2016-04-06

Evidence of a metacognitive benefit to memory?

Hollins, TJ

10.1080/09658211.2016.1171363
Memory
Informa UK Limited

Running Head: METACOGNITIVE BENEFIT

Evidence of a metacognitive benefit to memory?

Timothy J. Hollins

Plymouth University

Nathan Weber

Flinders University

Date of acceptance: 21/3/2016

Embargo period: 12 months

Note, this is not the document of record. The final version of the manuscript is available at

10.1080/09658211.2016.1171363

Address for correspondence:

Professor Tim Hollins,
School of Psychology,
University of Plymouth,
Drake Circus,
Plymouth,
PL4 8AA,
U. K.
Email: thollins@plymouth.ac.uk
Phone: +44 (0)1752 584803
Fax: +44 (0)1752 584808

Author's Notes

The author would like to thank Lowenna Wills for her help with data collection, and the Flinders University Norman Munn travel fund for partial support of this project. Some of the data reported here were presented at the Metacog2014 workshop in Clermont-Ferrand in September 2014.

Abstract

Studies of the memory control framework have contrasted free-report and forced-report recall, with little regard to the order of these two tests. The present experiment sought to demonstrate that test order is crucial, and that this suggests a potential role for metacognitive monitoring on memory retrieval. Participants undertook tests of episodic and semantic memory in both free- and forced-report format, in one of the two potential response orders. This showed that free-report performance was more accurate if conducted prior to forced-report, rather than after it, with no cost to memory quantity. Additionally, there was a trend towards higher forced-report performance if it was preceded by an initial free-report test, a pattern revealed by a meta-analysis to be consistent with previous studies in the literature. These findings suggest a reciprocal relationship between metacognitive monitoring and early retrieval processes in memory that results in higher memory performance when monitoring is encouraged.

Evidence of a metacognitive benefit to memory?

When answering to a question, we may think of an answer we believe to be wrong. This may constitute our best guess but we can choose to avoid making an error by not reporting it. This process of response-editing between memory retrieval and output formed the heart of Koriat and Goldsmith's (1996) memory control framework. In this original framework, the ability to moderate the accuracy of memory output depends upon the ability to monitor the likely accuracy of the best response that comes to mind and to then control output accordingly. If monitoring (or control) is imperfect then incorrect answers may be reported and correct answers withheld, with the frequency of such errors related to the willingness to report. Thus, a fundamental property of an editing process with imperfect monitoring is that any gains in memory accuracy (the proportion of volunteered responses that are correct) seen in free-report are accompanied by reductions in memory quantity (the number of correct responses volunteered), compared to that what would happen if all best guess answers were reported.

The first investigation into the ability to control and monitor the accuracy of memory using free-report methodologies was reported by Koriat and Goldsmith (1994). They contrasted forced-report and free-report performance using both recall and recognition as the criterion tests, for both semantic and episodic memory. For both memory domains, and for both test-formats, participants given the free-report option demonstrated a higher level of memory accuracy, but a lower number of correct answers reported than was observed in the forced-report conditions. Those in the free-report conditions were also subject to a second test phase, in which they provided their best guess answer to questions they had opted out of responding to. In line with the idea that low quality answers had been withheld, the accuracy on these questions was much below the accuracy observed for questions that

had been volunteered in the free-report phase. However, Koriat and Goldsmith (1994) argued that this was not the result of a change in memory accessibility, because the total number of items correct in the forced-report conditions did not differ from the total number correct across both phases for those who answered with free-report first. We return to this point later.

In 1996, Koriat and Goldsmith provided a formal treatment of their memory control framework, and further developed the methodology for investigating it. Their model specified three processing steps that lead to a response in a free-report memory test. The first is a retrieval phase which results in a best-candidate response to a given memory cue. This best candidate response is then subject to metacognitive monitoring which is measured by confidence in the accuracy of the candidate. Finally, there is a metacognitive control process that enables the decision to report or withhold that answer given the current task demands. Crucially, in this original framework memory quantity is not under direct metacognitive control. In contrast, accuracy is under strategic control, by varying the report criterion, but the success of this strategy depends upon the accuracy of the monitoring phase. If monitoring is less than perfect, then gains in report accuracy can only come with an associated loss in memory quantity.

Koriat and Goldsmith (1996) explored the nature of this quantity-accuracy trade-off in a series of simulations that varied three metacognitive parameters: 1) the extent to which mean confidence matched mean accuracy (calibration), 2) the extent to which variations in confidence across items predicted variations in response output (resolution), and 3) the distribution of confidence ratings across the confidence scale (distribution). These simulations demonstrated that the shape of the quantity-accuracy trade-off functions varied across these factors, but all shared the same fundamental principle: Increases in report

threshold led to accuracy functions that were monotonically increasing, and quantity functions that were monotonically decreasing, between performance asymptotes of zero and one. When people can perfectly discriminate which of their answers is correct versus incorrect then it is possible for people to increase their accuracy with no loss to quantity, at least in free-recall when the base-rate success for forced-guesses is essentially zero. However, in forced-choice recognition, having no memory renders performance at chance and so, if no guessing-correction is applied, gains in accuracy will be accompanied by reductions in quantity that are associated with the loss of correct guesses. It is also possible to observe losses in quantity with no gain in accuracy. This occurs when resolution is zero, and so the likelihood of withholding a correct response is equal to the overall proportion correct. In this case, people simply report fewer answers but become no more accurate. However, between resolution of zero and one, increases in accuracy are always associated with decreases in quantity.

Koriat and Goldsmith (1996) tested their framework by comparing free- and forced-report performance for the same general knowledge items by the same individuals in a two-step procedure. Experiment 1 contrasted forced-report with free-then-forced-report, replicating the methodology of their previous studies (Koriat & Goldsmith, 1994). More pertinent to the current work, Experiment 2 was the first study to directly contrast memory control decisions in two orders: As well as participants providing answers to questions they had previously passed (free-then-forced-report), a second group of participants were given the option to withhold answers they had already provided (forced-then-free-report). In both instances, the criterion tests were recall-based. Koriat and Goldsmith (1996) reported that both orders produced a quantity-accuracy trade-off, and moreover, that response order had "little or no effect" (page 504), other than producing a single interaction that applied to

deceptive general knowledge questions, an issue which is not relevant here. As a consequence of this observation, both response orders have been used in the literature subsequently (e.g., forced-first, Kelley & Sahakyan, 2003; free-first, Higham, 2002) with little attempt to distinguish them.

However, recently we (Perfect & Weber, 2012) reported a study which appeared to challenge the assumption that the two test orders are equivalent. Participants saw a brief video depicting a minor criminal act, before taking a 6-person identification test (henceforth *lineup*) for the person committing the crime. These tests were conducted under free- and forced-report instructions, with half the participants in the study taking the test in each order (free-first, forced-first), and half the participants in each condition seeing the lineup in target-present format (the perpetrator was present) and half in target-absent format (the perpetrator image was replaced by another foil). Of the 109 participants who did the free-report condition first, 52 responded *don't know*, 25 selected the perpetrator 32 made an error either by picking the wrong person or rejecting the lineup. Of the 110 who were initially forced to make a decision (with no *don't know* option), 25 selected the perpetrator and 85 made an error. That is, free-report led to the same number of correct responses, but at a much higher accuracy rate.

Participants then made the second decision. Following a free-report decision, those who had initially selected *don't know* were required to indicate their best guess response, and a further 12 participants selected correctly, such that forced-report after an initial free decision led to a total of 37 correct identifications (out of 109), compared to only 25 out of 110 correct identifications for those who had made an initial forced choice. Those who had made an initial forced decision were given the opportunity to withdraw responses they were uncertain of. Fifty three did so, including 8 who had initially selected correctly. Thus a

free-report decision after an initial forced choice led to fewer correct responses (17) than an initial free-report decision alone (25), even though the number withholding a response was almost identical in each condition (forced-first 53, free-first 52). Thus, it appears that an initial free-report decision leads both to higher accuracy for equal quantity (for the first decision), and higher quantity for a forced decision (when response bias is equated). The trade-off for the two response-orders in Perfect and Weber (2012) is depicted in Figure 1, which shows that, contrary to expectations, the two trade-off functions do not overlap.

As we discussed above, within the original memory control framework the absence of a cost to quantity with an increase in accuracy is possible only if two conditions apply: 1) monitoring is perfect, and 2) performance in the absence of memory is zero (i.e. guessing will not boost the score). If these two conditions hold, all incorrect answers can be withheld (thereby increasing accuracy), but when forced to provide responses to all questions, no guesses are correct (thereby producing no change in quantity). Neither of these two conditions applied in our study. The lack of perfect monitoring was clearly demonstrated by the within-subjects quantity-accuracy trade-offs seen in both response orders. Additionally, the study used a recognition-based procedure in which guessing would lead to above chance performance. Consequently, in our paper we speculated that an initial free-report response may have caused people to have based their memorial decisions on different evidence. We also noted that this pattern had been observed in a complex task in which participants must simultaneously judge (i) whether or not to pick someone from the lineup; (ii) and, if so, who; as well as (iii) whether to volunteer a response or respond *don't know*. In contrast, all laboratory experiments testing the memory control framework have involved either recall or only target-present materials in which a rejection response is never required. However, we were not in a position to speculate further, and so here we present two

additional studies designed to test the same ideas and to extend them beyond lineups for faces.

The motivation for the present work was to replicate and extend the pattern reported in Perfect and Weber (2012) because the findings appeared to contradict the original Koriat and Goldsmith (1996) memory control framework. However, recently, that framework has been refined further to include the possibility that late metacognitive monitoring processes may indirectly influence the contents of the early retrieval process. This framework, labelled the *Metacognitively Guided Retrieval and Report (Meta-RAR)* framework (Goldsmith, 2016), acknowledges the possibility of a feedback loop from output monitoring to retrieval, such that one potential outcome of a rejection of a best candidate answer is to reengage the retrieval process, perhaps utilising different retrieval cues (Goldsmith, 2016). A feedback loop such as this would be compatible with the findings reported in Perfect and Weber (2012), but the necessity for this component in the model remains to be convincingly demonstrated. Consequently, our present experiment, although originally designed as a test of the original memory control framework, also serves as potential demonstration of the effects of metacognitive monitoring on memory retrieval.

The study reported in Perfect and Weber (2012) was motivated in large part by the applied question of how monitoring can influence eyewitness identification decisions in lineups. Consequently, the design involved memory for faces, tested using a single trial with multiple faces in a lineup, and which included the explicit option to reject all response options provided (i.e. to reject the lineup). Half the participants saw a lineup which included the suspect seen previously, and half saw a lineup in which the suspect had been replaced with another foil. This procedure varies in many ways from the standard laboratory-based tasks previously used to test the memory-control framework, in which participants are

tested on many trials, for verbal materials, always with the correct answer present amongst the response options, and without the option to reject the set of options provided. In order to bridge the gap between the original demonstration by Perfect and Weber (2012), the current studies tested verbal memory for events and general knowledge across multiple trials, half of which omitted the correct answer, and all of which included option to reject all the choices provided.

Experiments 1a and 1b

In Experiment 1a, participants were given two response sheets, one in a free-report format, and one in a forced-report format, for two memory domains (eyewitness memory and general knowledge). However, when coding the data we noticed some participants had changed answers other than those they had originally responded *don't know* to, which made these items impossible to code within the memory control framework. We dealt with this in two ways. First, following Koriat and Goldsmith (1996), for each participant we only included items that were consistent across test formats, and calculated their performance proportionately. Second, we ran a second experiment (1b) which was identical to the first, except that we prevented participants from changing their answers. Because the experiments were otherwise identical we present them together.

Method

Materials

A series of small-scale pilot studies were used to develop test materials for tests of general knowledge and eyewitness memory. An initial set of questions was generated by the experimenters for a set of general knowledge items, and questions relating to a short US news-bulletin. These were given to 20 participants to answer in free-report format. Questions were then selected for which there were at least 4 erroneous answers provided.

Target present versions of the questions were then created by presenting the correct answer along with the 2nd, 3rd, and 4th most common error, whilst target absent versions of each replaced the correct answer with the most commonly reported error. In all cases "none of the previous" was included as the final response option. This process resulted in 14 general knowledge items (e.g. "Which of Disney's 7 dwarfs wears glasses?" a) Grumpy, b) Dopey, c) Doc, d) Sleepy, e) None of the previous), and 14 eyewitness memory items (e.g. "Where had the woman returning home from work been working?" a) Hospital, b) Laundrette, c) Restaurant, d) Children's Home, e) None of the previous).

Participants

All participants were recruited from the undergraduate volunteer panel at the University of Plymouth who participated in lieu of an assignment. No demographic data were collected, but the overwhelming majority of this population is aged between 18 and 22, and the majority (c. 70%) are female. Experiment 1a tested 50 participants, and Experiment 1b tested 26 participants.

Procedure

Participants were tested individually in a small testing booth. After giving their informed consent, participants were told that they would be shown a 5 minute clip from a local U.S. news bulletin and that their memory for what they saw and heard would be tested. After watching the clip they were first tested with a short general knowledge test, and then they completed the eyewitness test for the clip they witnessed. Both tests consisted of 14 multiple-choice items, in pencil and paper format, which participants completed at their own pace. In the forced-report version of each test, questions were accompanied by 5 options: 4 potential answers and a *none of the previous* response. In the free-report version,

an additional *don't know* option was added to this set. The free-report version of the test was accompanied by instructions that it was important not to make errors, and that participants should use the *don't know* option to avoid such errors. For half of the items in each test, the correct answer was present, and for half it was absent (i.e. *none of the previous* was the correct response), although participants were not informed of this. Allocation of target-presence to question was counterbalanced across participants.

Participants completed both free- and forced-versions of each test, but the order of these was counterbalanced across participants. The second version of each test was only introduced after the participant had completed both tests in the original format. The only difference between Experiments 1a and 1b was the way in which the second test was administered.

In Experiment 1a, the response sheets for the first test were removed, and participants were given a second response sheet, containing the appropriate response options, and they were simply asked to complete the response sheet once again (guessing where necessary in the forced-report condition, or using *don't know* to avoid errors in the free-report condition). In Experiment 1b, participants retained their original response sheets. For the free-report (following an initial forced answer), participants were instructed that they could decide whether they wished to report, or withdraw, each of their previous responses. For the forced-report (following initial free-report), participants were required to go back through each of their previous *don't know* responses and make a best guess answer from the remaining options. All these tests were self-paced.

Results

First, we analysed whether experimental method made any objective difference to performance. Consequently we compared Experiments 1a and 1b on the number of correct

responses in free-report, the number of incorrect responses in free-report and the number of correct responses in forced-report, separately for the GK and EM tests. There was no evidence of any significant change in performance on any measure (all t[47] <.96, all p's>.34), all d's <.28, and so for the purpose of the following analyses we collapsed across experiments to maximise power. We retained Experiment as a factor in all analyses to ensure that it did not moderate any of the results reported here.

We began by looking at the rate at which participants withheld an answer. The rate of withholding was analysed using a 2 (Materials: General knowledge vs Eyewitness memory) x 2 (Order: Free first, Forced first) x 2 (Experiment: 1a vs 1b) mixed ANOVA with repeated measures on the first factor. A higher proportion of responses were withheld if free-report followed after an initial forced-report (M=0.43, SE = .036) than if free-report was first (M=0.29, SE = .036), F(1,45) = 8.53, p=.005, MSe = .059, partial eta^2=.159, but no other main effects or interactions were significant (all F's<1, all p's>.42, all partial eta^2<.014).

Forced-report performance: With forced-report, quantity is equivalent to accuracy, and so only a single ANOVA was conducted, following the approach taken with withholding rate. This showed that performance was higher for general knowledge (M = 0.51, SE = .025) than for eyewitness memory (M = 0.40, SE = .021), F(1,45) = 18.19, p<.001, MSe = .017, partial eta^2=.296. The order of the two tests was non-significant, F(1,45) = 1.05, p=.310, MSe = .031, partial eta^2= .023 and no other main effects or interactions were significant, (all F's <1.45, all p's >.23, all partial eta^2<.032).

Free-report performance: When people can opt to respond when they wish, quantity and accuracy can diverge, and so we analysed free-report quantity and accuracy separately. For quantity, more correct answers were provided for general knowledge (M = 0.39, SE = .026) than for eyewitness memory (M = 0.32, SE = .021), F(1,45) = 6.19, p=.017, MSe = .020,

partial eta^2=.121. Also, more were reported if free-report was first (M = 0.41, SE = .027), than if it followed forced-report (M = 0.31, SE = .026), F(1,45) = 7.15, p<.001, MSe = .034, partial eta^2=.137, but no other main effects or interactions were significant (all F's <1, all p's> .39, all partial eta^2<.016). Analysis of free-report accuracy showed only a single main effect: responses were more accurate for general knowledge (M = 0.62, SE = .027) than for eyewitness memory (M = 0.50, SE = .028), F(1,45) = 9.92, p=.003, MSe = .033, partial eta^2=.181. There was no main effect of order, F(1,45) = .20, p=.660, MSe=.041, partial eta^2 = .004, and no other main effects or interactions were significant, (all F's <1, all p's > .53, all partial eta^2<.009).

The effects of order on the quantity-accuracy trade-off are illustrated in Figure 2, which shows that test order impacted upon number of answers reported correctly (quantity), but not the accuracy of the answers reported. The net result resembles the pattern previously reported by Perfect and Weber (2012), who reported that initial free-report performance was more accurate than initial forced-report performance, with no evidence of loss of quantity. This pattern was repeated here. Initial free-report led to answers that were more accurate than initial forced-report, F(1,45) = 13.40, p=.001, MSe = .03, partial eta^2 = .229, but there was no evidence of a significant, nor meaningfully sized, reduction in the number of correct answers, F(1,45) = .387, p=.537, MSe = .032, partial eta^2 = .009. Neither of these effects were qualified by any interactions with experiment, or materials, all Fs<1, all p's> .45. We return to this pattern below where we report the outcomes of analyses using Bayes Factor analysis.

Target presence vs absence: We also ran all the analyses above with the additional factor of target-presence vs absence as a within-subjects effect. This factor emerged as a simple main effect on all outcomes. People withheld an answer less often for target present

questions (M = .33, SE = .026) than for target absent questions, (M = .39, SE =.029), F(1,45) = 6.64, p=.013, MSe = .049, partial eta^2 = .129. More forced answers were correct for target present questions (M = .53, SE = .025) than for target absent questions (M = .35, SE = .023), F(1,45) = 33.80, p<.001, MSe = .021, partial eta^2 = .429. The same pattern is seen in free-report quantity, (target present, M = .44 , SE =.025; target absent, M = .25 , SE =.022) , F(1,45) = 55.75, p<.001, MSe = .028, partial eta^2 = .642, and free-report accuracy (target present, M = .68, SE = .028; target absent, M = .39 , SE = .035), F(1,45) = 53.2, p<.001, MSe = .054, partial eta^2 = .559. However, across these analyses, there was only a single interaction between target presence/absence and the other factors in the study (there was a significant 3 way interaction between target presence, materials and Experiment, F[1,45] = 8.67, p=.005, MSe = .054, partial eta^2 = .171) on free-report accuracy, but this made no theoretical sense, and so this is not pursued further. Crucially, target presence vs absence never interacted with report order in any analysis.

Bayes-Factor analyses of the report order effects on quantity and accuracy:

Concluding that report order influences quantity but not accuracy is problematic because it rests upon the contrast between a significant effect and a null effect. Consequently, we ran Bayes factor analyses using JASP (Love, Selker, Marsman *et al*., 2015) to determine the extent to which each of the effects supported or refuted the null hypothesis. Because materials had no impact upon the previous patterns reported, we collapsed across general knowledge and eyewitness memory tests. The significant order effect for free-report quantity produced a Bayes factor of 5.8, which is moderately strong evidence in favour of the hypothesis that the two orders are different. Specifically, the Bayes factor indicates the manner in which we should update our beliefs based on the observation of these data. Hence, our estimated odds that the orders are different (versus the same) should be 5.8

times as strong as they were before this observation. In contrast, the analysis of test order on free-report accuracy produced a Bayes Factor of 0.31, which is moderate evidence in favour of the null hypothesis. Specifically, our estimated odds of no difference (versus a difference depending on order) in free-report accuracy should be 3.2 times stronger than before observing these data. The net result of this pattern for the two report orders was that initial free-report performance compared favourably with initial forced-report performance: initial free-report performance was more accurate than initial forced-report performance (BF = 4.71), but there was evidence of no difference in the quantity of correct answers provided (BF = 0.35).

Discussion

Contrary to prior reports that response order has "little or no effect" (Koriat and Goldsmith, 1996, p.504), on memory or metamemory performance, the current study replicated the pattern first reported in Perfect & Weber (2012) in showing clear performance differences between the two orders. Free-report performance was superior when tested prior to forced-report rather than after it; more correct answers were reported, but there was no evidence of a change in the overall accuracy of responding. There are four candidate explanations for this pattern, two of which we believe to be incorrect, and two that remain plausible accounts in need of further investigation. We will begin with the two accounts we reject.

Account 1: The order effect is due to willingness to provide an answer.

One behavioural difference between the two response orders was that participants withheld more responses during free-report following forced-report than they did during initial free-report. It is therefore possible that a differential willingness to report items might

have caused the response-order effects illustrated in Figure 2. However, we believe that this is not the case, for two separate reasons.

First, we note that Perfect and Weber (2012) found a similar pattern, illustrated in Figure 1, despite no objective difference in willingness to report. Thus, it is possible to demonstrate order-effects without a change in willingness to respond across orders. However, we acknowledge that this is a relatively weak argument, and does not rule out the possibility that the experimental effects observed here are the result of such a shift. Consequently we decided to model the effects of withholding on quantity and accuracy of responses. The top panel of Figure 3 illustrates the separate effects of withholding rate on free-report quantity and accuracy.

In constructing the model we assumed no differences in either memory strength or monitoring across response orders. We further assumed that answers to questions were sampled from one of two distributions that differ only in their strength. If the answer is not known, then the strength of the chosen answer comes from a standard normal distribution (mean = 0, SD = 1), and if the answer is known, it comes from a standard normal distribution of mean strength *d*. That is, the parameter *d* is a measure of monitoring accuracy (cf. type-2 signal detection theory discriminability, e.g., Higham, 2002), because it is a measure of the discrimination between correct and incorrect responses, and this was fixed across orders. Participants were assumed to either know an answer, or to make a guess, with the proportion of trials known modelled by parameter *k*. The probability of a guess being correct was set as the reciprocal of the number of choices (i.e., 0.2). These parameters were fixed across response orders. In line with the Koriat and Goldsmith (1996) model, the strength of the selected answer is compared to a response criterion (Prc) and is then reported or withheld. We assumed that there were different response criteria (Prc_{Low} and

Prc_{High}) corresponding to the two test orders. We then ran an Excel Solver model to determine the best fitting model to the observed data, allowing these 4 parameters to vary, and seeking to predict the lowest sum of squared deviations from the observed measures of withholding rate, forced-choice quantity, free-report quantity, and free-report accuracy for each of the two conditions (i.e. 8 outcome variables). We ran this model multiple times from different randomised starting values between 0 and 1 for all parameters, and all runs converged on the same solution (d = 0.99, k = 0.30, Prc_{Low} = -0.22, Prc_{High} = .27). The solution predicted withholding rates of .28 (free then forced) and .41 (forced then free), closely matching the observed rates of .29 and .43. The resultant Q-A plot from the model is shown in the bottom panel of Figure 3.

It is immediately apparent that this figure does not resemble those shown in Figure 2, despite it being the best-fitting attempt to model those data. This is because this model assumes no differences in memory or monitoring accuracy and so variations in withholding rate result only in different extensions of the same underlying quantity-accuracy trade-off. That is, the two lines are always coincidental, and both are fixed at an identical performance on forced-report. Note that this is not because the model is insensitive to monitoring or memory; if we vary parameters d or k, then it is possible to produce Q-A plots with differing slopes (i.e., different trade-offs between quantity and accuracy). However, none of these variations can produce the observed dissociation between the two orders. Thus, we believe that withholding rate alone cannot explain the patterns reported by Perfect and Weber (2012) and replicated for both episodic and semantic memory tests in the current study.

Account 2: The order effect is a Type 1 error.

One possible explanation of these patterns is that the deviations observed in our studies represent Type 1 error, and that there is truly no difference in memory performance

across the two test orders. Again, we acknowledge this possibility, but we note that the 3 tests discussed to date - Perfect and Weber (2012) and the two tasks tested in the current work - all show a superiority for free-then-forced-reporting. The observation of this consistent pattern caused us to re-examine the original study by Koriat and Goldsmith (1996) which contrasted the two test orders using a paired-associate memory test. Intriguingly, although the original report found no significant differences across test orders, it did demonstrate a similar pattern, as can be seen in Figure 4 which is based upon the data from that study. [1]

In addition to the two experiments[2] we report here, the literature contains six other datasets that allow a direct comparison of forced-report performance between free-then-forced and forced-then-free-report orders in adult participants. Specifically, Koriat and Goldsmith (1994) reported three relevant comparisons in a recognition task and two in recall, and Koriat and Goldsmith (1996) reported the data depicted in Figure 4. For each of these, we calculated the effect size (Cohen's *d*) of the difference in forced-report proportion correct between forced-first and free-first. Figure 5 presents these effect sizes[3], and their 95% confidence intervals, with positive values reflecting a free-first advantage. Despite none of these differences being statistically significant (all *p*s > .2), the overall pattern is striking. While small, the direction of the difference is consistent across all 8 datasets and the

[1] We thank Morris Goldsmith for providing the original data file that enabled us to conduct this analysis. The figure shown reports only performance on standard items, and excludes deceptive items and difficult items for which recall performance was close to floor.

[2] We keep our Experiments 1a and 1b separate for the meta-analysis and Figure 5 as, despite no significant difference between them, the former displays one of the smallest effect sizes of the 8 studies and we do not want this to be masked by averaging with experiment 1b. Using only a single effect size for the combined data produces the same meta-analytic average effect size and only slightly changes the CI: $d = 0.26$ [0.05, 0.48].

[3] The *d*s were calculated using the MBESS (Lai & Kelley 2012) package in R, an open-source language and environment for statistical computing (R Core Team, 2013). Cumming's (2012) ESCI software was used to create the figure and for the meta-analytic calculations.

magnitude is similar. [4] Collectively they show a small memory advantage for forced-report after an initial free-report phase, d = 0.26 [0.04, 0.47]. Thus, the extant data clearly display evidence of a small, positive effect and we can confidently (at α = .05) reject the null hypothesis that the true effect size is zero.

Account 3: A change in resolution between response orders[5].

Increased accuracy of free-report relative to forced-report is a function of the ability to discriminate correct answers from incorrect answers. It is possible that is that this ability – resolution – is impaired if free-report follows forced-report relative to free-report carried out first. An alternate way of formulating the same idea is that following a forced-choice test, confidence in incorrect answers more closely resembles confidence in correct answers. This could happen for two potential reasons. Confidence in a correct answer could be reduced. This might happen if the foils appear more familiar because they were primed during the first test. A second possible mechanism is that confidence in erroneous choices is increased. This might occur due to confirmation bias. That is, having initially been forced to select an option may lead participants to have greater belief in that response on the subsequent test, and so to report errors that they might otherwise have withheld. Whilst this might be anticipated to effect correct and incorrect choices equally, this need not be so if correct choices are already close to ceiling in confidence. That is, increases in confidence due to confirmation bias may be greater for initially low-confidence responses, which are more likely to be errors. Either or both of these two processes could plausibly lead to more overlapping confidence distributions for correct and incorrect responses following an initial forced-report test. Consequently, any attempt to control accuracy by withholding responses

[4] There was no evidence of significant heterogeneity, Q(7) = 0.91, p = .996.
[5] We thank Maciej Hanczakowski for suggesting this potential account during the review process.

is likely to have a greater impact upon quantity for free-report after forced-report. That is, more withholding would be needed to achieve the same level of accuracy, with a greater cost to quantity, consistent with the pattern observed in the free-report data.

However, while the free-report pattern is consistent with this resolution account, the difference observed in the meta-analysis of the forced-report data is not. As forced-report performance does not depend upon resolution (although, cf. Perfect & Stollery's, 1993, demonstration that resolution does depend on memory quality), a resolution difference cannot account for consistently superior free-report performance when free precedes forced-report. The difference in forced-report quantity observed in the two research orders suggests the operation of a different process.

Account 4: A feedback loop from monitoring to retrieval

In the original memory control framework, if the best-candidate response is judged to be too weak, it is withheld, and no response is provided. However, another possibility is that when people initially generate a poor answer to the question, instead of opting to simply withhold that answer, they may instead seek a better answer. (Alternately, one can reformulate this argument by saying that forced-report responding encourages output of the first answer generated, rather than the best answer possible). This view is effectively one formulation of the idea of the feedback loop postulated in the Meta-RAR framework (Goldsmith, 2016). If this were the case, then it is possible that for a subset of items a second retrieval attempt may lead to a better answer being retrieved, and so a boost to memory quantity being observed. Crucially, this resampling account can only work if free-report is tested first, because if forced-report is the first test, then the only choice for the participant is to confirm acceptance of the forced-report option, or withhold it. Resampling is not an option. This view thus predicts three effects. It predicts that free report will show

superior performance when tested prior to forced-report rather than after it, and it predicts that free-report can actually lead to performance that is as good as, or perhaps even better than, forced-report. Both these predictions are supported by the work presented here, and in Perfect and Weber (2012). Also, it predicts superior forced-report performance when forced follows an initial free-report attempt rather than occurring first. Whilst no individual study has shown such a significant pattern, the meta-analysis shown in Figure 5 shows that this is the pattern that is generally observed in the literature.

Growing evidence for metacognitive influences on memory processes.

In summary, the present work provides further evidence for the role of metacognitive monitoring and control in memory performance. In particular, it confirms that the pattern of findings reported in Perfect and Weber (2012) was not an artefact of the particular design choices of that eyewitness study, but rather an example of a pattern that has been observed consistently. Although the theoretical basis for this pattern is not yet firmly established, these data are generally more supportive of the Meta-RAR framework (Goldsmith, 2016) than the memory control framework that preceded it. Additionally, the pattern of findings is more completely explained by a resampling hypothesis – consistent with the Meta-RAR framework – than an account based solely upon resolution, though we cannot rule out a role for resolution in the current pattern of data.

Our conclusions are in line with a number of studies that have recently demonstrated that metacognitive judgements can impact upon memory in a reactive manner, rather than simply providing an index of the strength of memory. For instance, Naveh-Benjamin and Kilb (2012) compared recognition memory performance by younger and older adults, with or without metacognitive monitoring in the form of recollection judgements. They found that metacognitive monitoring lead to better performance on the

recognition memory test, as measured by hits minus false positives, particularly for older adults. Soderstrom, Clark, Halamish, and Bjork (2015) showed that making judgements of learning during study lead to better subsequent memory performance for some but not all kinds of memory items. Likewise, Mitchum, Kelley and Fox, (2016) also found selective effects of making judgements of learning on subsequent test performance.

In sum, our results add to a growing body of research that metacognitive judgements should not be considered as simple epiphenomenal decisions made after retrieval has been attempted. Rather, they indicate that metacognitive decisions are an intrinsic part of the retrieval process itself, and can influence the quality of what is retrieved. Further, our work shows that different test formats differentially engage such metacognitive processes, and can result in differences in objective performance. As well as being theoretically important, these findings may be of some applied value when accurate recall is crucial.

References

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge

Goldsmith, M. (2016). Metacognitive quality-control processes in memory retrieval and reporting. In J. Dunlosky and S. K. Tauber (Eds.) *The Oxford Handbook of Metamemory* (pp 357-385), Oxford, Oxford University Press. doi: 10.1093/oxfordhb/9780199336746.013.28

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition, 30*, 67-80. doi: 10.3758/BF03195266

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory & Language, 48*, 704-721. doi: 10.1016/S0749-596X(02)00504-1

Koriat, A. & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General, 123*, 297–315. doi 10.1037/0096-3445.123.3.297

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490-517. doi: 10.1037/0033-295X.103.3.490

Lai, K. & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals. *British Journal of Mathematical & Statistical Psychology, 65*, 350-370. doi: 10.1111/j.2044-8317.2011.02029.x

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., Ly, A., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Rouder, J. N., Morey, R. D. & Wagenmakers, E.-J. (2015). JASP (Version 0.7) [Computer software].

Mitchum, A. L. Kelley, C., & Fox, M. C. (2016). Asking the question changes the ultimate answer: metamemory judgments change memory. *Journal of Experimental Psychology: General*, 145, 200-219. doi: 10.1037/a0039923

Naveh-Benjamin, & Kilb, A. (2012). How the measurement of memory processes can affect memory performance: the case of remember/know judgements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 194-203. doi: 10.1037/a0025256

Perfect, T. J. & Stollery, B. T. (1993). Memory and metamemory performance in older adults: One deficit or two? *Quarterly Journal of Experimental Psychology, 46A*, 119-135. Doi: 10.1080/14640749308401069

Perfect, T. J. & Weber, N. (2012). How should witnesses regulate the accuracy of their identification decisions: one step forwards, two steps back? *Journal of Experimental Psychology: Learning, Memory and Cognition, 38, 1810-1818*. doi: 10.1037/a0028461

R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Soderstrom, N. C., Clark, C. T., Halamish, V. & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory and Cognition, 41, 553-558. doi: 10.1037/a0038388*

Figure Legends

Figure 1: The quantity-accuracy trade-off reported by Perfect and Weber (2012), separately for participants tested with both orders of free- and forced-report testing. The dotted line labelled one-step shows the comparison of the two groups on the first test taken.

Figure 2: The quantity-accuracy trade-offs for Eyewitness Memory (top panel) and General Knowledge (bottom panel) tests, separately for participants tested with both orders of free- and forced-report testing. The dotted line labelled one-step shows the comparison of the two groups on the first test taken.

Figure 3: The top panel shows Quantity accuracy trade-off observed solely due to a change in withholding rate (rate of use of don't know option). Dashed vertical lines represent hypothetical criteria for responding for free-then-forced responding (low withholding rate), and forced-then-free responding (high withholding rate). The bottom panel shows the best-fitting Quantity-Accuracy plot for the two response orders differing only in withholding rate (Forced then free= 26.8% withheld. Free then forced = 41.2% withheld).

Figure 4: The quantity-accuracy trade-off for the non-deceptive items reported by Koriat and Goldsmith (1996) Experiment 2, separately for participants tested with both orders of free- and forced-report testing. The dotted line labelled one-step shows the comparison of the two groups on the first test taken.

Figure 5. Effect size (*d*) with 95% CI error bars for the forced-performance advantage following free report for adults. From the top, data are taken from: (a) Koriat & Goldsmith

(1994), Experiment 1, recognition; (b) Koriat & Goldsmith (1994), Experiment 2, recognition; (c) Koriat & Goldsmith (1994), Experiment 3, recognition; (d) Koriat & Goldsmith (1994), Experiment 1, recall; (e) Koriat & Goldsmith (1994), Experiment 3, recall; (f) Koriat & Goldsmith (1996), Experiment 2; (g) new data, Experiment 1a; (h) new data, Experiment 1b; and (i) meta-analytic average.

Figure 1

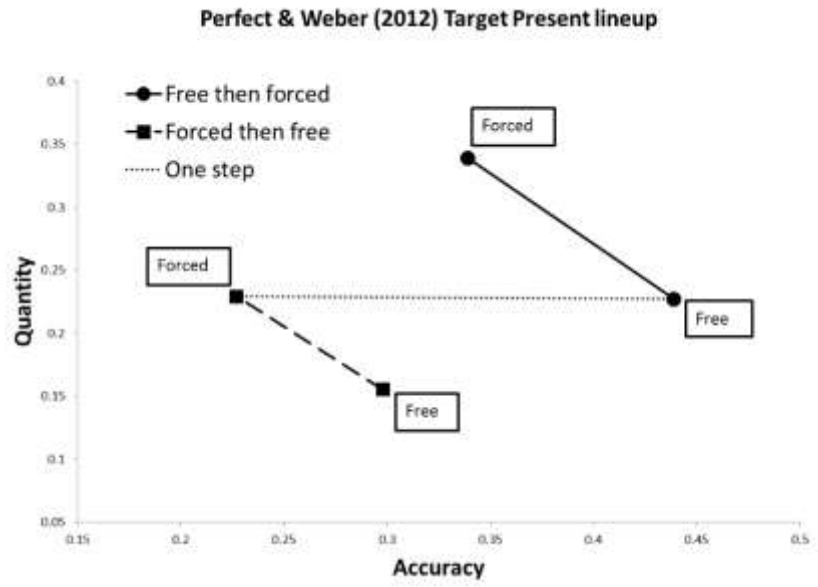

Perfect & Weber (2012) Target Present lineup
Free then forced
Forced then free
One step
Forced
Forced
Free
Free
Quantity
Accuracy
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5

Figure 2

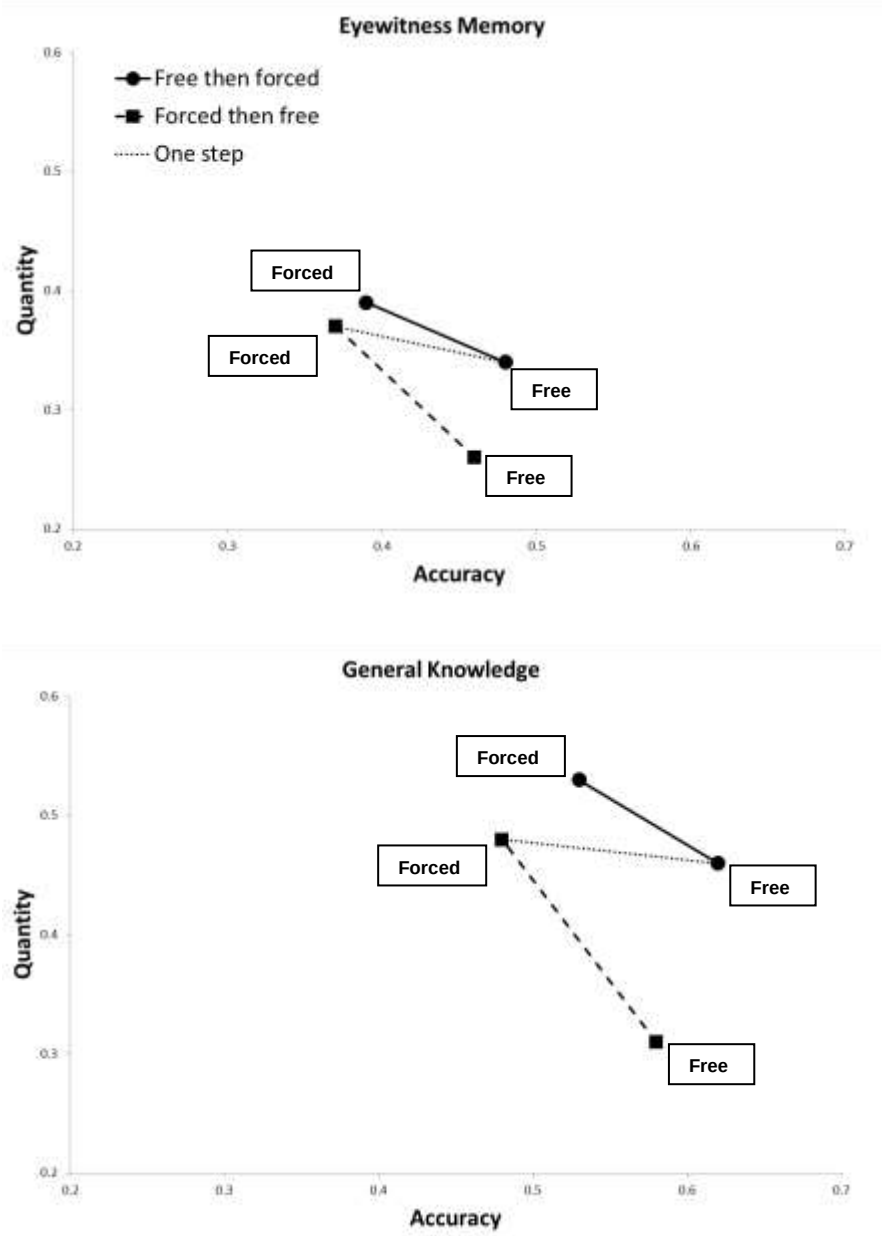
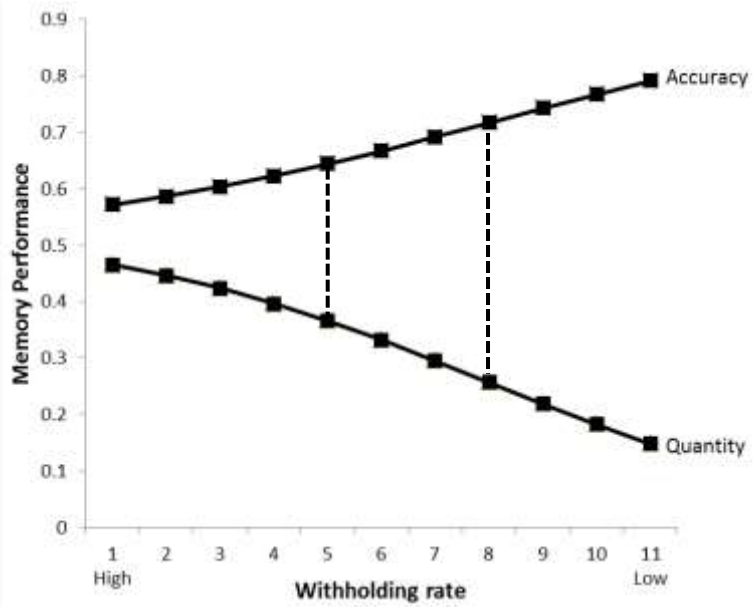

Eyewitness Memory
Free then forced
Forced then free
One step
Quantity
Accuracy
Forced
Forced
Free
Free
General Knowledge
Quantity
Accuracy
Forced
Forced
Free
Free

Figure 3


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Memory Performance
Accuracy
Quantity
1
2
3
4
5
6
7
8
9
10
11
High
Low
Withholding rate

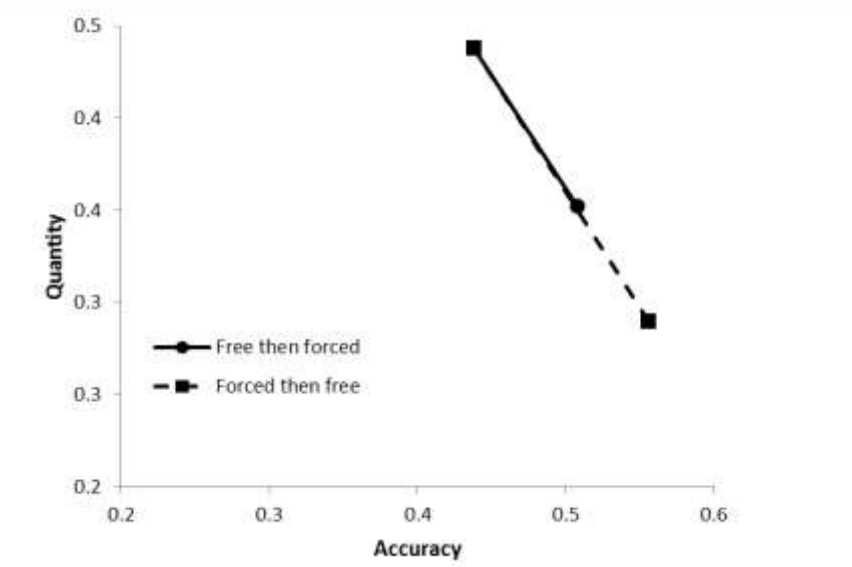

0.5
0.4
0.4
0.3
0.3
0.2
Quantity
Free then forced
Forced then free
0.2
0.3
0.4
0.5
0.6
Accuracy

Figure 4

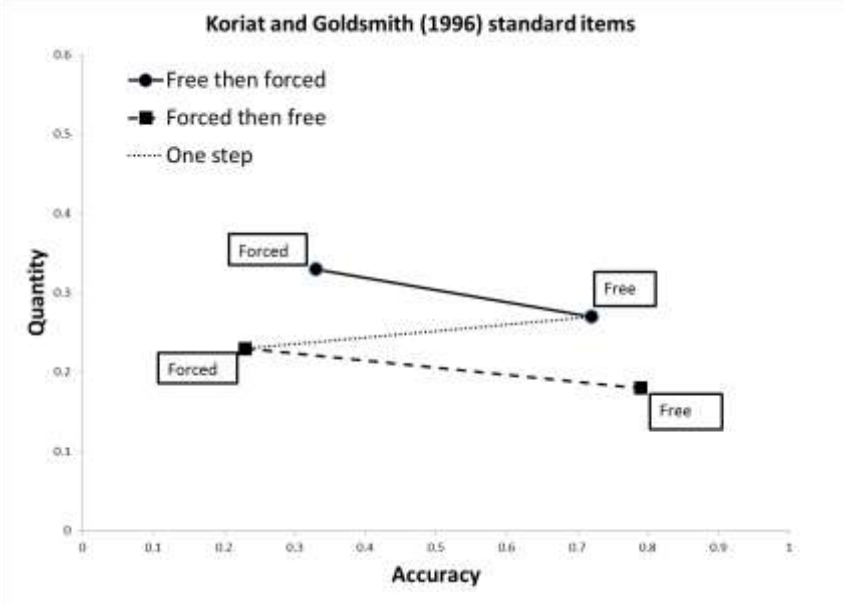

Koriat and Goldsmith (1996) standard items
Free then forced
Forced then free
One step
Forced
Free
Forced
Free
Quantity
Accuracy
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 5