Article:
Zawadzka, K. and Hanczakowski, M. (2018) Two routes to memory benefits of guessing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 45 (10). pp. 1748-1760. ISSN: 0278-7393

https://doi.org/10.1037/xlm0000676

Two routes to memory benefits of guessing

Katarzyna Zawadzka[1] and Maciej Hanczakowski[2]

[1] University of Sheffield, UK

[2] SWPS University of Social Sciences and Humanities, Poland

Author Note

Katarzyna Zawadzka, Department of Psychology, University of Sheffield; Maciej Hanczakowski, Interdisciplinary Center for Applied Cognitive Studies, SWPS University of Social Sciences and Humanities.

Correspondence should be addressed to Katarzyna Zawadzka, Department of Psychology, University of Sheffield, Cathedral Court, 1 Vicar Lane, Sheffield S1 2LT, UK, email: k.zawadzka@sheffield.ac.uk, or Maciej Hanczakowski, Interdisciplinary Center for Applied Cognitive Studies, SWPS University of Social Sciences and Humanities, ul. Chodakowska 19/31, 03-815, Warszawa, Poland, email: maciej.hanczakowski@gmail.com

All data are available online on https://osf.io/k6v3b

Abstract

Attempting to guess an answer to a memory question has repeatedly been shown to benefit memory for the answer as compared to merely reading what the answer is, even when the guess is incorrect. In this study, we investigate two potential explanations of this effect in a single experimental procedure. According to the semantic explanation, the benefits of guessing require a clear semantic relationship between the cue, the guess, and the target, and arise at the stage of guessing. The attentional explanation places the locus of the effect at the stage of feedback presentation and ignores the issue of semantic relatedness. To disentangle the two mechanisms, we used homograph cues with at least two different meanings (e.g., *arms*) and asked participants to either study an intact cue-target pair or guess a word related to each cue before being presented with the target. This allowed us to compare memory performance on trials on which participants' guesses tapped the same meaning of the cue as the later presented target (e.g., a guess *legs* for a pair *arms-hug*), versus a different meaning (e.g., *weapons*). In four experiments, we demonstrate that both the semantic and the attentional mechanism operate in the guessing task, but their roles are different: semantic relatedness supports memory for cue-to-target associations, while increased attention to feedback benefits memory for targets alone. We discuss these findings in the context of educational utility of errorful learning.

Keywords: errorful learning, testing, feedback, judgments of learning, education

Two routes to memory benefits of guessing

Enhancing students' learning is one of the key aims of any educator. It is thus unsurprising that over the last decades considerable effort has been made to shed light on the principles of effective learning. Many learning strategies, more or less commonly used by students, have been subjected to extensive study in order to establish their usefulness in educational contexts (see, e.g., Ariel & Karpicke, 2018; Grimaldi, Poston, & Karpicke, 2015; Kornell & Bjork, 2008; Szpunar, Khan, & Schacter, 2013; see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Weinstein, Madan, & Sumeracki, 2018, for reviews). One such strategy that has attracted particular attention is the use of tests as learning tools (Roediger & Karpicke, 2006; Yang, Potts, & Shanks, 2018). It has repeatedly been shown that testing oneself on some information – as contrasted with merely restudying this information by reading – produces greater memory improvement which persists even after a long delay (e.g., Roediger, Agarwal, McDaniel, & McDermott, 2011).

An issue of particular interest when tests are used for learning rather than assessment is one of errors. When students re-read the to-be-learned information, they are exposed to correct information only. When they test themselves instead, however, they are likely to answer at least some of the questions incorrectly. These errors can then persist and be repeated in future tests (Butler, Marsh, Goode, & Roediger, 2006; Marsh, Roediger, Bjork, & Bjork, 2007). It is thus crucial to establish conditions under which incorrect responding would not be detrimental to future memory performance. One obvious way to remedy this is the provision of feedback. When a student is presented with the right answer after making an error, an additional learning opportunity occurs. It has been shown that such feedback is effective for learning correct answers in the place of errors (e.g., Butler, Karpicke, & Roediger, 2008, Finn & Metcalfe, 2010).

Recently, a paradigm has been developed by Kornell, Hays, and Bjork (2009) that clearly shows beneficial aspects of retrieval attempts accompanied by corrective feedback. In this paradigm, participants are exposed to a test on new, never-studied materials. Crucially, this test is created so that all or almost all answers to the questions are guesses, with the vast majority of them being incorrect. After each guess, participants are presented with the correct answer and given some time to learn this answer in the context of the question. This guessing coupled with exposure to corrective feedback improves performance as compared

to simply studying the new materials, even when the time of exposure to correct information is much shorter in the guess than in the read-only condition. This counterintuitive finding shows that the benefits of testing extend also to the domain of errorful learning.

The benefit of attempting to guess what the answer is – as compared to simply being shown the answer – has been reported to occur for adults as well as children as young as five years old (Carneiro, Lapa, & Finn, 2018), and across a range of tasks and materials. It has been found for trivia questions (Kornell, 2014), fictional questions for which a real answer does not exist (e.g., *Who shot a fig out of a tree with a crossbow in the 11th century*; Kornell et al., 2009), educational materials (Richland, Kornell, & Kao, 2009), as well as learning foreign language vocabulary and meanings of rare English words (Potts, Davies, & Shanks, 2018; Potts & Shanks, 2014). However, the materials that have been used most extensively in previous studies are related word pairs (e.g., *pond-frog*). It has been demonstrated repeatedly in many laboratories (e.g., Bridger & Mecklinger, 2012; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell et al., 2009; Yang, Potts, & Shanks, 2017) that when participants are asked to guess an associate of a presented word, and this guess is immediately followed by feedback in the form of an answer designated to be the 'correct' to-be-learned associate, memory for word pairs is improved as compared to a condition in which the pair is presented intact.

The benefit of incorrect guessing for word pairs has been shown to be reasonably robust. It is present for weakly and strongly associated pairs (Yan, Yu, Garcia, & Bjork, 2014), persists regardless of the likelihood of making an error during guessing (Bridger & Mecklinger, 2014), and is independent of whether the cues and targets are prefamiliarized in an initial phase of the experiment, or are presented for the first time in the study/guess phase (Knight et al., 2012). However, some boundary conditions have also been discovered. One important limitation is that the benefits of guessing disappear when unrelated pairs are used as study materials (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012; see also Slamecka & Fevreiski, 1983, for a related finding). In fact, the opposite pattern can sometimes be found for unrelated pairs: better performance for pairs that were read in the study phase (Knight et al., 2012). It also seems that the correct answer needs to be presented immediately after the guessing attempt in order for the benefits of guessing to occur (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Vaughn & Rawson, 2012;

although note that delayed feedback seems to be equally effective as immediate feedback in promoting learning through guessing when trivia questions are used as study materials; Kornell, 2014). These failures to detect the benefits of guessing for memory performance under certain constrained conditions have shed some light on what the potential mechanism or mechanisms behind these benefits might be.

Two potential explanations of the benefits of guessing stress the same constraint on the discussed phenomenon: semantic relatedness between the cue and the target, which serves as a precondition for observing the benefits of guessing. The first semantic explanation points to the activation of the semantic network that arises when a guess is made (e.g., Bridger & Mecklinger, 2014; Grimaldi & Karpicke, 2012; see also Carpenter, 2009). If a participant is presented with a cue and attempts to predict what the target might be, one word might be provided as a response, but other related concepts also benefit from some degree of activation spreading through the semantic network. When the target is ultimately presented as the to-be-encoded word, encoding is facilitated due to this initial semantic activation. This increase in semantic activation is likely to last only for a relatively brief period of time (Neely, 1977; Raaijmakers & Shiffrin, 1981), which would also be consistent with the absence of the benefits of guessing when feedback is delayed.

A second explanation that stresses the role of semantic relatedness concentrates on the role of initially guessed words in creating the benefits of guessing. Studies have demonstrated that participants remember their initial guesses reasonably well (Knight et al., 2012; Vaughn & Rawson, 2012, Yan et al., 2014), and yet these guesses tend not to interfere with retrieving the correct target when tested (although see Hays et al., 2013). It has been proposed that the guesses can instead be used as mediators, related to both the target and the cue and linking the two to-be-remembered words (Huelser & Metcalfe, 2012; Vaughn & Rawson, 2012; see also Carpenter, 2011; Pyc & Rawson, 2010): if retrieved at test, these mediators can serve as additional cues for target retrieval. Again, when feedback is delayed, the association between the guess and the target might not be established, leading to the lack of memory improvement at test. These two accounts of benefits of guessing – semantic activation and guesses as mediators – will be henceforth referred to as *semantic accounts*.

Although the observation that the benefits of guessing do not emerge for semantically unrelated pairs of words has been the main driver in developing theories accounting for the benefits of incorrect guessing, it is important to note that there are some instances in which

such benefits were observed even under conditions in which there was no pre-existing relationship between the cues and associated target responses. In their Experiments 1, 2a, and 3, Potts and Shanks (2014) presented participants with rare English words (e.g., *frampold*) and asked to guess their more common synonyms (e.g., *quarrelsome*). Experiment 2b used as cues words from the Euskara language which has no known related languages. With both types of materials, the presentation of an unknown word was unlikely to generate any semantic activation, and similarly the probability to generate an effective mediator linking the cue to the target was presumably markedly lower than for related pairs. Nevertheless, Potts and Shanks observed in multiple-choice tests the same benefits of generating a candidate response as compared to learning by reading or guessing from a set of two or four provided alternatives. This suggests the explanations of the benefits of guessing in terms of semantic processing are at best incomplete, which should give impetus to developing alternative accounts. Potts and Shanks have proposed that when the task of guessing the correct answer is perceived by participants as particularly difficult – as in the case of never seen before English or foreign words – more attention is directed to the correct answer when it is finally presented as compared to the read-only condition.

There are two mechanisms through which the enhanced processing of feedback might be responsible for the benefits of guessing, henceforth referred to as the *attentional accounts* of guessing benefits. First, as postulated by Potts et al. (2018; see also Potts & Shanks, 2014), being asked to guess what the answer might be – especially if the guess is unlikely to be correct – can lead to a discrepancy between the participant's current state of knowledge and what they would like to know. This discrepancy drives curiosity, which then translates into increased attention to feedback when it is presented. To provide evidence for this explanation, Potts et al. asked their participants to provide curiosity ratings at study – assessments of how curious participants were to learn the meaning of an unfamiliar English word. These ratings were higher after generating a guess than either before its generation or when no guess was required. This suggests that the action of generating indeed does lead to increased curiosity to learn the correct answer, which might then translate into enhanced encoding of that answer.

The second and somewhat related explanation, also mentioned by Potts and Shanks (2014) as a potential account of their results, builds on findings by Butterfield and Metcalfe (2001) who demonstrated that feedback that is surprising is particularly likely to draw

participants' attention, making encoding of that feedback more effective (see also Fazio & Marsh, 2009; Griffiths & Higham, 2018). This account assumes that feedback that defies expectations – diverges from what participants consider to be a likely candidate answer, as evidenced by their initial guess – generates a feeling of surprise, resulting in greater attention being allocated to the correct answer as compared to items merely presented for reading. Thus, surprise is the second, alongside curiosity, variant of the attentional account of the benefits of errorful learning.

One limitation of the current discussion concerning the mechanisms of the benefits of guessing is that, so far, the semantic and attentional accounts have not been systematically assessed within the same experimental design. When pairs are related, as in the studies supporting the semantic explanation, no feedback seems particularly surprising: participants might reasonably expect that the correct answer will be not far removed from their initial guess. Similarly, the curiosity to discover what the correct answer is might be diminished if participants do not expect it to differ too much from their initial guess. By contrast, when the pairs include a novel, not known before cue, as in the Potts and Shanks' (2014) study supporting the attentional explanations, there are no pre-existing semantic associations that can be activated before the feedback is shown. At present, therefore, a direct comparison of different postulated mechanisms of the benefits of guessing is hindered by the fact that these mechanisms try to explain results obtained in different paradigms. In order to properly describe the contributions of semantic processing and attention to the benefits incorrect guessing confers, a procedure is necessary that will preserve the relatedness between the cue and the target, allow for activating a semantic network at the guessing stage, and yet promote curiosity to learn the correct answer and often result in feedback that might be considered surprising. To this end, we modified the procedure using related pairs by using only homograph cues.

In the present study, participants underwent the standard procedure introduced by Kornell et al. (2009). In the study phase, they were asked to learn weakly related cue-target pairs for a future test. Crucially, each cue had at least two different meanings and thus was embedded in two pairs. Each of these pairs tapped a different meaning of the cue (e.g., *arms – hug* and *arms – nuclear*), and in the course of the experiment participants were exposed to one of these pairs only. In the learning phase of the experiment, half of the studied pairs were presented intact: in this case, the meaning of the cue was determined

from the onset by the presentation of the target. For the other half, a cue was first presented, and participants' task was to guess a related word associated with that cue. Due to the nature of the cues, participants' guesses could either be related or unrelated to the meaning of the cue determined by the target, which – crucially – was presented only after an eight-second delay.[1] Related guesses served as an indication that the initial interpretation of the cue was related to the meaning of the later presented target. This, according to the semantic accounts, creates conditions that should later support correct responding at a later test. Unrelated guesses suggested that participants' initial interpretation of the cue was different from that suggested by the target. In this case, there should be far less semantic activation of concepts related to the target and, similarly, the generated guess should also remain unrelated to the target, precluding its effective use as a mediator. However, the actual target should often defy participants' expectations, leading to a feeling of surprise. In addition to that, the fact that participants could never be certain whether the correct answer would be related to their guess should also boost curiosity for all test pairs. This, on the grounds of the attentional accounts, should draw participants' attention to the feedback, producing benefits in terms of final-test performance.

The present study thus serves to simultaneously demonstrate the contributions of both the semantic and attentional mechanisms to the benefits of incorrect guessing for subsequent memory performance. First, the comparison of final-test cued-recall performance between targets from pairs that were presented intact (the *read* condition) and those from pairs for which the guessed word matched the meaning of the target (the *guessed-same* condition) should reveal the benefits of guessing the target as compared to being merely presented with it, as predicted by the semantic accounts, and providing a replication of previous findings from studies that used related pairs (e.g., Bridger & Mecklinger, 2014; Knight et al., 2012; Kornell et al., 2009). Second, the comparison of final-test performance between targets that mismatched the meaning of the guess (the *guessed-different* condition) and those from the read condition should speak to the applicability of

[1] Unrelated guesses could either be related to the non-presented meaning of the cue, or not related to any of the two meanings; for example, to the experimenters' initial puzzlement, when cued with the word *mole* – paired in the experiment with the words *face* and *hole* – many participants guessed the words *Sheffield* or *lecture*, as MOLE is the online learning platform at the University of Sheffield.

the attentional explanations to learning semantically related pairs when the relationship between the cue and the target is incorrectly predicted.

Experiment 1

Method

Participants. Thirty-two students of the University of Sheffield (seven male; age range: 18-63) who reported native-level proficiency in English took part in the experiment in exchange for course credit or monetary compensation. Sample size was determined on the basis of previous studies using similar methods, in which participant numbers are commonly in the range of 20-30 when within-participant designs are used (e.g., Kornell et al., 2009; Potts & Shanks, 2014; Yang et al., 2017), and allowed for an equal number of participants in each of the counterbalancing conditions. The study has been approved by the Department of Psychology Ethics Committee at the University of Sheffield.

Materials and Design. One hundred and eighty words were chosen from the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998). Sixty of the words were used as cues in the experiment, and the remaining 120 as targets. All cues were homographs with at least two different meanings (e.g., *arms*). Two sets of targets were created, each consisting of 60 words. Each set of targets consisted of weak associates of all homographs used as cues. On that basis, two lists of word pairs were formed. The lists contained the same homograph cues, but differed in terms of the targets associated with those cues. Each of the two targets paired with a given homograph tapped a different meaning of the cue word (e.g., *arms* – *hug* and *arms* – *nuclear*). The mean forward association between cues and targets was comparable between the lists (.05 versus .06, meaning a 5% versus 6% probability of producing the target when presented with the cue, $p = .21$), and ranged between .013 (for both *ash* – *volcano* and *ash* – *tree*) and .157 (*pole* – *flag*).

At study, each participant was presented with 60 pairs from one list only. The cue was presented on each trial for 13 seconds. Target presentation depended on the condition to which a given pair was assigned, manipulated within participants. In the *read* condition, consisting of 30 pairs, the target was presented together with the cue for 13 seconds. The remaining pairs were used for the *guess* condition in which the cue was presented next to a blank text field for eight seconds, after which the target appeared in the text field and

remained on the screen for the following five seconds. For each participant, the guesses were further subdivided on the basis of their responses. If the guess matched the meaning of the homograph that was the same as the one tapped by the (later presented) target, the trial was classified as *guessed-same* (e.g., a guess *legs* for the pair *arms – hug*). When the guess did not match the meaning of the homograph cue, the trial was classified as *guessed-different* (e.g., the guess *legs* for the pair *arms – nuclear*). Examples of trial classifications are presented in Figure 1, which also provides an overview of the experimental design. The assignment of pairs to the read versus guess conditions was counterbalanced across participants, and the order in which the pairs were presented at study and the cues at test was randomized anew for each participant and each experiment phase.

Procedure. Before the study phase, participants were instructed that they would be presented with pairs of related words to learn, and that there would be two different types of trials they would encounter. Whenever two words were presented, they should spend 13 seconds trying to memorize the presented pair. Whenever a single word was presented, they were told that first they should try to guess what the second word might be and type that word into the blank field next to the cue within eight seconds, and then learn the actual full word pair (rather than the cue and their guess) when the second word appears. These instructions were followed by a brief practice task consisting of both trial types. After completing the practice task, participants were informed that the study phase would be followed by a test, on which first words would be presented and second words would need to be recalled. The study phase consisted of all 60 word pairs from the list, with 30 pairs assigned to the read condition and 30 to the guess condition. After the study phase, the instructions for the test were restated and the test began immediately afterwards. The first words from each pair were presented one at the time and participants' task was to type in the word corresponding to this cue or press a "Continue" button to advance to the next pair if no response was provided. The time for completing the test was not limited. The full set of instructions for all phases of this experiment, as well as for Experiments 2-4, can be found in the Supplementary Materials.

Results and Discussion

Study phase.[2] The meaning of guesses was classified by the authors as either matching the meaning of the cue and the (later presented) target, or mismatching that meaning. For example, the guess *military* would be classified as matching the meaning of the cue *arms* if the target paired with that cue was *nuclear*, but mismatching if the target was *hug*. If the word was not fully typed in, it was classified as a guess only in rare cases in which it was possible to unambiguously identify what the full word would mean (e.g., *militar*); any ambiguity resulted in the exclusion of the trial from analyses. Spelling mistakes (e.g., *militayr*) were dealt with according to the same rule. All contentious cases were resolved through a discussion of both authors.

On guess trials, 55.2% of participants' guesses tapped a different meaning from the target and so were classified as *guessed-different*. The remaining guesses, provided on 39.2% of the trials, tapped the same meaning and were classified as *guessed-same*. On 5.6% of trials, no response was provided within the eight-second timeframe. Guesses matching the meaning of the targets were classified as correct if they either matched the target word exactly (bar any obvious spelling mistakes) or both the guess and the target shared the same word stem and had closely related meanings (e.g., *dance* – *dancer*, but not *ball* - *baseball*). On that basis, responses to 3.4% of all guess trials were classified as correct. We also conducted a latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998) to investigate the relatedness between guesses and targets in both conditions. LSA allows for establishing the degree of similarity in meaning between pairs of words, with relatedness being represented on a 0-1 scale. This analysis confirmed that incorrect guesses were less related to the targets in the guessed-different condition (M = .08, SD = .02) than in the guessed-same condition (M = .25, SD = .06), $t(31) = 18.114$, $p < .001$, d = 3.20.

Test phase. Given the focus of the present study, of main interest in the guess condition are trials on which participants' guesses did not match the target word. Correct guesses were thus excluded from the analyses, following Kornell et al. (2009). All descriptive statistics are provided in Table 1.

A one-way repeated-measures Analysis of Variance (ANOVA) that analyzed the influence of trial type (read, guessed-same, guessed-different) on cued-recall accuracy revealed a significant difference between the conditions, $F(2,62) = 5.633$, $MSE = 0.017$, $p = .006$, $\eta_p^2 =$

[2] As mentioned in the Author Note, data files (including stimuli) for all experiments can be found on https://osf.io/k6v3b

.154. To further investigate this difference, follow-up *t*-tests were conducted. A comparison of the read and guess-same conditions should provide a replication of previous results investigating the memory benefits of guessing using a paired-associates paradigm (e.g., Bridger & Mecklinger, 2014; Knight et al., 2012; Kornell et al., 2009). When performance on the guessed-same trials on which incorrect guesses were provided was compared to that from the read trials, a benefit of guessing was revealed, $t(31) = 2.941$, $p = .006$, d = 0.52. This result remains consistent with the semantic accounts of the benefits of guessing.

The comparison between cued-recall performance for read versus guessed-different pairs failed to reveal a significant difference, $t(31) = 1.325$, $p = .20$, d = 0.23. Although performance in the guessed-different conditions was numerically higher, this difference was very small compared to the benefits of guessing that accrued for guessed-same pairs. For completeness, we also compared directly these two types of pairs for which guesses were generated, confirming higher performance for the guessed-same rather than guessed-different condition, $t(31) = 2.350$, $p = .025$, d = 0.42.[3] These results are seemingly inconsistent with the attentional accounts of the benefits of guessing.

Together, the results offer two insights into why the benefits of guessing emerge. The addition of the novel guessed-different condition allowed us to simultaneously assess the semantic and attentional accounts of the benefits of guessing in a single experimental design. The results for the guessed-same pairs confirmed the predictions of the semantic accounts of guessing: whenever there is a pre-existing semantic association between a cue and a target, guessing incorrectly confers benefits for encoding, either because it serves to semantically activate the target even before it is presented or because it provides a mediator – an incorrect guess – that links the cue to its target. These semantic accounts predict less of a benefit when cues and targets remain unrelated, or – as in our paradigm – initial guesses do not match the actual meaning of the target, in which case there is no

[3] For all experiments, we also provide analyses in which data from all guess trials were included, independent of whether the guess was correct or not. For these analyses, correct guesses were included in the same-guessed condition. The ANOVA once again revealed a significant difference between the conditions, $F(2,62) = 6.788$, $MSE = 0.016$, $p = .002$, $\eta_p^2 = .180$. Performance was higher for the guessed-same than read items, $t(31) = 3.320$, $p = .002$, d = 0.59. Memory was also better for guessed-same than guessed-different items, $t(31) = 2.602$, $p = .014$, d = 0.46. As guesses could be excluded on the basis of correctness only from the guessed-same category, the comparison between read and guessed-different items was necessarily identical to that reported in the main text and thus is not reported here.

semantic activation of the target at the time of guessing, and any generated guess also remains unrelated to the target and thus is unlikely to serve as an effective mediator. The lack of benefits of guessing for the guessed-different pairs remains thus consistent with the semantic accounts.

The results seem to contradict the predictions derived from the attentional accounts, which argue that the benefits of guessing arise when feedback draws attention to the correct answer. Feedback in the form of a target associated with the alternative meaning of a cue should contradict expectations expressed by participants in their initial guesses, creating a feeling of surprise. One could even argue that surprise generated by feedback pointing to an alternative meaning of a homograph should be particularly strong, creating fertile grounds for the attentional mechanisms to operate, although one should also acknowledge that with all experimental cues being homographs, this feeling of surprise may diminish in the course of the procedure. In addition to that, the homograph task might also increase the curiosity to learn the correct answer as compared to the standard weak-associates task. This is because at the time of guessing it is impossible to predict whether the guess would even be related to the target. For those participants who are aware of the nature of the cues it might be particularly interesting to learn whether their guess was 'right' (i.e., tapped the meaning of the cue suggested by the later presented target) or not. Yet despite the conditions seemingly promoting increased attention to feedback, for guessed-different pairs the benefits of guessing failed to emerge. The present results could thus be taken to suggest that there is no need for postulating multiple mechanisms for explaining the benefits of guessing, with semantic explanations providing satisfactory account of the empirical patterns – an issue to which we return in Experiments 3 and 4.

Before moving forward with the discussion of the mechanisms responsible for the benefits of guessing, it is certainly worth ensuring that the results of Experiment 1 are robust. Given the novelty of the findings concerning guessed-different pairs, Experiment 2 was conducted in order to ascertain their replicability. We also took this opportunity to provide more in-depth insight into how participants perceive the different pair types at the time of study. To this aim, straight after the presentation of each pair participants were required to provide a judgement of learning (JOL) for this pair – that is, rate their confidence in recalling the second word from that pair at test when presented with the first word. As previous studies have shown (Huelser & Metcalfe, 2012; Potts & Shanks, 2014; Yang et al.,

2017), participants are generally unaware that guessing boosts performance as compared to being presented with the intact pair (although see Potts et al., 2018, for indications that this may change with task experience). Our aim was to reveal whether this finding extends to our homograph-cue task, with a particular interest in participants' memory predictions for guessed-different pairs.

Experiment 2

Method

Participants. Thirty-two students of the University of Sheffield (eight male; age range: 18-30) participated in exchange for course credit or monetary compensation.

Materials and Design. The materials were the same as those used in Experiment 1. The design was similar to that of Experiment 1, with the addition of immediate JOLs to the study phase being the only major change.

Procedure. The procedure closely followed that of Experiment 1, with one exception. After the presentation of each full word pair, participants were asked to rate their confidence in recalling the second word from that pair when cued with the first word, on a scale from 1 ("Not confident at all") to 5 ("Absolutely certain"). The time for providing the JOL was not limited.

Results

Study phase. On 54.4% of all guess trials, participants provided guesses mismatching the meaning of the target, and on 41.1% of trials guesses matching the meaning of the target. Guesses provided on 4.6% of all trials were classified as correct. No guesses were provided on 4.5% of trials.

Test phase. Descriptive statistics are presented in Table 1. Memory performance results provide a replication of the findings of Experiment 1. A one-way repeated-measures ANOVA performed on cued-recall data again revealed a significant difference between the conditions, $F(2,62) = 18.520$, $MSE = 0.013$, $p < .001$, $\eta_p^2 = .374$. Performance was better for guessed-same than for read items, $t(31) = 5.016$, $p < .001$, $d = 0.89$, again demonstrating the memory benefit of incorrect guessing. Most importantly from the perspective of this study, the difference in recall between guessed-different and read items again failed to reach statistical significance, $t(31) = 1.234$, $p = .23$, $d = 0.22$, and this time it numerically favored read over guessed-different items. The difference in performance between the guessed-

same and guessed-different items was again present, with performance for guessed-same items higher by 16 percentage points, $t(31) = 5.638$, $p < .001$, d = 1.00. Together, these results are again seemingly inconsistent with the attentional accounts, but provide further support for the semantic accounts.

Another one-way ANOVA was conducted on JOL data for read, guessed-same, and guessed-different items. This ANOVA revealed a significant difference between the conditions, $F(2,62) = 21.85$, $MSE = 0.086$, $p < .001$, $\eta_p^2 = .413$. JOLs were higher for read than guessed-same items, $t(31) = 3.600$, $p = .001$, d = 0.64. This constitutes a reversal of the pattern found in the memory performance data which showed lower performance when no attempt at guessing the target was made. It is also consistent with previous studies (Huelser & Metcalfe, 2012; Potts & Shanks, 2014; Yang et al., 2017) in showing participants' lack of awareness of the guessing benefits. JOLs for guessed-different items were also lower than for read items, $t(31) = 6.295$, $p < .001$, d = 1.11. This pattern also does not match memory performance which was comparable across these conditions. Generally thus, participants predicted lower performance when guessing rather than reading. It is perhaps worth noting, however, that JOLs for guessed-same items were higher than for guessed-different items, $t(31) = 3.256$, $p = .003$, d = 0.58. This particular pattern mirrored the one revealed in recall data, indicating that even though participants may underappreciate guessing as a learning strategy they still correctly grasp that items for which targets are related to their guesses will be remembered better.[4]

Together, the results of Experiments 1 and 2 offer a consistent story. They demonstrate that the existence of a semantic relationship between the cue and the target is not sufficient to improve recall performance when guessing the target is required: the right relationship needs to be assumed by participants at the time of guessing. This is consistent

[4] When all guesses were included in the analyses, the results for memory performance mirrored those for incorrect guesses. The ANOVA was once again significant, $F(2,62) = 20.82$, $MSE = 0.013$, $p < .001$, $\eta_p^2 = .402$. As in Experiment 1, two follow-up t-tests were conducted: $t(31) = 5.330$, $p < .001$, d = 0.94, for the comparison of memory performance for read and guessed-same items, and $t(31) = 6.129$, $p < .001$, d = 1.08 for the guessed-same to guessed-different comparison. For JOLs, the ANOVA again revealed a significant difference depending on item type, $F(2,62) = 24.61$, $MSE = 0.093$, $p < .001$, $\eta_p^2 = .443$. The difference between JOLs for guessed-same and guessed-different items was again significant, $t(31) = 5.435$, $p < .001$, d = 0.96. However, that between read and guessed-same items was not, $t < 1$, $p = .547$, d = 0.11. This is due to the fact that this time, guessed-same items included those that were guessed correctly by participants in the study phase, which were almost invariably assigned high JOLs.

with the semantic accounts which posit that it is the processing of the meaning of the cue at the guessing stage – either through activating the associated semantic network, or by generating a specific related guess that can later serve as a mediator – that matters. The presentation of corrective feedback alone failed to produce significant changes in later cued-recall performance. This stands in contrast to the assumption that feedback that is inconsistent with participants' initial guesses becomes better encoded as a result of being more attention-drawing and the operation of compensatory mechanisms that minimize the discrepancy between the expected and actual answer (e.g., Carrier & Pashler, 1992).

It is this latter finding of lack of guessing benefits for guessed-different (as compared to read) pairs that Experiments 3 and 4 are addressing. Experiments 1 and 2 were consistent in detecting no benefits of guessing for guessed-different pairs, which we view as speaking against the attentional accounts of the benefits of guessing in our procedure. Nevertheless, it again needs to be noted that attentional mechanisms have previously been argued to provide the best explanation for the advantage of guessing over reading in a related, although not identical task (Potts & Shanks, 2014), and there is evidence that guessing both increases curiosity to know the answer and boosts memory performance for that answer (Potts et al., 2018). It thus stands to reason not to fully dismiss the attentional accounts before our initial findings are shown to be generalizable across experimental conditions.

In the study that introduced the attentional account of the benefits of guessing, Potts and Shanks (2014; see also Potts et al., 2018) demonstrated how guessing boosts memory for two different types of materials: rare English words coupled with their more common synonyms (Experiments 1, 2a, & 3) as well as Euskara-English pairs (Experiment 2b). These materials made their learning and testing tasks differ on several levels from those used in the standard related-pairs paradigm. In the study phase, the cue (e.g., *frampold*) was close to meaningless before the feedback was presented, meaning that there was no semantic network to be activated at the time of guessing. This, coupled with the fact that the cue and the target by definition shared the same meaning, also limited the ways in which cue-to-target associations could be established at study. In this case, it is possible that guessing in the procedures developed by Potts and Shanks augmented – via the mechanism of attentional boost – not the cue-to-target associations but memory for targets themselves, as they become the focus of attention when feedback is provided. Memory for targets alone might be far more crucial in the Potts and Shanks' paradigm than in the standard related-

pairs paradigm for two reasons. First, Potts and Shanks used multiple-choice rather than cued-recall tests. Second, the paradigms also differ in terms of strategies facilitating correct responding at test. For semantically related pairs, the task is to identify for each cue one word – from a unique subset of related items that is restricted by the meaning of a given cue – that has the strongest link to this cue. When a cue engenders no particular semantic associations, it is likely that there are no cue-specific search sets; instead, the search set should simply be the pool of targets remembered from the study phase. It is thus good memory for targets that seems to be a necessary prerequisite for successful cue-to-target matching at test.

The same attentional mechanism could also augment memory for targets in our Experiments 1 and 2, but any such benefit might have gone unnoticed due to the memory task we employed. By administering the cued-recall test with original cues, we created testing conditions that tapped cue-to-target associations – the presumed locus of the benefits of guessing according to semantic accounts (Hays et al., 2013) – but at the expense of memory for targets alone, thus limiting potential contribution of the attentional mechanisms to successful memory performance. At the same time, on guessed-different trials the dominant meaning of the cue – that is, the one related to the guess but not the target – might have interfered at test with target retrieval.[5] In order to test for the benefits from attentional boost to target memory, a more sensitive test is thus needed – one that would limit the use of associations between original cues and their targets as well as the interference generated by the dominant meaning of the cue, while at the same time keeping other demands of the memory task comparable to the conditions utilized in Experiments 1 and 2. To this end, in Experiments 3 and 4 we employed a cued-recall test but modified the test materials, substituting the original cues with *extra-list* (or *independent*) cues, which have been used previously in a variety of paradigms to isolate target memory and limit interference (e.g., Anderson, 2003; Anderson & Green, 2001; Nelson, Kitto, Galea, McEvoy, & Bruza, 2013). The new cues were semantically related to both the original cue and the target (e.g., *bomb* for the studied pair *arms-nuclear*), but were not used before in the experiment. If guessing incorrectly results in an attentional boost which benefits

[5] We would like to thank an anonymous reviewer for this suggestion.

memory for targets, we would expect better performance for the guessed-different than the read pairs.

The change in the nature of the final test also raises questions about performance for guessed-same pairs. First, let us consider the predictions stemming from the two semantic accounts. The mediator account clearly predicts no benefits of guessing when extra-list cues are used at test because only with the original cues can the whole original cue-mediator-target path be activated. The semantic activation account is more ambiguous because although it has been argued that semantic activation primarily benefits memory for cue-to-target associations (Hays et al., 2013), in which case no benefits should emerge in the extra-list-cue test, it is also possible that targets themselves benefit from such activation, producing benefits across various tests. Second, considering the attentional accounts, both the surprise and curiosity account seem to predict better target encoding for guessed-same as compared to read pairs. However, the role of surprise for guessed-same pairs should be relatively limited, less pronounced than for guessed-different pairs, as there should be greater discrepancy between the guess and the correct answer for guessed-different as compared to guessed-same pairs. Thus, although the surprise account seems to predict the benefits of guessing for guessed-same pairs as compared to read pairs, it would also seem to predict these benefits to be smaller than the benefits accruing for guessed-different pairs. By contrast, the curiosity account – which assumes that the benefits of errorful learning stem from processes starting even before feedback is provided – seems to predict equivalent benefits for guessed-same and guessed-different pairs.

In order to provide a stronger test of the attentional explanations of guessing benefits, we conducted two experiments using the same design as Experiments 1 and 2, but changing the nature of the cues employed in the final cued-recall test. Again, both experiments utilized exactly the same design and procedure, with the exception of the additional JOL task which was included in Experiment 4 but not Experiment 3.

Experiment 3

Method

Participants. Thirty-two students of the University of Sheffield (eight male; age range: 18-30) participated for course credit or monetary compensation.

Materials and Design. The lists used for study were the same as those used in Experiment 1. An additional 120 words were sourced from the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998). These were associated with the targets, with a mean forward association of 0.23. These words were used for the creation of two new test lists. The lists comprised the same targets that were used for the study lists, but the cues were replaced. Each cue now had a single meaning that was related to one of the targets (e.g., targets *hug* and *nuclear*, both studied with the cue *arms*, were now cued by words *embrace* and *bomb*, respectively; see Figure 1 for a schematic representation of the new test phase). The design of the experiment was the same as that of Experiment 1.

Procedure. The procedure resembled that of Experiment 1, with three important changes. First, test instructions provided after the practice phase and again before the test made it clear that the participants' task at test would be to recall second words from each studied pair when cued with a *new* word, related to both the original cue and the target, and an example was provided to participants to ensure the right understanding. Second, at test homograph cues from the study phase were replaced by new cues, related to both the original cues and targets. Third, first letters of the targets were provided at test in order to decrease task difficulty, as pilot data revealed too low recall performance when extra-list cues were used with no additional memory aid.

Results

Study phase. Participants' guesses mismatched the meaning of the target word on 50.3% of guess trials, and matched it on 41.2% of trials. Responses to 3.9% of all guess trials were classified as correct. No guesses were provided on 8.5% of trials. Given the slight change in task instructions which might have affected how participants approached the study and guessing phase, we again conducted an LSA to ensure that participants' guessing patterns have not changed. This analysis revealed a very similar pattern to the one from Experiment 1: incorrect guesses in the guessed-different condition were again less related to targets (M = .08, SD = .02) than those in the guessed-same condition (M = .22, SD = .06), $t(31) = 15.17$, $p < .001$ d = 2.68.

Test phase. Descriptive statistics are presented in Table 1. A one-way repeated-measures ANOVA revealed significant differences in cued-recall performance depending on trial type, $F(2,62) = 5.914$, $MSE = 0.014$, $p = .004$, $\eta_p^2 = .160$. Despite providing participants with extra-

list rather than original cues in the test phase, the benefit of incorrect guessing was again revealed in the comparison of memory performance for read and guessed-same item, $t(31) = 3.076$, $p = .004$, d = 0.54. This time, however guessed-different items were also remembered better than read items, $t(31) = 3.805$, $p < .001$, d = 0.67. This finding contradicts those from Experiments 1 and 2 in which no significant difference in recall performance was found between these two classes of items. Also in contrast to Experiments 1 and 2, the difference in recall between guessed-same and guessed-different items was not significant in this experiment, $t < 1$, $p = .902$, d = 0.02.[6] Together, the results suggest that guessing what the target might be – as compared to merely reading it – improves memory performance, independent of whether this guess is related to the meaning of the actual target or not. The results provide the support for the attentional accounts that was missing from Experiments 1 and 2. In particular, they seem to be consistent with the curiosity variant of the attentional account given the lack of difference in memory performance for targets learnt on guessed-same versus guessed-different trials.[7]

Experiment 4

Method

Participants. Thirty-two students of the University of Sheffield (12 male; age range: 18-54) participated in exchange for course credit or monetary compensation.

Materials, Design, and Procedure. The materials were the same as those used in Experiment 3. The study phase was the same as in Experiment 2, with JOLs being elicited after each word pair presentation. The JOL prompt was adapted to accommodate the difference in the test phase: participants were told to rate their confidence in recalling the second word from each pair when cued with a new word related to both the first and second words. The test phase along with the test instructions were the same as in Experiment 3, with extra-list cues being presented instead of homograph cues from the study phase.

---

[6] When all guesses were analyzed, the results remained the same: $F(2,62) = 7.751$, $MSE = 0.012$, $p < .001$, $\eta_p^2 = .200$ for the one-way ANOVA, $t(31) = 3.788$, $p < .001$, d = 0.67, for the comparison of memory performance for read and guessed-same items, and $t < 1$, $p = .784$, d = 0.04, for the guessed-same to guessed-different comparison.

[7] To provide further support this claim, we conducted a Bayesian *t*-test using the JASP software (JASP Team, 2018) to quantify the evidence for the null hypothesis. This analysis revealed that the null hypothesis was 5.26 times as likely as the non-directional alternative hypothesis.

Results

Study phase. Participants' guesses mismatched the meaning of the target word on 50.4% of guess trials, and matched it on 44.1% of trials. Responses to 5.3% of all guess trials were classified as correct. No guesses were provided on 5.5% of trials.

Test phase. Descriptive statistics are presented in Table 1. The results for cued-recall performance fully replicated those from Experiment 3. A one-way repeated-measures ANOVA on cued-recall data was again significant, $F(2,62) = 3.557$, $MSE = 0.014$, $p = .034$, $\eta_p^2 = .103$. The difference in cued-recall performance between the read and guessed-same targets was again significant, $t(31) = 2.417$, $p = .022$, d = 0.43, and so was the difference between read and guessed-different items, $t(31) = 2.885$, $p = .007$, d = 0.51, demonstrating the positive impact of guessing the target on memory performance when memory for targets was assessed. As in Experiment 3, the difference in recall between guessed-same and guessed-different targets failed to reach significance, $t < 1$, $p = .782$, d = 0.05. Once again, these results provide support for the attentional account of benefits of guessing, and especially so for the curiosity-based explanation.[8]

A one-way ANOVA performed on JOL data again revealed a difference between the three item types, $F(2,62) = 46.06$, $MSE = 0.053$, $p < .001$, $\eta_p^2 = .598$. As in Experiment 2, JOLs failed to predict cued recall performance. JOLs were higher for read than for guessed-same items, $t(31) = 2.511$, $p = .017$, d = 0.44 – a pattern opposite to that found for memory performance. Also, JOLs were higher for read than for guessed-different items, $t(31) = 9.735$, $p < .001$, d = 1.72, which – given that in the present experiments the benefits of guessing were obtained even for those items for which the initial guess referred to a different meaning of a homograph – also failed to track memory performance. These results join the results of Experiment 2 in showing that people generally predict lower performance when guessing. JOLs for guessed-same items were also higher than those for guessed-different items, $t(31) = 7.055$, $p < .001$ d = 1.247, whereas the recall data showed equivalent memory

[8] A Bayesian *t*-test revealed that the null hypothesis assuming no difference between the guessed-same and guessed-different condition was 5.11 times more likely than the non-directional alternative hypothesis.

performance for these two classes of items. [9] In this case, JOL results are the same as in Experiment 2, while memory performance is not.

General Discussion

In four experiments, we have elucidated several of the conditions under which attempting to guess the correct answer at study leads to benefits in terms of memory performance at test as compared to merely reading what the correct answer is. Experiments 1 and 2 introduced a procedure with homograph cues that was designed to test predictions stemming from two theoretical accounts of guessing benefits. What we termed the semantic accounts – semantic activation arising at the time of guessing or the use of initial guesses as mediators – predicted that activating a meaning of the cue that matches the meaning of the later presented target is necessary for the benefits of guessing to be found, as it is the semantic processing of the *cue* before the target is even presented that prepares the grounds for better encoding. By contrast, the attentional accounts – curiosity evoked by the process of guessing and resolved by the presentation of the target or surprise elicited by detecting an inconsistency between the initial guess and the target – predicted that the benefits of guessing should be dependent on the processing of the *target*, and so the initial interpretation of the cue at the time of guessing is less relevant. The results of Experiments 1 and 2 were consistent in showing support for the semantic accounts of the benefits of guessing as these benefits were found only when – as revealed by the initial guesses – the meaning derived from the homographic cues was consistent with the meaning of a subsequently presented target. We further speculated that the unique support for the semantic accounts over the attentional accounts in Experiments 1 and 2 may derive from the conditions of testing we employed. By using original cues in the final test, we effectively maximized the role of cue-to-target associations, thus limiting the role of memory for

[9] As in previous experiments, when all guesses (including the correct ones) were analyzed, the results remained relatively unchanged. The ANOVA on cued-recall data was significant, $F(2,62) = 3.849$, $MSE = 0.013$, $p = .027$, $\eta_p^2 = .110$. Read items were recalled less often than those in the guessed-same condition, $t(31) = 2.536$, $p = .016$, $d = 0.45$. There was no difference in cued recall performance between guessed-same and guessed-different items, $t < 1$, $d = 0.05$. For JOLs, the ANOVA was significant, $F(2,62) = 50.48$, $MSE = 0.060$, $p < .001$, $\eta_p^2 = .620$. JOLs were higher for guessed-same than for guessed-different items, $t(31) = 8.456$, $p < .001$, $d = 1.50$, but there was no difference between those assigned to read versus guessed-same items, $t < 1$, $d = 0.02$. As in Experiment 2, this is most likely because correctly guessed items almost exclusively were assigned JOLs in the upper range of the scale.

targets that could benefit from additional attention triggered by the initial guessing process. In addition to that, interference from the originally guessed meaning of the cue might have impaired access to targets at test, further obscuring any potential effects on target memory. Accordingly, in Experiments 3 and 4 we attempted to eliminate the use of cue-to-target associations and mitigate interference at the time of test – by using the extra-list cue technique – in order to gain insight into the potential benefits of guessing for target memory alone. This time the results were consistent with predictions stemming from the curiosity-based attentional account, as guessing improved memory performance as compared to reading independent of whether the initial interpretation of the cue was consistent with the later presented target or not. Together, the present results demonstrate that the benefits of guessing are multifaceted, drawing on different mechanisms depending on the nature of a memory task.

Our study underscores the importance of considering both the conditions of learning and testing when predicting the effects initial guessing should have on memory. In terms of learning, our results highlight the key role of the relationship between the meaning of the cue at the time of guessing and the meaning of the target in determining how incorrect guessing benefits memory. The results of Experiments 1 and 2 indicate that guessing alone has little impact on memory for the cue-to-target associations unless the meaning of the cue activated at the time of guessing is consistent with the meaning of the later presented target. These results remain consistent with numerous reports demonstrating no benefits of guessing when participants are asked to study unrelated pairs of words (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012). However, as Experiments 3 and 4 demonstrate, the act of trying to guess the correct answer is not without consequences for memory even if the meaning of the target does not become activated during guessing. Under these conditions, guessing is sufficient to improve target memory as compared to reading the full pair intact. This improvement of target memory occurs in spite of a much shorter target presentation time in the guessing as compared to the reading condition (5 vs. 13 seconds). This stronger encoding of targets seems to be responsible for previous reports of benefits of guessing under conditions of no pre-existing semantic relationship between cues and their targets (Potts & Shanks, 2014; Potts et al., 2018).

In terms of testing, our study demonstrates how different testing conditions are sensitive to different aspects of memory traces arising as a consequence of initial guessing and the

feedback that follows it. To date, studies on the benefits of guessing have mostly focused on cued-recall tests in which original cues were used to tap memories for the studied pairs (although see Potts & Shanks, 2014, and Potts et al., 2018, who used recognition tests with and without cues). We argue that such tests are sensitive mostly to the strengthening of cue-to-target associations at the time of encoding and thus they can be preferentially used to assess the benefits of guessing that depend on the pre-existing association between cues and their targets (see Hays et al., 2013). In order to tap into other aspects of memory traces, different memory tests are necessary. We have shown here how a cued-recall test employing extra-list cues can tap into memory for targets (see Anderson & Spellman, 1995; Hanczakowski & Mazzoni, 2013, for a related logic) as well as potentially overcome interference caused by the (subjectively) dominant meaning of the cue (Anderson, 2003), revealing benefits of guessing that extend to situations in which associations between cues and targets may play a much smaller role. It stands to reason that other tests can serve to dissociate the influence of various mechanisms responsible for the benefits of guessing. A widely used method for distinguishing between item and associative memory is, for example, to contrast findings from item and associative recognition tests (e.g., Naveh-Benjamin, 2000), an avenue that awaits investigation in the context of the benefits of guessing.

Our study provides new insights into the mechanisms by which guessing benefits memory, but the theoretical work concerning this phenomenon is still far from finished. First, our study made no attempt at differentiating between semantic activation and 'guesses as mediators' accounts which we referred to jointly as semantic mechanisms. Further work is thus necessary to elucidate how one or both of these mechanisms affect encoding of cue-to-target associations.

Second, our study still leaves some questions open regarding the way guessing affects target memory. We argue that our results indicate that a target is strengthened as a result of guessing independently of whether the meaning of this target is activated before its presentation, as seen in Experiments 3 and 4 and consistent with the curiosity-based attentional mechanism of memory boost. However, it remains an open question whether semantic activation contributes to this strengthening. On the one hand, the similarity of results for guessed-same and guessed-different pairs in these experiments suggests a common mechanism – attentional boost resulting from the curiosity experienced after a

guess is made and before the correct answer is presented. On the other hand, similar patterns across two conditions do not rule out the possibility that these patterns reflect two different mechanisms – attentional boost for guessed-different pairs and semantic activation for guessed-same pairs, or perhaps two different combinations of mechanisms – curiosity and surprise for guessed-different pairs and curiosity and semantic activation for guessed-same pairs. For now, it seems that the principle of parsimony should suggest that semantic accounts are sufficient to describe the benefits of guessing when the meanings of the cues and targets are related and the curiosity-based attentional account is sufficient to describe the benefits of guessing when they are unrelated, but more nuanced versions of these theories are possible and should be subjected to further empirical tests.

Furthermore, it is also unclear what – if any – the role of surprise might be in driving the attention-based benefits of guessing. As discussed earlier, it can be assumed that – on average – feedback should more often be seen as surprising for guessed-different pairs. If surprise were to boost attention to feedback in our paradigm, in Experiments 3 and 4 it should have been revealed in even higher performance for guessed-different as compared to guessed-same pairs. This was not the case, which led us to conclude that curiosity is sufficient to explain the pattern of results with extra-list cues. However, it needs to be noted that effect sizes in Experiments 3 and 4 for comparisons with the read condition were higher for guessed-different than for guessed-same pairs (Cohen's ds of 0.67 vs. 0.54 in Experiment 3 and 0.51 vs. 0.43 in Experiment 4), suggesting that surprise might after all contribute to the benefits of guessing.[10] This issue also awaits further empirical investigations.

Finally, it is important to note that the main methodological innovation introduced in our study – the use of homographic cues – comes at a price, especially when it comes to elucidating the role of guessing in strengthening cue-to-target associations. Homographic cues provide an opportunity to disentangle situations in which semantic activation is either likely (guessed-same pairs) or unlikely (guessed-different pairs) to spread to subsequently presented targets. Their use also enables the examination of the benefits of errorful learning under conditions that maximize both curiosity – for all experimental pairs – and surprise in the case of guessed-different pairs. However, the use of homographic cues also limits to some extent the clarity of theoretical conclusions that can be derived from our

[10] We would like to thank an anonymous reviewer for this observation.

data. When homographic cues are used, experimental conditions are not created by experimenter-controlled manipulations but instead are defined by participants' guesses. This creates a situation in which at least some of the obtained results can be discussed in terms of item-selection artifacts, as from participants' perspective the cues from guessed-same and -different pairs might have been subjectively different in the first place.

We also noted that the procedure used in Experiments 1 and 2 could produce at test interference from the initial guess: it could be that for guessed-different pairs this interference was strong enough to overshadow the memory benefits of errorful learning. While we acknowledge that interference might well have contributed to the lack of memory benefits for guessed-different as compared to read pairs, we would like to underscore that there is strong evidence suggesting that guessing benefits stemming from the strengthening of cue-to-target associations should not be seen merely as a by-product of interference-prone testing conditions, as they can be detected even when interference is controlled for at test (Hays et al., 2013). We do believe the semantic activation account – which provides an intuitive account for the overall pattern of results when original cues are used at test and which remains consistent with a bulk of previous studies on errorful learning (e.g., Bridger & Mecklinger, 2014; Grimaldi & Karpicke, 2012; Hays et al., 2013) – receives additional support from our findings, but the limitations inherent to the homographic cue methodology underscore the importance of approaching theoretical issues from a multitude of empirical angles in order to home in on sound theoretical underpinnings of the phenomena of interest.

The research on various effects of testing – including the specific case of retrieval attempts preceding the actual study – has gained importance in recent years due to an increased interest in how the science of memory can inform educational practice. Before guessing could be used in classrooms as an educational tool for improving student learning, however, it is necessary to understand how and under what conditions the benefits of guessing emerge in different learning situations – an issue that should be addressed by future applied-oriented research. An optimistic conclusion from the present study is that the benefits of guessing are quite general – facilitating memory for associations and to-be-remembered items – even though the specific memory improvements may not be detectable under all testing conditions. Importantly, none of the presented experiments found a pattern in which memory performance would suffer as a result of guessing when

compared to the reading condition. Thus, guessing clearly seems to be a viable learning strategy. Guessing in the absence of pre-existing associations between cues and targets – such as in the case of vocabulary learning – is likely to improve memory for targets, although additional work will probably need to be invested into creating associations between these targets and their cues. However, when these associations are finally established, guessing should still benefit performance because semantic processes engaged in guessing can serve to strengthen these associations even further.

Given that guessing can serve as an effective learning technique, one obvious question concerns students' willingness to employ this technique when they engage in self-regulated study. The perception that a given technique is effective in supporting learning should serve as a precondition – not always sufficient but certainly necessary – for using it. Regarding the benefits of guessing, previous studies employing JOLs have shown no appreciation on participants' part of the memory benefits of guessing (Potts & Shanks, 2014; Yang et al., 2017; see also Huelser and Metcalfe, 2012, for a different method for arriving at the same conclusion). In Experiments 2 and 4 we have replicated these results, demonstrating that although memory benefits from guessing, participants actually predict higher performance when they engage in reading to-be-remembered material. A novel finding concerning JOLs emerging from our study is that participants consistently provided the lowest JOLs to guessed-different pairs. This shows that presenting feedback that defies expectations seems to undermine participants' belief in future retrievability of that feedback. This is a potentially important consideration given that the attentional accounts of the benefits of guessing assigns these benefits precisely to the fact that feedback remains inconsistent with the initial predictions.

Finally, it is worth comparing the correspondence between JOLs and memory performance between Experiments 2 and 4 (see Table 1). In Experiment 2, even though JOLs failed to predict the guessed-same > read difference, consistently with previous studies, they nevertheless accurately tracked the memory benefit for guessed-same as compared to guessed-different items. On that basis alone, it could be claimed that participants utilized cues that were diagnostic of their future memory performance when assigning a JOL to a guessed pair. A different story emerges, however, when Experiment 4 is taken into account. Even though the JOL pattern remains almost identical across the two experiments in which JOLs were collected, the same does not apply to memory performance: in Experiment 4, in

which independent cues were used, there was no difference in cued recall between the two guessed conditions. As a result, there was a gross mismatch between participants' metacognitive assessments and the actual memory scores that these assessments were expected to predict.[11] These findings serve to underscore a recent point made by Zawadzka, Simkiss, and Hanczakowski (2018) regarding cue diagnosticity in the JOL task. Zawadzka et al. defined a diagnostic cue as one that feeds into JOLs and memory performance simultaneously and in the same direction; conversely, a non-diagnostic cue is one which affects JOLs without a corresponding effect on performance. They noted that a cue for metacognitive judgments is never diagnostic or non-diagnostic in itself: variations in the task, including changes to testing conditions, can create or break correspondence between JOLs and future memory performance. The pattern revealed in the present study is a clear instantiation of this issue.

Conclusion

In four experiments using paired associates with homograph cues, we have revealed that the benefits of incorrect guessing for memory performance cannot be attributed to a single mechanism. Experiments 1 and 2 demonstrated that the guessing benefits that emerge in the commonly used paradigm in which participants attempt to predict the target when presented with a related cue, and later are tested with the original cue, require tapping the right semantic relationship between the cue and the target at the time of guessing and before the presentation of feedback. Experiments 3 and 4 revealed an additional mechanism through which the benefits of guessing may arise: an improvement to target memory. This benefit is masked in the standard paradigm which relies on the memory for cue-to-target associations, but emerges when a test designed specifically to tap target memory and limit interference is employed. It also does not require participants to correctly interpret the homograph cue at the time of guessing: in this case, it is the presentation of corrective feedback that matters. Together, the findings open new avenues for educationally-oriented research on employing errorful learning in classrooms.

[11] Note that the JOL patterns in Experiments 2 and 4 were almost identical despite the different tests employed; in each experiment, the test had been explained to participants, with examples, before the study/JOL phase commenced, and the JOL prompt presented on each trial clearly indicated what cues would be used at test.

References

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49,* 415-445. doi: 10.1016/j.jml.2003.08.006

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature, 410(6826),* 366-369. doi: 10.1038/35066572

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied, 24,* 43-56. doi: 10.1037/xap0000133

Bridger, E. K., & Mecklinger, A. (2012). Errorful and errorless learning: The impact of cue-target constraint in learning from errors. *Memory & Cognition, 42,* 898-911. doi: 10.3758/x13421-014-0408-z

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 918-928. doi: 10.1037/0278-7393.34.4.918

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology, 20*, 941-956. doi: 10.1002/acp.1239

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1491–1494. doi: 10.1037/0278-7393.27.6.1491

Carneiro, P., Lapa, A., & Finn, B. (2018). The effect of unsuccessful retrieval on children's subsequent learning. *Journal of Experimental Child Psychology, 166,* 400-420. doi: 10.1016/j.jecp.2017.09.010

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633–642. doi: 10.3758/BF03202713

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1563–1569. doi: 10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal*

of Experimental Psychology: Learning, Memory, and Cognition, 37, 1547–1552. doi: 10.1037/a0024140

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest, 14,* 4-58. doi: 10.1177/1529100612453266

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin and Review, 16,* 88–92. doi: 10.3758/ PBR.16.1.88

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory and Cognition, 38*, 951–961. doi: 10. 3758/MC.38.7.951

Griffiths, L., & Higham, P. A. (2018). Beyond hypercorrection: Remembering corrective feedback for low-confidence errors. *Memory, 26,* 201-218. doi: 10.1080/09658211.2017.1344249

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40,* 505– 513. doi: 10.3758/s13421-011-0174-0

Grimaldi, P. J., Poston, L., & Karpicke, J. D. (2015). How does creating a concept map affect item-specific encoding? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 1049-1061. doi: 10.1037/xlm0000076

Hanczakowski, M., & Mazzoni, G. (2013). Contextual match and cue-independence of retrieval-induced forgetting: Testing the prediction of the model by Norman, Newman, and Detre (2007). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 953-958. doi: 10.1037/a0030531

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,* 514– 527. doi: 10.3758/s13421-011-0167-z

JASP Team (2018). JASP(Version 0.8.4) [Computer software]

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,* 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. & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*, 585-592. doi: 10.1111/j.1467-9280.2008.02127.x

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 989–998. doi: 10.1037/a0015729

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14,* 194-199. doi: 10.3758/BF03194051

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1170-1187. doi: 10.1037/0278-7393.26.5.1170

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibition less spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*, 226–254. doi: 10.1037/0096-3445.106.3.226

Nelson, D. L., Kitto, K., Galea, D., McEvoy, C. L., & Bruza, P. D. (2013). How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & Cognition, 41,* 797-819. doi: 10.3758/s13421-013-0312-y

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida free association, rhyme, and word fragment norms*. Retrieved from http://w3.usf.edu/FreeAssociation/

Pyc, M. A., & Rawson, K. A. (2010, October 15). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi: 10.1126/science.1191465

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143,* 644–667. doi: 10.1037/a0033194

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*, 93–134. doi: 10.1037/0033-295X.88.2.93

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

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. R. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing**.** *Journal of Experimental Psychology: Applied**, 17,*** **382-395.** doi: 10.1037/a0026252

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi: 10.1111/j.1745-6916.2006.00012.x

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior, 22*, 153–163. doi: 10.1016/S0022-5371(83)90112-3

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences, 110,* 6313-6317. doi: 10.1073/pnas.1221764110

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review, 19,* 899–905. doi: 10.3758/s13423-012-0276-0

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications, 3:2.* doi: 10.1186/s41235-017-0087-y

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition, 42,* 1373–1383. doi: 10.3758/s13421-014-0454-6

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 1073-1092. doi: 10.1037/xlm0000363

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: a review of the forward testing effect. npj Science of Learning, 3, 8. doi: 10.1038/s41539-018-0024-y

Zawadzka, K., Simkiss, N., & Hanczakowski, M. (2018). Remind me of the context: Memory and metacognition at restudy. *Journal of Memory and Language, 101,* 1-17. doi: 10.1016/j.jml.2018.03.001