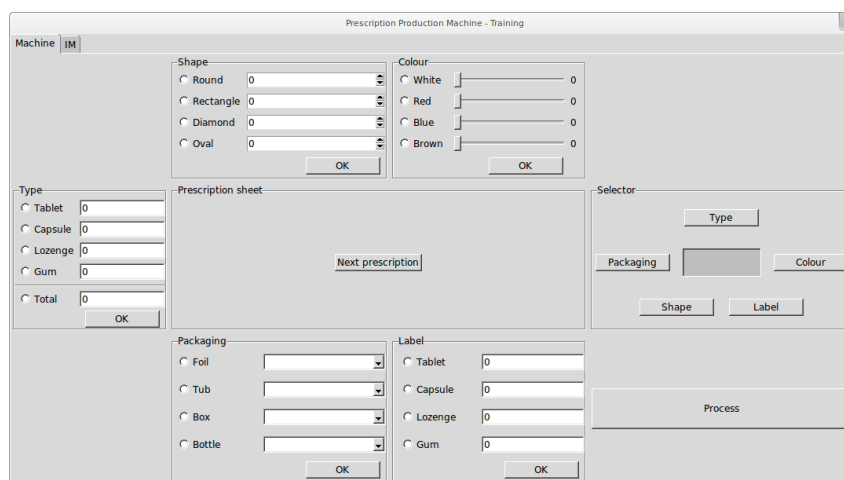

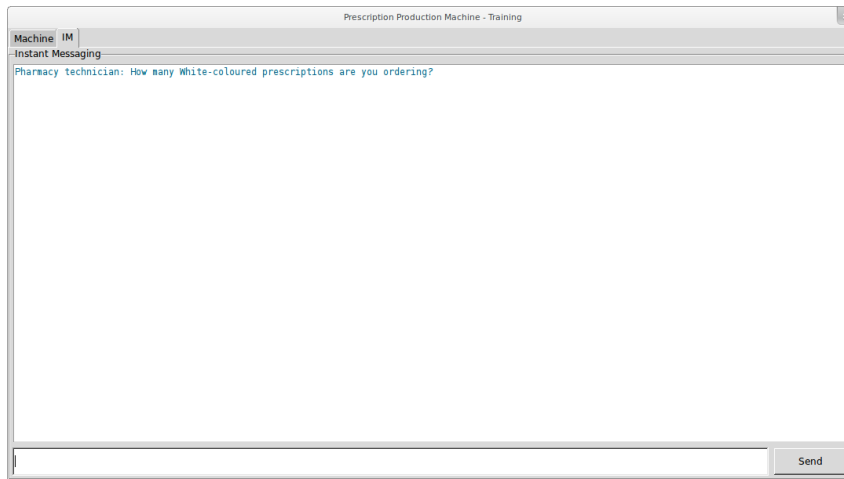**Figure 6.1:** A modified version of the Pharmacy Task was used as the primary task in Experiment 6.

Figure 6.2: In Experiment 6, the interrupting task was a simple IM window. Participants gave single-word responses to questions which either came from a technician (relevant) or from 'Phil' (irrelevant).

Participants had the opportunity to defer tasks indefinitely by ignoring the messages. No feedback was given to participant responses in the instant messaging window because it was important to maintain the notion that the interruptions were deferrable — that speed or accuracy of response was not a component of successful task completion. Progress in the primary task was monitored and participants were notified of mistakes and had to correct them before proceeding.

Post-experiment questionnaire   A six-item questionnaire was prepared for the purpose of gauging participants' perceptions of the interruptions they encountered. Participants were asked to rate how relevant, important and urgent they perceived the interruptions to be for both relevant and irrelevant interruptions. Responses were made on a five-point scale from 'Strongly disagree' to 'Strongly agree'. Participants were told explicitly to treat the two kinds of interruptions as being equally important; no mention was made of their relevance or urgency in the materials.

6.2.2.4   Procedure

The procedure for the study was very similar to that of the previous experiments. On arriving, participants were given an information sheet describing the task. After confirming they understood the information sheet and reading and signing a standard consent form, they were shown a demonstration trial.

After an opportunity for clarification, participants performed a minimum of three training trials, and continued to the experimental trial if two of the three training trials were performed without error. If participants failed to achieve two out of three successful trials, they were

required to perform an extra training trial. There were 12 experimental trials, with a break after completion of the sixth trial. After participants had completed the experimental trials, they completed the six-item questionnaire and were paid and debriefed.

There were two interruptions of the same kind per trial and these interruptions were cued at the first click of a widget inside subtasks two to five. All trials were interrupted: over a total of twelve trials, there were six trials with relevant interruptions and six trials with irrelevant interruptions for a total of twelve relevant and irrelevant interruptions per participant.

6.2.3 Results

One participant was unable to complete the training trials successfully and did not take part in the experiment. They were excluded from all subsequent analyses.

6.2.3.1 Interruption lag

Interruption lag refers to the period of time between an interruption being cued and work being started on it. In cases where participants switched instantly to the interrupting task, interruption lag measures the thinking time before switching. In cases where participants deferred interruptions, interruption lag measures the time it took participants to finish the subtask plus any thinking time. Here interruption lag is considered in both continuous and categorical sense. The continuous sense mere reflects the duration of time that elapses between the appearance of a cue and participants switching to an interruption. The categorical sense bins interruption lags into three categories: attend, where participants switch quickly to interruptions; defer, where participants perform other activities before switching; and ignore, were interruptions are not attended to.

It was predicted that relevant interruptions would result in lower interruption lag because they would have a lower switching cost. On average, participants were faster to attend when a relevant interruption was cued (M=4968ms, SD=3252) than they were when an irrelevant interruption was cued (M=5203ms, SD=3816). However, this difference was not significant; a Wilcoxon signed-rank test revealed that there was no effect of interruption relevance on interruption lag (z=1.10, p=.30, r=.22).

In addition to strategies that were common across the sample, individual deferral strategies could also be analysed through interruption lags. Participants who completed several steps after an interruption had been cued before switching — thereby deferring an interruption for longer — would have longer interruption lags. Average interruption lags ranged from 1594ms to 10599ms, with a mean of 5086ms (SD=3323ms). The data presented in

Figure 6.3 demonstrate there was significant individual variation in strategies; some participants switched very quickly, while others deferred interruptions for considerable periods of time.
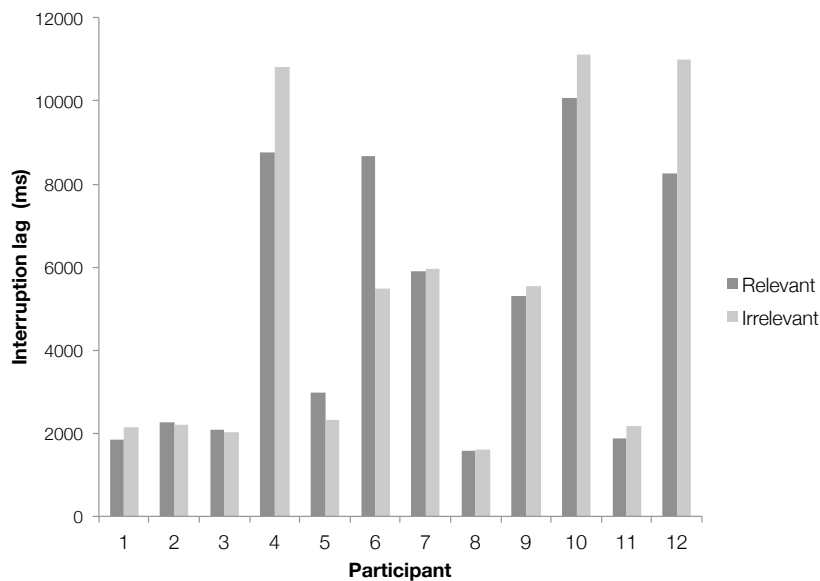

Figure 6.3: Average difference between an interruption being cued and it being attended to for each participant and by condition. The distribution of interruption lags across participants is bimodal. This suggests that participants adopted one of two discrete strategies; either to attend quickly to interruptions or to defer interruptions.

	Attend	Defer	Ignore
Relevant	53%	46%	1%
Irrelevant	54%	42 %	4%

Table 6.1: Interruption deferral strategies aggregated over all participants. Participants could either *attend* to an interruption immediately, *defer* it until later in the task, or *ignore* it entirely.

Responses to interruptions were categorised as attend, defer, or ignore. If participants attended, they would switch immediately on arrival of the interruption cue without any further primary task activity. If they deferred, they continued to interact with the primary task before switching. If participants ignored, they did not attend to the notification before it disappeared when the next subtask was selected.

Twelve participants took part in twelve trials and each trial had two interruptions; over the whole experiment, a total of 288 interruptions were presented. Across both relevant and irrelevant conditions, eight of the interruptions that were cued were ignored (3%). Of these eight interruptions, seven were ignored by a single participant. Six of the eight ignored interruptions were irrelevant.

Across all trials, participants utilised the 'attend' strategy and switched to the interruption with no further subtask interaction on 154 of the interruptions that were cued (53%, 78 irrelevant). The remaining 126 interruptions that were cued (44%) were deferred for some number of interactions after the arrival of the interruption cue.

Of the 126 interruptions that were deferred, 93 (48 irrelevant) were deferred until subtask boundaries. These boundaries occurred when the subtask had been completed but before the interruption disappeared when the next task was selected. Instances of participants deferring interruptions until a subtask boundary represent 73% of all deferred interruptions. The remaining 27% of deferred interruptions were attended to before subtask completion. Rates of attend and defer strategies were similar across conditions, but varied by participant. A by-participant analysis of deferral strategies is elucidated in the analysis of interruptions lags below.

6.2.3.2 Resumption lags

The resumption lag was the period of time it took participants to resume working on the primary task after returning from the instant messaging task. The period was defined as the moment of selecting the Pharmacy Task tab to interacting with an interface element on the primary task.

On average, participants were faster to resume after relevant interruptions (M=2311ms, SD=439) than they were after irrelevant interruptions (M=2610ms, SD=676). A Wilcoxon signed-rank test revealed that resumption lags differed significantly between relevant and irrelevant trials (z=2.04, p<.05, r=.42). This finding shows that even when participants were given a choice over when they dealt with interruptions, they were unable to mitigate the higher disruptiveness of irrelevant interruptions. However, caution is required when interpreting this effect because achieved power was low (.33).

6.2.3.3 Errors

Overall error rates were too low to perform any statistical analysis of their occurrence. In total, participants made 30 sequence errors. Thirteen of these errors occurred during twenty seconds of a single trial, indicating that the participant had lost their place in the task and had started to guess what they needed to do next. Consequently, all but the first of these errors were discounted, leaving 18 sequence errors for analysis.

Of these 18 errors, three occurred after switching back to the primary task. With an error-to-opportunity ratio of 1%, the resumption errors did not meet the 5% threshold of

systematicity, which is the recognised threshold in error research (Byrne & Bovair, 1997; Li et al., 2008). Of the 15 that did not follow an interruption, seven occurred when a participant either forgot to use the selector or forgot that they had already clicked the selector before starting the next (correct) subtask and so went back to the selector a second time. The other eight errors occurred through other incorrect manipulations of the task environment. These error rates are much lower than in the previous experiments presented in this thesis. This was an expected finding given the control participants had over the moment at which they were interrupted.

6.2.3.4 Participant perceptions of interruptions

The data collected through the six-item post-experiment questionnaire were analysed. The questions probed perceptions of interruption urgency, importance and relevance. To examine these three dimensions, three Wilcoxon signed-rank tests with compensation for ties were conducted on the results to analyse differences in responses. There was a significant difference in participants' perceptions of interruption importance ($z=2.23$, $p<.05$, $r=.46$), with messages from the pharmacist technician (i.e., relevant interruptions) being rated as more important. There were no significant differences in perceptions of urgency ($z=1.36$, $p=.22$, $r=.28$) or relevance ($z=1.99$, $p=.06$, $r=.41$) between interruption types.

6.2.4 Discussion

The results of Experiment 6 show that participants resume more quickly after relevant interruptions than irrelevant interruptions, even if they have control over when they attend to interruptions. This suggests that the working memory effects identified in Chapters 4 and 5 are robust even when other factors, like strategy, come into play. This finding is in line with the those of Czerwinski et al. (2000) and Iqbal and Bailey (2008), who also show that relevant interruptions are less disruptive than irrelevant interruptions when it comes to resuming a task after interruption.

Although participants resumed more quickly after relevant interruptions, relevance had no effect on how quickly they attended to an interruption. One possible explanation for the lack of effect of relevance on interruption lags could be that participants were told that both tasks were important to them, and the perception of importance meant that participants were not concerned with the relevance of the task. However, results of the post-experiment questionnaire would seem to rule out this explanation; despite being told that both kinds of interruption were important, participants rated interruptions from 'Phil' as being significantly less important than interruptions from a colleague.

Perhaps of greater concern was the fact that there was no difference in participants' ratings of perceived interruption relevance, despite this being the experimental manipulation. This shows a degree of dissonance as resumption data show that participants were faster to resume on relevant trials than irrelevant trials. It is difficult to come to any firm conclusions about the cause of the dissonance observed. The difference between participants' perceptions of the task and their actual performance is something that should be considered in future experiments. It may simply be a demand characteristic.

Experiment 6 elicited far lower error rates than previous experiments. Indeed, errors that occurred in the course of completing the experiment did not even reach the 5% level required for systematicity. Previous experiments in this thesis and investigations with the Doughnut Task have produced systematic levels of error using forced interruptions (e.g., Back et al., 2010; Brumby et al., 2013).

The simple explanation is that when given the freedom to defer interruptions, participants always had time to prepare themselves for an interruption before working on it. Trafton et al. (2003) showed that when participants were given warning of an impending interruption, they were able to prepare themselves and mitigate some of the disruptive effects. In Experiment 6 participants had complete control over the duration of the interruption lag; they had as long as they wanted to prepare and rehearse their progress.

Across all participants, there was an even split between attend and defer strategies for dealing with interruptions. Plotting individual interruption lags for each participant revealed that participants tended to choose one strategy over another; there were large individual differences in strategy selection.

That participants made different choices about how they dealt with interruptions is not surprising; the effect of individual differences on multitasking performance has been well documented. Several studies have shown that individual differences in spatial memory capacity predict resumption accuracy (e.g., Meys & Sanderson, 2013; Werner et al., 2011). A positive relationship between general intelligence and ability to deal with interruptions has also been demonstrated (Cades, McKnight, Kidd, King, & Boehm-Davis, 2010).

One explanation for the dichotomy in strategies observed in Experiment 6 is that participants were cognisant of limitations on their working memory capacity and adapted their strategy. Participants with poor working memory deferred interruptions, whereas participants with high working memory capacity were able to manage interruptions as they appeared. However, the plausibility of this theory is questionable in light of results from

Ophir, Nass, and Wagner (2009). Their results suggest that people do not always adapt their task-switching behaviour to match prevailing conditions and their limitations. In their study they found that those who were worst at multitasking were mostly likely to engage in it. This is perhaps not surprising given that frustration with a lack of progress can induce people to engage in distractions (Adler & Benbunan-Fich, 2012).

The individual differences in interruption management that are manifest in the result of Experiment 6 present something of a challenge for interruption management systems. They serve to further highlight that understanding interruptions management as task management is not sufficient. All people have limitations and predispositions that affect both their ability to deal with interruptions (e.g., working memory capacity) and the strategies they adopt (e.g., personality differences; Mark et al., 2008). Designers of interruption management systems need to be mindful of the potential for individual differences to radically affect the way that such systems are appropriated in use.

Overall, the results of Experiment 6 demonstrate that the effect of relevance on resumption lag might be reliable, even when participants have control over when they deal with interruptions. This is significant because to be useful an interruption management system must offer users the ability to defer interruptions until later. The results of this study show that even in a scenario where people can choose when they deal with interruptions, working memory effects still play a significant role in determining the disruptiveness of interruptions. This provides further evidence that interruption management systems could be more effective if they were to model the relationship between working memory and interrupted performance, although caution should be exercised in interpreting the results due to a small sample size.

Of course, this study was still laboratory-based; everyday life is often busy and uncontrolled. Previous work has demonstrated that transporting laboratory measures of interrupted performance into the wild is challenging (Grundgeiger et al., 2010). It is unclear whether the memory effects observed here would still be significant in real world scenarios or whether other factors that determine disruptiveness would 'crowd-out' memory effects. The next experiment in this thesis explores these effects outside of the lab in a less controlled setting.

6.3 Experiment 7: The effects of relevance and timing in an online environment

Part of this study and a subset of the results described here were published in Gould, Cox, Brumby, and Wiseman (2013) and presented at the AAAI Meeting on Human Computation & Crowdsourcing.

6.3.1 Motivation

The previous experiment demonstrated that even when participants have control over when they attend to interruptions, there was still an effect of interruption relevance on participants' resumption performance. Although this was a demonstration of the reliability of the effect, the experiment still took place in a laboratory.

How well do the effects observed translate to less controlled environments? Are the effects still reliable, or are they masked by confounding factors? These are critical questions when deciding what to model in an interruption management system. To better understand the robustness of these effects in the face of confounds, the experiment reported here was conducted online.

Online experimentation has become increasingly common in recent years as devices and connections have rapidly improved (e.g., Kittur, Chi, & Suh, 2008; Heer & Bostock, 2010; Sampath, Rajeshuni, Indurkhya, Karanam, & Dasgupta, 2013). Many researchers have questioned whether results from online studies are comparable to those derived in the lab, but a number of studies have demonstrated that data collected in the two contexts is usually statistically equivalent (e.g., Dandurand, Shultz, & Onishi, 2008; Paolacci, Chandler, & Ipeirotis, 2010; Germine et al., 2012; Komarov, Reinecke, & Gajos, 2013).

Although these studies have demonstrated that experimental data collected online are usually reliable, there are a number of confounding factors that could affect performance. For instance, Gould, Cox, and Brumby (2013a) showed that participants frequently interrupt themselves during online experiments for periods of time that would easily confound any experiment using response time data. Kapelner and Chandler (2010) have shown that participants engage in satisficing behaviour in online studies. Other work has shown that participants' attention during online studies is short and frequently wavers (Rzeszotarski, Chi, Paritosh, & Dai, 2013; Mao, Kamar, & Horvitz, 2013).

Given that prior work has shown that online experimentation is workable but not without uncertainties, it would seem to be an ideal venue for testing the robustness of the effects previously uncovered in this thesis. Although an online experiment is clearly no surrogate for real-world testing, it does have the potential to elicit many of the confounds that are characteristic of real-world settings. To provide a baseline for expected performance, Experiment 7 replicated Experiment 2 from Chapter 4 in an online setting.

As the experiment was already implemented in the browser, replicating Experiment 2 online provided an efficient way of replicating the experiment without a significant time investment. Due to the ease of running participants online, the experiment also offered the potential to run a higher-powered experiment. Experiment 2 demonstrated a significant effect of interruption relevance and a marginal effect of timing. Given this marginal effect it was decided that the experiment would be replicated with sample size computed using effect sizes from Experiment 2 as a guide.

The effect sizes observed in Experiment 2 differed, with a smaller effect of interruption timing (η_P^2=0.15) than there was for interruption relevance (η_P^2=0.18). The smaller effect of interruption timing was used to compute the required sample because it was this manipulation that produced the margin effect. Based on the η_P^2 value for the timing effect, a minimum sample size of 35 was computed.

It was expected that the effect of interruption relevance would be replicated; relevant interruptions would result in faster resumptions than irrelevant interruptions. It was also expected that, due to the use of within-subtask interruptions in Experiment 7, resumptions would be faster after within-subtask interruptions than between subtask interruptions.

6.3.2 Method

6.3.2.1 Participants

Seventy-two participants (42 female) with a mean age of 26 years (*SD*=8 years) took part in the study. Participants were solicited through the online study section of the UCL participant pool. Participant sign-up and participation were automated. Each participant was given a unique link when they signed up and a deadline to start the experiment. After completing the study they were paid £7 in Amazon vouchers for approximately one hour of their time.

6.3.2.2 Design

The experiment used the same 2×2 design as Experiment 2. There were two independent variables, interruption relevance and timing. Interruption relevance had two levels, relevant and irrelevant. Interruption timing also had two levels, within-subtask and between-subtask. Resumption lag, the time it took participants to resume after an interruption, was the primary measure. Other measures of participants' performance on the primary and interrupting tasks were also recorded.

6.3.2.3 Materials

The materials used were almost identical to those employed in Experiment 2, with the exception of some modifications to enable the experiment to be run online without an experimenter present. A set of instructions specifically for online participants was inserted before the general experimental instructions. These enumerated the browsers that could be used to participate in the experiment, instructions not to perform other tasks during participation, and an area for testing sound output before starting the experiment. In previous studies, debriefing was verbal. As this was infeasible for this study, a text-based debriefing was added to the end of the experiment.

Some background code changes were made to enhance the stability and usability of the experiment 'in the wild'. A scaling algorithm was added to ensure that the experiment was displayed properly regardless of the screen-size that participants were using. Participants' unique identifiers were passed from the recruitment to the experiment to make the tracking of participation simple and reliable.

6.3.2.4 Procedure

The experiment was listed in a section for online studies in the UCL psychology subject pool. Participants signed up for slots that were listed by latest time at which participation could begin. After this time the slot would expire. Participants could complete the study at any point between signing-up and the expiration of the slot.

When participants were ready to complete the study, they selected a link in the subject pool system which took them to an external website. Participants were given a set of general instructions relating to online participation, followed by a set of experiment-specific instructions accompanied by an instructional video.

The rest of the study proceeded in the same manner as Experiment 2; participants completed four training trials followed by twelve experimental trials. After completing the experimental trials, participants completed a post-experiment questionnaire.

On completing the study, participants followed a link to the subject pool administration system and their participation was automatically recorded. Amazon vouchers were emailed to participants soon after.

6.3.3 Results

Before the data were analysed, it was necessary to process participants' data. One of the challenges of running experiments online is that there is no opportunity to correct any misconceptions or allow participants to ask questions. Additionally, any technical issues that participants encountered during task execution were non-recoverable.

Of the 72 participants who took part in the experiment, six were excluded because they did not complete the study or because they used an incompatible browser and their data were not received. Data from a further three participants was discarded because there were strong indications that they had taken part in the experiment a second time under a duplicate pseudonymous accounts created in the subject pool system. This left 63 participants whose data could be analysed.

Each of these participants completed twelve trials, encountering six interruptions in each condition. Some participants struggled to correctly resume; thirteen participants were unable to manage two or more correct resumptions in one or more conditions. This suggested they had not fully understood how the task worked. These thirteen participants were dropped from all analyses. This left data from a total of 50 participants (29 female) for analysis.

6.3.3.1 Resumption lag

As in previous experiments, resumption lag was the primary measure of performance. Before analysing the resumption lag data, incorrect and outlying resumptions were removed from the data. Incorrect resumptions occurred either when participants tried to work on the wrong subtask or, after within-subtask interruptions when participants chose the wrong subtask element. Of a total of 1200 resumptions, there were 242 incorrect resumptions (20%). The distribution of these errors is analysed later.

After dropping incorrect resumptions from the sample, outliers were removed from the remaining 958 resumptions. Experiment 7 used the same criterion as Experiment 2: resumptions that were ±1.96 standard deviations from a participant's mean resumption time for a particular condition were removed. This accommodation of individual differences and variation across conditions meant that outlier removal was conservative: a total of 18 trials were removed from the analysis of resumption times (i.e., 2% of the sample). Of a total of 1200 resumptions (50 participants, 12 trials, 2 interruptions per trial), 242 were incorrect and 18 were outliers. This left 940 resumptions for analysis.

Resumptions were generally faster after within-subtask interruptions than after between-subtask interruptions. Resumptions were usually faster after relevant interruptions than after irrelevant interruptions. Table 6.2 shows the mean resumption times for each combination of interruptions.

A factorial repeated measures ANOVA was used to determine whether there were any significant differences in performance arising from the experimental manipulations. There was a significant main effect of interruption relevance on resumption performance ($F(1,49)=4.15$, $p<.05$, $\eta_P^2=.08$). There was also a significant main effect of interruption timing ($F(1,49)=12.67$, $p<.001$, $\eta_P^2=.20$). Achieved power was high for both relevance (.94) and timing (>.99). There was no interaction between the relevance and timing of interruptions ($F(1,49)=0.48$, $p=.49$, $\eta_P^2=.001$).

These results show that participants were generally faster to resume after within-subtask interruptions that were relevant to the task at hand. This supports the hypotheses for Experiment 7 and replicates the trends observed in Experiments 2 and 3.
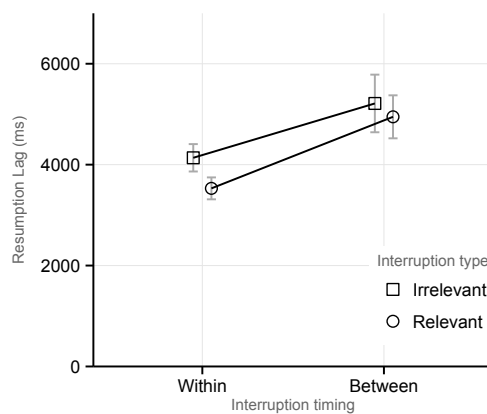
	Within	Between
Relevant	3530 (*SD*=1536)	4948 (*SD*=3013)
Irrelevant	4136 (*SD*=1919)	5213 (*SD*=4038)

Figure 6.4 and Table 6.2 show aggregate resumption lags for Experiment 7. All times in milliseconds.

6.3.3.2 Resumption error frequency and typology

The rates of different errors across conditions were also measured in Experiment 7. Participants resumed after an interruption a total of 1200 times. Of these 1200 resumptions, participants tried to resume on the wrong subtask or subtask element on 242 occasions (20%). These errors were broken down further by condition.

After resuming from a between-subtask interruption, participants selected the wrong subtask 126 times (21%). Fifty-eight of these errors were made resuming after relevant interruptions (19%), and the remaining 68 were made resuming after irrelevant interruptions (23%).

The other 116 resumption errors occurred after within-subtask interruptions. After within-subtask interruptions, participants could make errors in two ways. Either they could choose the wrong subtask or they could choose the correct subtask but the wrong subtask element. When resuming after a within-subtask interruption, participants chose the wrong subtask on seven occasions after relevant interruptions (2%) and 17 occasions after irrelevant interruptions (6%).

As well as selecting the wrong subtask, participants could choose the wrong subtask element when resuming after a within-subtask element. For instance, a participant might correctly identify that they needed to work on the *Colour* subtask on resumption, but might incorrectly try to work on the *Red* subtask element when they should have worked on the *Purple* subtask element. Participants chose the incorrect subtask element on 55 occasions when resuming after relevant within-subtask interruptions (18%). They made the same kind of errors on 47 occasions when resuming after irrelevant interruptions (16%).

Overall, the error rates in Experiment 7 paint a similar picture to those in previous experiments. In particular, Experiment 7 also showed that participants are more likely to choose the correct subtask when resuming after within-subtask interruptions than they are after between-subtask interruptions. A breakdown of resumption errors is shown in Table 6.3.

	Within		Between
	Element	Subtask	Subtask
Relevant	18% (55)	2% (7)	19% (58)
Irrelevant	16% (47)	6% (17)	23% (68)

Table 6.3: Rates of resumption errors by condition and error type. Figures in grey are frequency.

	Within	Between
Relevant	60%	52%
Irrelevant	63%	65%

Table 6.4: Audit response accuracy by condition. Expected rate produced by guessing would be 20%.

6.3.3.3 Interrupting task performance

Given the loss of control entailed by moving the experiment online, it was important to understand participants' performance on parts of the task which had no measures to prevent satisficing. Without an experimenter present, participants might have been more inclined to rush through whatever they could. One part of the task that might have been susceptible to satisficing was the audit tasks. Participants were not given feedback on their responses to audit questions. As such, there was no time-based disincentive to guessing as there was when resuming the primary task.

On any given audit, participants could choose from one of five options, so a 20% correctness rate would be indicative of guessing behaviour. Table 6.4 details audit accuracy for each condition. Participants were most accurate when answering irrelevant questions between subtasks (65%) and accuracy in all conditions suggests that although participants found the task difficult, they were not guessing.

6.3.4 Discussion

Experiment 7 was run entirely online and without experimenter supervision. The objective was to see how robust the effects observed in previous experiments were in a noisy online environment. The results revealed significant main effects of relevance and timing on performance. These effects were in the same direction as in Experiments 2, 3 and the relevant conditions in Experiments 4 and 5. This demonstrates that the effects previously observed are robust even when experimental control is significantly diminished.

Although the main effects were the same, the consequences of losing control manifested in a number of ways. Of the 72 participants that took part, 22 were discarded because they completed the experiment more than once, did not complete the task, or had difficulty understanding the instructions. Discarding 30% of the sample is a cause for concern, but ultimately strengthens the argument that the effects observed were robust despite a variety of ways in which participants could get side-tracked.

Some of the deviations in performance were also quite high: the standard deviation for resumptions after irrelevant between-subtask interruptions was over four seconds — on average, resumption only took around five seconds. One possible cause for this high variability is that participants were distracted by other things in their environment on their computers. Previous work by Gould, Cox, and Brumby (2013a) showed that when participants completed the Pharmacy Task online they interrupted themselves to perform non-experimental tasks 12 times on average and that these self-interruptions lasted for an average of 16 seconds.

The uncontrolled nature of the experiment might seem concerning at first. However, it is one of the most underutilised and underappreciated features of crowdsourcing and online experimentation. Online experiments allows for large numbers of participants to be run quickly and cheaply. This is one of the main selling points of online studies (Snow, O'Connor, Jurafsky, & Ng, 2008). That online experiments are fast and cheap is usually mitigation for the loss of control they entail. Indeed, much research effort has been expended on trying to understand how lost control can be compensated for either during (e.g., Kapelner & Chandler, 2010; Mason & Watts, 2010) or after (e.g., Rzeszotarski & Kittur, 2011; Snow et al., 2008) experiments. These efforts ignore the potential upsides of running experiments online; in particular the potential for such studies to investigate phenomena in more naturalistic settings.

Indeed, the purpose of running Experiment 7 online was to test the robustness of an effect. Rather than attempt to re-establish control in online settings, there is much to be gained by leveraging the loss of control that comes when moving experiments online to develop novel insights. In this sense it provides a useful halfway house between laboratory experiments and fully naturalistic studies of behaviour.

The results of Experiment 7 provide further evidence that the effects of memory on resumption performance are significant and robust. They show that the relevance of an interruption should be thought of as the extent to which an interruption reinforces or interferes with place-keeping goals. Relevant interruptions result in more rehearsal and faster resumptions. Irrelevant interruptions result in active interference and slower resumption. Further investigations with a wider variety of tasks would strengthen this argument, but the experiments presented in this chapter show that the effect of relevance on performance is reliable when people are given control over when they deal with interruptions and in environments where task performance might be confounded as a result of reduced control.

6.4 Summary

The objective of this chapter was to investigate the robustness of the memory-driven effects of interruption relevance that were revealed in previous chapters. To do this, two experiments were run, each of which introduced an element that could confound the effects observed in Chapters 4 and 5.

The first experiment in this chapter, Experiment 6, investigated how giving control over interruptions to participants would affect performance. In previous experiments, interruptions appeared spontaneously without participant control. In Experiment 6, participants were able to defer interruptions until a more convenient moment. The question was whether participants would exercise strategic control to minimise disruption to working memory. The results of the experiment showed that even with discretionary control, participants still resumed more quickly after relevant interruptions than they did after irrelevant interruptions.

Experiment 6 was still laboratory based; although it demonstrated that the effect of relevance on disruptiveness is robust when participants are given control over interruptions, the experiment itself was still highly controlled. Interruption management systems are designed to be deployed in complex, uncontrolled environments. Thus it made sense to investigate the robustness of the relevance effect in a less controlled setting. For this purpose, the Internet was chosen and an online study conducted.

Experiment 7 replicated the earlier Experiment 2 in an online environment. The question was whether moving to a less controlled environment would introduce other influences on performance such that the memory-driven effect of relevance on disruptiveness would be confounded. The experiment was run online in its entirety using the UCL participant pool. Although there were difficulties with data quality - training participants sufficiently was a challenge - the results were consistent with expectations. There were significant effects of relevance and timing on resumption performance. This suggested that even in environments with few constraints on behaviour, memory effects are still responsible for the disruptiveness of interruptions and need to be modelled by interruption management systems.

Taken together, the results of Experiments 6 and 7 demonstrate that the effect of relevance that was detected in previous chapters is reliable and robust. This is an important step towards the implementation of these ideas in models because it demonstrates that these

results contribute significantly to the disruptiveness of interruptions. Interruption management systems should try to anticipate the effects of interruptions on working memory because this determines task performance consistently and to a significant degree.

Chapter 7

General discussion

7.1 Outline

This chapter consists of three sections. The first section concisely reiterates the empirical findings of the experiments in this thesis. The second section is a synthesis of the empirical findings into a narrative account of the contribution to knowledge that this thesis makes. Finally, the third section enumerates several outstanding questions raised by the investigations and presents a prospective research agenda to answer them.

7.2 Empirical findings

The objective of this section is not to provide an exhaustive revisitation of the scores of results reported in this thesis. Nor does it attempt to contextualise the contribution of empirical results. Instead it simply aims to concisely reiterate the hypotheses and most important findings from each of the seven experiments in this thesis. In doing so, this section provides an easy-to-assimilate precis of the underlying logic for the experiments that were conducted. The contribution of these results and the questions that they raise are addressed separately in their own sections. The graphs attached to each section act as an aide-memoire for the original results.

Experiment 1

Experiment 1 investigated the relationship between interruption relevance and working memory. It was hypothesised that the relevance — and therefore disruptiveness — of interruptions would vary with the contents of working memory of task execution. To test this hypothesis, the relevance (relevant, irrelevant) and timing (within-subtask, between-subtask) of interruptions was manipulated. A significant interaction of relevance on timing on resumption lag was found. A simple effects analysis showed that relevance affected post-interruption resumption after between-subtask interruptions but not after within-subtask interruptions. Participants resumed more quickly after relevant between-subtask interruptions and more slowly after irrelevant between-subtask interruptions. Resumption speeds after within-subtask interruptions were slower than after relevant between-subtask interruptions but faster than after irrelevant between-subtask interruptions. This suggested that when participants were interrupted between subtasks they were using a different kind of place-keeping representation than when they were interrupted within subtasks.
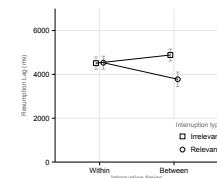
Experiment 2

Experiment 2 questioned whether the interaction effect observed in Experiment 1 was an artefact of the interrupting task used. It was hypothesised that replacing the interrupting task in Experiment 1 with an interrupting task that focused on within-subtask place-keeping would see a reversal of this effect. To test this, relevance (relevant, irrelevant) and timing (within-subask, between-subtask) were once again manipulated. The hypothesis was not supported: tests on the results showed a significant effect of interruption relevance on resumption lag. There were no other significant effects. Participants generally resumed faster after relevant than irrelevant interruptions. This result suggested that the hierarchical nature of subtasks and subtask elements contributes to interfering and reinforcing effects on the contents of working memory.
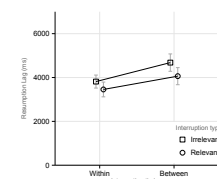
Experiment 3

Experiment 3 examined the role that the presentation of interrupting tasks plays in disruptiveness. It was hypothesised that negating the use of spatial memory in the completion of interrupting tasks would reduce the hierarchical effects of spatial organisation on resumption performance. This was not supported. Rather, the effects were the same as in Experiment 2; there was a significant effect of interruption relevance on post-interruption resumption lag. There was no main effect of
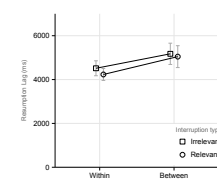
timing and no interaction. Participants resumed more quickly after relevant interruptions than after irrelevant interruptions. This result suggested that the presentation of interrupting tasks has little effect of the disruptiveness of interruptions as long as their effects on representations in working memory are consistent.

Experiment 4

Experiment 4 looked at whether the concepts of interruption relatedness and relevance need to be treated as independent. It was hypothesised that the extent to which an interruption was to do with the task at hand would have less of an effect on disruptiveness than the interfering and reinforcing effects associated with relevance. Interruption relevance (element-reinforcing, element-interfering, subtask-reinforcing, subtask-interfering, irrelevant) and timing (within-subtask, between-subtask) were once again manipulated. This hypothesis was supported. Tests on the results revealed a significant interaction between relevance and timing. The results showed that interruptions that are entirely unrelated to the task at hand can be less disruptive than interruptions that are related to the task at hand but irrelevant. This finding suggests that assuming an interruption is helpful simply because it is related to the task at hand is a poor strategy. The potential for active interference from related but irrelevant interruptions means that such interruptions could be unexpectedly disruptive.
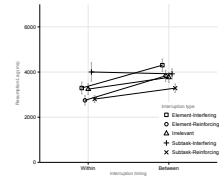
Experiment 5

Experiment 5 sought to understand whether the resumption results obtained in the previous four experiments were an artefact of the Pharmacy Task resources available to participants when resuming after interruption. It was hypothesised that a proportion of the resumption lag measured in previous studies was made up of pre-resumption operations. This hypothesis was supported. The experiment used the same variables with the same levels as Experiment 4. The resumption lags showed that for this task, participants performed fewer operations after within-subtask interruptions and so resumed more quickly after them than after between-subtask interruptions. This result suggested that previous accounts of subtask boundaries being 'good moments' for interruptions may not be applicable when working memory representations are simple and there is little encoding required.
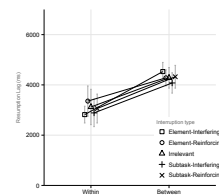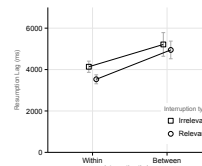
Experiment 6

Experiment 6 investigated whether the effects of relevance observed in previous experiments would persist when participants were given discretion over when they attended over interruptions. It was hypothesised that participants would defer interruptions until subtask boundaries and that this would mitigate the effect of relevance on interruption disruptiveness. This hypothesis only held true for some participants. Results of the experiment showed there were large variations in the strategies that participants used to deal with interruptions. Despite participants having the opportunity to adjust their strategy to minimise the negative effects of irrelevant interruptions, resumptions were still significantly faster after relevant interruptions. Although the small sample means that caution is required in interpreting the findings of this experiment, the results suggest that even with strategic control over interruption timing it is difficult to overcome the reinforcing effects of relevant interruptions.
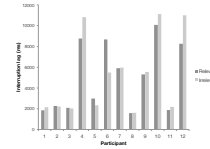
Experiment 7

Experiment 7 replicated Experiment 2 in an online environment. The objective was to determine whether the effects previously observed were robust in an environment with significant potential for confounds. It was hypothesised that these confounds would disrupt task execution, and resumption in particular. This hypothesis was not supported — there was little evidence to suggest participants behaved differently. Tests revealed significant main effects of relevance and timing on resumption lag. Participants generally resumed more quickly after within-subtask interruptions. Resumption was often quicker after relevant interruptions. The results of Experiment 7 demonstrate the reliability of the effects obtained through laboratory investigation.

7.3 Contribution to knowledge

Designing effective interruption management systems requires a clear understanding of the factors that influence the disruptiveness of interruptions. The goal of the experiments that comprise this thesis has been to understand how the relevance and timing of interruptions affects their disruptiveness. This section sets out the contribution to knowledge made by these experiments.

Each of the three empirical chapters makes a different contribution to knowledge. They are considered in turn here. Chapter 4 explores the relationship between interruption relevance and working memory. Chapter 5 demonstrates that the relevance and relatedness

of interruptions are distinct concepts with different effects on performance. Finally, Chapter 6 shows that the effects revealed in the other chapters are robust, even in environments where confounding factors are likely to influence performance.

7.3.1 Interruption relevance is a memory effect

Prior work has examined the relationship between the relevance of an interruption and its disruptiveness. Czerwinski et al. (2000) showed that people were disrupted less by interruptions that were relevant to what they were trying to achieve. Iqbal and Bailey (2008) also investigated the effect of relevance, finding that relevant interruptions were often less disruptive than interruptions of 'general interest'.

One of the limitations of this prior work is the lack of specificity given in the definition of relevance. As no firm definition of relevance is provided, no theoretical explanation for the role of relevance in disruptiveness can be tendered. Chapter 4 addresses these limitations by developing a theoretically-grounded account of the relationship between interruption relevance and disruptiveness.

Memory for Goals theory (Altmann & Trafton, 2002) has been successful in explaining how people resume after interruption; during the execution of tasks, goals ('control codes', Trafton et al., 2011) are created to represent task progress. These goals have a particular activation level that can be boosted through goal rehearsal. If a goal is not rehearsed it will lose activation until its recall becomes unreliable or impossible.

The interaction of various interruption properties with these goals is what makes particular interruptions more or less disruptive. For instance, interruptions that are more cognitively demanding leave less time for goal rehearsal and are thus more disruptive (Cades et al., 2007; Adler & Benbunan-Fich, 2014). Longer interruptions mean that goals are likely to decay further, also increasing disruptiveness (Li et al., 2006; Monk et al., 2008).

This thesis demonstrates that interruption relevance can also be viewed through the lens of interfering and reinforcing effects on the contents of working memory. An interruption that is relevant will require that current goals are rehearsed as part of its completion. An irrelevant interruption will prevent rehearsal of current goals or actively interfere with them. The relevance of an interruption is therefore contingent on its demands but also on the contents of working memory at the moment of interruption.

The experiments in Chapter 4 showed that relevance is the result of the interaction of an interruption with the contents of working memory. They also showed that as the content of working memory varies over the course of task execution, so the relevance of interruptions also varies.

This rigorous formulation of relevance is novel and stands in contrast to prior work which has used loose definitions of relevance. These definitions have also ignored the role that the contents of working memory plays in determining whether an interruption is relevant. Having a strong definition of relevance is an important step toward its incorporation into interruption management systems as a measure.

7.3.2 Interruption relatedness is not interruption relevance

Developing a strong working definition of interruption relevance invites questions about the definition of other concepts that might influence the disruptiveness of interruptions. One such question developed from the relevant-irrelevant dichotomy that was investigated in the first three experiments of this thesis.

In these experiments, interruptions were either relevant or irrelevant. However, both classes of interruption still asked participants questions that pertained to the task at hand — the Pharmacy Task. In reality, interruptions come from a wide variety of sources. Can interruptions that are related to the task at hand truly be irrelevant? Are all related interruptions relevant? Experiments 4 and 5 definitively answer these questions by demonstrating that relevance and relatedness must be treated as separate — though not entirely independent — factors.

Previous work had not considered the relationship between interruption relevance and relatedness. The aim of this thesis was to build on the definition of relevance that had already been developed to understand what it meant for an interruption to be related, rather than relevant. As the definition of relevance was based on the working memory described by Memory for Goals theory (Altmann & Trafton, 2002), it was important that relatedness was explored in the same manner.

The theory of relevance that was developed in this thesis holds that relevant interruptions reinforce working memory representations, whilst irrelevant interruptions interfere with these same representations. Experiment 4 explored how interruption relatedness fits into this model. The results presented in Chapter 5 showed that interruptions that are related to the task at hand but are irrelevant (i.e., cause interference) to the contents

of working memory have the potential to be more disruptive than interruptions that are entirely related to the task at hand. This is because interruptions that are related to the task at hand but are irrelevant have the potential to cause significantly more disruption to working memory representations because of proactive interference to place-keeping goals. The effect of similarity on disruption to working memory is well documented in the psychological literature (e.g., see Bunting, 2006).

This characterisation of relatedness and relevance stands in contrast to previous attempts to define the relationship between interruptions and tasks. It therefore has implications for the design of interruption management systems. Previous systems, such as the one developed by Arroyo and Selker (2011), judge whether interruptions are relevant by examining the content of an incoming interruption against the content of a primary task. If there is a high degree in the similarity of the content of both, then an interruption is deemed to be relevant to the task at hand. The results of the investigations presented in this thesis show that this approach does not adequately protect against the possibility of related but irrelevant interruptions causing significant disruption to place-keeping in working memory.

7.3.3   The role of memory in interruption disruptiveness is robust

The final contribution of this thesis is to demonstrate that the effects of working memory on interruption disruptiveness are robust. Two experiments at the end of the thesis show that the effects of relevance observed in the earlier experiments of this thesis occur in scenarios where other factors had the potential to confound them.

Experiment 6 shows that the effect of relevance persists even when participants are given discretionary control over when they deal with interruptions. Even when participants have the option to defer interruptions until more convenient moments, irrelevant interruptions remain more disruptive than relevant interruptions.

Experiment 7 shows that the effect of relevance is robust in an online setting. This suggests that the effects observed in previous studies are reliable even in the presence of factors that have the potential to drown out the effects of memory on disruptiveness.

These findings are significant because they provide evidence to support the contention that memory effects should be considered when developing interruption management systems. Previous interruption management systems (e.g., Iqbal & Bailey, 2010; Arroyo & Selker, 2011; Shrot, Rosenfeld, Golbeck, & Kraus, 2014) have been influenced by the psychological aspects of post-interruption resumption. However, these systems have

generally used behavioural measures to determine good and bad moments for interruptions. Although they are informed by psychological principles, the extent to which these measures provide quality proxies that can account for the working memory processes involved in dealing with interruptions is not clear.

An argument for using proxy measures might be that memory effects make up a small proportion of the process of resumption. Indeed, other researchers have suggested that the relative contribution of working memory to the time it takes to resumption after interruption might be small (Salvucci, 2010). A miscellany of factors contribute to the disruptiveness of interruptions; stress (Levy, Wobbrock, Kaszniak, & Ostergren, 2011), face-to-face interactions (Szóstek & Markopoulos, 2006) and an over-abundance of tasks (González & Mark, 2004), to list a few.

Given the effects of these factors, it is easy to see how one could argue against relying on memory effects to understand interrupted performance in real work. However, the results presented in this thesis suggest that models of memory might be valuable for interruption management systems in some environments. Settings where routine procedural tasks make up the bulk of activity would be most likely to benefit from a memory-focused intervention. In these environments memory load is likely to be lower and sources of interruption often more predictable. Previous work (Ratwani, McCurry, & Trafton, 2008; Trafton, Jacobs, & Harrison, 2012) has already shown using models of memory to predict performance can be practically useful. The empirical results from this thesis show that memory effects are still important, even as waning experimental control introduces confounds.

7.4   Prospective research agenda

This thesis makes a strong contribution to our knowledge of the effects of interruption timing and relevance on working memory during interrupted work. It also develops lines of questioning that cannot be fully addressed in the scope of the thesis.

This section explores four of the most significant issues raised by this thesis. The first issue that is pondered is the utility of resumption lag as a measure of interruption disruptiveness. Although resumption lag has been widely used as a measure of interruption disruptiveness, findings presented in this thesis suggest that there may be problems with resumption lag that need further attention.

The second issue that is considered is the effect of primary task structure on interruption disruptiveness. This thesis has mainly focused on how changing the properties of inter-

ruptions affects disruptiveness. Here the potential for changes to primary tasks to affect disruptiveness will be considered.

The third issue addressed in this section is understanding interruptions in online and crowdworking environments. Although this topic is investigated empirically in this thesis, there are potential contributions to be made in this area that go beyond replications of lab experiments.

The final issue considered in this prospective research agenda is the implementation of the findings that comprise this thesis in an interruption management system. Throughout this thesis, results have been discussed extensively in the context of interruption management systems. The focus here is on how the results from this thesis could practicably be deployed in an interruption management system.

7.4.1 Resumption lag as a measure of interruption disruptiveness

Post-interruption resumption lag is commonly used in the interruptions literature (e.g., Altmann & Trafton, 2004; Brumby et al., 2013; Hodgetts & Jones, 2006a; Trafton et al., 2003) as a measure of interruption disruptiveness. The assumption that underlies its use is that the longer it takes to resume after an interruption the more that goals in working memory must have decayed.

One of the problems with operationalizing disruptiveness by using resumption lag is that it also includes any re-encoding and re-orientation time associated with resumption (Salvucci, 2010). This means that the time it takes to resume after an interruption does not simply reflect the memory retrieval processes that Memory for Goals seeks to explain. A large proportion of the resumption process might be made up of looking to see if the state of a task has changed during an interruption, re-encoding any information that has been forgotten and encoding any new information that is required to make progress in the task.

This effect is apparent in resumption lags reported in each of the seven experiments that make up this thesis. In Experiment 1 a baseline measure of performance is developed to give some insight into how much of the resumption lags in the Pharmacy Task are due to activities that would have taken place even if there was no interruption. This baseline showed that the majority of the time costs incurred when resuming after interruptions in the Pharmacy Task would have been incurred anyway as part of task execution. The contribution of non-memory activities to resumption lag was explored systematically in Experiment 5. By disallowing the normal activities that would occur in task execution

149

immediately after resumption, the results showed that even accounting for re-orientation costs on resumption significant activity occurs before there is a measurable resumption signal.

The fact that resumption lag can be 'muddied' by activities that do not arise from the context switches forced by interruptions. This puts some limits the generalizability of the findings reported in this thesis. The additional activities that take place during the resumption lag are highly task bound. For instance, in the Pharmacy Task participants must retrieve attribute-value pairs from the Prescription Sheet before they can do anything after resumption. Each task will have its own set of post-resumption activities that must be executed after goals have been retrieved. Differences in these activities confound resumption lags making it more difficult to draw comparisons across studies and tasks. This thesis uses the same task throughout which maintains internal commensurability but limits broader generalizability.

To allow us to compare results across studies and tasks we need to have better ways of measuring resumption lag. Currently the literature is mostly clear on the retrieval, encoding and re-encoding activities that need to take place on resumption. What existing work does not do is provide practical ways of dissociating these factors in experimental measures. It would be useful, for instance, to be able to understand how long the memory retrieval processes take for a given task. Computational cognitive models provide one avenue for exploring resumption processes further but they are not always practical for the types of tasks and environments that are studied in Human-Computer Interaction and associated fields.

Tasks can be restructured to partition activities and this option was explored in Experiment 5 but this limits the external validity of tasks. It might be that inferential approaches using gaze detection or pupil dilation provide more direct access to the components of resumption but these approaches also come with technical and practical constraints. It is a challenge that will need to be overcome, however, if resumption lag is to be a measure of interruption disruptiveness that can be practically and usefully deployed in interruption management systems.

7.4.2   Primary task as mediator of interrupted performance

This thesis has largely focused on how manipulations to the properties of interruptions affect subsequent disruptiveness. In particular, there has been a focus on how the relevance and timing of interruptions affects their disruptiveness. Clearly this only represents

part of the picture; the disruption caused to a task by an interruption is partly contingent on the structure of the task itself.

As Salvucci (2010) points out, the process of resumption is not simply a recognition and recall task. On resumption, it is also necessary to re-encode any new information, decide what must be done next and encode any information that is required to make further progress. Understanding the role that the design of tasks plays in this process is critical to furthering our understanding of why interruptions are disruptive.

In this thesis, Experiment 5 presents an initial investigation of the contribution of a primary task to the process of resumption. The study breaks down resumption lag into discrete components; a recall component and a re-encoding and restarting component. The time it took participants to re-encode information and restart the primary task suggested that where participants were in the task at the moment of interruption significantly affected how quickly they could resume.

The results of Experiment 5 demonstrated that the characteristics of primary tasks can significantly influence the disruptiveness of interruptions as measured by resumption lag. However, the manipulations were still of interruption properties, rather than properties of the task. Experiment 5 did not investigate how variations in task interfaces can mediate interruption disruptiveness.

Prior work has taken a similar approach to the work presented in this thesis, having focused almost exclusively on the effects of various properties of interruptions on subsequent resumption. Researchers have investigated the effects of interruption duration (e.g., Monk et al., 2008), complexity (e.g., Magrabi et al., 2010), and relevance (e.g., Czerwinski et al., 2000). There has been far less focus on how the structure of tasks themselves determines resumption performance.

The fact that only one task, the Pharmacy Task, has been explored in the preceding experiments is one of the limitations of this thesis. There were good reasons for using one task consistently throughout this thesis. In particular, the use of a single task affords straightforward comparisons across the experiments that have been presented. For instance, resumption lags in Experiment 1 can be compared directly to those in 2. The limitation of using one task is that less is discovered about how the particular design and implementation of task affects its relationship with interruptions.

For instance, resumption lag is the product of both memory processes and post-resumption re-encoding processes. Any activities that occur post-resumption are determined in part by the design of the task. For example, the Pharmacy Task could be redesigned to increase the time it took to extract target information after resumption. This would increase the effective resumption lag. Likewise, what this thesis has discovered about the effects of interruption relevance and timing on performance is to some extent bounded by the particular task that was used. In Experiments 1, 2 and 7, interruptions are relevant when they relate to progress through the task and irrelevant when they test general knowledge about the operation of the task. The content of both types of interruption is necessarily determined by the structure of the task. Given both the findings of this thesis as well as its limitations, the question is how should work to understand the effects of primary tasks on interruption handling.

Some progress has been made in looking at how specific additions to tasks can support resumption. Studies that have used cues have shown that they significantly reduce the disruptiveness of interruptions (e.g., Jones et al., 2012; Hodgetts & Jones, 2006a). The focus of these studies has been on introducing cues specifically to aid post-interruption resumption; the focus has not been on how the design of the primary task can itself aid or inhibit resumption. Borst et al. (2013) have examined how tasks can support resumption by adding and removing elements that maintain place-keeping information. However, this does not constitute a systematic investigation of interface differences on their interruption-mitigating properties.

For instance, it might be possible to design tasks in ways that encourage or discourage task-switching behaviour. It has been suggested that people often switch tasks at sub-task boundaries (e.g., Bailey & Iqbal, 2008; Back, Cox, & Brumby, 2012). This suggests that it might be possible to 'nudge' people toward interrupting at particular moments by restructuring subtask boundaries. Likewise, changing the availability of information in a task can affect interruption-handling performance (Morgan et al., 2013).

Understanding whether the design of tasks can mitigate the negative effects of interruptions provides an interesting counterpoint to interruption management systems that have been a focus of development. These systems have taken an interruption-centric approach, capturing notifications and holding them until 'good moments'. Given the obstacles involved in developing systems that are sufficiently flexible and adaptive to handle complex environments, a better approach might be to shift some of the workload to those building tasks. Designers have a good understanding of the tasks that they design; with

appropriate training and knowledge, they are well-positioned to develop tasks and systems that are resilient to the negative effects of interruptions. Future work should aim to systematically investigate task-supported interruptions and to develop tools that support interruption-aware applications.

7.4.3 Understanding interruptions in online environments

Experiment 7 deployed the Pharmacy Task online in order to understand how a less controlled environment might affect relevance and timing. The results suggested that while the effects of relevance and timing observed in prior experiments were reliable, there are also several challenges with online experimentation that need to be overcome, particularly with participant training and data quality.

The challenges associated with developing and deploying interruptions in online environments present both empirical and methodological research opportunities. That is, the online settings provide the chance to learn something new about interruptions but also to better understand other phenomena through the effect of interruptions.

Online studies of interruption have the potential to tell us something about interruptions in a domain that has yet to be satisfactorily investigated. Interruptions and multitasking have been thoroughly studied in a number of environments: The effects of interruptions on office work (e.g., González & Mark, 2004), in hospitals (e.g., Chisholm et al., 2000), at steering wheels (e.g., Iqbal, Horvitz, Ju, & Mathews, 2011) and in cockpits (e.g., Loukopoulos et al., 2001) have all been extensively investigated.

Despite coverage of myriad domains, there has been little said about how interruptions might affect online work and almost no concrete research. Kittur et al. (2013) have argued strongly that online and crowdsourced work will only grow over time. Just as it has been necessary to understand traditional working environments to get the best out of people in the past, to get the best out of crowdworkers we will need to understand their working habits and environments much better than we currently do.

Understanding how interruptions affect working performance — and what can be done to mitigate any negative effects that accompany them — has been one of the key contributions of interruptions work in recent decades. It is essential that, as new ways of working develop, our knowledge of how interruptions affect performance remains applicable. Future studies should investigate what kinds of interruptions crowdworkers attend to and how this affects their performance.

As well as addressing purely empirical questions about interruptions during online crowd-work, online interruptions also offer the potential to make methodological improvements to crowdsourcing studies. Studies of interruptions can be improved by using participant-initiated interruptions. The reliability of all crowdsourcing experiments could be improved by recording participants' self-interrupting behaviour.

Interruption studies can be improved by making use of interruptions that naturally occur during the completion of online experiments. Previous work has shown that during online experiments, participants frequently interrupt themselves to do other things like contribute to social networks or make phone calls (Gould, Cox, & Brumby, 2013a).

Experiment-generated interruptions are usually conceived with a particular effect in mind that may or may not reflect the kinds of interruptions people normally deal with. Participant-initiated interruptions are, de facto, representative of the kinds of interruptions that people deal with in reality. The ecological validity of online studies of interrupted behaviour would increase by letting participants generate their own interruption; if participants generate their own interruptions, those interruptions are necessarily closer to replicating natural behaviour. In this way, the loss of control that accompanies online experiments can be turned to a researcher's advantage.

In addition to the benefits to studies of interruption produced by participant-generated interruption, keeping track of interruptions during online experiments might also make a useful covariate in all online experiments. Concerns over the quality of data produced in online studies have motivated researchers to develop preventative (e.g., Kapelner & Chandler, 2010) and remedial (e.g., Rzeszotarski & Kittur, 2011) approaches to rectifying issues with data quality. Rather than attempt to exert greater control, researchers could be more thorough with their measures. Recording task switching data would allow for some variability in performance to be accounted for, increasing the reliability of results.

7.4.4 Practical application

The final issue for future work to consider is the practical implementation of the ideas enumerated in this thesis. One of the themes of this thesis has been the relationship between the results presented and interruption management systems. Some of the results reported in this thesis suggest that existing approaches to the development of interruption management systems might need more refinement in order to be effective.

Previous attempts to develop interruption management systems have been based on strong empirical evidence. Iqbal and Bailey's *Oasis* system (2010) is based on extensive investigations of user behaviour in the face of interruptions (e.g., Iqbal & Bailey, 2007, 2008; Bailey & Iqbal, 2008). Their theoretically-informed and practically-driven exploration of interrupted behaviour led to the development of a system based on behavioural measures. Likewise, Shrot et al.'s *CRISP* system (2014) also makes use of behavioural measures of performance. These measures are aggregated across a large number of users and fed into a rule-based model that decides whether it is a good or bad moment to interrupt a user. Empirical testing of these systems, and others (e.g., Arroyo & Selker, 2011; McFarlane, 1999), has demonstrated that they are effective in identifying less disruptive moments for interruptions.

Despite the success of interruption management systems in controlled settings, their dependence on behavioural measures necessarily limits their utility in unusual or non-routine scenarios. Without supporting behavioural data, these systems are left reliant on heuristics. To develop systems that cope with the gamut of scenarios users might encounter, a model of users' cognition of interruptions might be required. There have not been any efforts to develop interruption management systems that use models of cognition to predict the disruptiveness of interruptions.

A large body of prior work suggests that modelling the mental phenomena surrounding interruptions and multitasking is feasible. Salvucci, Monk, and Trafton (2009) embedded a model of the encoding of place-keeping goals in a larger model of behaviour for a routine task. They were able to successfully model the changes in working memory over the course of task execution. In another example, researchers were able to model the process of resumption using the ACT-R cognitive architecture (Trafton et al., 2012). By accurately modelling the cognitive processes involved in post-interruption resumption, the researchers were able to design a robot that could predict whether people could remember where they were in a story after an interruption. These and other findings show that cognitive models can be put to practical use in realtime systems in a way that benefits users.

The challenge then is to integrate models of cognition into interruption management systems in a way that makes them useful for end users. An open question is whether it is actually necessary to model cognition to achieve this goal or whether more refined behavioural measures are able to provide an adequate proxy. More work is required to understand the potential benefits of cognition-led models of interrupted performance over models based

on behavioural metrics. Undertaking computational modelling might provide an effective way of thoroughly exploring this space before engaging in situated studies of behaviour.

7.5 Summary

The aim of this chapter was to elucidate the contribution to knowledge made by this thesis. First, it summarised the empirical findings of this thesis. Secondly, it established the contribution of the results obtained with regard to prior work. Finally, it sketched three potential directions for future research.

These sections make the argument that the results of this thesis represent a novel and important contribution to our knowledge of what makes interruptions disruptive. Of particular significance is the contribution of a clear definition of interruption relevance. Previous work has failed to provide such a definition, limiting the utility of findings.

There are number of theoretical and practical issues that need to be addressed before the findings in this thesis can be usefully employed in the design of interruption management systems. The goal of future work should be to determine whether and how these hurdles can be overcome.

References

Adamczyk, P. D., & Bailey, B. P. (2004). If not now, when?: the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 271–278). New York, NY, USA: ACM.

Adler, R. F., & Benbunan-Fich, R. (2012). The effects of positive and negative self-interruptions in discretionary multitasking. In *Proceedings of the 2012 ACM annual conference extended abstracts on human factors in computing systems extended abstracts* (pp. 1763–1768). New York, NY, USA: ACM.

Adler, R. F., & Benbunan-Fich, R. (2014, March). The effects of task difficulty and multitasking on performance. *Interacting with Computers*.

Altmann, E. M., Trafton, G. J., & Hambrick, D. Z. (2013). Momentary interruptions can derail the train of thought. *Journal of Experimental Psychology: General*, *143*(1), 215–226.

Altmann, E. M., & Trafton, J. G. (2002, January). Memory for goals: an activation-based model. *Cognitive Science*, *26*(1), 39–83.

Altmann, E. M., & Trafton, J. G. (2004). Task interruption: Resumption lag and the role of cues. In *Proceedings of the 26th annual conference of the cognitive science society* (p. 43-48). Mahwah, NJ, USA: Erlbaum.

Altmann, E. M., & Trafton, J. G. (2007, December). Timecourse of recovery from task interruption: Data and a model. *Psychonomic Bulletin & Review*, *14*(6), 1079–1084.

Ament, M. G., Cox, A. L., Blandford, A., & Brumby, D. P. (2010). Working memory load affects device-specific but not task-specific error rates. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 91–96). Austin, TX, USA: Cognitive Science Society.

Ament, M. G., Cox, A. L., Blandford, A., & Brumby, D. P. (2013). Making a task difficult: Evidence that device-oriented steps are effortful and error-prone. *Journal of Experimental Psychology*, *19*(3), 195–204.

Andrews, A., Ratwani, R. M., & Trafton, J. G. (2009). Recovering from interruptions: Does alert type matter? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(4), 409.

Anthony, K., Wiencek, C., Bauer, C., Daly, B., & Anthony, M. K. (2010, June). No interruptions please: Impact of a no interruption zone on medication safety in intensive care units. *Critical Care Nurse*, *30*(3), 21–29.

Arroyo, E., & Selker, T. (2011). Attention and intention goals can mediate disruption in human-computer interaction. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2011* (Vol. 6947, pp. 454–470). Berlin, Heidelberg: Springer Berlin Heidelberg.

Back, J., Brumby, D. P., & Cox, A. L. (2010). Locked-out: investigating the effectiveness of system lockouts to reduce errors in routine tasks. In *CHI '10 extended abstracts on human factors in computing systems* (pp. 3775–3780). New York, NY, USA: ACM.

Back, J., Cox, A., & Brumby, D. P. (2012). Choosing to interleave: human error and information access cost. In *Proceedings of the 2012 ACM annual conference on human factors in computing systems* (pp. 1651–1654). New York, NY, USA: ACM.

Bailey, B. P., & Iqbal, S. T. (2008, January). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction*, *14*, 1–28.

Bailey, B. P., & Konstan, J. A. (2006, July). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, *22*(4), 685–708.

Beckford, M. (2011, August). Nurses wear 'do not disturb' signs during drug rounds. *Telegraph.co.uk*.

Biron, A. D., Lavoie-Tremblay, M., & Loiselle, C. G. (2009, December). Characteristics of work interruptions during medication administration. *Journal of Nursing Scholarship*, *41*(4), 330–336.

Bogunovich, P., & Salvucci, D. D. (2011). The effects of time constraints on user behavior for deferrable interruptions. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3123–3126). New York, NY, USA: ACM.

Borst, J. P., Buwalda, T. A., van Rijn, H., & Taatgen, N. A. (2013, October). Avoiding the problem state bottleneck by strategic use of the environment. *Acta Psychologica*, *144*(2), 373–379.

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.

Botvinick, M. M., & Bylsma, L. M. (2005, December). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychonomic Bulletin & Review*, *12*(6), 1011–1017.

Botvinick, M. M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*(2), 395–429.

Brixey, J. J., Tang, Z., Robinson, D. J., Johnson, C. W., Johnson, T. R., Turley, J. P., … Zhang, J. (2008, April). Interruptions in a level one trauma center: A case study. *International Journal of Medical Informatics*, *77*(4), 235–241.

Brumby, D. P., Cox, A. L., Back, J., & Gould, S. J. J. (2013). Recovering from an interruption: Investigating speed-accuracy trade-offs in task resumption behavior. *Journal of Experimental Psychology: Applied*, *19*(2), 95–107.

Brumby, D. P., Salvucci, D. D., & Howes, A. (2009). Focus on driving: how cognitive constraints shape the adaptation of strategy when dialing while driving. In *Proceedings of the 27th international conference on human factors in computing systems* (pp. 1629–1638). New York, NY, USA: ACM.

Bunting, M. (2006). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology*, *32*(2), 183–196.

Byrne, M. D., & Bovair, S. (1997, January). A working memory model of a common procedural error. *Cognitive Science*, *21*(1), 31–61.

Cades, D. M., Boehm-Davis, D. A., Trafton, J. G., & Monk, C. A. (2011). Mitigating disruptive effects of interruptions through training: What needs to be practiced? *Journal of Experimental Psychology: Applied*, *17*(2), 97–109.

Cades, D. M., Davis, D. A. B., Trafton, J. G., & Monk, C. A. (2007, October). Does the difficulty of an interruption affect our ability to resume? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(4), 234–238.

Cades, D. M., McKnight, P. E., Kidd, D. G., King, E. B., & Boehm-Davis, D. A. (2010, September). Factors affecting interrupted task performance: Effects of adaptability, impulsivity and intelligence. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 458–462.

Carey, B. (2009, March). Liked the show? maybe it was the commercials. *The New York Times*.

Chisholm, C. D., Collison, E. K., Nelson, D. R., & Cordell, W. H. (2000). Emergency department workplace interruptions are emergency physicians "interrupt-driven" and "multitasking"? *Academic Emergency Medicine*, *7*(11), 1239–1243.

Chisholm, C. D., Weaver, C. S., Whenmouth, L., & Giles, B. (2011). A task analysis of emergency physician activities in academic and community settings. *Annals of Emergency Medicine*, *58*(2), 117–122.

Chung, P. H., & Byrne, M. D. (2008, April). Cue effectiveness in mitigating postcompletion errors in a routine procedural task. *International Journal of Human-Computer Studies*, *66*(4), 217–232.

Colligan, L., Guerlain, S., Steck, S. E., & Hoke, T. R. (2012, August). Designing for distractions: a human factors approach to decreasing interruptions at a centralised medication station. *BMJ Quality & Safety*, *21*(11), 939–947.

Committee on Quality of Health Care in America, & Institute of Medicine. (2000). *To err is human: Building a safer health system* (L. T. Kohn, J. M. Corrigan, & M. S. Donaldson, Eds.). National Academies Press.

Cutrell, E. B., Czerwinski, M., & Horvitz, E. (2000). Effects of instant messaging interruptions on computing tasks. In *HI '00 extended abstracts on human factors in computing systems* (pp. 99–100). The Hague, The Netherlands: ACM.

Czerwinski, M., Cutrell, E. B., & Horvitz, E. (2000). Instant messaging: Effects of relevance and timing. In *British computer society* (Vol. 2, pp. 71–76).

Dabbish, L., Mark, G., & González, V. M. (2011). Why do I keep interrupting myself?: environment, habit and self-interruption. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3127–3130). New York, NY, USA: ACM.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008, May). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434.

Duggan, G. B., Johnson, H., & Sørli, P. (2013). Interleaving tasks to improve performance: Users maximise the marginal rate of return. *International Journal of Human-Computer Studies*, *71*(5), 533–550.

Fischer, J. E., Greenhalgh, C., & Benford, S. (2011). Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* (pp. 181–190). New York, NY, USA: ACM.

Fordyce, J., Blank, F., Pekow, P., Smithline, H., Ritter, G., Gehlbach, S., … Henneman, P. (2003). Errors in a busy emergency department. *Annals of Emergency Medicine*, *42*(3), 324–333.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012, October). Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857.

Gillie, T., & Broadbent, D. (1989, April). What makes interruptions disruptive? a study of length, similarity, and complexity. *Psychological Research*, *50*(4), 243–250.

González, V. M., & Mark, G. (2004). "Constant, constant, multi-tasking craziness": managing multiple working spheres. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 113–120). New York, NY, USA: ACM.

Gould, S. J. J., Brumby, D. P., & Cox, A. L. (2013, September). What does it mean for an interruption to be relevant? an investigation of relevance as a memory effect. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 149–153.

Gould, S. J. J., Cox, A. L., & Brumby, D. P. (2013a). Frequency and duration of self-initiated task-switching in an online investigation of interrupted performance. In *Human computation and crowdsourcing: Works in progress and demonstration abstracts AAAI technical report CR-13-01* (pp. 22–23).

Gould, S. J. J., Cox, A. L., & Brumby, D. P. (2013b, January). Using graphical representations to support the calculation of infusion parameters. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2013* (pp. 721–728). Springer Berlin Heidelberg.

Gould, S. J. J., Cox, A. L., Brumby, D. P., & Wiseman, S. (2013). Assessing the viability of online interruption studies. In *Human computation and crowdsourcing: Works in progress and demonstration abstracts AAAI technical report CR-13-01* (pp. 24–25). AAAI.

Grandhi, S., & Jones, Q. (2010, May). Technology-mediated interruption management. *International Journal of Human-Computer Studies*, *68*(5), 288–306.

Gray, W. D. (2000, April). The nature and processing of errors in interactive behavior. *Cognitive Science*, *24*(2), 205–248.

Grundgeiger, T., & Sanderson, P. (2009). Interruptions in healthcare: Theoretical views. *International Journal of Medical Informatics*, *78*(5), 293–307.

Grundgeiger, T., Sanderson, P., MacDougall, H. G., & Venkatesh, B. (2010, December). Interruption management in the intensive care unit: Predicting resumption times and assessing distributed support. *Journal of Experimental Psychology: Applied*, *16*(4), 317–334.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 203–212). New York, NY, USA: ACM.

Hess, S. M., & Detweiler, M. C. (1994, October). Training to reduce the disruptive effects of interruptions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *38*(18), 1173–1177.

Hillel, G., & Vicente, K. J. (2003, October). Nursing interruptions in a post-anesthetic care unit: A field study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(12), 1443–1447.

Hiltz, K., Back, J., & Blandford, A. (2010, September). The roles of conceptual device models and user goals in avoiding device initialization errors. *Interacting with Computers*, *22*(5), 363–374.

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

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

Hodgetts, H. M., & Jones, D. M. (2007). Reminders, alerts and pop-ups: the cost of computer-initiated interruptions. In *Proceedings of the 12th international conference on human-computer interaction: interaction design and usability* (pp. 818–826). Berlin, Heidelberg: Springer-Verlag.

Horrey, W. J., & Lesch, M. F. (2009, January). Driver-initiated distractions: Examining strategic adaptation for in-vehicle task initiation. *Accident Analysis & Prevention*, *41*(1), 115–122.

Iqbal, S. T., & Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *CHI '05 extended abstracts on human factors in computing systems* (pp. 1489–1492). New York, NY, USA: ACM.

Iqbal, S. T., & Bailey, B. P. (2007). Understanding and developing models for detecting and differentiating breakpoints during interactive tasks. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 697–706). New York, NY, USA: ACM.

Iqbal, S. T., & Bailey, B. P. (2008). Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 93–102). New York, NY, USA: ACM.

Iqbal, S. T., & Bailey, B. P. (2010, December). Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *17*(4), 15:1–15:28.

Iqbal, S. T., Horvitz, E., Ju, Y.-C., & Mathews, E. (2011). Hang on a sec!: effects of proactive mediation of phone conversations while driving. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 463–472). New York, NY, USA: ACM.

Janssen, C. P., Brumby, D. P., & Garnett, R. (2012, March). Natural break points the influence of priorities and cognitive and motor cues on dual-task interleaving. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 5–29.

Jin, J., & Dabbish, L. A. (2009). Self-interruption on the computer: a typology of discretionary task interleaving. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1799–1808). New York, NY, USA: ACM.

Jones, S. A., Gould, S. J. J., & Cox, A. L. (2012). Snookered by an interruption?: use a cue. In *Proceedings of the 26th annual BCS interaction specialist group conference on people and computers* (pp. 251–256). Swinton, UK, UK: British Computer Society.

Kalisch, B. J., & Aebersold, M. (2010, March). Interruptions and multitasking in nursing care. *Joint Commission Journal on Quality and Patient Safety*, *36*, 126–132.

Kapelner, A., & Chandler, D. (2010). Preventing satisficing in online surveys. In *CrowdConf.*

Katidioti, I., & Taatgen, N. A. (2014, June). Choice in multitasking how delays in the primary task turn a rational into an irrational multitasker. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(4), 728-736.

Kern, D., Marshall, P., & Schmidt, A. (2010). Gazemarks: gaze-based visual placeholders to ease attention switching. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2093–2102). New York, NY, USA: ACM.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453–456). New York, NY, USA: ACM.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., … Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 1301–1318). New York, NY, USA: ACM.

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013). Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 207–216). New York, NY, USA: ACM.

Larkin, J. H., & Simon, H. A. (1987, March). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*(1), 65–100.

Latorella, K. A. (1998, October). Effects of modality on interrupted flight deck performance: Implications for data link. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(1), 87–91.

Levy, D. M., Wobbrock, J. O., Kaszniak, A. W., & Ostergren, M. (2011). Initial results from a study of the effects of meditation on multitasking performance. In *CHI '11 extended abstracts on human factors in computing systems* (pp. 2011–2016). New York, NY, USA: ACM.

Li, S. Y. W., Blandford, A., Cairns, P., & Young, R. M. (2008, December). The effect of interruptions on postcompletion and other procedural errors: An account based on the activation-based goal memory model. *Journal of Experimental Psychology: Applied*, *14*(4), 314–328.

Li, S. Y. W., Cox, A. L., Blandford, A., Cairns, P., & Abeles, A. (2006). Further investigations into post-completion error: the effects of interruption position and duration. In *Proceedings of the 28th annual meeting of the cognitive science conference* (pp. 471–476). Vancouver, BC, Canada: Cognitive Science Society.

Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2001). Cockpit interruptions and distractions: A line observation study. In *Proceedings of the 11th international symposium on aviation psychology.*

Magrabi, F., Li, S. Y. W., Day, R. O., & Coiera, E. (2010). Errors and electronic prescribing: a controlled laboratory study to examine task complexity and interruption effects. *Journal of the American Medical Informatics Association*, *17*(5), 575–583.

Mao, A., Kamar, E., & Horvitz, E. (2013, March). Why stop now? predicting worker engagement in online crowdsourcing. In *First AAAI conference on human computation and crowdsourcing* (pp. 103–111). AAAI.

Mark, G., Gonzalez, V. M., & Harris, J. (2005). No task left behind?: examining the nature of fragmented work. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 321–330). New York, NY, USA: ACM.

Mark, G., Gudith, D., & Klocke, U. (2008). The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 107–110). New York, NY, USA: ACM.

Mason, W., & Watts, D. J. (2010, May). Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, *11*(2), 100–108.

McFarlane, D. C. (1999). Coordinating the interruption of people in human-computer interaction. In A. Sasse & C. Johnson (Eds.), *Human-computer interaction – INTERACT 2013* (pp. 295–303). Springer Berlin Heidelberg.

McGillis Hall, L., Pedersen, C., & Fairley, L. (2010, April). Losing the moment: understanding interruptions to nurses' work. *The Journal of Nursing Administration*, *40*(4), 169–176.

Meys, H. L., & Sanderson, P. M. (2013, September). The effect of individual differences on how people handle interruptions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 868–872.

Monk, C. A. (2004, September). The effect of frequent versus infrequent interruptions on primary task resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(3), 295–299.

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

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

Morgan, P. L., & Patrick, J. (2010, September). Designing interfaces that encourage a more effortful cognitive strategy. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 408–412.

Morgan, P. L., & Patrick, J. (2013). Paying the price works: Increasing goal-state access cost improves problem solving and mitigates the effect of interruption. *The Quarterly Journal of Experimental Psychology*, *66*(1), 160–178.

Morgan, P. L., Patrick, J., & Tiley, L. (2013, January). Improving the effectiveness of an interruption lag by inducing a memory-based strategy. *Acta Psychologica*, *142*(1), 87–95.

Morgan, P. L., Patrick, J., Waldron, S. M., King, S. L., & Patrick, T. (2009, December). Improving memory after interruption: Exploiting soft constraints and manipulating information access cost. *Journal of Experimental Psychology: Applied*, *15*(4), 291–306.

Ophir, E., Nass, C., & Wagner, A. D. (2009, September). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, *106*(37), 15583–15587.

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

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010, August). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pape, T. M. (2003, April). Applying airline safety practices to medication administration. *Medsurg Nursing: Official Journal of the Academy of Medical-Surgical Nurses*, *12*(2), 77–93; quiz 94.

Parnin, C., & DeLine, R. (2010). Evaluating cues for resuming interrupted programming tasks. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 93–102). New York, NY, USA: ACM.

Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General*, *136*(3), 370–388.

Peebles, D., & Cheng, P. C.-H. (2001). Graph-based reasoning: From task analysis to cognitive explanation. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 762–767). Mahwah, NJ, USA: Erlbaum.

Ratwani, R. M., Andrews, A. E., Sousk, J. D., & Trafton, J. G. (2008). The effect of interruption modality on primary task resumption. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *52*, 393–397.

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2008). Predicting postcompletion errors using eye movements. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 539–542). New York, NY, USA: ACM.

Ratwani, R. M., & Trafton, J. G. (2006). *Now, where was I? examining the perceptual processes while resuming an interrupted task* (Tech. Rep.). George Mason University.

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

Ratwani, R. M., & Trafton, J. G. (2010, August). An eye movement analysis of the effect of interruption modality on primary task resumption. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 370–380.

Relihan, E., O'Brien, V., O'Hara, S., & Silke, B. (2010, October). The impact of a set of interventions to reduce interruptions and distractions to nurses during medication administration. *Quality and Safety in Health Care*, *19*(5), 1–6.

Rouncefield, M., Hughes, J. A., Rodden, T., & Viller, S. (1994). Working with "constant interruption": CSCW and the small office. In *Proceedings of the 1994 ACM conference on computer supported cooperative work* (pp. 275–286). New York, NY, USA: ACM.

Rzeszotarski, J. M., Chi, E., Paritosh, P., & Dai, P. (2013, March). Inserting micro-breaks into crowdsourcing workflows. In *Human computation and crowdsourcing: Works in progress and demonstration abstracts AAAI technical report CR-13-01* (pp. 62–63). AAAI.

Rzeszotarski, J. M., & Kittur, A. (2011). Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on user interface software and technology* (pp. 13–22). New York, NY, USA: ACM.

Salvucci, D. D. (2010). On reconstruction of task context after interruption. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 89–92). New York, NY, USA: ACM.

Salvucci, D. D., & Bogunovich, P. (2010). Multitasking and monotasking: The effects of mental workload on deferred task interruptions. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 85–88). New York, NY, USA: ACM.

Salvucci, D. D., Monk, C. A., & Trafton, J. G. (2009, October). A process-model account of task interruption and resumption: When does encoding of the problem state occur? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(12), 799–803.

Salvucci, D. D., & Taatgen, N. A. (2010). *The multitasking mind*. New York, NY, USA: Oxford University Press.

Sampath, H. A., Rajeshuni, R., Indurkhya, B., Karanam, S., & Dasgupta, K. (2013, March). Effect of task presentation on the performance of crowd workers — a cognitive study. In *Human computation and crowdsourcing: Works in progress and demonstration abstracts AAAI technical report CR-13-01* (pp. 64–65). AAAI.

Shrot, T., Rosenfeld, A., Golbeck, J., & Kraus, S. (2014). CRISP: An interruption management algorithm based on collaborative filtering. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3035–3044). New York, NY, USA: ACM.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263). Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Sykes, E. R. (2011, August). Interruptions in the workplace: A case study to reduce their effects. *International Journal of Information Management*, *31*(4), 385–394.

Szóstek, A. M., & Markopoulos, P. (2006). Factors defining face-to-face interruptions in the office environment. In *CHI '06 extended abstracts on human factors in computing systems* (pp. 1379–1384). New York, NY, USA: ACM.

Thompson, C. (2005, October). Meet the life hackers. *The New York Times*.

Trafton, J. G., Altmann, E. M., & Brock, D. P. (2005, September). Huh, what was I doing? how people use environmental cues after an interruption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(3), 468–472.

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

Trafton, J. G., Altmann, E. M., & Ratwani, R. M. (2011, June). A memory for goals model of sequence errors. *Cognitive Systems Research*, *12*(2), 134–143.

Trafton, J. G., Jacobs, A., & Harrison, A. (2012, March). Building and verifying a predictive model of interruption resumption. *Proceedings of the IEEE*, *100*(3), 648–659.

Tucker, A. L., & Spear, S. J. (2006). Operational failures and interruptions in hospital nursing. *Health Services Research*, *41*(3p1), 643–662.

Tugend, A. (2008, October). Multitasking can make you lose ... um ... focus. *The New York Times*.

Wainer, J., Dabbish, L., & Kraut, R. (2011). Should I open this email?: inbox-level cues, curiosity and attention to email. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3439–3448). New York, NY, USA: ACM.

Wearden, G. (2010, August). Multi-tasking media consumption on rise among britons, says ofcom study. *The Guardian*.

Werner, N. E., Cades, D. M., Boehm-Davis, D. A., Chang, J., Khan, H., & Thi, G. (2011, September). What makes us resilient to interruptions? understanding the role of individual differences in resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 296–300.

Westbrook, J. I., Coiera, E., Dunsmuir, W. T. M., Brown, B. M., Kelk, N., Paoloni, R., & Tran, C. (2010). The impact of interruptions on clinical task completion. *Quality and Safety in Health Care*, *19*(4), 284–289.

Westbrook, J. I., Woods, A., Rob, M. I., Dunsmuir, W. T. M., & Day, R. O. (2010, April). Association of interruptions with an increased risk and severity of medication administration errors. *Archives of Internal Medicine*, *170*(8), 683–690.

Appendix A

Materials

A.1   Experiment 1

Figure A.1 Participant information sheet

Figure A.2 Post-experiment questionnaire

Figure A.3 Screenshot of interface

Information Sheet

Thank you for participating in this study. We're interested in the way that people perform routine tasks. Participants have been randomly selected, and this study will involve approximately 15 participants in total.

Participation in this study is entirely voluntary. There is no penalty if you decide not to take part. You can withdraw from the study at any point without further explanation.

This experiment involves inputting data into a computer-based system.

There are a minimum of four training trials and 12 experimental trials in total. The experiment will last for around one hour.

There are five subtasks in the Pharmacy Task – Type, Shape, Colour, Packaging and Label. They need to be completed in this order.

Every now and again, the main task will disappear and will be replaced with a screen featuring a 'Progress Audit' or a 'Knowledge Audit'.

Each audit is composed of two audit tasks and two barcode tasks. Barcode tasks require you to copy a sequence of numbers from the list into the text boxes. Audit tasks require you to indicate something either about the subtasks in the main task (Knowledge Audits), or your current position in the main task (Progress Audits). At the end of the four audit subsections, you are returned to the main task where you continue with your activities.

If you make an error on the main task, the interface with lock for 8 seconds. After 8 seconds you will be able to continue with the task.

You will be paid £7 for your participation in this study.

There is one rest period in the experiment at the halfway point. You may rest for up to two minutes, or you may proceed at your own discretion.

After the experiment finishes, you will be asked to complete a six-item questionnaire and answer six free-response questions. You will then be given a debriefing and paid.

All data generated as a result of your participation in this study are confidential and not personally identifiable. Future research may utilise this dataset.

If you have any questions about this study (or would like to be informed of publications based on this study), please contact s.gould@cs.ucl.ac.uk

Thank you for taking part in this study; your time is much appreciated.

This study has been approved by the UCL Research Ethics Committee, number: Staff/1011/004

Figure A.1: Participant information sheet for Experiment 1

Post study questionnaire

Participant code: ……………………

To what extent do you agree with the statements:

"I felt that progress audits were important."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that knowledge audits were important."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that progress audits were urgent."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that knowledge audits were urgent."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that progress audits were relevant."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that knowledge audits were relevant."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

Questions to be discussed verbally:
I)   Did you find the task difficult? If so, what was difficult about it?
II)   Which of the two audit tasks did you find easiest?
III)   Do you think that you made more mistakes after one audit than after another?
IV)   Do you have any other comments?

Figure A.2: Post-experiment questionnaire for Experiment 1

Figure A.3: Screenshot of task interface for Experiment 1

A.2 Experiment 2

Figure A.4 Participant information

Figure A.5 Post-experiment questionnaire

Figure A.6 Screenshot of the Pharmacy Task

Figure A.7 Screenshot of irrelevant interrupting audit task

Figure A.8 Screenshot of transcription task

Instructions for participants

Thank you for participating in this study. Our main interest in this study is the way that people perform routine tasks. Participants have been randomly selected, and this study will involve around 24 participants in total.

Participation in this study is entirely voluntary. There is no penalty if you decide not to take part. You can withdraw from the study at any point without further explanation. The experiment comprises four training trials and twelve experimental trials. It will take approximately one hour to complete and you will be paid £7 for your time.

This experiment involves working through a computer-based data-entry task. Your job is to copy information from the prescription sheet into the five subtasks; Type, Shape, Colour, Packaging and Label.

Every now and again, the main task will disappear and will be replaced with an audit. Each audit is composed of two audit tasks and two transcription tasks. Transcription tasks require you to copy a sequence of numbers from the list into the text boxes. Audit tasks require you to answer questions about the current subtask or the one you had just completed. At the end of the four audit subsections, you are returned to the main task where you continue with your activities from exactly where you left off. If you choose the wrong subtask, you will incur an 8-second penalty. If you choose the correct subtask but choose a field you have already completed, you will forfeit your progress on the subtask and will have to enter the values again.

There are two rest periods in the experiment which come one third and two thirds of the way through the experimental trials. You may rest for up to one minute during each rest period, or you may proceed at your discretion. After the experiment finishes you will be asked to complete a few questions about the experiment. You will then be given a debriefing and paid.

All data generated as a result of your participation in this study are confidential and not personally identifiable. Future research may utilise this dataset. If you have any questions about this study (or would like to be informed of publications based on this study), please contact s.gould@cs.ucl.ac.uk

Thank you for taking part in this study; your time is much appreciated.This study has been approved by the UCL Research Ethics Committee, number: Staff/1011/004

Figure A.4: Participant information for Experiment 2

Please indicate the extent to which you agree with the statements below

Q1. The primary task was hard
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q2. The audit task was hard
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q3. The audits were harder when they came in the middle of a subtask
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q4. The audits were distracting
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q5. My goal was to be as accurate as possible
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q6. The audits were harder when they came after finishing one subtask but before starting the next
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q7. It was a struggle to remember what I had been doing in the primary task after finishing an audit
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q8. I had to change the way I did things because of the audits
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q9. I struggled to remember what I had been doing in the primary task after finishing an audit
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q10. I found the lockouts frustrating
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q11. My goal was to go as quickly as possible
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q12. I had a strategy for coping with the audits
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Figure A.5: Post-experiment questionnaire for Experiment 2

Figure A.6: Screenshot of Pharmacy Task used in 2
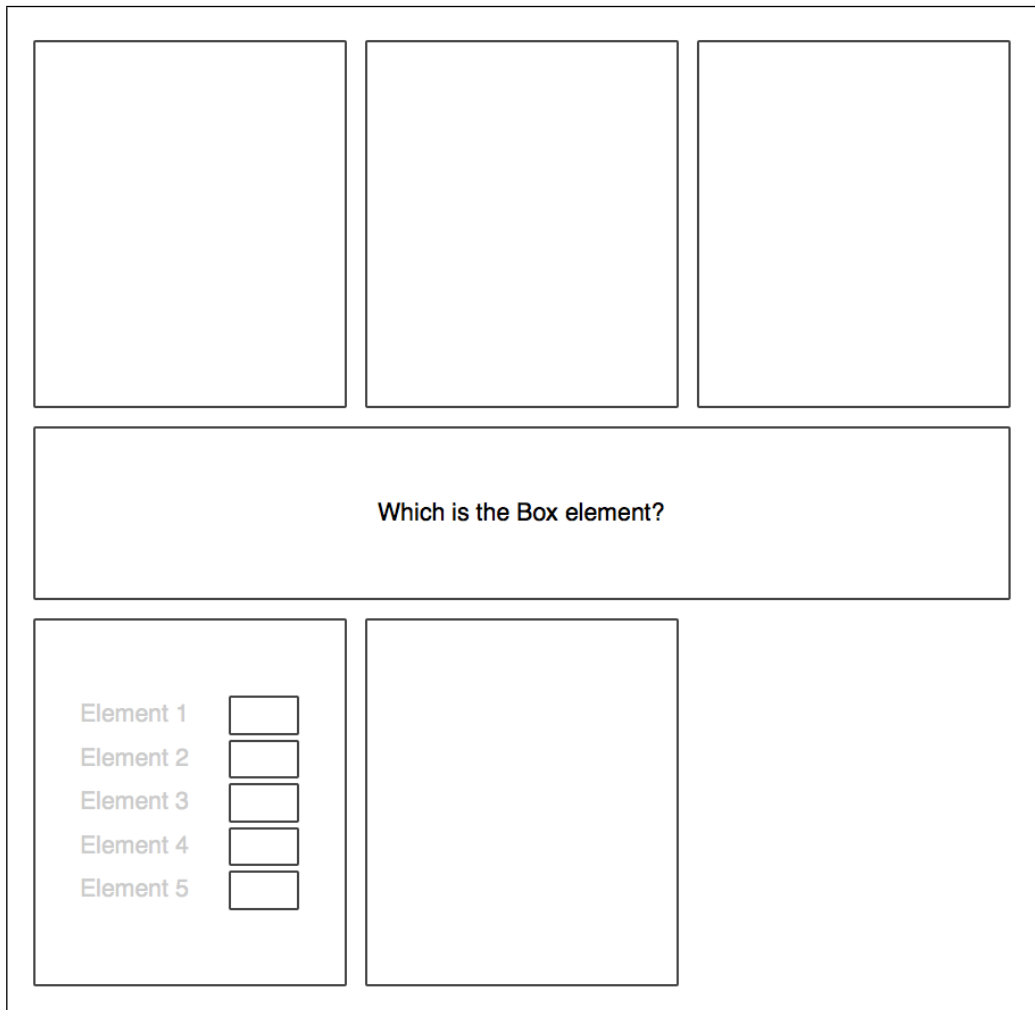
Figure A.7: Screenshot of irrelevant interrupting audit from Experiment 2

Figure A.8: Screenshot of an interrupting transcription task from Experiment 2

A.3   Experiment 3

Figure A.9 Participant information sheet

Figure A.10 Post-experiment questionnaire

Information Sheet

Thank you for participating in this study. We're interested in the way that people perform routine tasks. Participants have been randomly selected, and this study will involve 24 participants in total.

Participation in this study is entirely voluntary. There is no penalty if you decide not to take part. You can withdraw from the study at any point without further explanation.

This experiment involves inputting data into a computer-based system.

There are a minimum of four training trials and 12 experimental trials in total. The experiment will last for around one hour.

There are five subtasks in the Pharmacy Task – Type, Shape, Colour, Packaging and Label. They need to be completed in this order.

If you make an error on the main task, the interface with lock for 8 seconds. After 8 seconds you will be able to continue with the task.

Every now and again, the main task will disappear and will be replaced with an audit. Each audit is composed of two audit tasks and two barcode tasks. Barcode tasks require you to copy a sequence of numbers from the list into the text boxes. Audit tasks require you to answer questions about the current subtask or the one you had just completed. At the end of the four audit subsections, you are returned to the main task where you continue with your activities from where you left off. If you choose the wrong subtask, you will incur an 8-second penalty. If you choose the correct subtask but choose a field you have already completed, you will forfeit your progress on the subtask and will have to enter the values again.

You will be paid £7 for your participation in this study.

There is one rest period in the experiment at the halfway point. You may rest for up to two minutes, or you may proceed at your own discretion.

After the experiment finishes, you will be asked to complete a six-item questionnaire and answer six free-response questions. You will then be given a debriefing and paid.

All data generated as a result of your participation in this study are confidential and not personally identifiable. Future research may utilise this dataset.

If you have any questions about this study (or would like to be informed of publications based on this study), please contact s.gould@cs.ucl.ac.uk

Thank you for taking part in this study; your time is much appreciated.

This study has been approved by the UCL Research Ethics Committee, number: Staff/1011/004

Figure A.9: Participant information sheet for Experiment 3

Post study questionnaire

Participant code: ……………………

To what extent do you agree with the statements:

"I felt that the primary task was hard."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that the audits were hard."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that the audits were hard when they came in the middle of a subtask."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that the audits were hard when they came between two subtasks."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that the audits were distracting."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

"I felt that I had to change the way I did things because of the audits."
Strongly disagree   Disagree   Neither agree nor disagree   Agree Strongly Agree

Questions to be discussed verbally:
I)   Did you find the task difficult? If so, what was difficult about it?
II)   What did you think of the audit task? Did you find it difficult?
III)   Did you find that there were some occasions when the audits were more inconvenient than others?
IV)   Did you notice that sometimes the audits asked about what you were doing at the moment it switched?
V)   If yes, how do you think that this affected your performance when you resumed the main task?
VI)   Did you have any strategies to help you remember where you were in the task?

Figure A.10: Post-experiment questionnaire for Experiment 3

A.4 Experiment 4

Figure A.11 Participant information

Figure A.12 Post-experiment questionnaire

Figure A.13 Screenshot of irrelevant interrupting audit task

Instructions for participants

Thank you for participating in this study. Our main interest in this study is the way that people perform routine tasks. Participants have been randomly selected, and this study will involve around 36 participants in total.

Participation in this study is entirely voluntary. There is no penalty if you decide not to take part. You can withdraw from the study at any point without further explanation. The experiment comprises four training trials and twenty experimental trials. It will take approximately one and a half hours to complete and you will be paid £10 for your time.

This experiment involves working through a computer-based data-entry task. Your job is to copy information from the prescription sheet into the five subtasks; Type, Shape, Colour, Packaging and Label.

Every now and again, the main task will disappear and will be replaced with an audit. Each audit is composed of two audit tasks and three transcription tasks. Transcription tasks require you to copy a sequence of numbers from the list into the text boxes. Audit tasks require you to answer questions about the current subtask or the one you had just completed. At the end of the five audit subsections, you are returned to the main task where you continue with your activities from exactly where you left off. If you choose the wrong subtask, you will incur an 8-second penalty. If you choose the correct subtask but choose a field you have already completed, you will forfeit your progress on the subtask and will have to enter the values again.

There is one rest period which comes half way through the experimental trials. You must rest for at least one minute, after which you may rest for a further three minutes or you may proceed at your discretion. After the experiment finishes you will be asked to complete a few questions about the experiment. You will then be given a debriefing and paid.

All data generated as a result of your participation in this study are confidential and not personally identifiable. Future research may utilise this dataset. If you have any questions about this study (or would like to be informed of publications based on this study), please contact s.gould@cs.ucl.ac.uk

Thank you for taking part in this study; your time is much appreciated.This study has been approved by the UCL Research Ethics Committee, number: Staff/1011/004

Figure A.11: Participant information for Experiment 4

Please indicate the extent to which you agree with the statements below

Q1. The primary task was hard
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q2. The audit task was hard
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q3. The audits were harder when they came in the middle of a subtask
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q4. The audits were distracting
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q5. My goal was to be as accurate as possible
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q6. The audit questions about the names of subtasks and elements was hard
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q7. The audits were harder when they came after finishing one subtask but before starting the next
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q8. It was a struggle to remember what I had been doing in the primary task after finishing an audit
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q9. I had to change the way I did things because of the audits
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q10. The audit questions about the smallest and highest numbers were hard
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q11. I tried my best not to make any errors
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q12. Having to wait at the end of trial made me think harder on the audits
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q13. I struggled to remember what I had been doing in the primary task after finishing an audit
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q14. I found the lockouts frustrating
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q15. My goal was to go as quickly as possible
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q16. The audit questions about the task I was on were hard
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q17. I had a strategy for coping with the audits
○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Submit

Figure A.12: Post-experiment questionnaire for Experiment 4

Figure A.13: Screenshot of irrelevant interrupting audit from Experiment 4

A.5   Experiment 5

Figure A.14 Participant information about the experiment itself

Figure A.15 Post-experiment questionnaire

Figure A.16 Screenshot of pre-resumption choice screen

Figure A.17 Participant debriefing

Instructions for participants

Thank you for participating in this study. Our main interest in this study is the way that people perform routine tasks. Participants have been randomly selected, and this study will involve around approximately 35 participants in total.

Participation in this study is entirely voluntary. There is no penalty if you decide not to take part. You can withdraw from the study at any point without further explanation. The experiment comprises four training trials and twenty experimental trials. It will take approximately one and a half hours to complete and you will be paid £10 for your time.

This experiment involves working through a computer-based data-entry task. Your job is to copy information from the prescription sheet into the five subtasks; Type, Shape, Colour, Packaging and Label.

Every now and again, the main task will disappear and will be replaced with an audit. Each audit is composed of two audit tasks and three transcription tasks. Transcription tasks require you to copy a sequence of numbers from the list into the text boxes. Audit tasks require you to answer questions about the current subtask or the one you had just completed. At the end of the five audit subsections, you are asked to indicate where you'll resume the primary task. If you choose the wrong subtask or subtask element, you will incur an 8-second penalty. After this, you will be returned to the main task where you continue with your activities from where you left off.

There is one rest period which comes half way through the experimental trials. You must rest for at least one minute, after which you may rest for a further three minutes or you may proceed at your discretion. After the experiment finishes you will be asked to complete a few questions about the experiment. You will then be given a debriefing and paid.

All data generated as a result of your participation in this study are confidential and not personally identifiable. Future research may utilise this dataset. If you have any questions about this study (or would like to be informed of publications based on this study), please contact s.gould@cs.ucl.ac.uk

Thank you for taking part in this study; your time is much appreciated. This study has been approved by the UCL Research Ethics Committee, number: Staff/1011/004

Figure A.14: Participant information for Experiment 5

Please indicate the extent to which you agree with the statements below

Q1. The primary task was hard
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q2. The audit task was hard
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q3. The audits were harder when they came in the middle of a subtask
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q4. The audits were distracting
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q5. My goal was to be as accurate as possible
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q6. The audit questions about the names of subtasks and elements was hard
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q7. The audits were harder when they came after finishing one subtask but before starting the next
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q8. It was a struggle to remember what I had been doing in the primary task after finishing an audit
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q9. I had to change the way I did things because of the audits
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q10. The audit questions about the smallest and highest numbers were hard
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q11. I tried my best not to make any errors
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q12. Having to wait at the end of trial made me think harder on the audits
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q13. I struggled to remember what I had been doing in the primary task after finishing an audit
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q14. I found the lockouts frustrating
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q15. My goal was to go as quickly as possible
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q16. The audit questions about the task I was on were hard
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Q17. I had a strategy for coping with the audits
  ○ Strongly disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly agree

Submit

Figure A.15: Post-experiment questionnaire for Experiment 5

Figure A.16: Screenshot of pre-resumption choice screen from Experiment 5

Figure A.17: Participant debriefing from Experiment 5

A.6   Experiment 6

Figure A.18 Participant information sheet

Figure A.19 Post-experiment questionnaire

Figure A.20 Primary task screenshot

Figure A.21 Interrupting task screenshot

Information Sheet

Thank you for showing an interest in participating in this study. We're interested in the way that people perform routine tasks. Participants have been randomly selected, and this study will involve approximately 20 participants in total.

Participation in this study is entirely voluntary. There is no penalty if you decide not to take part. You can withdraw from the study at any point without further explanation.

This experiment involves practising prescription fulfilment using a computer-based system. In each experimental trial, two prescriptions arrive, and your task will be to fulfil the prescription by modifying the values at each of the five steps required to complete one prescription. There is a minimum of three training trials and twelve experimental trials in total. The experiment will last for around one hour. The experimenter will provide a demonstration of the system, and you will have a few trial runs before the experiment-proper starts.

From time-to-time during a trial, an instant message will arrive, signified by a flashing bar and the name of the sender of the instant message. Messages will come either from a fellow pharmacy technician, or from a friend. Message alerts continue for a short period until the message is viewed.

Responding to queries from the pharmacy technician is an important part of your job. Your friend Phil is searching for potential flat-shares for you, and what he has to say is important to you.

Responses to the instant messages should be kept to single words (8, Yes, No, 23, OK, etc)

There is one rest period in the experiment at the halfway point.

After the experiment concludes, you will be asked to complete a very short questionnaire and you will be given a debriefing.

All data generated as a result of your participation in this study are private and confidential. The data stored will not be personally identifiable. Depending on the results of the current study, further studies in the same area may draw this dataset.

If you have any questions about this study (or would like to be informed of publications based on this study), please contact s.gould@cs.ucl.ac.uk. If you have any concerns about the way that this study has been run please contact r.benedyk@ucl.ac.uk.

Thank you for taking part in this study; your time is much appreciated.

This study has been approved by the UCL Research Ethics Committee, number: Staff/1011/004

Figure A.18: Participant information sheet for Experiment 6

Post study questionnaire

To what extent do you agree with the statements:

"I felt that messages from the pharmacy technician were important."
Strongly disagree   Disagree   Neither agree nor disagree Agree Strongly Agree

"I felt that messages from Phil were important."
Strongly disagree   Disagree   Neither agree nor disagree Agree Strongly Agree

"I felt that messages from the pharmacy technician were urgent."
Strongly disagree   Disagree   Neither agree nor disagree Agree Strongly Agree

"I felt that messages from Phil were urgent."
Strongly disagree   Disagree   Neither agree nor disagree Agree Strongly Agree

"I felt that messages from the pharmacy technician were relevant."
Strongly disagree   Disagree   Neither agree nor disagree Agree Strongly Agree

"I felt that messages from Phil were relevant."
Strongly disagree   Disagree   Neither agree nor disagree Agree Strongly Agree

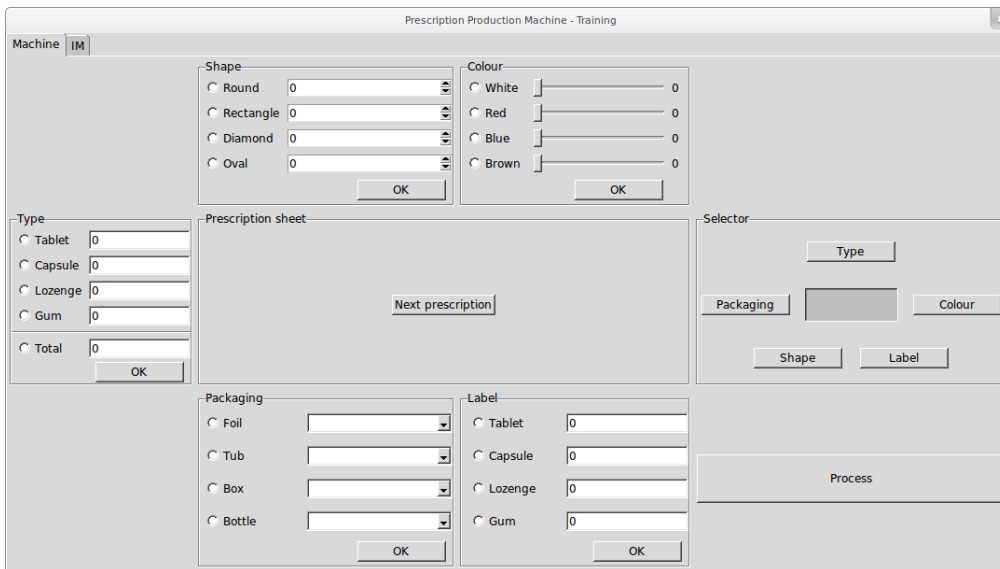**Figure A.19:** Post-experiment questionnaire for Experiment 6

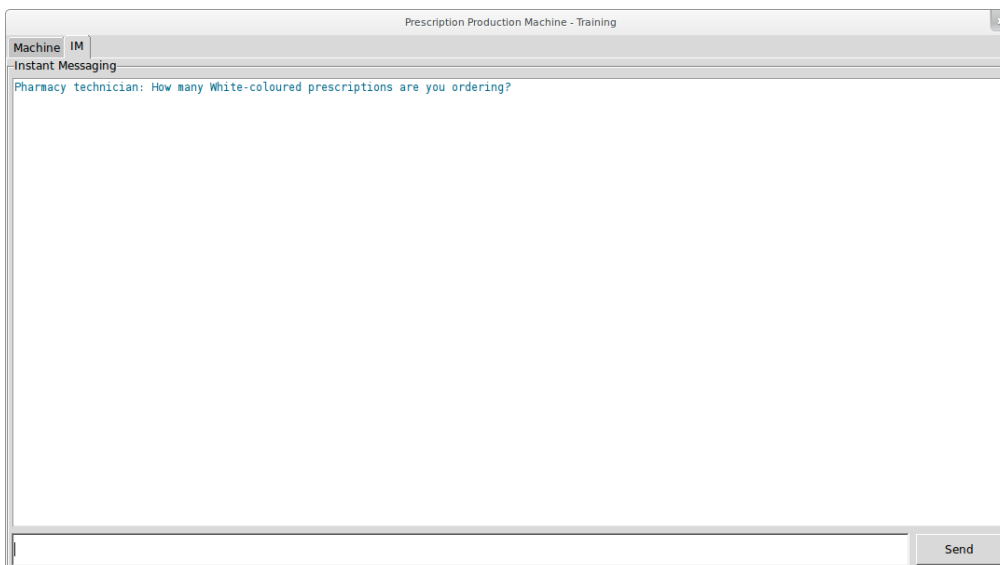**Figure A.20:** Screenshot of the Pharmacy Task interface for Experiment 6

Figure A.21: Screenshot of interrupting task for Experiment 6

A.7 Experiment 7

Figure A.22 Participant information about online experiments

Figure A.23 Participant information about the experiment itself

Figure A.24 Post-experiment questionnaire

Instructions for online participants

Please read this page carefully - it contains important instructions. Not following them may affect your payment at the end of the study.

This study is conducted entirely online, and we're not recording your screen. This means we're trusting you to do some things:

- Please do this study on a desktop computer with a keyboard and mouse or a keyboard and trackpad (i.e. a laptop)
- Be using an up-to-date web-browser, at least IE9+, Firefox, Chrome or Safari.
- Please do not do this study on a phone, tablet or other touch screen device
- Please make sure you are able to spend an hour undisturbed on the experiment
- There's sound in this study (see below) so make sure you have speakers or headphones
- Please don't do the study when you've been drinking, if you're particularly tired or if you're otherwise impaired.

There is sound involved in this study. You can check the volume of these sounds by clicking this button - please set your volume low, I don't want you to burst your ear drums.

[Test sound]

The most important thing is that you are able to set aside up to one hour to complete the study where you won't be disturbed by other people, whether in the same room as you, on your phone or via instant messaging. We're interested in how long different aspects of the task take so if you go and make yourself a cup of tea at some point, that will cause problems. The study needs to be completed all in one go. Don't worry if you feel you are working slowly - that's fine. All participants who complete the study will be paid £3 in Amazon gift vouchers. If you complete the study without wandering off to make dinner or take a phone call, you'll be paid a total of £7 of Amazon gift vouchers - hopefully this sounds fair to you.

[I understand. Proceed.]

Figure A.22: Participant information about online participation for Experiment 7

Instructions for participants

Thank you for participating in this study. Our main interest in this study is the way that people perform routine tasks. Participants have been randomly selected, and this study will involve around 24 participants in total.

Participation in this study is entirely voluntary. There is no penalty if you decide not to take part. You can withdraw from the study at any point without further explanation. The experiment comprises four training trials and twelve experimental trials. It will take approximately one hour to complete and you will be paid either £3 or £7 in Amazon vouchers depending on your performance (as explained previously - completing the study is worth £3, there is a £4 bonus for not wandering off in the middle of it!).

This experiment involves working through a computer-based data-entry task. Your job is to copy information from the prescription sheet in the middle of the screen into the five subtasks; Type, Shape, Colour, Packaging and Label. Once the values have been entered into a subtask, click the 'OK' button, this clears the subtask you were working on, and you can continue to the next one. If you make any typos you will be prompted to correct them before you can continue.

Every now and again, the main task will disappear and will be replaced with an audit. Each audit is composed of two audit tasks and two transcription tasks. Transcription tasks require you to copy a sequence of numbers from the list into the text boxes. Audit tasks require you to answer questions about the current subtask or the one you had just completed. At the end of the four audit subsections, you are returned to the main task where you continue with your activities from exactly where you left off. If you choose the wrong subtask, you will incur an 8-second penalty. If you choose the correct subtask but choose a field you have already completed, you will forfeit your progress on the subtask and will have to enter the values again.

There are two rest periods in the experiment which come one third and two thirds of the way through the experimental trials. You may rest for up to one minute during each rest period, or you may proceed at your discretion. After the experiment finishes you will be asked to complete a few questions about the experiment. You will then be given a debriefing.

All data generated as a result of your participation in this study are confidential and not personally identifiable. Future research may utilise this dataset. If you have any questions about this study (or would like to be informed of publications based on this study), please contact s.gould@cs.ucl.ac.uk

Thank you for taking part in this study; your time is much appreciated. This study has been approved by the UCL Research Ethics Committee, number: Staff/1011/004

Figure A.23: Participant information for Experiment 7

Please indicate the extent to which you agree with the statements below

Q1. The primary task was hard

Q2. The audit task was hard

Q3. The audits were harder when they came in the middle of a subtask

Q4. The audits were distracting

Q5. My goal was to be as accurate as possible

Q6. The audits were harder when they came after finishing one subtask but before starting the next

Q7. It was a struggle to remember what I had been doing in the primary task after finishing an audit

Q8. I had to change the way I did things because of the audits

Q9. I struggled to remember what I had been doing in the primary task after finishing an audit

Q10. I found the lockouts frustrating

Q11. My goal was to go as quickly as possible

Q12. I had a strategy for coping with the audits

Figure A.24: Post-experiment questionnaire for Experiment 7