In press: Consciousness and Cognition

Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability

Sam J. Gilbert

Institute of Cognitive Neuroscience, University College London, UK

Keywords: distributed cognition; intentions; prospective memory; internet; metacognition; confidence.

Running head: Offloading Intentions

Address correspondence to:

Sam Gilbert
Institute of Cognitive Neuroscience
17 Queen Square
London WC1N 3AR
UK
Email: sam.gilbert@ucl.ac.uk
Tel: +44 (0)20 7679 1149
Fax: +44 (0)20 7813 2835

Acknowledgement: SJG is supported by a Royal Society University Research Fellowship.

How do we decide whether to use external artifacts and reminders to remember delayed intentions, versus relying on unaided memory? Experiment 1 (N=400) showed that participants' choice to forgo reminders in an experimental task was independently predicted by subjective confidence and objective ability, even when the two measures were themselves uncorrelated. Use of reminders improved performance, explaining significant variance in intention fulfilment even after controlling for unaided ability. Experiment 2 (N=303) additionally investigated a pair of unrelated perceptual discrimination tasks, where the confidence and sensitivity of metacognitive judgments was decorrelated from objective performance using a staircase procedure. Participants with lower confidence in their perceptual judgments set more reminders in the delayed-intention task, even though confidence was unrelated to objective accuracy. However, memory confidence was a better predictor of reminder setting. Thus, propensity to set reminders was independently influenced by a) domain-general metacognitive confidence; b) task-specific confidence; and c) objective ability.

1. Introduction

Suppose you have decided to make a phone call at 2pm tomorrow. To ensure that you do so, you might make a note in a diary, or leave a reminder taped to a noticeable place. If you have a smartphone or digital watch, you might set an alarm for the appropriate time. These are all ways of 'offloading' intentions into the external environment. Alternatively, you might simply rely on your unaided memory. How will you decide which strategy to use?

The question of how we decide whether or not to offload intentions is of both practical and theoretical significance. It is of practical significance because external artifacts and reminders can be effective tools for remembering delayed intentions (Allen, 2002; Henry, Rendell, Phillips, Dunlop, & Kliegel, 2012; Heylighen & Vidal, 2008; Maylor, 1990). If we are to understand how intentions are fulfilled in everyday life it is therefore important to understand how we decide whether or not to set reminders, and what interventions might affect these decisions. This is likely to become increasingly important as smartphones and wearable technologies become more commonplace (Migo et al., 2014; Svoboda, Rowe, & Murphy, 2012). It also has practical relevance to compensation for memory difficulties in the context of ageing, disease, and brain injury (Fish, Wilson, & Manly, 2010; Thöne-Otto & Walther, 2008; Wilson, Emslie, Quirk, & Evans, 2001).

On a theoretical level, this question is important particularly due to its relationship with metacognition, i.e. the ability to represent and reason about our own mental states and processes (Metcalfe, 1996). In deciding whether or not to offload an intention, it seems plausible that we reflect on the difficulty of its fulfillment, and our unaided ability, before deciding whether to use external support. This assumption will be investigated

empirically below. Thus, investigating participants' decisions whether or not to offload intentions can help us to understand the relationship between metacognition, memory ability, and strategy use (Kvavilashvili & Ford, 2014; Meeks, Hicks, & Marsh, 2007; Meier, von Wartburg, Matter, Rothen, & Reber, 2011; Metcalfe, 2009; Rummel & Meiser, 2013; Schnitzspahn, Zeintl, Jäger, & Kliegel, 2011). While metacognition is often assumed to be an important factor for triggering strategic use of reminders for prospective memory (e.g. Arango-Muñoz, 2013; Knight, Harnett, & Titov, 2005; Phillips, Henry, & Martin, 2008), this assumption has received little empirical attention, due to the lack of experimental paradigms that allow participants' use of reminders to be systematically investigated.

Experimental paradigms for investigating prospective memory vary on a number of characteristics (McDaniel & Einstein, 2007). One dimension on which they vary is the way in which intentions are cued. In event-based tasks an intended behavior is cued when we notice a relevant stimulus (as in remembering to post a letter when we notice a mailbox). In time-based tasks an intended behavior is cued at a particular moment (as in remembering to attend a meeting at the appropriate time). Activity-based tasks, where an intention is triggered by the completion of a prior activity, are also possible. Another dimension on which paradigms vary is the delay between encoding an intention and acting on it. Outside the laboratory this can range from periods of seconds (e.g. going to the staff room to pick up one's mail, and intending to collect a milk bottle from the fridge on one's way back), to days, weeks, or longer (e.g. intending to attend a meeting scheduled months in advance). For practical reasons, laboratory studies typically use a seconds-to-minutes timeframe. A third dimension, at least in event-based tasks, is the salience of the cue to which an intention is attached; this relates to the distinction made by some theorists between "focal" and "non-focal" cues (McDaniel & Einstein, 2000).

Despite these variations in experimental paradigms, one characteristic shared by virtually all studies is that participants are prevented from creating external reminders. This is in sharp contrast to everyday life, in which such strategies are commonplace.

In order to examine cognitive processes involved in setting external reminders, Gilbert (in press) investigated a web-based "intention offloading" task, in which participants fulfilled an intention on each trial, after a brief filled delay (see also Landsiedel & Gilbert, 2015, for a neuroimaging investigation of this task). On each trial, participants sequentially removed a set of numbered circles from a box, by dragging each circle in turn to the bottom edge (Figure 1). They were instructed to remove target circles by dragging them to an alternative edge (e.g. "please drag 7 to the top instead"). Participants were given the option of setting an external reminder, by placing the target circles next to the intended edge at the start of the trial. When they did this, the intention was offloaded in the sense that it was now directly cued by the location of the target circle. Consequently, there was no need to mentally rehearse the intended behavior. An everyday analogy might be leaving an object in a noticeable position so that it cues an intended behavior, for example leaving something by the front door so that we remember to bring it with us when leaving the house.

Gilbert (in press) found that accuracy on this task predicted fulfillment of a real-world intention of visiting a specified weblink up to 7 days later to receive a small bonus payment, with greater predictive validity than more traditional time- and event-based prospective memory tasks. Thus, the task had significant external validity as a measure of fulfillment of delayed intentions (albeit with a weak effect size). Furthermore, when this task was matched in accuracy with a control task that did not permit external reminders but was otherwise identical, only the version permitting external reminders significantly

predicted real-world behavior. Investigating performance of the intention-offloading task revealed three main findings. 1) Intention offloading promoted intention fulfillment: permitting this strategy boosted performance, and participants who set more reminders fulfilled more intentions. 2) Participants were more likely to offload intentions when they had three targets to remember rather than one. Thus, the decision to offload intentions was influenced by memory load. 3) Participants were more likely to offload intentions when they encountered an interruption in the ongoing task. Thus, the decision to offload intentions was also influenced by the ongoing task in which the memory load was embedded. The latter two findings show that participants offload intentions adaptively based on the difficulty of the task, seeing as a higher memory load and the presence of interruption both reduced intention fulfillment in a matched version of the task that disallowed intention offloading.

The results outlined above show that task-specific factors influenced participants' decisions whether to set reminders. However, even when the task was held constant, there was considerable variation between participants in this behavior. The present study aims to explore the factors that might underlie such individual differences in intention offloading. Why do some participants tend to rely on unaided memory, while others are more likely to set reminders? Note that the experimental paradigm under investigation involves intentions delayed for brief periods on the order of 5-15 seconds. Of course, intentions in everyday life operate over a variety of timescales, from brief interruptions (e.g. delaying a pending task for a few seconds, during periods of high workload) to hours, days, or longer. Although Gilbert (in press) found that the present intention-offloading task predicted fulfillment of real-world intentions up to one week later, it is not claimed in the present article that the processes involved in fulfilling delayed intentions over such different time periods necessarily overlap. Rather, it is hoped that

studying the ways in which participants offload briefly-delayed intentions might serve as a useful first step to understanding metacognitive processes related to the use of reminders over a longer timescale.

2. Experiment 1

Participants performed the intention-offloading task described above in two phases. In phase one, they performed the unaided task, without being able to set reminders. Subsequently, in phase two, they were permitted to offload intentions, and their use of reminders was investigated. Before and after each phase, participants provided a subjective rating of how well they expected to perform, or how well they thought they had performed. This allowed participants' use of reminders to be related to a) objective performance measures and b) metacognitive judgments of their own ability.

Two hypotheses might be contrasted here. On the one hand, participants who assess their unaided ability to be poor, or those who do indeed perform poorly in the first phase of the task, might subsequently set reminders as a compensatory strategy. In this case, objective and/or subjective performance measures in phase one should correlate *negatively* with reminder setting in phase two. On the other hand, it might be that the very participants who already perform well on the task and/or have high confidence in their performance will be the ones who subsequently make the most use of reminders. This would be a 'rich get richer' scenario. It might occur because the most able participants would also have the metacognitive insight to realize the utility of an externalizing strategy, and hence make use of it. Alternatively, participants who are highly motivated to perform well, which should boost performance in phase one, might also make the most use of intention offloading in phase two. These hypotheses predict a *positive* relationship

between objective and/or subjective performance measures in phase one and externalizing behavior in phase two.

[Figure 1 about here]

2.1 Methods

2.1.1 Experiment 1a

The experiment followed a similar procedure to Gilbert (in press). Participants were recruited from the Amazon Mechanical Turk website (http://www.mturk.com), an online marketplace in which participants receive payments for completion of web-based tasks (Crump, McDonnell, & Gureckis, 2013). Ethical approval was received from the UCL Research Ethics Committee and informed consent was obtained from all participants. As in the earlier study by Gilbert (in press), participation was restricted to volunteers living in the USA, to reduce heterogeneity; furthermore, participants who had taken part in the earlier studies reported by Gilbert (in press) were blocked to ensure a fresh sample of participants.

Participants completed the task via their computer's web browser. On each trial, ten yellow circles numbered 1-10 were positioned randomly within a box (Figure 1). Participants were instructed to drag the circles in turn (1, 2, 3 etc.) to the bottom of the box, using their computer mouse. When a circle was dragged to the bottom of the box it disappeared, leaving the other circles on the screen. After the 10th circle had disappeared, the screen was cleared and the next trial began. For a demonstration, please visit "http://www.ucl.ac.uk/sam-gilbert/demos/circleDemo.html".

[Figure 1 about here]

Alongside this ongoing task, participants were provided with a delayed intention on each trial: they were instructed to remove one circle by dragging it to a specific alternative location (i.e. left, right, or top) when it was reached in the sequence. Thus, they formed delayed intentions to perform particular actions when they encountered pre-specified cues, although they could produce a standard ongoing response (i.e. dragging the circle to the bottom of the box) if they forgot. However, if they attempted to drag a target circle to an incorrect target location (left, top, or right), or if they attempted to drag a nontarget circle to any of these locations, it remained on the screen, allowing participants to realize that they had made a mistake.

In some conditions intention offloading was permitted, meaning that participants could drag the target circle toward its intended location at the beginning of the trial. By placing a circle next to the instructed edge, participants created a perceptual trigger that reminded them to remove that circle by dragging it to the left, right, or top of the box, instead of the bottom. This reminder then cued the appropriate behavior when the target circle was reached in the sequence. The *externalizing proportion* was defined as the proportion of target circles for which participants set an external reminder in this way (calculated in an identical manner to Gilbert, in press; accuracy measures were also calculated in an identical manner to the earlier study). In other conditions, intention offloading was not allowed. This was achieved by fixing the position of the circles on the screen so only the upcoming circle in the sequence could ever by dragged (i.e. circle 2 could only be moved after circle 1 had been dragged to the bottom of the screen, circle 3 could only be moved after circle 2 had been removed, etc).

The sequence of events for each participant, following their provision of consent, was as follows:

A) Phase 1 prediction: Following two practice trials of the experimental task, participants were told that the experiment was about to start and asked to provide a prediction of what percentage of targets they thought they would drag to the instructed locations. They provided this prediction by dragging a circle along a slider, which allowed them to select any value between 0% and 100%. The slider was initially placed at the 50% mark, and participants were only allowed to proceed to the next part of the task after they had clicked the circle to provide a response.

B) Phase 1: Participants performed 10 trials of the experimental task.

C) Phase 1 postdiction: Participants were then asked to report what percentage of target circles they thought they had correctly dragged to the instructed location, using the same slider as in part A.

D) Offloading instructions and practice: Participants were then informed of the intention offloading strategy. They were told that this was an optional strategy that they could use if they wished, but it was up to them whether to use it. They then performed two practice trials with offloading permitted.

E) Phase 2 prediction: Participants provided a performance prediction for phase 2, using the same slider as in part A.

F) Phase 2: Participants performed 10 trials of the experimental task, with offloading permitted.

G) Phase 2 postdiction: Participants reported their performance in phase 2, using the same slider as in part A.

2.1.2 Experiment 1b

An additional sample of participants was recruited who performed a more difficult version of the task. On each experimental trial, participants were instructed to drag three targets to alternative locations instead of one. Furthermore, participants additionally received a distracting arithmetic question during each trial, as in the 'Interruption' condition of Gilbert (in press; Experiment 1). This occurred immediately after dragging one of the nontarget circles to the bottom of the box, at a position in the sequence randomly selected between the first circle and the first target. In this experiment, participants performed three initial practice trials before commencing the experimental trials, rather than two. Apart from these changes, the procedure of Experiment 1b was identical to Experiment 1a.

2.2 Participants

200 participants were recruited in Experiment 1a (mean age 32, range 18-66, 95 male) and a further 200 participants in Experiment 1b. 5 participants were replaced in Experiment 1b due to arithmetic-verification accuracy below 80% (final sample: mean age 33, range 18-65, 80 male). Participation took approximately 20 minutes, for which participants were paid $2.

2.3 Results

In Experiment 1a the mean retention interval (i.e. time from the start of each trial until the first target circle was reached in the sequence) was 8.8s in phase 1 and 9.3s in phase 2. The equivalent figures for Experiment 1b were 11.0s and 14.0s respectively. Mean arithmetic-verification accuracy in Experiment 1b was 99%. Table 1 lists the mean prediction and postdiction judgments for Experiments 1a and 1b, alongside measures of objective performance and externalizing proportion.

[Table 1 about here]

In both experiments, predictions were more pessimistic than postdictions ($F(1,199) > 42$, $p < 10^{-10}$, $\eta^2 > .19$) and phase 2 ratings were higher than phase 1 ratings ($F(1,199) > 43$, $p < 10^{-9}$, $\eta^2 > .18$). These two factors interacted significantly in Experiment 1a ($F(1,199) = 61$, $p < 10^{-12}$, $\eta^2 = .24$) and marginally significantly in Experiment 1b ($F(1,199) = 3.3$, $p = .07$, $\eta^2 = .016$). In Experiment 1b, phase 2 objective accuracy, when intention offloading was permitted, was higher than phase 1 ($F(1,199) = 56$, $p < 10^{-12}$, $\eta^2 = .24$). There was no significant difference in Experiment 1a ($F < 1$; Experiment x Phase interaction: $F(1,398) = 26$, $p < 10^{-7}$, $\eta^2 = .06$). Thus, availability of the intention offloading strategy improved performance in Experiment 1b but not Experiment 1a (where unaided performance in phase 1 was already near ceiling). Participants were significantly more likely to set reminders in Experiment 1b than Experiment 1a ($t(391.8) = 7.3$, $p < 10^{-12}$, $d = .74$). In order to evaluate their interrelations, the correlations between these measures are presented in Table 2.

[Table 2 about here]

In both experiments, the four subjective judgments were positively correlated with each other. Measures of objective performance in the two phases were positively correlated with all four subjective judgments in Experiment 1b, with phase 1 accuracy most strongly correlated with phase 1 postdiction, and phase 2 accuracy with phase 2 postdiction. A more complex pattern was evident in Experiment 1a. In this experiment, the two postdiction judgments were positively correlated with objective performance in the relevant phase of the task. Participants were therefore able to retrospectively evaluate their task performance with some degree of accuracy. However, predictions in the two phases were uncorrelated with subsequent objective performance (phase 1: $r = -.007$; phase 2: $r = -.020$). Thus, in Experiment 1a participants were able to evaluate their recent performance, but not to predict forthcoming accuracy. Moreover, their retrospective judgments of phase 1 performance significantly predicted subsequent accuracy in phase 2, even though their actual predictions for phase 2 did not. This suggests that in this experiment, with a very simple task, participants had little or no prospective insight into their likelihood of making errors, whereas they were able to detect errors after they were committed, which reliably predicted future errors.

Turning now to the externalizing proportion (i.e. propensity to set reminders), the simple correlations between this measure and the other measures may be difficult to interpret, due to multiple causal pathways that need to be disentangled. For example, setting reminders in phase 2 might be expected to improve performance, leading to a positive correlation between externalizing proportion and phase 2 accuracy. On the other hand, if reminders are most likely to be set by participants with poor ability and do not fully compensate for this, then a higher externalizing proportion might be associated with poorer phase 2 performance. Thus, two opposing effects might cancel out in the simple

correlation between externalizing proportion and phase 2 performance. Another example would be that the different prediction and postdiction ratings may consist of various components, such as a general motivation to perform well and a general sense of metacognitive confidence, a preference for using one part of the rating scale, and metacognitive evaluations that apply in a task-specific manner to the individual ratings. These effects can be distinguished using a multiple regression approach. Thus, considering the strict temporal ordering of the measures collected in these experiments, a path analysis was conducted to investigate their multiple possible relationships (Figure 2). This can be seen as a form of structural equation modeling in which every variable is a directly observed measure (Garson, 2012). In order to calculate the path weights in this model, a linear regression is calculated for each measure in turn, including all antecedent measures as predictors. Path weights reflect standardized betas and significance testing reflects the significance of each predictor in the relevant linear regression. In other words, a significant path weight from an antecedent to a consequent variable indicates that the antecedent variable accounts for significant variance in the consequent variable, after controlling for all of the other antecedent variables connected to that consequent variable.

A model was specified including the following sequentially ordered measures: 1) Phase 1 prediction; 2) Phase 1 performance; 3) Phase 2 prediction; 4) Phase 2 intention offloading behavior; 5) Phase 2 performance. For simplicity, the two postdiction measures were not included (Table 2 already shows that the postdiction measures were in every case significantly correlated with the relevant objective measure). Paths were included linking each measure to all of its antecedent measures. This yielded a total of 10 paths, labeled A-J. Figure 2 shows the results of these analyses; the path weights are discussed in turn below. Note that every significant path weight was either replicated in

the two independent experiments, significant after Bonferroni correction for 20 paths (10 paths x 2 experiments), or both.

[Figure 2 about here]

Path A shows that participants were able to predict their forthcoming phase 1 performance in Experiment 1b but not Experiment 1a. Path B indicates that in both experiments phase 1 predictions were correlated with subsequent phase 2 predictions, perhaps reflecting 1) individual participants' tendency to use a restricted part of the response scale; 2) their general confidence; 3) their general motivation to perform well in the task, or a combination of these factors. Path C indicates that in Experiment 1b participants updated their performance predictions from phases 1 to 2 on the basis of objective phase 1 performance, but this was not the case in Experiment 1a. Path D indicates that participants' metacognitive judgment of their unaided ability to perform the task was negatively related to their use of an intention offloading strategy in phase 2. In other words, participants who predicted their unaided ability to be relatively poor were more likely to offload intentions when they could. Path E indicates that participants' objective performance in phase 1 was negatively related to intention offloading in phase 2, even after controlling for the influence of their prospective metacognitive performance predictions. Thus, the experience of making errors in the unaided phase of the task may have triggered an increase in phase 2 intention offloading, and / or participants with lower cognitive ability, reflected in their phase 1 performance, may have learned to utilize cognitive strategies more often even without metacognitive awareness of this. Path F indicates that in Experiment 1b, participants' predictions of phase 2 accuracy when intention offloading was available predicted their actual use of an intention offloading strategy, even after controlling for their predicted unaided ability.

There was no such relationship in Experiment 1a. Path G indicates that accuracy in phase 1 predicted accuracy in phase 2, perhaps reflecting cognitive factors that could influence accuracy both with and without the opportunity to offload intentions, and / or motivation to perform well in the experiment. Path H showed that in both studies, intention offloading was effective in boosting performance. Finally, paths I and J showed that the phase 1 and 2 predictions did not account for any additional variance in phase 2 performance, after controlling for other antecedent variables. For the three paths that were significant in Experiment 1b but not 1a (paths A, C, and F), weights were significantly different between the two experiments ($p < .02$; Paternoster, Brame, Mazerolle, & Piquero, 1998). None of the other path weights differed significantly between the experiments.

In a final analysis, stepwise regression was conducted to evaluate the additional variance in Phase 2 performance that could be explained by intention offloading, after first controlling for unaided ability represented by Phase 1 performance. A model of Phase 2 performance including just Phase 1 performance as a regressor yielded an adjusted r^2 of 0.31 in Experiment 1a and 0.36 in Experiment 1b. This rose to 0.36 and 0.52 respectively when externalizing proportion was included as an additional regressor. Thus, the model including both regressors accounted for over a third of the variance in Experiment 1a's Phase 2 performance, and over half in Experiment 1b. Of this variance, 12% was uniquely attributable to intention offloading in Experiment 1a, and 30% in Experiment 1b.

2.4 Discussion

Metacognitive judgments of unaided ability, along with objective measures of actual unaided performance, independently predicted the use of intention offloading. Both of these relationships were negative, indicating that participants offload intentions in a compensatory manner, i.e. when they believe their unaided ability to be relatively poor and, independently, when their unaided ability objectively is relatively poor. Strikingly, in Experiment 1a performance predictions were uncorrelated with objective performance measures, but both subjective and objective measures predicted intention offloading. Furthermore, intention offloading facilitated task performance: when this strategy was available, its increased use led to better performance.

While the two experiments were consistent in showing that intention offloading was independently predicted by subjective confidence and objective ability, as well as showing that intention offloading boosted performance, there also appeared to be divergence between the experiments. In particular, Experiment 1b revealed that A) phase 1 predictions significantly predicted objective phase 1 performance; B) phase 2 predictions were significantly updated as a result of objective phase 1 performance; and C) phase 2 predictions significantly predicted phase 2 intention offloading. None of these effects was significant in Experiment 1a. A possible explanation of this discrepancy would be that the task in Experiment 1a was extremely simple: participants were only required to remember a single intention, and they were not interrupted when they did so. Accuracy was high in this task (93%). Under such conditions, participants may lack the metacognitive insight to predict rare errors in the task. Furthermore, individual differences in general motivation to perform well might not translate reliably into objective performance differences, seeing as the task was so simple, and participants' intention to use an offloading strategy may not have strongly influenced their predictions of how well they would do at the task. In this case, any relationship between

performance predictions and objective measures of behavior would be expected to be reduced or absent.

3. Experiment 2

Results of Experiment 1 showed that intention offloading was predicted by participants' confidence in their ability to perform a matched task in which reminders were not permitted. However, the question remains whether strategy use is also influenced by domain-general metacognitive confidence, i.e. a belief in one's ability to effectively carry out diverse tasks, regardless of the specific domain. This might reflect a general factor relating to beliefs about one's abilities, much as the hypothesized 'g' factor of general intelligence relates to performance of diverse tasks (Carroll, 2003; Spearman, 1904).

In order to investigate this question, participants in Experiment 2 performed the same tasks and provided the same metacognitive judgments as Experiment 1. They also performed a pair of perceptual discrimination tasks that were entirely unrelated to the delayed intention task. On each trial, participants made a difficult perceptual judgment, using a staircase procedure to maintain accuracy at about 70%. They also provided a metacognitive evaluation on each trial of how confident they were that they had made the correct judgment. Two relevant measures can be derived from these evaluations. A measure of metacognitive *confidence* can be derived from the mean confidence rating across trials. This reflects each participant's bias towards higher or lower confidence, even though objective performance was fixed at around 70% by the staircase procedure. Additionally, a measure of metacognitive *sensitivity* can be derived by evaluating the relationship between confidence and accuracy on a trial-by-trial basis, i.e. how reliably accuracy varies with confidence. This reflects how well participants can discriminate

between correct responses and guesses. A similar approach was used by Song et al. (2011), who also administered a pair of perceptual tasks, using a staircase procedure and with a confidence judgment after each trail. Song et al. (2011) found that both confidence and sensitivity were correlated across the two perceptual tasks, suggesting the existence of common metacognitive processes operating across different perceptual decisions. However, confidence and sensitivity were not correlated with each other. They also suggested that confidence ratings partly reflect a task-independent component of general confidence, and partly a task-specific component. The present study aims to replicate these findings in a web-based task. It also goes further by investigating 1) whether any putative domain-general confidence signal correlates not only across perceptual tasks, but also relates to confidence in the quite different domain of the intention offloading task; and 2) whether domain-general confidence relates to strategic offloading of intentions.

3.1 Methods

3.1.1 Procedure

The experiment took place in two parts. One part was identical to Experiment 1b, with three targets and interruption from an arithmetic question on each intention offloading trial (this was because there was more variability in performance with this version of the task). As in Experiment 1b, participants performed the task in two phases, the first phase disallowing intention offloading and the second phase permitting it, with pre- and post-dictions before and after each phase. The other part of Experiment 2 consisted of a pair of perceptual discrimination tasks. Participants were randomly allocated to perform the intention offloading tasks before the perceptual tasks, or vice versa.

[Figure 3 about here]

There were two perceptual tasks. In each task, participants viewed a fixation point with a pair of grids to its left and right, each formed of 20 horizontal and 20 vertical pale green lines, to yield a total of 400 internal squares (Figure 3). In the Number task, a randomly selected 200 squares on one side (i.e. 50%) were filled in a pink colour; on the other side more than 200 were filled. Participants were asked to judge which side had more filled squares. The starting difficulty for this task was set so that 300 squares were filled on one side; in order to increase difficulty this could be gradually reduced so that the side with more filled squares approached 200. In the Contrast task, every square was filled a different shade of gray. Participants were asked to judge on which side the contrast between the different shades was greater. The starting difficulty for this task was set so that the shades on one side varied from 15% maximum brightness to 85% and the other from 35% to 65%. The difference between the sides could be gradually reduced to approach 25% to 75% on both sides. In this task, the brightness of the 400 squares was uniformly spaced from the brightest to the darkest, with each shade placed in a random position within the grid. For a demonstration of the perceptual tasks, please visit "http://www.ucl.ac.uk/sam-gilbert/demos/staircaseDemo.html".

In order to train participants on each task, they first viewed an example stimulus with minimum difficulty, which remained on screen until they made a response. If their response was incorrect, another example stimulus was presented. Once they had made a correct response, they were presented with five more trials, with each stimulus presented for 800ms. If they made any mistakes on these five trials, five more were presented. Next they were presented with five trials with each stimulus presented for 250ms, requiring at

least four correct responses to continue. After this point, 80 practice trials were presented. From this point onwards, difficulty was adjusted with a "two-down one-up" staircase procedure, so that difficulty was increased one step after two consecutive correct responses and decreased one step after any incorrect response. This has the effect of stabilising accuracy at a level of approximately 70.7% (Levitt, 1971). After these practice trials, the instructions for metacognitive judgements were presented as follows: "After each trial we would like you to indicate how confident you are that you responded correctly. Please answer on a scale of 1-4, where 1 would indicate that you were purely guessing, and 4 would indicate that you were confident you got the right answer." Participants were also instructed: "You must use the full extent of the scale, so that you use every answer from 1-4 at least sometimes. Please remember that this scale indicates relative confidence, and you will rarely be absolutely certain that you were correct. So even if you select 4, you may still have some doubt". Following each trial, participants used the same response keys as they used for the perceptual judgements to move an arrow between the four confidence options, with 1 labelled "guessing" and 4 labelled "confident". They pressed the 'm' key to submit this judgement. The final timings for each trial were as follows: 1) unfilled grids for 400ms; 2) stimulus presentation for 250ms; 3) unfilled grids until a response was made; 4) confidence response until the 'm' key was pressed; 5) 500ms delay. Participants performed a further 10 practice trials with the confidence judgements. They then performed two experimental blocks of 75 trials. After completing the experimental trials, the other perceptual task (Number or Contrast) was then administered in an identical manner. Ordering of the two perceptual tasks was counterbalanced between participants. After each of these tasks, participants gave a 'postdiction' judgement of what percentage of trials they thought they had answered correctly in that task, using a slider in an identical manner as in the pre- and post-dictions given in the intention offloading task.

3.1.2 Data analysis

Measures from the intention offloading task were extracted in an identical manner to Experiment 1. Additional measures were extracted from each of the perceptual tasks as follows. 1) Mean *accuracy* provides a measure of task performance, calculated as the percentage of correct responses. 2) Metacognitive *sensitivity* indexes the strength of the relationship between confidence and accuracy on each trial. It was calculated using the type II receiver operating characteristic (ROC) curve (Fleming & Lau, 2014). This is the same measure as was used in the related study by Song et al. (2011; see p. 1789 for further details). It is calculated by constructing a ROC curve reflecting the cumulative probability of being correct for each level of confidence. The area under this curve provides a robust estimate of metacognitive sensitivity, independent of response bias (Kornbrot, 2006). 3) Metacognitive *confidence* reflects each participant's bias towards higher or lower confidence; it is calculated simply as the mean confidence level across trials. 4) Metacognitive *postdiction* provides an additional measure of confidence, from the single rating given by participants at the end of each task reflecting what proportion of trials they thought they had answered correctly.

3.1.3 Participants

317 participants were recruited (mean age: 32; range: 18-68; 146 male) from Amazon Mechanical Turk. Participation in the experiment took approximately 40 minutes, for which participants were paid $4.

3.2 Results and Discussion

In order to ensure reliable data, any participant scoring below 80% on the arithmetic verification task or below 50% (i.e. below chance) on either of the perceptual discrimination tasks was discarded. This led to the removal of 9 participants (2.8%). Furthermore, inspection of the data from the perceptual tasks showed that a small number of participants performed close to chance, suggesting a failure of the staircase procedure and/or frequent guessing or random responses. Thus, any participant with accuracy outside 3 standard deviations of the group mean for either task was additionally discarded, leading to the removal of a further 5 (1.6%) participants. The final sample therefore consisted of 303 participants.

[Table 3 about here]

Performance measures are given in Table 3. Results from the intention offloading task were similar to those obtained in Experiment 1b (phase 1 accuracy: mean = .81, SD = .17; phase 2 accuracy: mean = .87, SD = .16; phase 2 externalizing proportion: mean = .66, SD = .47). Accuracy was around 71% in the perceptual tasks, suggesting that the staircase procedure was effective at avoiding floor and ceiling effects, at least for the participants included in the final sample. Accuracy was slightly higher in the Number task than the Contrast task ($t(302) = 3.9$, $p = .0001$); however, both the accuracy postdiction ($t(302) = 2.5$, $p = .01$) and the mean confidence rating ($t(302) = 6.0$, $p < 10^{-8}$) were higher for the Contrast task than the Number task. Thus, the small difference in mean accuracy between the two tasks was not reflected in the confidence ratings. Metacognitive sensitivity did not differ between the tasks ($t(302) = 1.4$, $p = .18$).

[Figure 4 about here]

Mean accuracy levels for each confidence rating are shown in Figure 4; this figure also shows that the difficulty level in each task had plateaued by the time of the experimental trials as a result of the staircase procedure. Table 4 shows a correlation matrix between the measures of accuracy, sensitivity, confidence, and postdiction, separately for each task (see Methods for definition). Every measure correlated significantly with the analogous measure in the other task. Furthermore, the two types of confidence measure (mean confidence from the trial-by-trial ratings, and the single postdiction of accuracy at the end of each task) were also significantly intercorrelated. However, none of the correlations across different types of measure was significant. These results replicate the findings of Song et al. (2011), who showed that 1) measures of metacognitive sensitivity were significantly correlated across tasks, 2) measures of metacognitive confidence were significantly correlated across tasks, and 3) measures of sensitivity and confidence were uncorrelated with each other. However, unlike the findings of Song et al. (2011), the measures of accuracy were also significantly correlated across tasks in the present study.

[Table 4 about here]

In order to investigate the relationship between the measures derived from the perceptual tasks and those derived from the intention offloading task, first the measures of accuracy, sensitivity, confidence, and postdiction were collapsed over the Number and Contrast tasks. The correlations between these measures and those derived from the intention-offloading task were then calculated (Table 5).

[Table 5 about here]

Correlations between the intention offloading measures were qualitatively similar to those obtained in Experiment 1b, i.e. significant correlations from Experiment 1b were replicated and nonsignificant correlations remained nonsignificant. There were five exceptions to this: the correlation between phase 1 prediction and phase 2 performance and the correlation between the measure of externalizing proportion and the four pre/post-dictions. In each of these cases, results were qualitatively similar to Experiment 1a rather than Experiment 1b. It was suggested above that path analysis may be a more appropriate approach for investigating results from the present paradigm. Thus, a path analysis analogous to the one shown in Figure 2 was conducted. This showed that every significant path from Experiment 1b remained at least marginally significant in the present dataset ($t(302) > 1.68$, $p < .094$). Thus, there was good consistency of results across experiments.

Turning now to the relationship between measures from the perceptual tasks and those from the intention-offloading task, results were as follows. The measure of metacognitive sensitivity did not correlate with any other measure, even though this measure was significantly correlated across the two perceptual tasks, demonstrating its reliability. Likewise, the perceptual accuracy measure was uncorrelated with all other measures. However, the perceptual confidence and postdiction measures were positively correlated with phase 1 prediction in the intention-offloading task, and negatively correlated with the externalizing proportion, i.e. propensity to set reminders. Thus, participants with higher confidence in their perceptual judgements also tended to predict better performance in phase 1 of the intention offloading task, as well as setting fewer reminders when they were permitted to do so (Figure 5). This occurred even though perceptual confidence was unrelated to any objective measure of performance, and even though the perceptual and intention offloading tasks were unrelated to each other. In

other words, results suggested a contribution of domain-general confidence to both the perceptual and delayed intention tasks, which was related to the propensity to set reminders. Higher perceptual confidence was not related to objective performance, but it was related to strategic use of reminders in an unrelated task.

[Figure 5 about here]

In order to evaluate whether task-specific confidence was also related to the use of reminders, along with the domain-general effect demonstrated above, a multiple regression was conducted to predict externalizing proportion from both the phase 1 prediction (i.e. prediction of unaided memory ability) and perceptual confidence. The perceptual confidence measure was generated by transforming the mean confidence and postdiction measures into Z scores and then collapsing across them. In the regression model, the perceptual confidence measure was not significant (standardised beta = -.08, $t(302) = 1.37$, $p = .17$), but significant additional variance was explained by the memory confidence measure (standardised beta = -.19, $t(302) = 3.2$, $p = .001$). Results were similar when using the mean confidence or postdiction measures alone, rather than the combined perceptual confidence measures. Thus, propensity to set reminders was influenced both by domain-general confidence (shown by the correlation between perceptual confidence and externalizing proportion in the intention offloading task) and a task-specific effect (shown by the additional predictive validity conferred by the memory confidence measure).

Seeing as the order of the perceptual and intention-offloading tasks was counterbalanced between participants, it might be predicted that participants performing the perceptual tasks first would show a stronger relationship between their perceptual confidence

ratings and subsequent offloading behaviour than participants performing the tasks in the reverse order. There was some limited support for this hypothesis. For participants performing the perceptual tasks first, the correlation between the collapsed perceptual confidence measure and propensity to set reminders in the intention-offloading task was -.21 ($p = .009$). For participants performing the tasks in the reverse order, this correlation was -.09 ($p = .29$). However, the difference between these correlation coefficients was not significant ($p = .28$).

4. General Discussion

The present study investigated individual differences in 'intention offloading', i.e. the process of setting up external reminders for delayed intentions. Intention offloading was effective: participants who set more reminders fulfilled more intentions. Furthermore, individual differences in offloading behaviour explained significant variance in intention fulfilment, even after controlling for unaided ability. Promoting the use of external artifacts may therefore be an effective means for improving real-world prospective memory, one that is likely to be more amenable to behaviour-change interventions than interventions targeted at individuals' unaided ability.

Intention offloading was predicted by 1) low metacognitive predictions of unaided ability and 2) objectively poor unaided ability. These influences were independent of one another, and occurred even in Experiment 1a, where subjective confidence was unrelated to objective accuracy. In addition, Experiment 2 showed that confidence in an unrelated perceptual task also predicted intention offloading in the delayed intention task, indicating a contribution of domain-general metacognitive confidence. Thus, results

indicate that strategic use of reminders is influenced by multiple metacognitive confidence signals, and, independently, objective ability.

4.1 Metacognition and prospective memory

Previous studies investigating the relationship between metacognition and prospective memory have reported three main findings. First, participants generally have modest but significant metacognitive insight into their prospective memory performance. Second, participants tend to be underconfident about their performance (Meeks et al., 2007; Schnitzspahn et al., 2011). Results of the present study are consistent with both of these findings. Third, participants' metacognitive beliefs influence prospective memory strategies, for example the allocation of attention towards the prospective memory task versus the ongoing task in which it is embedded (Rummel & Meiser, 2013). The present study extends this finding, by showing that metacognitive judgments, in addition, influence propensity to use strategies involving external reminders.

4.2 Intention offloading and metacognitive confidence

The three experiments reported here were consistent in showing an effect of metacognitive confidence on intention offloading behaviour. This suggests that lowering participants' confidence might increase the use of offloading strategies, even without improving their insight into their unaided ability. Interestingly, in a study by Knight et al. (2005), participants with traumatic brain injury made overconfident predictions of their prospective memory ability, whereas control participants were underconfident. This suggests that in the context of ageing or rehabilitation, it could be more helpful to target overconfidence than unaided ability in order to promote the fulfilment of delayed

intentions. Furthermore, the influence of domain-general confidence on intention offloading in Experiment 2 suggests that a single intervention that has the effect of reducing confidence might influence strategic behaviour across diverse domains.

4.3 Intention offloading and objective ability

There are several possible explanations of the relationship between objective unaided accuracy and intention offloading, independent of metacognitive confidence. First, participants with poor unaided ability may have learned the benefit of intention offloading to a greater extent than those with better ability, even without metacognitive awareness of this. Second, participants' experience of making errors in the unaided first phase of the delayed intention task may have prompted subsequent use of intention offloading, independent of their prior predicted ability. This would suggest an additional contribution of a metacognitive process of online performance monitoring, as well as prior metacognitive confidence. Third, participants with poor unaided ability may have perceived the task as more difficult, and hence requiring the use of compensatory strategies, even without insight into their level of ability. Of course, these possibilities are not mutually exclusive.

4.4 Domain-specificity of metacognition

On a theoretical level, the present results are relevant to the delineation of domain-general and task-specific metacognitive processes, and their relation to strategic behaviour. Experiment 2 replicated the finding of Song et al. (2011) that both metacognitive sensitivity and metacognitive confidence contain a domain-general component that operates across diverse perceptual tasks, although the two measures

seem independent of one another. This is consistent with the suggestion made by de Gardelle and Mamassian (2014) of a common metacognitive 'currency' across perceptual tasks. Furthermore, Experiment 2 showed that metacognitive confidence can be generalized not only across perceptual tasks but also into the mnemonic domain, and that this signal relates functionally to strategic behaviour. In related studies, both Baird et al. (2013) and McCurdy et al. (2013) showed a correlation between metacognitive confidence in perceptual and mnemonic tasks, consistent with a domain-general confidence signal. This suggests the existence of a 'g' factor of metacognitive confidence, that operates across diverse domains.

Turning now to metacognitive sensitivity, i.e. the ability to discriminate correct responses from guesses on a trial-by-trial basis, Baird et al. (2013) found no correlation between perceptual and mnemonic tasks (though see McCurdy et al., 2013, for a different result), and Fleming et al. (2014) provided neuropsychological evidence for a dissociation between sensitivity in the two domains. In the present study, although metacognitive sensitivity generalised across the two perceptual tasks, it was unrelated to any measure from the delayed intention task. These results suggest that metacognitive sensitivity is less likely to generalise across domains than metacognitive confidence. It also suggests that metacognitive confidence, rather than metacognitive sensitivity, may be more likely to influence strategic behaviour.

4.5 Web-based data collection

The use of online web-based experiments, rather than traditional laboratory-based studies presents both advantages and disadvantages. The chief advantage is that it is an effective means for the collection of large volumes of data. This helps to mitigate against

the chief disadvantage, which is the reduction in experimental control over participant behaviour and the precise experimental parameters (e.g. due to the use of different computer systems from one participant to the next). Although measurements derived from web-based experiments will likely be noisier than those derived from laboratory studies, sensitivity to experimentally-induced effects can be greater due to larger samples (often orders of magnitude larger than those that are practical with more traditional approaches). Furthermore, the participant pool is likely to be more representative of the general population than the undergraduate convenience samples typical in behavioural studies, albeit not perfectly representative (Mason & Suri, 2012). Another advantage of web-based experiments is that they provide a relatively simple means of following up participants after the experimental session via further online testing. For example, Gilbert (in press) showed that performance of the present intention-offloading task was correlated with successful fulfilment of a naturalistic intention: visiting a specified web link up to one week after the experimental session. Thus, the experimental paradigm has significant external validity with respect to participants' fulfilment of real-world intentions embedded in their everyday routines, over periods of up to one week. Nevertheless, understanding the generalizability of the effects reported in the present study would benefit from the enhanced experimental control made possible by testing participants in a laboratory setting. It would also benefit from examination of the ways in which participants use external reminders to support memory for intentions delayed over periods of hours, days, and weeks. In everyday life, we use external artefacts to remind us of intentions delayed by periods of seconds (e.g. physically holding a task-relevant item to cue an intended behaviour after a brief interruption in a nursing setting; Grundgeiger, Sanderson, MacDougall, & Venkatesh, 2010) as well as over longer periods (e.g. creating a diary entry for a planned hospital appointment). However, it is unclear how far the processes involved in setting external reminders for intentions delayed by a few seconds,

as in the present study, overlap with those involved in long-term intentions. Indeed, it is debatable whether "prospective memory" (as opposed to e.g. "working memory") is the best label to describe the present paradigm, given the brief retention interval (see Gilbert, in press, for further discussion of this point). Thus, an important question for further research is the extent to which similar mechanisms play a role in the use of reminders for intentions delayed over different timescales.

4.6 Conclusion

Due to technological advances such as time-, location-, and person-based smartphone reminders and wearable devices, our ability to fulfil delayed intentions may increasingly be supported by external artefacts. The present results indicate that an individual's choice to use such external artefacts may be based on diverse metacognitive processes, alongside an influence of objected unaided ability. In particular, the present results highlight the importance of both domain-general and task-specific metacognitive confidence signals, suggesting that such signals could be important targets for therapeutic interventions to improve individuals' ability to fulfil delayed intentions. Understanding the contribution of these factors, and the ways in which they interact, will help to ensure that such artefacts are designed efficiently and used effectively, in order to promote behavioural independence.

References

Allen, D. (2002). *Getting Things Done: How to Achieve Stress-free Productivity*. Piatkus.

Arango-Muñoz, S. (2013). Scaffolded Memory and Metacognitive Feelings. *Review of Philosophy and Psychology*, *4*, 135–152.

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and Lateral Networks in Anterior Prefrontal Cortex Support Metacognitive Ability for Memory and Perception. *Journal of Neuroscience*, *33*, 16657–16665.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In *The scientific study of general intelligence* (pp. 5–21).

Crump, M. J. C., McDonnell, J. V, & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*, e57410.

De Gardelle, V., & Mamassian, P. (2014). Does Confidence Use a Common Currency Across Two Visual Tasks? *Psychological Science*, *25,* 1286-1288.

Fish, J., Wilson, B. a, & Manly, T. (2010). The assessment and rehabilitation of prospective memory problems in people with neurological disorders: a review. *Neuropsychological rehabilitation*, *20*, 161–79.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 1–9.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137,* 2811-2822.

Gilbert, S. J. (in press). Strategic offloading of delayed intentions into the external environment. *Quarterly Journal of Experimental Psychology.* doi: 10.1080/17470218.2014.972963

Garson, G. D. (2012). *Path Analysis*. Asheboro, NC: Statistical Associates Publishers.

Grundgeiger, T., Sanderson, P., MacDougall, H. G., & Venkatesh, B. (2010). Interruption management in the intensive care unit: Predicting resumption times and assessing distributed support. *Journal of experimental psychology: Applied*, *16*, 317–34.

Henry, J. D., Rendell, P. G., Phillips, L. H., Dunlop, L., & Kliegel, M. (2012). Prospective memory reminders: a laboratory investigation of initiation source and age effects. *Quarterly journal of experimental psychology*, *65*, 1274–87.

Heylighen, F., & Vidal, C. (2008). Getting Things Done: The Science behind Stress-Free Productivity. *Long Range Planning*, *41*, 585–605.

Knight, R. G., Harnett, M., & Titov, N. (2005). The effects of traumatic brain injury on the predicted and actual performance of a test of prospective remembering. *Brain Injury*, *19*, 19–27.

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: model-based and distribution-free measures and evaluation. *Perception & psychophysics*, *68*, 393–414.

Kvavilashvili, L., & Ford, R. M. (2014). Metamemory prediction accuracy for simple prospective and retrospective memory tasks in 5-year-old children. *Journal of Experimental Child Psychology*, 1–17.

Landsiedel, J., & Gilbert, S. J. (2015). Creating external reminders for delayed intentions: Dissociable influence on "task-positive" and "task-negative" brain networks. *NeuroImage*, *104*, 231–240.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*, Suppl 2:467+.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, *44*, 1–23.

Maylor, E. A. (1990). Age and prospective memory. *The Quarterly Journal of Experimental Psychology Section A*, *42*, 471–493.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, *33*, 1897–906.

McDaniel, M. A., & Einstein, G. O. (2000). Strategic and automatic processes in prospective memory retrieval: a multiprocess framework. *Applied Cognitive Psychology*, *14*, S127–S144.

McDaniel, M. A., & Einstein, G. O. (2007). *Prospective Memory: An Overview and Synthesis of an Emerging Field*. Los Angeles: Sage Publications Ltd.

Meeks, J. T., Hicks, J. L., & Marsh, R. L. (2007). Metacognitive awareness of event-based prospective memory. *Consciousness and cognition*, *16*, 997–1004.

Meier, B., von Wartburg, P., Matter, S., Rothen, N., & Reber, R. (2011). Performance predictions improve prospective memory and influence retrieval experience. *Canadian journal of experimental psychology*, *65*, 12–8.

Metcalfe, J. (1996). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.

Metcalfe, J. (2009). Metacognitive Judgments and Control of Study. *Current directions in psychological science*, *18*, 159–163.

Migo, E. M., Haynes, B. I., Harris, L., Friedner, K., Humphreys, K., & Kopelman, M. D. (2014). mHealth and memory aids: levels of smartphone ownership in patients. *Journal of mental health (Abingdon, England)*, *8237*, 1–5.

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, *36*, 859–866.

Phillips, L. H., Henry, J. D., & Martin, M. (2008). Adult aging and prospective memory: the importance of ecological validity. In M. Kliegel, M. A. McDaniel, & G. O. Einstein (Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives.* (pp. 161–186). Mahwah: Erlbaum.

Rummel, J., & Meiser, T. (2013). The role of metacognition in prospective memory: Anticipated task demands influence attention allocation strategies. *Consciousness and Cognition*, *22*, 931–943.

Schnitzspahn, K. M., Zeintl, M., Jäger, T., & Kliegel, M. (2011). Metacognition in prospective memory: are performance predictions accurate? *Canadian journal of experimental psychology*, *65*, 19–26.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and cognition*, *20*, 1787–92.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Svoboda, E., Rowe, G., & Murphy, K. (2012). From Science to Smartphones: Boosting Memory Function One Press at a Time. *Journal of Current Clinical Care*, *2*, 15–27.

Thöne-Otto, A. I. T., & Walther, K. (2008). Assessment and treatment of prospective memory disorders in clinical practice. In M. Kliegel, M. A. McDaniel, & G. O. Einstein (Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives.* (pp. 321–345). Mahwah: Erlbaum.

Wilson, B. A., Emslie, H. C., Quirk, K., & Evans, J. J. (2001). Reducing everyday memory and planning problems by means of a paging system: a randomised control crossover study. *Journal of neurology, neurosurgery, and psychiatry*, *70*, 477–82.

Table 1. Experiment 1 performance measures.

	Exp 1a (one target)		Exp 1b (three targets + interruption)	
	M	*SD*	*M*	*SD*
Phase 1 prediction	0.83	0.13	0.72	0.21
Phase 1 postdiction	0.95	0.11	0.79	0.21
Phase 2 prediction	0.93	0.12	0.83	0.20
Phase 2 postdiction	0.95	0.10	0.87	0.17
Phase 1 objective target accuracy	0.93	0.13	0.83	0.15
Phase 2 objective target accuracy	0.93	0.13	0.90	0.14
Phase 2 externalizing proportion	0.29	0.43	0.63	0.49

Table 2. Correlations between metacognitive and objective measures. Shaded cells: Experiment 1a; White cells: Experiment 1b.
* p < .05; ** p < .01; *** p < .001

	Phase 1 prediction	Phase 1 postdiction	Phase 2 prediction	Phase 2 postdiction	Phase 1 accuracy	Phase 2 accuracy	Phase 2 externalizing
Phase 1 prediction	-	.58***	.64***	.50***	.27***	.23***	-.05
Phase 1 postdiction	.15*	-	.65***	.62***	.59***	.32***	-.11
Phase 2 prediction	.47***	.16*	-	.74***	.41***	.45***	.23**
Phase 2 postdiction	.13	.22**	.41***	-	.50***	.59***	.21**
Phase 1 accuracy	-.01	.56***	-.07	.04	-	.60***	-.05
Phase 2 accuracy	-.02	.28***	-.02	.14*	.56***	-	.37***
Phase 2 externalizing	-.18**	-.18*	-.02	-.04	-.20**	.10	-

Table 3. Performance measures from the perceptual tasks in Experiment 2.

	Number task		Contrast task	
	Mean	SD	Mean	SD
Accuracy (proportion correct)	0.72	0.02	0.71	0.02
Metacognitive sensitivity (aROC)	0.62	0.07	0.63	0.07
Metacognitive confidence	2.49	0.51	2.61	0.48
Accuracy postdiction	0.67	0.11	0.68	0.12

Table 4. Correlations between measures from the perceptual tasks in Experiment 2.
* p < .05; *** p < .001

		Number task				Contrast task			
		Accuracy	Sensitivity	Confidence	Postdiction	Accuracy	Sensitivity	Confidence	Postdiction
Number task	Accuracy	-	-.10	-.02	-.03	.13*	.00	-.04	.04
	Sensitivity		-	.00	-.01	-.07	.35***	-.04	-.02
	Confidence			-	.55***	.02	-.05	.75***	.45***
	Postdiction				-	-.01	-.10	.40***	.64***
Contrast task	Accuracy					-	-.01	-.02	.00
	Sensitivity						-	-.03	.03
	Confidence							-	.52***
	Postdiction								-

Table 5. Correlations between perceptual and intention offloading measures in Experiment 2.
* p < .05; ** p < .01; *** p < .001

	Phase 1 prediction	Phase 1 postdiction	Phase 2 prediction	Phase 2 postdiction	Phase 1 accuracy	Phase 2 accuracy	Phase 2 offloading	Perceptual accuracy	Perceptual sensitivity	Perceptual confidence	Perceptual postdiction
Phase 1 prediction	-	.44***	.50***	.32***	.14*	.04	-.21***	.00	-.03	.24***	.28***
Phase 1 postdiction		-	.45***	.61***	.46***	.20***	-.22***	-.02	.00	.09	.14*
Phase 2 prediction			-	.57***	.24***	.26***	.03	.09	.07	.11	.21***
Phase 2 postdiction				-	.38***	.46***	.05	.04	.04	.07	.16**
Phase 1 accuracy					-	.68***	-.09	.00	.03	-.11	-.02
Phase 2 accuracy						-	.19***	.06	.04	-.11	.00
Phase 2 offloading							-	.09	.05	-.13*	-.12*
Perceptual accuracy								-	-.07	-.02	.00
Perceptual sensitivity									-	-.04	-.03
Perceptual confidence										-	.57***
Perceptual postdiction											-

Figure captions

Figure 1. Schematic illustration of the intention offloading task.

Figure 2. Path analysis model of the relationships between measures collected in Experiments 1a and 1b. Significant paths in both experiments are shown in black. Significant paths in Experiment 1b only are shown in red (note: these path weights were significantly different between the two experiments). Nonsignificant paths in both experiments are shown in gray.

Figure 3. Schematic illustration of the perceptual tasks. Note: each trial was preceded by an unfilled grid for 400ms (not shown).

Figure 4. A: Difficulty level in the perceptual tasks. Lower numbers indicate that the stimuli were more similar, hence a more difficult discrimination. Green shading indicates standard error of the mean. Results indicate that difficulty had plateaued by the time that experimental trials begin, with a slight reduction in difficulty before onset of experimental trials in the contrast task, potentially due to the additional demands of the metacognitive judgment. B: Accuracy in the two tasks, for the four confidence levels. Error bars indicate standard error of the mean.

Figure 5. Relationship between confidence in perceptual judgments and propensity to set reminders in the intention-offloading task. Participants were ranked in perceptual confidence by converting mean perceptual confidence and mean perceptual postdiction judgments into Z scores, then collapsing across these measures. Individuals were then allocated to five quintiles, from lowest to highest confidence, and the mean externalizing proportion in the intention-offloading task (i.e. propensity to set reminders) was calculated for each quintile. Error bars indicate standard error of the mean.

Figure 1

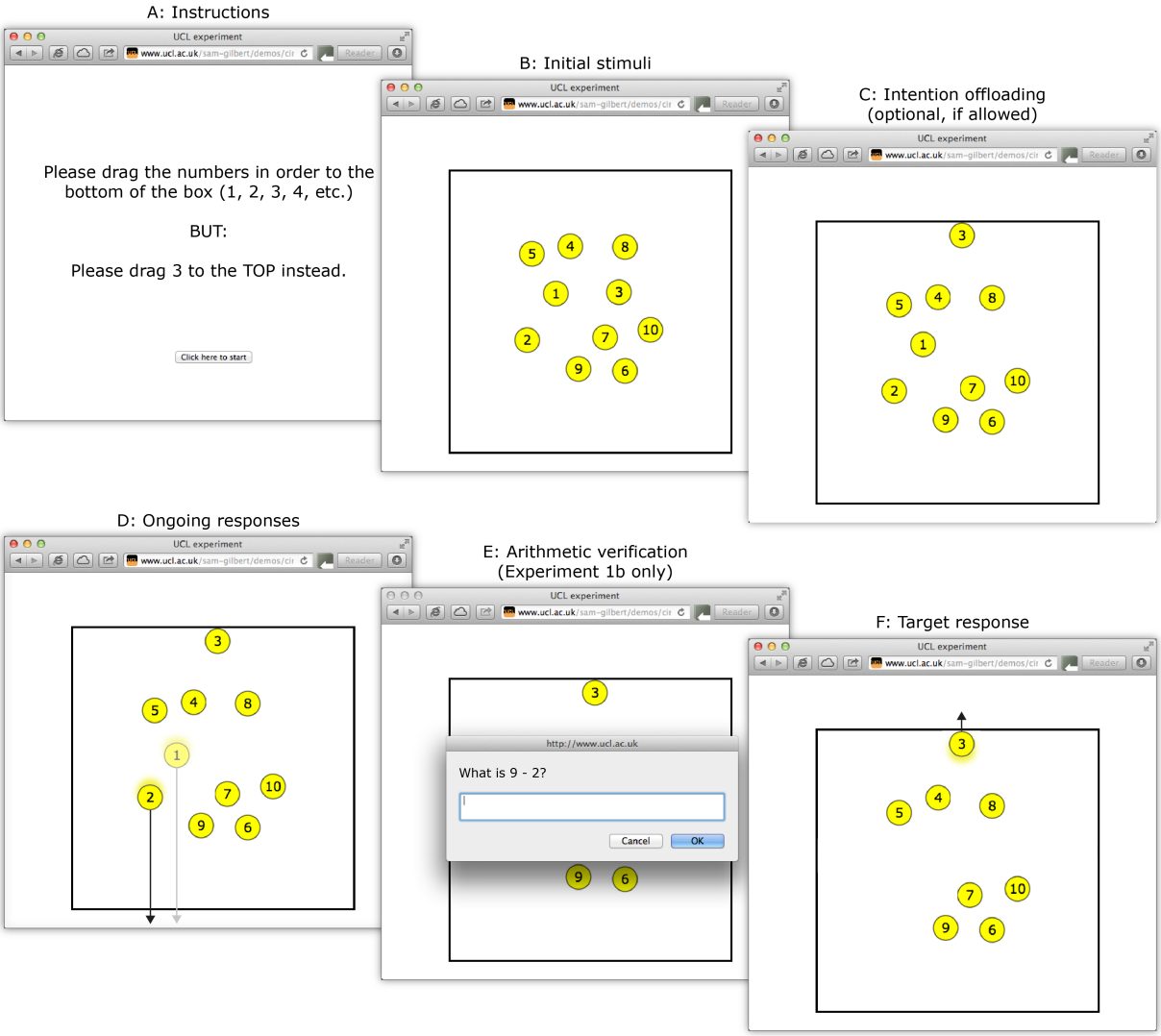

A: Instructions
UCL experiment
Please drag the numbers in order to the bottom of the box (1, 2, 3, 4, etc.)
BUT:
Please drag 3 to the TOP instead.
Click here to start
B: Initial stimuli
C: Intention offloading (optional, if allowed)
D: Ongoing responses
E: Arithmetic verification (Experiment 1b only)
http://www.ucl.ac.uk
What is 9 - 2?
Cancel
OK
F: Target response

Figure 2

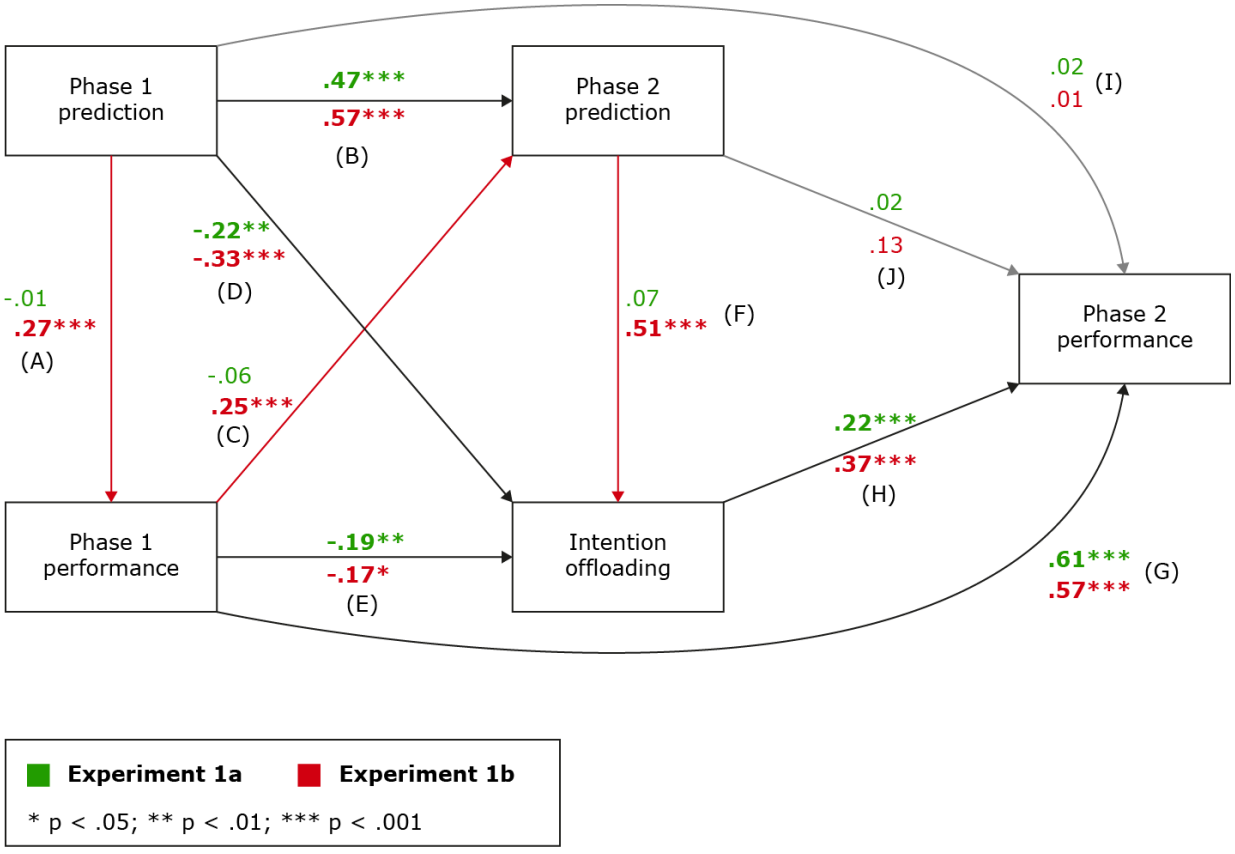

Phase 1 prediction
Phase 2 prediction
Phase 2 performance
Phase 1 performance
Intention offloading
.47***
.57***
(B)
.02
.01
(I)
-.22**
-.33***
(D)
.02
.13
(J)
-.01
.27***
(A)
.07
.51***
(F)
-.06
.25***
(C)
.22***
.37***
(H)
-.19**
-.17*
(E)
.61***
.57***
(G)
Experiment 1a
Experiment 1b
* p < .05; ** p < .01; *** p < .001

Figure 3

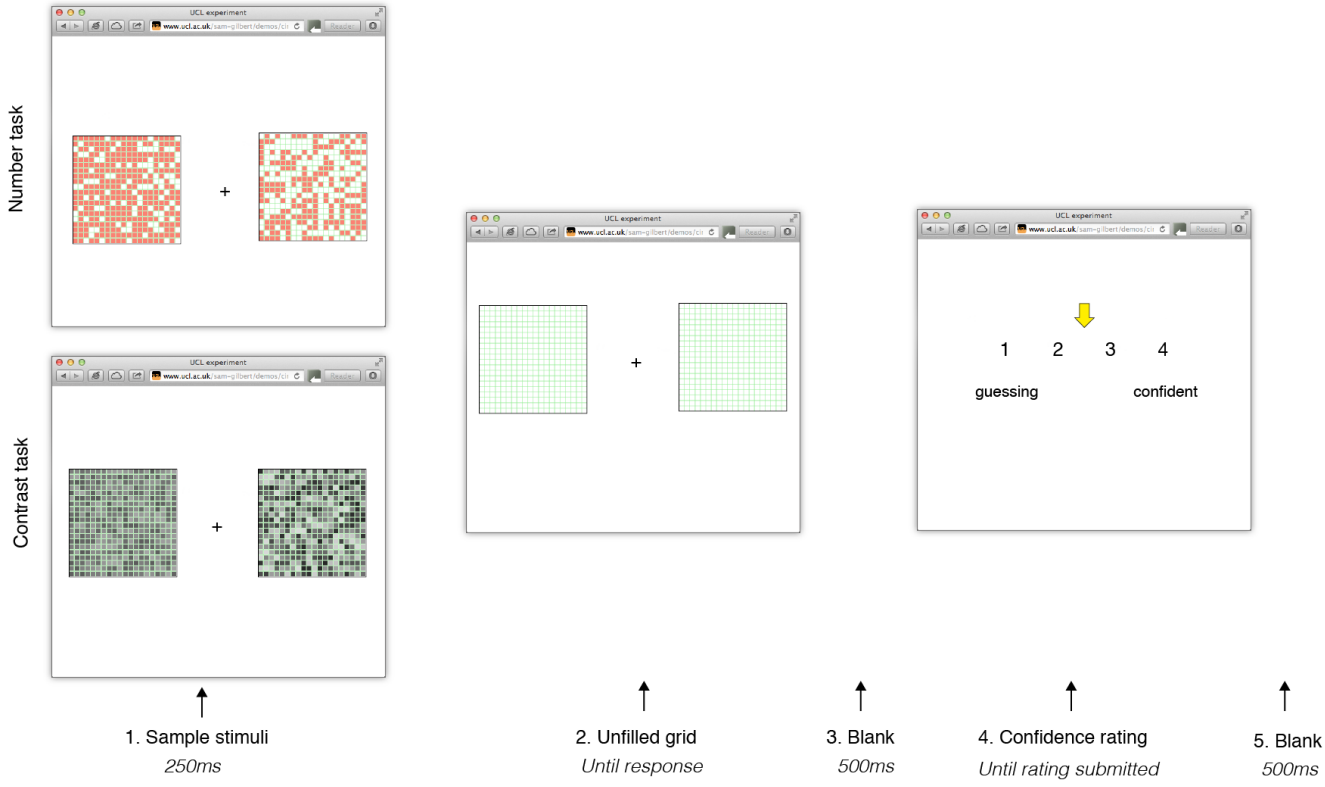

Number task
Contrast task
1 2 3 4
guessing
confident
1. Sample stimuli
250ms
2. Unfilled grid
Until response
3. Blank
500ms
4. Confidence rating
Until rating submitted
5. Blank
500ms

Figure 4

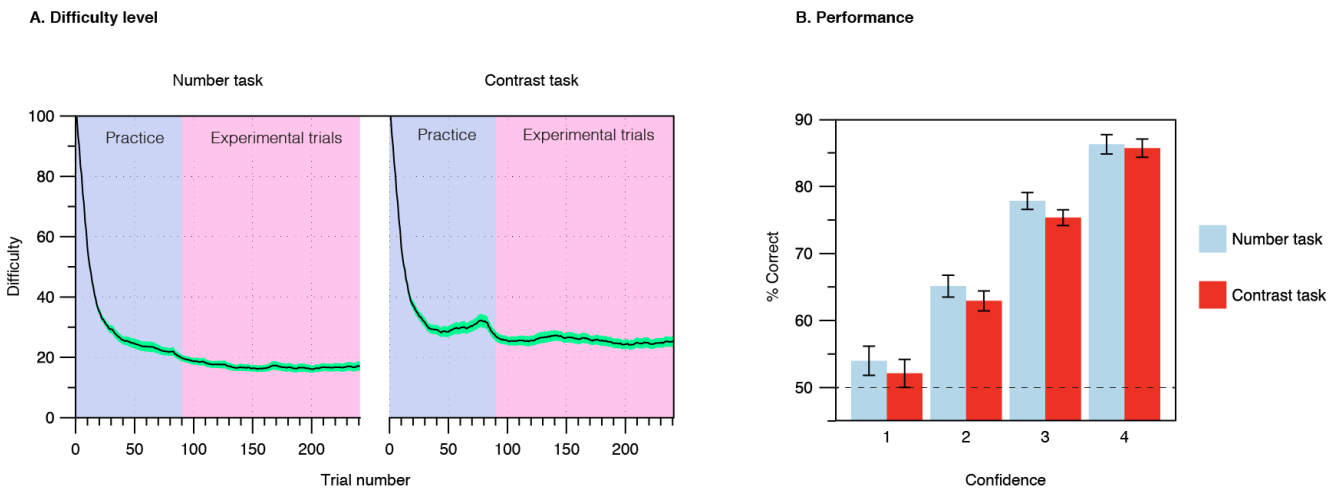
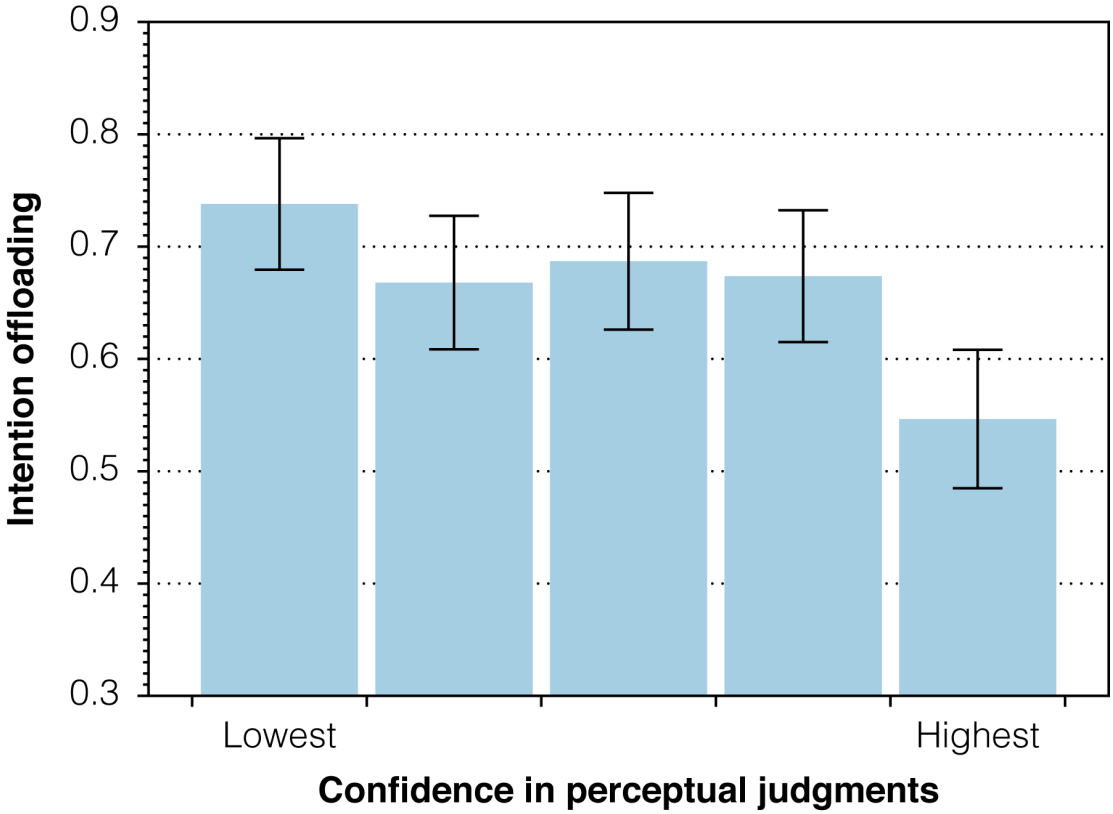

A. Difficulty level
Number task
Contrast task
Practice
Experimental trials
Practice
Experimental trials
Difficulty
100
80
60
40
20
0
0
50
100
150
200
0
50
100
150
200
Trial number
B. Performance
% Correct
90
80
70
60
50
1
2
3
4
Confidence
Number task
Contrast task

Figure 5


0.9
0.8
0.7
0.6
0.5
0.4
0.3
Intention offloading
Lowest
Highest
Confidence in perceptual judgments