

Pre-testing Effects Are Target-Specific and Are Not Driven by a Generalised State of Curiosity

Timothy. J. Hollins¹, Tina Seabrooke², Angus Inkster³, Andy Wills¹ & Chris J. Mitchell¹

¹ University of Plymouth

² University of Southampton

³ University of Cardiff

Author note

Correspondence should be addressed to Prof Tim Hollins, School of Psychology, University of Plymouth, Plymouth, PL4 8AA, U.K. Email t.hollins@plymouth.ac.uk. This work was supported by the Economic and Social Research Council [grant number ES/N018702/1]. We are grateful to Jack McColl for help with data collection. Upon acceptance of the manuscript, the data [and materials](#) from all experiments will be publicly archived at <https://osf.io/jw7yg/>.

Abstract

Guessing an answer to an unfamiliar question prior to seeing the answer leads to better memory than studying alone (the *pre-testing effect*), which some theories attribute to increased curiosity. A similar effect occurs in general knowledge learning: people are more likely to recall information that they were initially curious to learn. Gruber and Ranganath (2019) argued that unanswered questions can cause a *state* of curiosity during which encoding is enhanced for the missing answer, but also for incidental information presented at the time. If pre-testing similarly induces curiosity, then it too should produce better memory for incidental information. We tested this idea in three experiments that varied the order, nature and timing of the incidental material presented within a pre-testing context. All three experiments demonstrated a reliable pre-testing effect for the targets, but no benefit for the incidental material presented before the target. This pattern suggests that the pre-testing effect is highly specific and is not consistent with a generalised state of curiosity.

Keywords. Pre-testing, curiosity, episodic memory

Pre-testing Effects Are Target-Specific and Are Not Driven by a Generalised State of Curiosity

If asked an intriguing question, such as “*Which chilli is officially rated the hottest in the world?*” you may have two reactions. One, unless you know nothing at all about chillies, is that you might try to generate potential answers, perhaps even settling upon a single guess. You might also experience a degree of curiosity to learn the true answer. Recent research has suggested that both of these factors – 1) guessing the answer to the question¹ and 2) being highly curious to learn that answer – boost subsequent learning. In the present work, we examine the extent to which guessing and curiosity boost memory through a shared mechanism. Recent research suggests that curiosity boosts memory in a very general way; the state of curiosity is argued to result in better memory not only for the target answer, but also for other incidentally presented information. Our approach here is to see whether making a guess similarly boosts memory in this general way; will we see a memory benefit of guessing for incidental stimuli presented following the guess?

The Specific Effects of Guessing on Subsequent Memory

Studies investigating the impact of guessing on subsequent learning have used one of two broad methodologies. One of these, the *test-set* method, has consistently shown that the guessing benefit is specific to the target answer; it does not generalise to information about which participants did not make a guess. In this *test-set* methodology, participants first make all of their guesses in a pre-test phase. They are then presented with the study-phase in the form of an expository text, video-taped lecture or similar (e.g. Carpenter & Toftness, 2017; James & Storm, 2019; Little & Bjork, 2016; Richland et al, 2009; St. Hilaire et al, 2019). All of the answers to the questions asked in the pre-test phase are presented in this study phase, along with further information about which participants did not guess. The finding that the guessing benefit is specific to the guessed information cannot be explained by an increase in a general state of curiosity aroused by guessing – otherwise, memory would be boosted for all information presented at study.

¹ For those particularly curious, at the time of writing, Guinness World Records (2022) lists the Carolina Reaper as the hottest, but growers are continuously developing new strains of chilli.

However, it should be noted that the *test-set* method, in which all guesses are made and then all answers are given, is quite different to the studies in which the benefits of a state of curiosity have been observed. In these studies curiosity-inducing question and their associated answers are presented on a trial-by-trial basis. –One possibility, that we discuss later, is that curiosity is a relatively short-lived state. In this case, curiosity might produce effects within a trial that do not survive the longer, filled delays used in the *test-set* method. We turn now to the second method that has been used to study the effects of guessing – *the item-based* approach – which is much more similar to the work conducted to examine the effects of curiosity.

In contrast to the *test-set* method, the *item-based* approach uses a trial-by-trial task. Hence, on each trial, participants are asked a question (the pre-test) that targets a single fact. They then make their guess, following which, the answer is presented as feedback. These guess trials are contrasted with study-only trials. Here, the question and answer are presented together with no pre-test (e.g. Clark et al., 2021; Cyr & Anderson, 2015; 2018; Grimaldi & Karpicke, 2012; Hays et al., 2013; Huelser & Metcalfe, 2012; Knight et al., 2012; Metcalfe & Huelser, 2020; Potts et al., 2019; Potts & Shanks, 2014; Seabrooke, Hollins et al., 2019, Seabrooke, Mitchell et al., 2019; Seabrooke, Mitchell & Hollins, 2021; Seabrooke, Mitchell, Wills et al., 2021; Seabrooke, Mitchell, Wills, Inkster et al., 2021; Zawadzka, & Hanczakowski, 2019).

In contrast to the *test-set* method, the question of generalisation of the guessing benefit to incidental material – which is central to our current concerns – has been little studied using the *item-based* method. This is because, in the *item-based* method, which requires presentation of the correct answer immediately after each question, it is less natural to present incidental material than it is in the *test-set* method (which might use expository text). However, to foreshadow the discussion below, it is both methodologically possible and theoretically desirable to test generalisation using an *item-based* approach. One salient reason is, as we shall see below, that the *item-based* approach uses a very similar method to that used to successfully reveal generalized effects of curiosity on subsequent memory (Gruber & Ranganath, 2019). The curiosity paradigm

involves participants initially attempting to answer an unfamiliar question (i.e. guessing), before rating their curiosity, seeing incidental material, then finally seeing the answer to the question. In essence, this resembles a pre-testing trial (a guess followed by feedback), with the addition of a curiosity rating and incidental material. Thus, an obvious question to ask is whether the generalised effect seen in curiosity research is a result of the initial guess. Below, we begin with a provide a more detailed treatment of the *item-based* methodology that is the focus of the current study, and then turn our attention to the curiosity literature.

Kornell et al. (2009) first developed a version of the *item-based* methodology to explore the benefits of testing on the learning of information that was previously unknown to participants. In their study, participants were asked to guess potential associates of weakly-related pairs (e.g., *freckle* - ?) before subsequently seeing the “correct” associate that had been selected by the researchers (*mole*). Relative to studying the word pairs intact from the outset, initial guessing led to superior cued-recall for the target on a final criterion test. This *pre-testing* effect has subsequently been replicated repeatedly, with differing theoretical accounts developed to explain the patterns observed (e.g. Clark et al., 2021; Cyr & Anderson, 2015; 2018; Grimaldi & Karpicke, 2012; Hays et al., 2013; Huelser & Metcalfe, 2012; Knight et al., 2012; Metcalfe & Huelser, 2020; Potts et al., 2019; Potts & Shanks, 2014; Seabrooke, Hollins et al., 2019, Seabrooke, Mitchell et al., 2019; Seabrooke, Mitchell & Hollins, 2021; Seabrooke, Mitchell, Wills et al., 2021; Seabrooke, Mitchell, Wills, Inkster et al., 2021; Zawadzka, & Hanczakowski, 2019). Rather than discussing these different accounts, we instead highlight something that they share. With four exceptions that we cover below, all these *item-based* studies have tested memory specifically for the material presented as the “correct” answer after a guess, relative to memory for the same items studied without guessing.

Consequently, the theoretical accounts developed have all sought to explain why guessing improves memory for the *designated target item*. What has been less explored, and so less considered from a theoretical perspective, is whether *item-based* guessing might lead to a generalised change in learning state that would benefit *anything* encountered after a guess. Without

wishing to propose any particular theoretical position at this point, it may be that guessing is more interesting, motivating, or curiosity-inducing than merely studying the same information, and being in such a state would improve learning of any information, relative to the rather dull control condition of study-only. While some have theorised that such states might explain the pre-testing effect (e.g., Potts, et al., 2019; Potts & Shanks, 2014), they have done so only with respect to memory of the corrective feedback itself and have not explored whether there is a wider benefit to learning after a guess.

The hypothesis that guessing produces a generalised benefit to memory appears at first to be incompatible with research showing null effects for cue-target pairs that are semantically unrelated (Grimaldi & Karpicke, 2012; Seabrooke, Mitchell et al., 2019). However, this null effect occurs only when the final criterion test is cued recall. When the final test is item recognition, a pre-testing effect is observed even for unrelated materials (Seabrooke, Hollins et al., 2019, Seabrooke, Mitchell et al., 2019; Seabrooke, Mitchell & Hollins, 2021; Seabrooke, Mitchell, Wills et al., 2021; Seabrooke, Mitchell, Wills, Inkster et al., 2021). These findings are therefore compatible with the idea that guessing could result in a generalised state that boosts the subsequent recognition of any information encountered in that state, not just the target of the original guess.

Three of the four studies to explore generalisation of item-based pretesting have shown that pre-testing boosts memory for the question as well as the answer. Hays et al (2013) followed a standard pre-testing design using weakly-related word pairs. As well as showing the standard pre-testing effect for the target from the cue, they also reported that pre-testing boosts cued recall of the cue when presented with the target at test. A similar conclusion was reached by Seabrooke, Hollins et al., (2019) who used a pre-testing design to test memory for rare-word definitions (e.g. *roke – mist*), and found that pre-testing boosted recognition separately for both the target and the cue, but not their association. In Pan et al., (2019) participants either studied triplets of weakly-related words (e.g. *gift, rose, wine*), or initially tried to guess one of the triplet from the remaining

two cues (e.g. *gift, rose, ?*) prior to seeing the correct answer presented as feedback. At test, participants had to recall the missing item from a pair of cues that either matched the original pre-test format (e.g. *gift, rose, ?*) or mismatched it (e.g. *gift, ?, wine*). Relative to studying intact triplets, pre-testing boosted both the target that appeared as feedback after the pre-test (*wine*), and the non-tested member of the triplet (*rose*). Thus, collectively, these studies demonstrate that pre-testing using the item-method generalises to improved memory for the cue, as well as the target of the guess.

However, unlike studies using the *test-set* approach, none of these *item-based* studies has explored whether item-based pre-testing boosts memory for other potential answers presented after a guess. The only study to look at this question to date is Seabrooke, Mitchell et al. (2019). In their study, participants saw novel faces and attempted to learn four facts about each person. These facts consisted of exemplars from four fixed categories – occupation, hobby, favourite food, and best friend’s name. On Study-only trials, participants saw the four exemplars on screen together for a fixed period. On Pre-Test trials, participants guessed two of the exemplars before all four facts appeared simultaneously (for the same duration as the Study-only trials). Thus, on Pre-Test trials there were pre-tested targets, but also studied targets that were not pre-tested, but were associated with guesses that occurred on that trial. If guessing triggered a generalised learning state, then these items should have been better recognised than the equivalent items on Study-only trials. This did not happen, however, suggesting that the effects of pre-testing were highly specific to the answers presented as feedback, rather than providing a general boost to motivation or attention that improved learning in general. However, for reasons explained below, this conclusion may have been premature, and we revisit this issue in the present work.

The Generalised Effect of Curiosity on Subsequent Memory

There is strong evidence that curiosity predicts subsequent learning. In particular, people show better memory for the answers to general knowledge questions that they are more curious to

learn. For example, Kang et al. (2009, Experiment 2) presented participants with 40 general knowledge questions that had been pre-tested to evoke a range of curiosity levels. On each trial, participants saw the question, silently guessed an answer, and then indicated their curiosity about the correct answer and their confidence in their guess. Immediately afterwards, they saw the question and the correct answer for 10 seconds. After 11-16 days, participants returned for a surprise memory test. Participants were most likely to remember the answers to questions that they were most curious about (for similar results, see also Gruber et al., 2014; Swirsky, et al., 2021; Van de Cruys, et al., 2021).

One difficulty in interpreting these studies is that there was no experimental control over which items induced high or low curiosity, or what people did in order to make their curiosity judgements. Of particular relevance is the potential role that guessing may play in judgements of curiosity, and the impact that any guessing may have on subsequent memory. For example, Kang et al. (2009) explicitly instructed participants to silently guess an answer to every question, perhaps reflecting Loewenstein's (1994) observation that guesses provide a direct way to engage curiosity. Thus, it is possible that the improved memory for high-curiosity items reflected a form of the pre-testing effect (stemming from the guesses), rather than curiosity alone². This is likely to be true even in studies that did not explicitly require participants to guess. It is hard to imagine how anyone could judge their curiosity about a general knowledge question without first attempting to answer it, that is, to guess. Indeed, the typical approach in this literature is to exclude those questions for which the participants already know the answer, implying that each question must elicit an attempt at an answer. This leaves open the possibility that the subsequent memory benefit associated with high-curiosity may result from the initial guess, as reported in the *pre-testing effect* literature. In turn, this raises the question as to the extent to which the curiosity and pre-testing literatures have been

² One potential critique of this argument is that guesses were evoked for both high and low curiosity items. However, a counter-argument is that low-curiosity items may correspond to those questions for which the participant is unable to generate any plausible guesses, perhaps because the topic is so unfamiliar.

studying the same fundamental phenomenon, at least in respect to the subsequent boost to memory.

While this general question remains open, one recent theoretical account of the effects of curiosity on memory makes a unique prediction that is untested in the pre-testing paradigm. According to the PACE model (Gruber & Ranganath, 2019), curiosity represents a motivational state in which heightened hippocampal activation improves encoding of any information encountered whilst in a state of high curiosity, relative to low curiosity. This theoretical position accounts for several demonstrations that high curiosity improves memory for incidental information as well as the target of the curiosity. For example, in the second phase of Gruber et al.'s (2014) Experiment 1, participants saw general knowledge questions they had previously rated as inducing high or low curiosity. On each trial, participants then saw an incidental facial photograph before the answer appeared, and they judged whether they thought that the person depicted would be knowledgeable about the topic of the question³. They then saw the correct answer to the initial question. In the final criterion memory tests that occurred a day later, participants showed superior recall for the facts associated with high-curiosity questions, and superior recognition of the faces associated with high-curiosity questions, relative to low-curiosity questions. This pattern has since been replicated (Galli, et al. 2018; Murphy et al., 2021; Stare, et al., 2019), although one failure to replicate has also been reported (Fandakova & Gruber, 2020).

The benefit of curiosity on incidental memory is the crucial evidence in support of the claim that curiosity represents a generalised *state*, which drives learning via increased attentional processing and enhanced memory encoding and consolidation (the PACE framework: Gruber & Ranganath, 2019). A crucial aspect of this account is that the state only exists prior to the presentation of the target information that induced the curiosity. Once a person learns the correct answer, there is no longer an information gap, and so the state of curiosity no longer exists

³ A subsequent study by Murphy et al. (2021, Experiment 2) showed the same benefit for incidental faces without the requirement to judge whether the person depicted would know the target fact.

(Loewenstein, 2014). Consequently, if participants attend to the incidental material after the target information, then no benefit to the incidental material is expected.

Does Pre-Testing Induce a State of Curiosity that Benefits Incidental Material?

We are now in a position to explain why Seabrooke, Mitchell et al.'s (2019) conclusions may have been premature. In their study, participants guessed two answers before seeing all four facts simultaneously. Thus, it is possible that when the answers appeared, participants may have first attended to the details they were curious about – the answers they were pre-tested on – thereby boosting memory for these facts. Consequently, the incidental facts may have been encoded only once the participants were no longer in a state of curiosity. If this were the case, then the PACE model (Gruber & Ranganath, 2014) would not predict a memory advantage for the incidental facts relative to Study-only trials, in line with what Seabrooke, Mitchell et al. (2019) reported.

Consequently, in the present work, we set out to test whether pre-testing generates a general state which produces a benefit to memory for any item presented while the participant is in that state. In three experiments, participants either studied cue-target pairs (Study-only trials) or guessed the target from a cue (Pre-test trials). They then encountered incidental material prior to seeing the answer. Because we were keen to reduce the interval between the guess and the subsequent presentation of the incidental material, we did not ask participants to rate their curiosity, but instead rely on previous demonstrations that guessing increases curiosity. The key question for all experiments was whether pre-testing produces a generalised memory benefit for incidental information, in line with the prediction from the PACE model (Gruber & Ranganath, 2019).

Experiment 1

The aim of Experiment 1 was to determine whether the memory benefit associated with pre-testing extends to other material encountered prior to the corrective feedback. To achieve this, we adapted the procedure used by Seabrooke, Mitchell et al. (2019), by having participants study faces

associated with three facts (rather than four), one of which was guessed (rather than two). This allowed us to control the position of the guessed item in the study sequence that followed the guess, with the target appearing first, second, or third in the sequence. This enabled us to look at memory for the incidental items both before the presentation of target item (while any general state of curiosity should still be active) and after the presentation of the target item (when the curiosity would have dissipated). We also included Study-only trials, in which the participants did not provide a guess, to control for order effects across the sequence.

An incidental benefit of this paradigm is that it allowed us to explore another aspect of the guessing benefit for the *target* that has not yet been studied. In the pre-testing paradigm, participants see the correct answer directly after their guess, either alone (Grimaldi & Karpicke, 2012; Kornell, et al., 2009; Potts & Shanks, 2014; Seabrooke, Hollins et al., 2019; Seabrooke, Mitchell & Hollins, 2021; Seabrooke, Mitchell, Wills et al., 2021; Seabrooke, Mitchell, Wills, Inkster et al., 2021; Zawadska & Hancwakowski, 2018), or jointly with other facts (Seabrooke, Mitchell et al., 2019). In contrast, in the current design, participants made a guess about a specific target, and then encountered the answer either immediately, or following presentation one or two incidental facts, without making any further guesses. This enabled us to test whether the benefit of guessing for target memory survives the interpolation of irrelevant study material.⁴

Method

Participants

Forty undergraduates from the University of Plymouth took part for partial course credit, although one participant's data were lost because of a computer malfunction. No other demographic data were collected. This sample size has good power (90%) to detect a medium-sized

⁴ Previous research looking at the effect of delayed feedback (e.g. Grimaldi & Karpicke, 2012; Hayes, et al., 2013; Kornell, 2014; Vaughn & Rawson, 2012) involved delaying corrective feedback across multiple trials. They therefore confounded delay with the presentation of other cues, generation of other guesses, and learning of other facts. Additionally, all used cued recall as the criterion test. We omit discussion of this literature because of the many differences with the present work.

within-subject effect at Cohen's $d_z = 0.5$ ⁵. All participants spoke fluent English. Experiments 1 and 2 were approved by the University of Plymouth ethics committee, and Experiment 3 was approved by the University of Southampton ethics committee.

Materials

The materials were based upon those used in Seabrooke, Hollins, Waters et al. (2019). Thirty-six exemplars from each of three categories (jobs, hobbies, and foods) were used. One quarter of the exemplars from each category were randomly paired with a unique face to be presented at study, with the remaining exemplars serving as foils in the final multiple-choice recognition test. Three of these trials served as practice items, with the remainder used in the main experiment. Each set of exemplars was randomly paired with one of 27 photographs of unfamiliar people taken from DeBruine and Jones (2017). The experiment was programmed in PsychoPy (Peirce, et al., 2019) and presented on a 22-inch monitor in a laboratory environment. Figure 1 provides a schematic representation of how these materials were deployed through the experiment.

Procedure

Participants were told that they would be asked to learn three exemplar facts about a series of unfamiliar people (their job, their hobby and their favourite food), and that on some trials they would be asked make a single-word guess for one of the facts before all were revealed. The participants were also instructed to try to remember the facts as they appeared, in anticipation of a later test. The participants completed three practice encoding trials, consisting of two Pre-test trials and one Study-only trial. They then completed 24 main encoding trials, which consisted of 18 Pre-test trials and six Study-only trials. This ratio was used to ensure that there were equal numbers of

⁵ There is no straightforward way to calculate power for the interaction of within-subject factors (Povin & Schutz, 2000); our sample size provides 90% power to detect a medium-sized within-subjects difference between two conditions (the average effect size in psychology is approximately 0.5; Bakker et al., 2012) . Consequently, as well as reporting the omnibus ANOVA results, in all experiments we also report follow-up tests of our key hypotheses using within-subject t-tests, for which our power calculations are appropriate.

each item type at test. Study-only trials (SSS trials) were all equivalent, but Pre-test trials were split into 3 sequential orders; Guess 1st (GSS trials), Guess 2nd (SGS trials) and Guess 3rd (SSG trials).

Each trial began with a fixation cross for three seconds, followed by a photograph of an unfamiliar person for two seconds, which was presented centrally towards the top of the screen. On Study-only trials, the participant saw the face, together with one selected exemplar fact from each of the 3 categories, presented sequentially for 2 seconds each, with a 2 second interval between each fact, and the order of fact presentation randomised. On Pre-test trials, participants were initially cued by a photograph and category cue to guess the corresponding fact about the person, by typing in their response. Responding was self-paced, but was limited to 30 seconds maximum. After the guess, the three exemplar facts to be learned about the person were presented as in Study-only trials. The order of these details was counterbalanced, such that the target answer to the guess appeared equally either first, second or third in the sequence.

All four trial types were randomly intermixed through the list, with a 1 second gap between each trial. Immediately after the study phase, there was a practice old-new target recognition test phase using the targets from the 3 practice encoding trials, followed by the experimental test trials. All 72 studied exemplar facts (3 per face) were tested in random order. On each trial, the test photograph appeared with a question specifying which target fact was being tested (job, hobby, or favourite food). Four multiple choice answers were provided: the target and three unfamiliar foils not used elsewhere in the experiment, with allocation to location on the screen randomised on a trial-by-trial basis. Participants responded by clicking on their chosen answer with the mouse. This terminated the trial, and there was a 0.5 second gap before the next trial appeared. Responding was self-paced, and the test terminated once all trials had been completed.

Results

The data in this experiment, and all subsequent experiments, were analysed using R (R core team, 2021). Bayes Factors were calculated using the BayesFactor package (version 0.9.12-4.2:

Morey et al., 2015). Where appropriate, we report Greenhouse-Geisser adjusted degrees of freedom to correct for violation of the assumption of sphericity in our repeated-measures ANOVAs. In all cases, we had a strong a-priori hypothesis that Pre-test trials should lead to better recognition memory than Study-only trials, both for the target facts and the incidentally studied material, and so we report one-tailed test results throughout.

During the encoding phase, the participants provided a guess on almost all Pre-test trials ($M = 99.57\%$, $SEM = 0.32\%$). Across all participants on all trials, four participants correctly guessed one target fact each during the encoding phase. These targets were removed from the test dataset for those participants.

Figure 2 shows the mean percentage of correctly recognised targets at test, separated according to Encoding Condition (GSS, SGS, SSG, or SSS) and the Position (first, second, or third) that the target appeared at encoding. An Encoding Condition \times Position repeated-measures ANOVA revealed that Encoding Condition did not impact recognition accuracy, $F(2.86, 108.5) = 1.95$, $MSe = 266.1$, $p = .12$, $\eta_g^2 = .009$, $BF_{10} = 0.18$, while the effect of Position was significant, although the Bayes Factor was indeterminate, $F(1.96, 74.38) = 5.02$, $MSe = 210.9$, $p = .009$, $\eta_g^2 = .01$, $BF_{10} = 1.32$. There was a significant interaction between the two factors, $F(4.94, 187.9) = 4.95$, $p < .001$, $MSe = 268.7$, $\eta_g^2 = .04$, $BF_{10} = 219.09$.

To explore this interaction, we first considered the effect of pre-testing versus study on target recognition. To this end, we collated the pre-tested targets *only* on GSS, SGS, SSG trials and collectively compared them to targets that were presented on SSS (Study-only) trials. We also considered the position that each target occurred at encoding (i.e., the guessed target was presented first on GSS trials, second on SGS trials, and so forth). An Encoding Condition (Pre-tested or Studied targets) \times Position condition (first, second, or third) repeated-measures ANOVA revealed a significant main effect of Encoding Condition, $F(1, 38) = 21.17$, $p < .001$, $MSe = 250.3$, $\eta_g^2 = .07$, $BF_{10} = 20016$. That is, the participants showed superior recognition of targets that they had incorrectly

guessed at encoding ($M = 88.9\%$, $SEM = 1.42\%$) compared to targets that they had just studied on SSS trials ($M = 79.3\%$, $SEM = 1.86\%$). Position in the sequence did not influence recognition, $F(2, 76) = 1.05$, $p = .35$, $MSe = 183.2$, $\eta_g^2 = .005$, $BF_{10} = .076$, and there was no Encoding Condition \times Position interaction, $F < 1$, $BF_{10} = .08$. Although there was no interaction, we also carried examined the magnitude of the pre-testing effect at each position in the sequence. The pre-testing effect was significant at each position, with an indeterminate Bayes Factor at position 1, but with Bayes Factors favouring the experimental hypothesis at Positions 2 and 3 (Position 1, $t(38) = 2.40$, $p = .02$, $BF = 2.16$, Position 2, $t(38) = 3.00$, $p = .0048$, $BF = 7.73$, Position 3, $t(38) = 3.20$, $p = .0028$, $BF = 12.47$).

Finally, we examined recognition of studied targets on Pre-test trials when the guess trial appeared second in the sequence. This allowed us to contrast incidental material encountered *before* the guessed target with those presented *after* the guessed target, again using the equivalent items on Study-only trials as a control for order effects. An Encoding Condition (SGS vs. SSS) \times Position (first vs. third) repeated-measures ANOVA revealed no difference between incidental targets presented first ($M = 81.2\%$, $SEM = 2.22\%$) and third ($M = 79.9\%$, $SEM = 2.44\%$), $F(1, 38) = 3.62$, $MSE = 283.7$, $p = .065$, $\eta_g^2 = .02$, $BF_{10} = 0.91$, although the Bayes Factor was indeterminate. Additionally, there was no significant main effect of Encoding Condition, $F(1, 38) = 0.19$, $MSe = 331.0$, $p = .66$, $\eta_g^2 = .001$, $BF_{10} = 0.19$ and no Encoding Condition \times Position interaction, $F(1, 38) = 0.47$, $MSe = 241.9$, $p = .50$, $\eta_g^2 = .002$, $BF_{10} = 0.28$, with the Bayesian evidence supporting the null in both cases. That is, there was clear evidence for equivalent performance on incidental items appearing before and after the presentation of the correct answer. We sought to confirm this using a Bayesian t-test comparing the magnitude of the pre-testing effect at Positions 1 and 3. This confirmed that there was no difference, $t(38) = .69$, $p = .50$, $BF = 0.21$.

Discussion

The findings of Experiment 1 conceptually replicate those of Seabrooke, Mitchell et al. (2019) in showing a robust boost to recognition memory for pre-tested facts that appeared as corrective

feedback to a guess, whether compared to facts from Study-only trials or facts studied along with guesses on Pre-test trials. This was the case regardless of the sequential order of the test items. That is, targets presented as feedback to guesses were better recognised than equivalent study items, whether they were the first, second or third in the test sequence.

There ~~was also clear~~ evidence also suggested that the enhanced memory for the target did not generalise to other items. Study items that appeared first on Pre-test trials were no better recognised than equivalent study items from Study-only trials, despite appearing before the answer associated with the guess. There was a suggestion in the data from the Pre-test trials that incidental items studied first (before the target appeared) were recognised slightly better than those studied third (after the target answer had been shown). However, this comparison is confounded with presentation order, and the same pattern was observed for Study-only trials, indicating that the difference was due to order, not the temporal relationship with guessing. That is, there was no evidence for better memory associated with a generalised state that lasts until the information gap is closed.

In summary, these data support are consistent with the conclusion originally suggested by Seabrooke, Mitchell et al. (2019) that the benefits of guessing are specific to the sought-after answer. However, although these findings are clear, it is worth noting that there are potentially important methodological differences between Experiment 1 and the studies that have found support for the PACE model by showing an incidental memory benefit for faces seen before a target fact (Gruber et al., 2014; Murphy et al., 2021; Stare et al., 2018). Experiment 1 is also atypical with respect to the pre-testing effect literature, in that most experiments present a single fact per trial, rather than three facts as we did here. Consequently, Experiment 2 used a more typical pre-testing methodology with a single target per trial, and with the introduction of incidental faces to remember prior to the presentation of feedback, as in the studies by Gruber and colleagues.

Experiment 2

In Experiment 2, we made three substantive changes to the methodology used in Experiment 1, to make it closer to previous studies demonstrating an effect of curiosity on learning incidental information. The first is that, in Experiment 2, we used faces as our incidental material encountered between a guess and the corrective feedback, as has been used in all the studies of curiosity described above. Faces are unlikely to compete with memory for the answers to the general knowledge facts, because they are unrelated to the original cue, and they are a different class of stimuli. The second change is that we had participants guess or study only a single fact per trial, which is more typical of studies in both the pre-testing and curiosity literatures. The final change to Experiment 2 is that we set a tighter time restriction in which participants had to make a guess (7 s). We were concerned to give enough time for participants to generate and then type in their guess, while at the same time not creating too long an interval between their guess and the presentation of the face. We return to this issue in Experiment 3.

Method

Participants

Forty-four participants completed the experiment via Prolific (www.prolific.com). Participants were recruited on the basis that they spoke English as a first language and were aged between 18 and 60. The participants had to complete the experiment using a laptop or desktop computer. We asked, but did not insist, that participants completed the experiment using either Google Chrome, Microsoft Edge, or Firefox, because these were the browsers that we had developed and checked the experiment with. Before the experiment, we excluded participants that reported using another browser ($N = 1$), participants that said that they did not speak fluent English ($N = 0$), participants that failed our “bot check” question ($N = 1$; see below), and participants that admitted using additional memory aids during the experiment ($N = 1$). The final sample consisted of 41 participants (27 females, 14 males), who were aged between 18 and 53 years ($M = 30.10$ years, $SEM = 1.48$ years). This sample size has good power (88%) to detect medium-sized effects at $d_z = 0.5$. Each participant received £5 for completing the study.

Materials

Eighty-eight word pairs, consisting of rare English words and their common English definitions, appeared during the experiment. Each participant saw a random selection of 40 word pairs on either Study-only or Pre-test trials (20 word pairs each). A further 40 word pairs served as foils during the target recognition test. The remaining eight pairs appeared on practice trials (2 Guess, 2 Study-only and 4 foils).

A further 88 unique photographs of men and women were presented during the experiment and were selected from the same database as Experiment 1. Each participant saw 40 randomly selected photographs split equally across Guess and Study-only trials. A further 40 photographs served as foils, and the remaining eight photographs were presented on the practice trials (2 Guess, 2 Study-only and 4 foils). The experiment was programmed using jsPsych (de Leeuw, 2015).

Procedure

Participants were directed from the Prolific website to an online study, where they read a participant information sheet and provided online consent by clicking a button. They then gave their Prolific identification number, the browser they were using to view the study, and their age and gender. Participants also confirmed whether they spoke fluent English, and then answered a simple question that aimed to screen out bots and participants that were not paying attention to the task. Here, participants saw a 4 x 4 grid that contained a unique letter in each cell. Their task was to select the one letter that was presented in red (all other letters were black). Before starting the main task, the programme switched to full-screen, and participants were encouraged to a) turn off music, cell phones, and other devices that might be distracting, b) complete the experiment in one sitting, and c) keep the experiment in full-screen and avoid visiting other webpages during the study.

Before the encoding phase, participants were told that they would be presented with a series of rare English words, and that their task was to remember the common English definition of those words. They were told that they would be asked to guess the English definitions of some trials,

and that, while it was very important that they guessed the definitions, it did not matter whether their guesses were right or wrong. The participants were also told that, on all trials, a photograph of a person would be presented part-way through the trial, and that they should try to remember both the definitions and the photographs that were presented. All participants then had to agree that they would complete the study without using any memory aids (e.g., writing the word pairs down or recording the presentation). Participants who did not agree to this request were unable to proceed with the experiment.

The participants completed a practice demonstration of each phase of the experiment (encoding task, face recognition task, and target recognition task) before starting the proper encoding phase. The practice encoding task consisted of two Pre-test and two Study-only trials, presented in a randomly determined order for each participant. On Pre-test trials, participants were presented with a rare English word and an equality sign (e.g., *spoffish* =). An input box appeared to the right of the equality sign, below the request, "Please guess the definition". The participants had seven seconds to type in their guess as to the definition of the rare English word, before the display cleared. A photograph 200x200 pixels in size of an unfamiliar person with a neutral expression then appeared centrally for three seconds. Finally, the photograph was replaced by the complete word pair (e.g., *spoffish* = *fussy*), which was presented for a further three seconds. On Study-only trials, the complete word pair (e.g., *mechlin* = *lace*) was initially presented for seven seconds, to match the guessing phase of Pre-test trials. The photograph then appeared in an identical fashion to the Pre-test trials before the complete word pair appeared for a further three seconds, to match the feedback phase of the Pre-test trials. The trials were separated by one second intervals.

After the practice encoding phase, the participants completed practice rounds of the face and target recognition tests. The practice face-recognition test consisted of eight trials, the order of which was randomly determined for each participant. The four faces that were presented during the practice encoding phase, plus four novel faces, were presented centrally and individually, and

participants had to determine whether they had seen the photograph earlier in the study or not, by choosing between “Yes” and “No” options. A response was required on each trial before the participants were able to progress, with participants guessing if necessary. The practice target recognition test followed the same format as the face recognition test, except that the photographs were replaced by eight common English words (the four targets presented during the practice encoding phase, plus four novel words).

After completing the practice target recognition test, the participants completed the main encoding phase, which consisted of 20 Pre-test and 20 Study-only trials. They then completed the main face recognition test, followed by the main target recognition test. Both recognition tests consisted of the 40 photographs/targets that were presented during the main encoding phase, plus 40 novel foils (photographs/targets). In each phase of the experiment, the order of trials was randomly determined for each participant. Upon completion of the target recognition test, the participants were asked to confirm whether they had used any memory aids during the study (e.g., writing the words down, or recording the presentation), and received a written debrief. The entire experiment lasted approximately 30 minutes.

Results and Discussion

Across all participants, five targets were guessed correctly on Pre-test trials during the encoding phase. These targets, plus the photographs that were presented on those trials, were removed from the target and face recognition test datasets, respectively. On average, the participants submitted a guess on 80.85% ($SEM = 4.72\%$) of Pre-test trials.

During the target recognition test, the participants were very good at correctly rejecting the foils as “new” ($M = 90.49\%$, $SEM = 1.88\%$)⁶. Figure 3a shows the mean percentage of correctly identified hits to “old” Guess and Study-only targets. As expected, the participants correctly

⁶ It is not possible to associate new items with the pre-test vs study manipulation, and so we do not know the false-positive rate associated with each condition. Consequently it is not possible to calculate signal detection measures of d' or c across this manipulation. Instead, our analysis focuses on hit rates associated with pre-tested vs studied items.

recognised more targets from the Guess condition than the Study-only condition, $t(40) = 5.90$, $p < .001$, $d_z = 0.92$, $BF_{10} = 202.1$.

In the face recognition test, the participants were generally good at identifying the foils as “new” ($M = 77.80\%$, $SEM = 2.24\%$). As shown in Figure 3b, however, the participants correctly identified more Study-only targets as “old” than Pre-test trials, thereby contradicting our hypothesis, $t(40) = 2.69$, $p = .99$, $d_z = -0.42$, $BF_{10} = 0.08$. Thus, the Bayesian analysis strongly refutes our hypothesis that a pre-testing effect would benefit recognition for the faces.

In summary, Experiment 2 largely confirmed the pattern seen in Experiment 1. Once again, a robust pre-testing effect for the target was observed, which survived the interpolation of incidental material between the guess and the presentation of the corrective feedback. However, also consistent with Experiment 1, there was no evidence of a benefit to that incidental material, despite using faces. In fact, we observed the opposite effect to that predicted, with superior recognition of the faces associated with study-only items. Because this was unexpected, we reserve further discussion of this pattern for incidental memory for faces until we have presented Experiment 3, which was designed to address whether the null effect was caused by the delay between the guess and the subsequent presentation of the faces in our first two experiments.

Experiment 3

As mentioned earlier, the benefit to incidental material associated with high-curiosity questions first reported by Gruber et al. (2014) has been replicated by Galli et al. (2018) with younger and older adults and by Stare et al. (2019) using both immediate and delayed final tests. However, Fandakova and Gruber (2020) failed to replicate this pattern. To examine this discrepancy, Murphy et al. (2021) manipulated the duration of the interval between the presentation of the curiosity-inducing question and the incidental face. Their first experiment showed a boost to the recognition of the faces associated with high curiosity questions only if they had been presented 1s after the offset of the curiosity-inducing question, and not when the delay was 7s. Experiment 2

explored a range of intervals from 2s to 8s, and showed a monotonic decline with interval length, such that there was a robust recognition boost of 15.8% for a 2s interval, which fell to 5.5% for a 4s interval, and did not differ from zero at 6s or longer. This pattern is entirely consistent with the previous set of studies: those reporting reliable boosts to incidental memory for faces used intervals of 4s or less between the offset of the curiosity-inducing questions and the presentation of the incidental faces, whereas Fandakova and Gruber (2020) used a 7s interval⁷.

This new observation raises a potential concern with the methodology used by Seabrooke, Mitchell et al. (2019), and in Experiments 1 and 2 here. In Seabrooke, Mitchell et al. (2019), participants had to make two self-paced guesses, which introduced an uncontrolled delay for the first item, as participants made their second guess. Additionally, participants also indicated their motivation to learn each fact after generating guesses but before seeing the correct answers, thereby introducing a further delay between the guess and the onset of the incidental information, which was again self-paced. Consequently, in their work, it is hard to determine the interval between the guesses and the subsequent presentation of the incidental material, and there is a strong possibility that the intervals were more than 8 seconds.

The same issues apply, albeit to a lesser extent, to the first two experiments reported here.⁸ These were designed to test the idea that incidental material encountered *before* the target of the guess would benefit from a state of curiosity, but they did not ensure that this incidental material was encountered within 4 s of the initial question offset. Experiment 3 was designed to address this issue.

There are several differences in the timing of the incidental material between our procedure and that used in the curiosity paradigm (Fandakova & Gruber, 2020; Gruber et al., 2014; Murphy et al., 2021; Stare et al., 2018). Experiment 3 was based upon the methods of Murphy et al.'s (2021)

⁷ One curious aspect of Murphy et al. (2021) is that although they reported the impact of a delay between the guess and the incidental face on memory for the faces, they did not report whether this impacted memory for the target fact itself, or if there was an overall curiosity effect for those facts.

⁸ Experiments 1 and 2 were designed and run before the publication of Murphy et al. (2021).

Experiment 2, in which high- and low-curiosity questions appeared for four seconds, with no participant response required. After the offset of the question, there was a variable interval between two and eight seconds before the incidental face appeared.

Although Murphy et al. (2021) reported their timings with respect to the offset of the question, the precise time-course of the curiosity state is unclear because the question was on screen for four seconds. We do not know how long it takes to read the question and curiosity to be evoked. Consequently, Murphy et al.'s (2021) description that the incidental memory benefit is strongest 2 seconds after question *offset* could equally be described as being strongest 6 seconds after the question *onset*. That is, curiosity peaks at somewhere between 2 and 6 seconds after a question appears, and diminishes to zero in around 6 seconds thereafter (i.e., between 8 and 12 seconds after question onset). It is worth noting that in our Experiment 2, the faces appeared immediately after the 7 second guessing period, measured from question onset, and so Experiment 2 used a delay that may have fallen within the period that produced a curiosity effect in Murphy et al.'s (2021) Experiment 2. Nevertheless, we felt it wise to run a further experiment in which the interval between the question and incidental faces was more carefully controlled, and systematically targeted to fall in the critical periods identified by Murphy et al. (2021).

The way we achieved this was by reducing the time needed for participants to indicate their guess prior to the presentation of the incidental faces. In line with previous work on the pre-testing effect, our first two experiments involved participants typing their guess in full. In Experiment 1, responding was self-paced, and in Experiment 2, we set a maximum response time of 7 seconds. This allowed participants to generate and type their response, which allowed us to exclude correct Pre-test trials from further analysis. This means the gap between the initial guess coming to mind, and the subsequent presentation of the incidental face is unknown, and covers a wide time range. To overcome this problem in Experiment 3, we minimised the typing element of the guessing phase by asking participants only type the first letter of their guess. This then triggered the interval before the

face appeared. Participants then completed typing their guess, using the first letter they had already provided, which allowed us to exclude correct guesses, as before.

In addition, we manipulated the delay between the participants' guesses and the onset of the incidental faces. Once the first letter key had been pressed, the screen cleared and an incidental face appeared either two or eight seconds later, thereby targeting periods in which incidental benefits should be observed (2s delay) or not (8s delay). Ten seconds after the offset of the cue, participants completed typing in their guess, starting with the first letter cue they had typed earlier. The remainder of the Experiment replicated Experiment 2.

Method

Participants

Forty-seven participants were recruited as in Experiment 2, but three were excluded for using non-approved browsers. One further participant was excluded because, on Study-only trials in which the participants were asked to copy the first letter of the target that was presented to them, they only selected the correct answer on one out of 12 occasions. The final sample consisted of 28 females and 15 males, who were aged 19 and 60 years ($M = 35.23$ years, $SEM = 1.60$ years). This sample size provides over 90% power to detect the effect size reported by Murphy et al. (2021) for the incidental memory benefit at 2 second delay ($d_z = 0.6$). The participants received £2.20 for their participation.

Materials

Fifty-six rare English words and their definitions were selected as word pairs for the experiment. Fifty-six unfamiliar faces were also selected. Both the word pairs and the faces were a subset of those used in Experiment 2. For each participant, the word pairs and faces were randomly selected for presentation in either the practice or main experiment phases. They were also randomly

allocated to be presented on Pre-test or Study-only trials, or as foils in the target or face recognition test. Other aspects of the materials were as in Experiment 2.

Procedure

The initial set up for the participants (demographic questions and initial instructions) was the same as in Experiment 2. The participants completed a practice demonstration of each phase of the experiment (encoding task, face recognition task, and target recognition task), as in Experiment 2, before moving on to each phase in order. As in Experiment 2, the experiment pushed the task into full screen after the demographic questions, and the participants were asked to keep the task in full screen throughout the experiment.

On Pre-test trials, the participants were first shown the cue and a question mark (e.g., *imprecation = ?*) along with an instruction to “*Guess the definition and type the first letter of your guess*”. The participants had unlimited time to complete this initial task, and the task did not progress until a response was entered. This response triggered a delay of either two or eight seconds, during which time the screen was blank until a fixation cross appearing centrally for the last second of the delay. A face then appeared centrally for two seconds, followed by a further delay of either eight or two seconds to equate the total trial time across conditions. Finally, the cue was presented along with the first letter of the participant’s guess (e.g., *imprecation = c*) along with the instruction to “Please complete your guess now”. The participants had six seconds to complete their guess, before the complete word pair (e.g., *imprecation = curse*) was presented for three seconds.

On Study-only trials, the cue and target were initially presented together (e.g., *imprecation = curse*), along with an instruction to “*Type the first letter of the definition*”. The participants had unlimited time to complete that task, before a delay was presented (as on the Pre-test trials), followed by the presentation of the face (2 seconds) and a second delay. The timings of the second delay matched those of the Pre-test trials, with the addition of six seconds to equate for the time spent prior to typing the first letter of their guess. Thus, the length of second delay was either eight

seconds (if the first delay was eight seconds) or 14 seconds (if the first delay was two seconds). After the second delay, the participants studied the complete word pair for a further three seconds. The trials were presented in a random order for each participant and were separated by one-second intervals. The participants completed four practice trials (two Pre-test and two Study-only trials, one of each with a long and short delay before the face presentation). The main encoding phase consisted of 12 Pre-test and 12 Study-only trials. Half of the trials within each encoding condition had a long delay (eight seconds) before the face presentation, and the rest had a short delay (two seconds).

The face and target recognition tests followed the same format as those in Experiment 2. Each practice test consisted of eight trials, with the four targets/faces from the practice encoding trials intermixed with four foils. The main face and target recognition tests consisted of 48 trials consisting of the 24 targets/faces from the encoding phase randomly intermixed with 24 novel targets/faces. After the target recognition test, as in Experiment 2, the participants were asked to confirm that they had not recorded or looked up the word pairs or faces during the task. They also had an opportunity to provide feedback and they received a written debrief. The experiment lasted approximately 20 minutes.

Results and Discussion

During the encoding phase, the participants collectively correctly guessed a total of eight targets on Pre-test trials. These targets, and the faces presented on those trials, were removed from the target and face recognition data analyses. The experimental program did not progress until the participants had entered the first letter of their guess, so there were no trials in which the participants did not provide *any* guess (although they did not always enter a full word⁹). Participants almost always selected the correct first letter of the target on Study-only trials ($M = 98.60\%$, $SEM = 0.47\%$).

⁹ One participant failed to complete any guess: removing them from the analysis made no difference to the pattern reported here, and so they are retained in the current analyses.

In the target recognition test, the participants correctly recognised the foils as novel on most trials ($M = 94.96\%$, $SEM = 1.23\%$). Overall, Pre-tested items ($M = 77.3\%$, $SEM = 3.03\%$) were recognised significantly more often than Study-only items ($M = 71.5\%$, $SEM = 3.61\%$), $t(42) = 1.91$, $p = .032$, $d_z = 0.29$, $BF_{10} = 1.65$, although the Bayes factor for was indeterminate.

In the face recognition test, the mean percentage of foils that were correctly rejected as novel was 79.84% ($SEM = 1.96\%$). For face targets, in addition to the manipulation of Encoding condition, there was a manipulation of Delay, and so recognition performance was submitted to a 2 (Encoding condition: Pre-test vs Study-only) \times 2 (Delay condition: Long vs short) repeated-measures ANOVA. This revealed no significant main effects or interactions, $F_s < 1$ (see Table 1). We also conducted Bayesian t-tests to examine the effect of pre-testing at each delay to see whether the evidence merely failed to support the experimental hypothesis, or actively supported the null. The Bayesian evidence provided substantial evidence for the null in the long delay condition, $BF_{10} = 0.17$, and moderate evidence for the null in the short delay condition, $BF_{10} = 0.39$. We also ran one further analysis only on the Pre-test trials, to explore whether there was any evidence performance varied across the two delays. Consistent with the omnibus ANOVA, this supported the null hypothesis, $t < 1$, $BF_{10} = 0.20$.

In summary, the results of our third experiment largely confirm the pattern observed in the previous two experiments. There was no evidence of an incidental memory benefit to faces presented between an initial guess and subsequent feedback to that guess, regardless of the duration of the delay between the initial guess and the presentation of the face. The pattern of the findings is therefore incompatible with those reported in Murphy et al. (2021).

It is worth noting that lack of an incidental memory benefit associated with pre-testing replicates pattern observed in Experiment 1, rather than advantage for study-only trials seen in Experiment 2. We have no explanation for the unexpected pattern observed in Experiment 2, other than the possibility that it is a chance finding. We leave exploration of that issue for future research.

One potential concern is that our pre-testing effect for targets was smaller in this experiment relative to previous experiments, and this might have contributed to the failure to observe an effect on the incidental faces. To address this, we carried out an exploratory analysis looking only at those participants ($N = 19$) whose target recognition in the Pre-test condition exceeded that in the Study-only condition. The average pre-testing effect for the target words in this group was 24.1% ($SEM = 3.76\%$), which corresponds to an effect size of $d_z = 1.47$. Nevertheless, for this group, there was no evidence of any pre-testing benefit for the faces, with the evidence still favouring the null hypothesis ($BF_{10} = 0.18$ with a short delay, and $BF_{10} = 0.11$ with a long delay).

A second potential concern is that our guessing manipulation may not have had a sufficiently strong effect on curiosity to impact upon the incidental faces, notwithstanding the observed boost to target recognition. We concede that this is possible, but there is one potential counter-argument against it. In their recent study that showed a benefit for high-curiosity items at short delay, Murphy et al. (2021) also reported an interaction between delay and curiosity level on face recognition. Recognition of faces associated with high-curiosity items was high, regardless of delay between question offset and the face. However, for faces associated with low-curiosity items, face recognition was higher after a long delay compared to a shorter delay. If this rise in recognition with delay is a signature of low-curiosity, we did not observe it here. Comparison of recognition of the faces across delay favoured the null hypothesis, which resembles more closely the pattern reported by Murphy et al. (2021) for high-curiosity items, not low-curiosity items.

General Discussion

In three experiments, we demonstrated the benefits of a pre-test on subsequent recognition of the target of a guess, relative to study alone. Experiment 1 used arbitrary facts associated with faces, while Experiments 2 and 3 used the meanings of unfamiliar rare English words. These experiments therefore add to a growing literature demonstrating that pre-testing boosts subsequent recognition memory for targets, regardless of any pre-existing semantic association (Potts & Shanks,

2014; Potts et al. 2019; Seabrooke, Hollins et al., 2019, Seabrooke, Mitchell et al., 2019; Seabrooke, Mitchell & Hollins, 2021; Seabrooke, Mitchell, Wills et al., 2021; Seabrooke, Mitchell, Wills, Inkster et al., 2021).

In the current studies, we presented incidental material between the initial guess and the subsequent feedback. In Experiment 1, the additional material was facts associated with categories that were separate to the target. In Experiments 2 and 3, the incidental material was unfamiliar faces, entirely unrelated to the target facts. All three experiments showed reliable pre-testing effects for the target facts, despite the presentation of this incidental material, and Experiment 1 showed that the magnitude of the pre-testing effect was unaffected by whether there were zero, one or two intervening facts prior to the presentation of the target.

While the pre-testing effect was robust enough to survive intervening items between a guess and the target as feedback, there was no evidence to suggest that the beneficial effect of guessing generalises to the incidentally presented material encountered prior to the feedback, as predicted by a curiosity-state hypothesis (Gruber & Ranganath, 2019). In all three experiments, the evidence supported the null hypothesis regarding incidental memory benefits.

The current studies sought to test the idea that pre-testing effects are mediated by a general state of curiosity. Although the current data are incompatible with this view, they do not rule out a *specific* version of a curiosity account. It is possible that participants pay greater attention to those items that close the information gap created by the initial guess, which in turn boosts recognition of those targets (Potts, et al., 2019; Seabrooke, Hollins et al., 2019, Seabrooke, Mitchell, Wills, Waters et al., 2021).

We leave to future research the question of when curiosity produces specific or generalised memory benefits. However, the present work identified key constraints that any future account of the pre-testing effect must explain. As well as demonstrating that the recognition effect is not dependent upon prior cue-target associations, Experiment 1 demonstrated that the pre-testing

effect is both highly-specific, and robust enough to survive the presentation of other verbal material that is not the answer to the question guessed. In this respect, the effect resembles the *aha effect* (e.g., Auble & Franks, 1978; Auble, et al., 1979; Zaromb, et al., 2010), in which participants either encounter materials that are initially unclear (e.g. “the house was small because the sun came out”) and only later made clear through additional information (“igloo”), or they encounter the equivalent material that is already integrated coherently (“the igloo was small because the sun came out”). Here, there is a subsequent memory advantage associated with the cognitive reappraisal brought about by the key detail that was previously missing. There are clear parallels with the pre-testing effect, where participants initially try to generate meaning (i.e., guess), and only later receive the key information needed (the corrective feedback). Consistent with the *aha effect*, our work shows that the benefit of guessing is specific to the information needed to resolve the information gap, and does not benefit anything else encountered at the same time.

Our original aim was to determine whether the pre-testing and curiosity effects are manifestations of the same behavioural phenomenon. While the present work strongly suggests that this is not the case, there are still unanswered questions about the role of curiosity in the pre-testing effect, and vice versa. Our failure to observe an incidental memory benefit associated with pre-testing must be squared with the observation that people report being more curious about facts associated with guesses (Potts et al., 2019). It could be that the self-reported curiosity has no causal relation to the subsequent memory benefit, or it could be that the curiosity elicited by guessing somehow differs to the curiosity elicited by different general knowledge questions. Questions also remain about the role of guessing in the memory benefits associated with general knowledge questions that elicit different levels of curiosity. The present work rules out a role for guessing in the incidental memory benefit seen for high curiosity questions, but whether guessing underpins the memory benefit seen for the targets of curiosity must remain an open question. This point applies both to the item-based method used in the present work, but also to the guessing benefits seen using the test-set method discussed earlier-

In the present work, we have sought to test the idea that the pre-testing effect is driven by curiosity, and so would be expected to generalise to incidental material encountered whilst in that state. However, this is not the only theoretical account of the pre-testing effect. A recent review by Mera et al (2022) outlines four broad theoretical accounts that have sought to explain the range of findings observed in pre-testing studies (*Error Prediction Theory*¹⁰, *Mediator Effectiveness*, *Recursive Reminding* and *Search Set Theory*). The present work was not designed to adjudicate between these different theoretical accounts. While all theories would predict the pre-testing effects seen for the targets in the present work, only the curiosity-state version of Error Prediction Theory makes the additional prediction of benefits for incidental material that was tested here.

¹⁰ Here we use Mera et al's (2022) terminology. The curiosity based account discussed in the present work is one instantiation of an Error Prediction Theory, inasmuch that it is the pre-testing error that drives the state of curiosity.

References

- Auble, P. M., & Franks, J. J. (1978). The effects of effort toward comprehension on recall. *Memory & Cognition*, 6, 20-25. doi: 10.3758/BF03197424
- Auble, P. M., Franks, J. J. & Soraci, S.A. (1979). Effort toward comprehension: Elaboration or “aha”? *Memory & Cognition* 7, 426-434. doi: 10.3758/BF03198259
- Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, 44(1), 158-175. doi:10.3758/s13428-011-0123-7
- Bakker, M., van Dijk, A. & Wicherts, J. M. (2012) The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543-554. doi: 10.1177/1745691612459060
- Clark, C. M., Bjork, E. L., & Bjork, R. A. (2021). On the role of generation rules in moderating the beneficial effects of errorful generation. *Zeitschrift für Psychologie*, 229, 120-130. doi 10.1027/2151-2604/a000442
- Carpenter, S. & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition*, 6, 104-109. doi: 10.1016/j.jarmac.2016.07.014
- Cyr, A.-A. & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3), 841-850. doi:10.1037/xlm0000073
- Cyr, A.-A. & Anderson, N. D. (2018). Learning from your mistakes: does it matter if you're out in left foot, I mean field? *Memory*, 26(9), 1281-1290. doi:10.1080/09658211.2018.1464189
- DeBruine, L. & Jones, B. (2017). Face Research Lab London Set. *Figshare Dataset*. doi:10.6084/m9.figshare.5047666.v5
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1-12. doi:10.3758/s13428-014-0458-y.

- Dubey, R. & Griffiths, T.L. (2021). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, 127, 455-476. doi: 10.1037/rev0000175
- Fandakova Y, & Gruber M. J. (2020). States of curiosity and interest enhance memory differently in adolescents and in children. *Developmental Science*, 24, e13005. doi:10.1111/desc.13005.
- Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, 40(4), 505–513. doi:10.3758/s13421-011-0174-0
- Gruber M.J., Gelman, B.D., Ranganath C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron*, 84: 486–496.
doi:10.1016/j.neuron.2014.08.060
- Gruber M.J., & Ranganath C. (2019). How curiosity enhances hippocampus-dependent memory: the prediction, appraisal, curiosity, and exploration (PACE) framework. *Trends in Cognitive Science*, 23, 1014–1025. doi:10.1016/j.tics.2019.10.003
- Guinness World Records (2022, May 26). *Hottest chilli pepper*. Retrieved from <https://www.guinnessworldrecords.com/world-records/hottest-chili>
- Hays, M. J., Kornell, N. & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 39, 290-296. doi: 10.1037/a0028468
- Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, 40(4), 514–527. doi:10.3758/s13421-011-0167-z
- James, K. K. & Storm, B. C. (2019). Beyond the pretesting effect: what happens to the information that is not pretested. *Journal of Experimental Psychology: Applied*, 25, 576-587. doi: 10.1037/xap0000231
- Kang M.J., Hsu M., Krajbich I.M., Loewenstein G., McClure S.M., Wang J.T.Y., & Camerer C.F. (2009). The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances

- memory. *Psychological Science*, 20, 963–973. doi:10.1111/j.1467-9280.2009.02402.x
- Kidd C., & Hayden B.Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88, 449–460. doi:10.1016/j.neuron.2015.09.010
- Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, 66(4), 731–746. doi: 10.1016/j.jml.2011.12.008
- Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 106–114. doi: 10.1037/a0033699
- Kornell, N., Hays, M., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 989–998. doi:10.1037/a0015729
- Little, J. L. & Bjork, E. L. (2016). Multiple-choice pretesting potentiates learning of related information. *Memory & Cognition*, 44, 1085–1101. doi: 10.3758/s13421-016-0621-z
- Loewenstein, G. (1994). The psychology of curiosity: a review and reinterpretation. *Psychological Bulletin*, 116, 75–98. doi:10.1037/0033-2909.116.1.75
- Mera, Y., Rodríguez, G. & Marin-García, E. (2022). Unraveling the benefits of experiencing errors during learning: definition, modulating factors, and explanatory theories. *Psychonomic Bulletin & Review*, 29, 753–765. doi: 10.3758/s13423-021-02022-8
- Metcalfe, J., & Huelser, B. J. (2020). Learning from errors is attributable to episodic recollection rather than semantic mediation. *Neuropsychologia*, 138. doi 10.1016/j.neuropsychologia.2019.107296
- Morey, R. D., Rouder, J. N., & Jamil, T. (2015). Package 'BayesFactor.' Retrieved from <https://richarddmores.github.io/BayesFactor/>

- Murphy, C., Dehmelt, V., Yonelinas, A. P., Ranganath, C. & Gruber, M. J. (2021). Temporal proximity to the elicitation of curiosity is key for enhancing memory for incidental information. *Learning & Memory*, 28, 34-39. doi: 10.1101/lm.052241.120
- Pan, S. C., Lovelett, J., Stoeckenius, D. & Rickard, T. (2019). Conditions of highly specific learning through cued recall. *Psychonomic Bulletin & Review*, 26, 634-640. doi: 10.3758/s13423-019-01593-x
- Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J. & Rickard, T. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition*, 44, 24-36. doi: 10.3758/s13421-015-0547-x
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*, 51, 195-203. doi: 10.3758/s13428-018-01193-y
- Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning Memory and Cognition*, 45(6), 1023–1041. doi:10.1037/xlm0000637
- Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, 143(2), 644–667. doi:10.1017/CBO9781107415324.004
- Potvin, P. J. & Schutz, R. W. (2000). Statistical power for the two-factor repeated measures ANOVA. *Behavior Research Methods, Instruments & Computers*, 32, 347-356. doi: 10.3758/BF03207805
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002), 335. doi:10.1126/science.1191465
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Richland, L. E., Kornell, N. & Kao, L. S. (2009). The pretesting effect: do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, 15, 243-257. doi:

10.1037/a0016496

Seabrooke, T., Hollins, T. J., Wills, A. & Mitchell, C. J. (2019). Learning from failure: Errorful generation improves memory for items, not associations. *Journal of Memory and Language*, 104, 70-82. doi: 10.1016/j.jml.2018.10.00

Seabrooke, T., Mitchell, C. J. & Hollins, T. J. (2021). Pre-testing boosts item but not source memory. *Memory*, 29, 1245-1253. doi: 10.1080/09658211.2021.1977328

Seabrooke, T., Mitchell, C. J., Wills, A. J., & Hollins, T. J. (2021). Pre-testing boosts recognition, but not cued recall, of targets from unrelated word pairs. *Psychonomic Bulletin & Review*, 28, 268-273. doi:10.3758/s13423-020-01810-y

Seabrooke, T., Mitchell, C. J., Wills, A. J., Waters, J. L., & Hollins, T. J. (2019). Selective effects of errorful generation on recognition memory: The role of motivation and surprise. *Memory*, 27(9), 1250-1262. doi:10.1080/09658211.2019.1647247

Seabrooke, T., Mitchell, C. J., Wills, A. J., Inkster, A. & Hollins, T. J. (2021). The benefits of impossible tests: reconsidering the error correction hypothesis. *Memory and Cognition*, doi:10.3758/s13421-021-01218-6

St. Hilaire, K. J., Carpenter, S. K. & Jennings, J. M. (2019) Using prequestions to enhance learning from reading passages: the roles of question type and structure building ability, *Memory*, 27, 1204-1213, doi: 10.1080/09658211.2019.1641209

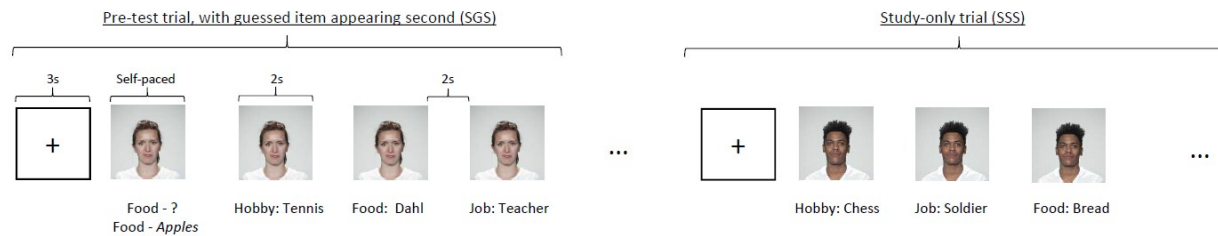
Stare C. J., Gruber M. J., Nadel L., Ranganath C., & Gómez R. L. (2018). Curiosity-driven memory enhancement persists over time but does not benefit from post-learning sleep. *Cognitive Neuroscience*, 9, 100-115. doi:10.1080/17588928.2018.1513399

Swirsky, L. T., Shulman, A. & Spaniol, J. (2021). The interaction of curiosity and reward on long-term memory in younger and older adults. *Psychology and Aging*, 36, 584-595. doi: 10.1037/pag0000623

- Van de Cruys, S., Damiano, C., Boddez, Y., Krol, M., Goetschalckx, L., & Wagemans, J. (2021). Visual affects: linking curiosity, aha-erlebnis, and memory through information gain. *Cognition*, 212, doi: 10.1016/j.cognition.2021.104698
- Vaughn, K. E. & Rawson, K. E. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin and Review*, 19, 899-905. doi: 10.3758/s13423-012-0276-0
- Wade S., Kidd C. (2019). The role of prior knowledge and curiosity in learning. *Psychonomic Bulletin & Review*, 26, 1377–1387. doi:10.3758/s13423-019-01598-6
- Zaromb, F. M., Karipicke, J. D., & Roediger, H. L. (2010). Comprehension as a basis for metacognitive judgments: effects of effort after meaning on recall and metacognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 36, 552-557. doi: 10.1037/a0018277
- Zawadzka, K., & Hanczakowski, M. (2019). Two routes to memory benefits of guesing. *Journal of Experimental Psychology: Learning Memory and Cognition*, 45(10), 1748–1760. doi:10.1037/xlm0000676

Figures

A: Study phase, intermixing pre-test trials (n = 18) and study-only trials (n=6).



B: Test phase involving multiple-choice recognition for targets intermixed with novel foils.



Figure 1: Schematic representation of procedure used in Experiment 1: In the study phase pre-test trials are interleaved with study-only trials, with pre-test trials involving a guess for one target fact, that subsequently appears 1st, 2nd or 3rd in the study sequence. In the test phase, target items are tested in a fully-randomised order.

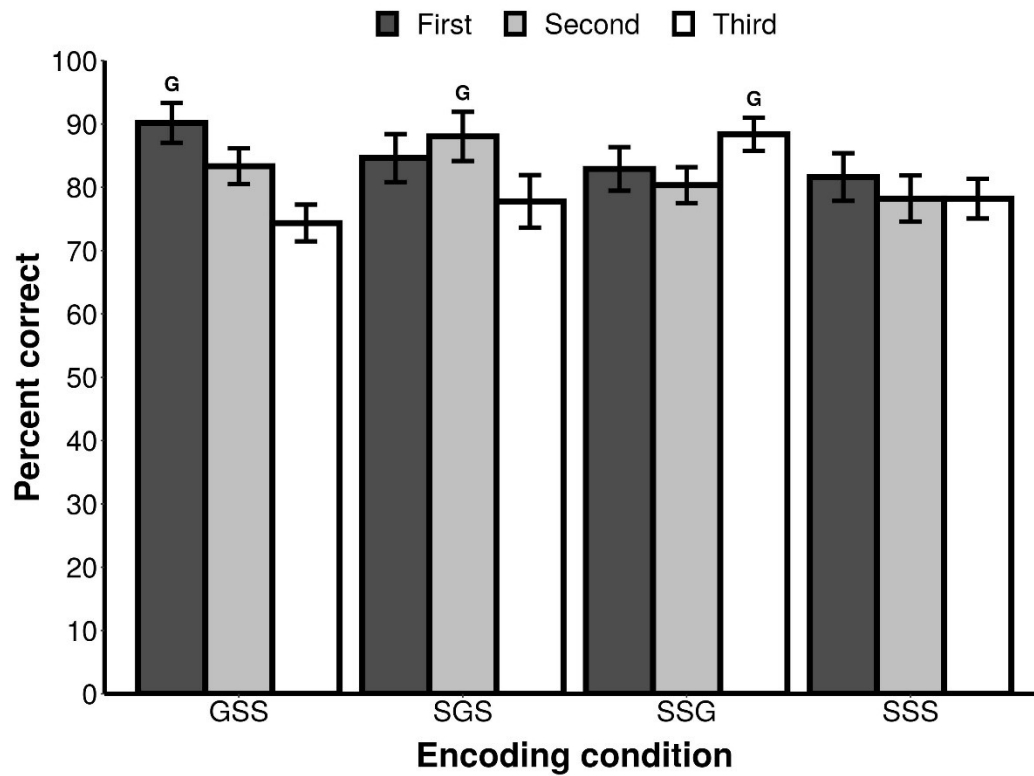


Figure 2. Mean percentage of correctly recognised targets in the multiple-choice test of Experiment 1, according to the encoding condition and position (first, second, or third) that the target was presented at encoding. The “G” symbol highlights the targets that were guessed at encoding. Error bars represent difference-adjusted, 95% within-subjects confidence intervals (Baguley, 2012).

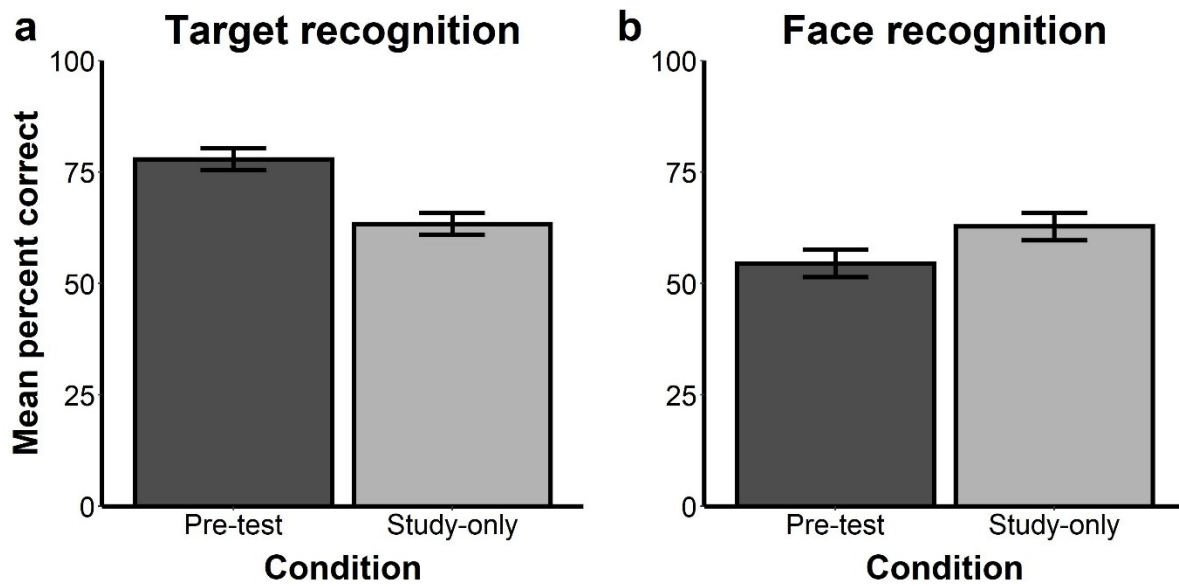


Figure 3. Mean percentage of correctly identified "hits" to "old" items in the (a) target and (b) face recognition tests of Experiment 2. Error bars represent difference-adjusted, 95% within-subjects confidence intervals (Baguley, 2012)

Table 1*Mean (SEM) Percentage of Hits to Incidental Faces in Experiment 3*

| | Delay | |
|------------|--------------|--------------|
| | Short (2 s) | Long (8 s) |
| Pre-test | 67.91 (3.53) | 70.78 (3.63) |
| Study-only | 72.87 (4.56) | 70.16 (3.26) |