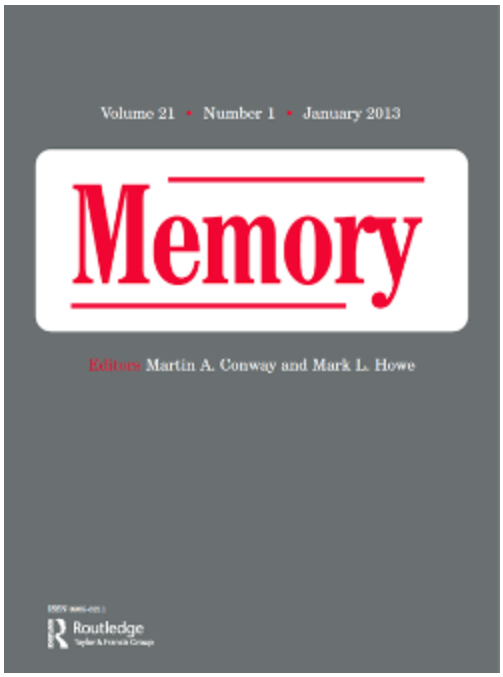


# Selective effects of errorful generation on recognition memory: The role of motivation and surprise.

| | |
|---|---|
| Journal: | *Memory* |
| Manuscript ID | MEM-OP 18-172.R2 |
| Manuscript Type: | Original Paper |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Seabrooke, Tina; University of Plymouth,<br>Mitchell, Christine; Plymouth University<br>Wills, Andy; University of Plymouth<br>Waters, Jessica; University of Plymouth<br>Hollins, Tim; University of Plymouth |
| Keywords: | errors, motivation, surprise, memory, education |
| | |

# Selective effects of errorful generation on recognition memory: The role of motivation and surprise.

Tina Seabrooke, Chris J. Mitchell, Andy J. Wills, Jessica L. Waters and Timothy J. Hollins

University of Plymouth



Running head: Errorful generation

Please address correspondence to:

Dr Tina Seabrooke

School of Psychology

University of Plymouth

Devon

PL4 8AA

United Kingdom

Email: tina.seabrooke@plymouth.ac.uk

**Abstract**

The current research examined the effects of errorful generation on memory, focusing particularly on the roles of motivation and surprise. In two experiments, participants were first presented with photographs of faces and were asked to associate four facts with each photograph. On Generate trials, the participants guessed two of the facts (Guess targets) before those correct facts, and another two correct facts (Study targets), were revealed. On the remaining Read trials, all four facts were presented without a guessing stage. In Experiment 1, participants also ranked their motivation to know the answers before they were revealed, or their surprise on learning the true answers. Guess targets were subsequently better recognised than the concurrently presented, non-guessed Study targets. Guess targets were also better recognised than Read targets, and recognition of Study and Read targets did not differ. Errorful generation also increased self-reported motivation, but not surprise. Experiment 2 showed that the results of Experiment 1 can outlive a 20-minute delay, and that they generalise to a more challenging recognition test. Together, the results suggest that errorful generation improves memory specifically for the guessed fact, and this may be linked to an increase in motivation to learn that fact.

*Keywords*: errors, motivation, surprise, memory, education

Tests play an important role in educational practices, in so much as they are routinely used to assess students' abilities, and to identify areas to focus on in subsequent study (Karpicke, Butler, & Roediger III, 2009; Kornell & Bjork, 2007). It is now well established, however, that tests are not only useful for identifying knowledge gaps; the very act of testing can also substantially improve memory on subsequent tests. This benefit of testing is known as the *testing effect* (for reviews, see Kornell & Vaughn, 2016; Roediger & Karpicke, 2006; for recent meta-analyses, see Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014). The testing effect suggests that tests are not neutral events that are only useful for assessing current knowledge. Rather, tests can also have powerful effects on subsequent learning and memory.

Although the overall effects of testing seem to be positive, there is some debate as to how memory is affected by any errors that are made when taking those tests. Advocates of the so-called *errorless learning* approach argue that errors will be reinforced during learning, leading to a perseveration of those errors on subsequent tests (e.g., Skinner, 1958; Terrace, 1963). Support for this view comes from studies showing that generating errors during learning sometimes impairs memory on subsequent tests (Baddeley & Wilson, 1994; Forlano & Hoffman, 1937; Kessels & De Haan, 2003; Squires, Hunkin, & Parkin, 1997). The more frequent finding in recent years, however, is that errors aid learning (Cyr & Anderson, 2015; Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Kane & Anderson, 1978; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell, 2014; Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009; Slamecka & Fevreiski, 1983; Tanaka, Miyatani, & Iwaki, 2019; Vaughn & Rawson, 2012; Yan, Yu, Garcia, & Bjork, 2014; Yang, Potts, & Shanks, 2017; Zawadzka & Hanczakowski, 2018). In these cases, failed tests are beneficial, and they can even be as beneficial as successful tests (Kornell, Jacobs Klein, & Rawson, 2015). Indeed, harder tests (which, by definition, are more likely to produce errors) appear to produce the greatest benefit to subsequent learning and retrieval (Carpenter & DeLosh, 2006; Kang, Mcdermott, & Roediger III, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Pyc & Rawson, 2009).

Much of the evidence for the benefits of failed tests comes from the so-called "unsuccessful retrieval" paradigm, which was first established by Kornell et al. (2009). In this task, participants are asked to learn weakly associated word pairs such as *pond-frog* and *whale-mammal*. On some trials – Generate trials – the participants are first presented with the cue (e.g., *pond*) and they are asked to guess the target (*frog*) before it is revealed. Since the word pairs are weakly related and have not been presented previously during the experiment, the participants have no way of knowing what the correct target is, and so their guesses are usually wrong. These Generate trials are then compared to control trials – Read trials – in which the participants are simply asked to study each word pair for the full trial duration, without first guessing the target. The typical finding is that, in a subsequent cued recall test, the participants recall more targets from the Generate condition than the Read condition (e.g., Kornell et al., 2009). Most importantly, this effect remains even when the Read trials are compared to just the Generate trials in which the participants incorrectly guessed the target at encoding. Thus, taking a test can improve subsequent memory when compared against an equivalent period of pure study time, even if that test results in failure. This effect is known as the benefits of *unsuccessful retrieval*.

Several recent experiments have used a variant of Kornell et al.'s (2009) unsuccessful retrieval task to provide especially strong support for the notion that failed tests can be beneficial (Potts, Davies, & Shanks, 2019; Potts & Shanks, 2014; Seabrooke, Hollins, Kent, Wills, & Mitchell, 2019). In these experiments, the cues are unfamiliar words – either very rare English words or foreign vocabulary – and the targets are the corresponding common English definitions (e.g., *roke-mist* or *gazta-cheese*). On Generate trials, then, the participants are given the task of guessing the definitions of words that they have never seen before. Their guesses are therefore inevitably incorrect on almost all trials. Nevertheless, the participants still show better memory for targets that they incorrectly guessed than targets that they simply studied at encoding. Following Potts and Shanks (2014), we refer to this novel vocabulary effect as an *errorful generation* effect.

Errorful generation effects appear to be robust, but it is also important to note that they only occur under certain test conditions. Errorful generation attempts have repeatedly been shown to improve performance on cue and target recognition tests that involve discriminating "old" items (those that were presented at encoding) from "new" foils that were not presented at encoding (Potts et al., 2019; Seabrooke et al., 2019). Comparable effects have also been observed with multiple-choice tests in which the participants have to identify the correct target for a given cue from among novel foils (Potts & Shanks, 2014; Seabrooke et al., 2019). These multiple-choice tests also assess target recognition because, since the foils are novel, the correct answer can be derived based on target familiarity. On the other hand, errorful generation attempts do not appear to improve performance on tests that assess *associative* memory, such as cued recall or associative recognition tests (Seabrooke et al., 2019). We note that this constraint contrasts with unsuccessful retrieval effects, where errors do typically improve subsequent cued recall of related word pairs such as *pond-frog* (e.g., Kornell et al., 2009). Thus, errorful generation effects concern the scenario where a participant generates an error when studying novel word pairs that do not have a pre-existing association. Such errors appear to improve subsequent recognition of the cues and targets on their own, but not the *association* between them (Seabrooke et al., 2019).

The present experiments were designed to examine the mechanism(s) that allow errorful generation attempts to boost target recognition. Our main aim was to explore *why* errorful generation attempts produce a benefit to recognition memory. Potts and Shanks (2014) suggested that attempting to guess the definitions of novel words in an encoding phase might boost motivation to learn those definitions, which could then improve processing of the correct answers. In initial support of this account, Potts et al.'s (2018) participants rated their curiosity to learn the definitions of rare English words more highly when they had previously (incorrectly) guessed the definitions than when they had just viewed the rare English words without guessing the definitions. In related experiments, Kang et al. (2009) found that participants' self-rated curiosity to learn the answers to trivia questions predicted how well they would recall those answers after a 1-2 week delay. These

results suggest that the process of generating guesses for questions increases curiosity to learn the correct answers, which could then improve processing of those answers.

Another possibility is that, after an errorful generation attempt, the presentation of the correct answer produces surprise. An increase in surprise might then serve to direct attention to the target more effectively than pure study trials (see also Brod, Hasselhorn, & Bunge, 2018; Carrier & Pashler, 1992; Kornell et al., 2009). This idea is consistent with influential theories of associative learning that suggest that learning is driven by prediction errors that arise from discrepancies between one's expectations and the actual outcome (e.g., Rescorla & Wagner, 1972).

Initial support for the "surprise" hypothesis comes from research showing that corrective feedback is processed more effectively following errors that are made with high confidence than errors that are generated with low confidence (the *hypercorrection effect*: e.g., Butterfield & Metcalfe, 2001, 2006; Fazio & Marsh, 2009; Griffiths & Higham, 2017). Brod et al. (2018) also recently reported that participants showed better memory for feedback that was inconsistent with their initial predictions than feedback that was consistent with their predictions. Inconsistent feedback also produced larger surprise-associated pupillary responses than consistent feedback, and the strength of these responses correlated positively with subsequent learning. Both of these findings are consistent with the idea that generating guesses that are followed by expectancy-violating feedback produces surprise, which then improves memory of the correct answer by increasing attention to the feedback.

A quite separate question addressed in the current experiments is whether errorful generation selectively boosts the processing (and consequently recognition) of the specific information that was guessed. One possibility is that errorful generation has a general effect, in that it may improve the processing of all information that is presented on Generate trials. Alternatively, errorful generation attempts could *impair* memory for other information. Previous research that has examined the specificity of errorful generation effects has typically used complex educational

materials, such as lecture videos or textbook passages. This research has produced mixed results. Some experiments have found that pretesting improves subsequent retrieval of the pretested material, but does not benefit other, non-pretested information (Carpenter, Rahman, & Perkins, 2018; Pressley, Tanenbaum, McDaniel, & Wood, 1990; Richland et al., 2009; Toftness, Carpenter, Lauber, & Mickes, 2018). Other studies, by contrast, have found that pretesting improves memory for both pretested *and* non-pretested information (Carpenter & Toftness, 2017), or that pretesting does not significantly benefit retrieval at all (Geller et al., 2017). Indeed, some experiments have even found that testing can *impair* the learning of new information when it is presented alongside the tested material (Davis & Chan, 2015; Finn & Roediger, 2013). Given the mixed status of the literature, a key aim of the present research was to examine the specificity of pretesting effects in a controlled, laboratory experiment with reasonably simple materials. It is also worth noting that most of these studies did not differentiate between the effects of answering pre-questions correctly and incorrectly. Richland et al. (2009) provided one exception, but there were mixed findings across the experiments with respect to the specificity of the pretesting effect. The present experiments specifically examined the selectivity of the *errorful generation* effect, in which participants are asked questions about novel materials, and their initial guesses are therefore almost always incorrect.

In sum, a number of recent experiments have shown that, in the context of novel vocabulary learning, errorful generation attempts can improve subsequent target recognition memory. A number of mechanisms have been proposed to explain the benefits of errorful generation. These mechanisms are not necessarily mutually exclusive, but their relative contributions are not well understood at present. In the current experiments, we therefore adapted the traditional errorful generation paradigm to assess the specificity of the errorful generation effect, and the role of the motivation and surprise.

## Experiment 1

Experiment 1 utilised a novel errorful generation procedure in which participants were first presented with photographs of unfamiliar people and were asked to learn four facts (a hobby, a name, a job and a food) relating to each person per trial. On Generate trials, two facts were designated as either "Guess" or "Study" facts each. The participants were asked to guess the two Guess facts before both the Guess and Study facts were revealed. On the remaining Read trials, the participants simply studied all four facts without guessing any of them. Some of the participants also ranked their motivation to learn the correct facts before they were revealed (Motivation group). The remaining participants ranked their surprise regarding the four facts after they were revealed (Surprise group). Memory of the Guess, Study and Read targets was then assessed in a final recognition test, in which the faces were presented with the correct facts and novel foils. Both groups had to select the correct facts for each face.

Based on the earlier work by Potts and colleagues (2014, 2019) and Seabrooke et al. (2019), we expected the participants to recognise more Guess targets than Read targets in the final recognition test. The primary question of interest was with respect to the Study targets (presented on Generate trials, but for which no guess was made). By pairing each cue (face) with multiple targets (facts), the task allowed us to assess the specificity of the errorful generation effect. If errorful generation produces a generalised increase in attention, then all facts that are presented on Generate trials (i.e., both Guess and Study items) should be better recognised than facts that are presented on Read trials. If errorful generation has a more specific effect, on the other hand, then the participants should show better recognition of just the Guess facts, and recognition of Study facts might even be impaired relative to the Read facts. Moreover, if errorful generation improves processing of the targets via an increase in motivation to learn the correct answers, or surprise when the answers are revealed, then Guess targets should receive higher motivation and/or surprise rankings than Study targets.

We adopted this novel errorful generation paradigm for several reasons. First, the use of novel faces as cues ensures that the cues do not have pre-existing associations with the targets (the facts). A picture of an unfamiliar person reveals nothing about their favourite food or hobby, for instance. Second, faces are ideal stimuli to use when exploring the specificity of errorful generation effects, because they can be paired with many unrelated facts. Learning that a person works as an accountant, for example, provides no information about their favourite food. Verbal cues such as *pond*, by contrast, have many associates (e.g., *frog*, *lily*, *water*), but they often have shared semantic features. By using faces that could be easily and realistically paired with multiple non-overlapping facts, it allowed us to examine the specificity of the errorful generation effect, without concern that the participants' guesses would be related to both the Guess and Study targets.

Third, pairing each cue with multiple unrelated facts allowed us to examine whether errorful generation attempts have specific or general effects on motivation and surprise. Errorful generation might have a general effect, in that it could increase motivation and/or attention to the cue and all related information. For instance, having to guess on a trial may reduce the likelihood of mind-wandering, or attentional lapses during a Generate trial, relative to a Read trial. If this is the case, then all facts that are presented on Generate trials (i.e. both Guess and Study items) should be remembered better than the facts that are presented on Read trials, where no guessing occurs. Alternatively, errorful generation might increase motivation to learn the specific item that the participant guessed, without concurrently increasing motivation to learn the other facts. In this case, the Guess facts should be remembered better than both the Study and Read facts. Finally, the pairing of each cue (face) with multiple targets (facts) gives us more data per trial, and therefore provides a more powerful test of the effects of errorful generation on self-reported motivation and surprise.

## Method

**Design**. Encoding condition was manipulated within-subjects, with Guess and Study facts presented on Generate trials, and Read facts presented on Read trials. One group of participants ranked their motivation to learn each fact during the encoding phase. A separate group of participants ranked their surprise regarding each fact at encoding. The key measures during the encoding phase were the number of correctly guessed facts on Generate trials, and participants' mean motivation/surprise rankings for Guess, Study and Read facts. The primary dependent measure for the recognition test was the percentage of Guess, Study and Read facts that were correctly recognised.

**Participants**. Forty University of Plymouth students (27 females, aged between 18 and 30 years, mean *[M]* = 20.53 years, standard error of the mean *[SEM]* = 0.45 years) took part in the experiment on either a voluntary basis or in exchange for course credit. The participants were randomly allocated to either the Motivation (*N* = 19) or Surprise (*N* = 21) group at the start of the experiment. Both of the experiments reported here were approved by the University of Plymouth Psychology Ethics Committee.

**Apparatus and materials**. The experiment was programmed in E-Prime 2.0 and was presented on a 22-inch computer monitor. Stimuli were presented on a white background, and responses were made using a standard keyboard and mouse. The cues consisted of 27 pictures of human faces (12 males, 15 females) from the Florida Department of Corrections (2002) database (http://www.dc.state.fl.us). For each participant, a randomly selected 24 photographs were presented during the main task, and the remaining three were presented on practice trials. The targets/foils were 54 facts from each of the four categories (occupations, hobbies, foods and names). In each category, 24 facts served as targets, another 24 facts served as foils, and the remaining six facts served as targets or foils for the practice trials.

**Procedure**. Both groups of participants were first told that they would see faces of different people, and that they would learn about their jobs, their favourite hobbies, their favourite foods, and their best friend's names. We used "best friend's name" as a category rather than the name of the person shown in the photograph so that the names could be randomly allocated to the photographs for each participant, without concern for gender. The participants were told that for some of the photographs, they would need to provide one-word guesses for two of the facts before they were revealed. They were also informed that they should try to remember all of the facts, because they would be tested later on. Finally, they were told that they would be asked to indicate how surprised they were by each fact (Surprise group), or how motivated they were to learn each fact (Motivation group).

***Encoding phase.*** Both groups first completed three practice encoding trials, which consisted of two Generate trials and one Read trial. They then completed 24 proper encoding trials, which consisted of 16 Generate trials and eight Read trials. We chose to present half as many Read trials as Generate trials so that there were the same number of facts in each condition (each Generate trial produces two Guess facts and two Study facts, while each Read trial produces four Read facts, for a total of 32 facts for each of the Guess, Study and Read conditions). The trials were randomly intermixed and were separated by 1500ms intervals. The category facts were randomly assigned to the faces for each participant, and the faces were randomly allocated to the Generate and Read trials. The 54 facts for each category were randomly allocated to serve as Guess, Study or Read items during the main or practice encoding phases, or as novel foils in the final recognition test. Figure 1 depicts an example Generate trial at encoding.

The Generate trials began with the presentation of a fixation cross for two seconds. A picture of a person's face was then presented on the top center of the screen, above the four different categories, which were overtly labelled. The location of the four categories was randomly determined on each trial. We randomised the location of the categories to discourage the

participants from attending to some categories (e.g., the top two categories) more than others. On Generate trials, the four categories were also randomly allocated to the two Guess and two Study items. The category options were presented in bold text, and the first Guess category (randomly chosen from the two Guess items) was presented in red. All other text was presented in black. The participants had to guess the corresponding fact by typing it, using the keyboard. Their guesses appeared beneath the category as they typed, and they were able to use the Backspace key to change their answer up until they pressed the Enter key. Responses were not time limited. After a 200ms inter-stimulus interval (ISI) in which the stimuli briefly disappeared, the second Guess category was presented in red. All other text was presented in black, including the participant's guess from the first Guess item. The participants had to guess the corresponding fact in the same way as for the first Guess item. After another 200ms ISI, the Motivation group were then asked to rank their motivation to learn the four facts. At this point, the participants' guesses disappeared and all of the category labels were presented in black. The question, "Which fact are you most motivated to learn?" was presented centrally, in between the picture of the face and the four category labels. The participants had to select a category label using the mouse, which then turned navy blue and was not available for further selection. The question changed to "Which fact are you [second/third] most motivated to learn?" for the second and third rankings, with the chosen categories changing to navy blue and becoming unavailable for further selection. The category that the participants were least motivated to learn was inferred. The participants had unlimited time to rank their motivation because (in contrast to the Surprise group), the motivation rankings were recorded before the correct facts were presented. After the ranking procedure was complete, the picture of the face was presented on the top-center of the screen, and the category labels were presented beneath, with the correct facts shown underneath each. The statement, "Please study the facts now" was presented centrally, between the face and the four facts. The category labels were all presented in navy blue; all other text was presented in black. The participants had eight seconds to study the facts. The Read trials were the same as the Generate trials, except that the participants did not

complete the generation component (the trials started with the motivation rankings). Hence, the facts were presented for eight seconds on both Generate and Read trials.

The guessing phase of the Generate trials for the Surprise group was identical to that of the Motivation group. In brief, participants were presented with a face and the four category labels, and they had to guess two of the category facts in turn. After the second ISI (see Figure 1), all four correct answers were presented beneath the corresponding category options (the participants' guesses were not presented hereafter). All text was now presented in black, and the participants had 12 seconds in total to rank their surprise about each of the four facts. We gave the Surprise group four seconds more than the Motivation group to view the facts, because the Surprise group were required to rank their surprise while the facts were presented, whereas the Motivation group were not. The question, "Which fact are you most surprised by?" was presented centrally, in between the picture of the face and the four category options. The participants had to click on one of the four category labels using the mouse. The chosen category was then presented in navy blue and was not available to be ranked thereafter. The question changed to "Which fact are you [second/third] most surprised by?", and the participants had to click on one of the three remaining category labels (which also turned navy blue and was no longer available for further rankings). The category that the participants were least surprised by was inferred. All of the category options were then presented in navy blue, with the corresponding facts presented beneath each category, for the remainder of the 12 seconds. If the participants failed to rank all of the categories within the 12 seconds, a "too slow" warning message and a reminder to rate their surprise by clicking on the categories was presented instead. No extra time was allowed if the participants failed to rank all of the items within the 12 seconds.

The Read trials for the Surprise group followed the same format as the Generate items, except that the participants did not guess any of the facts (and hence the trial began with the

surprise rankings). Thus, the four facts were presented for 12 seconds on both Generate and Read trials.

***Final recognition test.*** The encoding phase was followed by a two-choice recognition test, which was identical for both groups. Four practice trials were administered to begin with. These trials followed the format of the main recognition test (described below), but used the faces and targets (one from each category) from the practice encoding trials (as well as randomly selected foils from the same categories as the targets). In the main recognition test, each trial began with the presentation of one of the faces from the encoding phase, along with a correct fact from one of the four categories and a randomly selected novel foil from the same category. The face was presented in the top-center of the screen, the target and the foil were presented side-by-side on the bottom half of the screen, and the question, "What is this person's [category]?" was presented centrally. The text in brackets was replaced by the appropriate category ("occupation", "favourite hobby", "favourite food", or "best friend's name"). The location of the target and the foil was counterbalanced across trials. The participants had to select the target using the mouse (responding was not time limited). There were 96 trials, with 24 trials from each category. The faces from the encoding phase were presented four times each, once for each category. The trials were randomly ordered for each participant, and the facts associated with each face were randomly interleaved with other face-fact pairs. The trials were separated by one-second intervals.

**Results**

Upon publication of the manuscript, the trial-level raw data will be publicly archived at https://osf.io/tc976/.

**Encoding phase.** During the encoding phase, the participants were given unlimited time to guess the facts on Generate trials. On average, the participants spent 6656ms (*SEM* = 345ms) generating each guess in the Motivation group, and 6229ms (*SEM* = 346ms) in the Surprise group. A Welch two-sample *t*-test showed that the time taken to guess in each group did not significantly

differ, $t$ (37.9) = 0.87, $p$ = 0.39, $d$ = 0.28, although the Bayesian evidence for the null was inconclusive, $BF_{10}$ = 0.42. All Bayes Factors ($BF_{10}$) were calculated using version 0.9.12.4.2 of the *BayesFactor* package (Morey, Rouder, & Jamil, 2015) in R (R Core Team, 2018).

Across participants, 11 facts were correctly guessed on Generate trials. These consisted of six facts from the Motivation group, and five facts from the Surprise group. When analysing the ranking data, we removed the trials involving correct generations from the Surprise group, since the correct generations would be likely to influence the participants' ranking of the four facts. Trials involving correct guesses were not removed from the Motivation ranking dataset, however, because the rankings were recorded before the correct answers were presented. We also removed any trials in which the participants failed to rank all four categories in the Surprise group[1] (14 trials total). Figure 2a shows the mean rankings given Guess and Study facts on Generate trials for the Motivation and Surprise group[2]. Rankings of four reflect the highest degree of motivation/surprise; rankings of one reflect the lowest degree of motivation/surprise. Higher motivation rankings were given to the Guess facts than the Study facts, $t$ (18) = 6.26, $p$ < .001, $d_z$ = 1.44. Surprise rankings, by contrast, did not significantly differ for Guess and Study facts, $t$ (20) = 0.78, $p$ = .44, $d_z$ = 0.17, $BF_{10}$ = 0.30.

**Final recognition performance.** Figure 2b shows the mean percentage of correct responses per group in the final recognition test. All facts that were presented on trials in which a participant generated a correct guess at encoding were removed from the test dataset. Facts that were presented on trials in which the participants failed to rank all four categories in the Surprise group were also removed. An encoding condition × group mixed ANOVA[3] revealed a significant main effect

[1] The Motivation group had unlimited time to rank their motivation of the four facts, so there were no trials in which the Motivation group failed to rank all four facts.
[2] Although our primary interest was with respect to the motivation and surprise rankings that were given to Guess and Study facts, we also report the mean rankings given to each category (name, job, food, and hobby) on Read trials in the Supplementary Materials for completeness.
[3] Here and in all subsequent relevant cases, Greenhouse-Geisser corrected $p$-values are reported to correct for violations of sphericity.

of encoding condition, $F$ (2, 76) = 6.62, $p$ = .004, generalised eta squared ($\eta_g^2$) = 0.04, but not of group, $F$ (1, 38) = 1.61, $p$ = .21, $\eta_g^2$ = 0.03. The encoding condition × group interaction was not significant, $F$ (2, 76) = 0.80, $p$ = .44, $\eta_g^2$ = .005. The Guess targets were better recognised than both the Study targets, $t$ (39) = 3.36, $p$ = .002, $d_z$ = 0.53, and the Read targets, $t$ (39) = 3.45, $p$ = .001, $d_z$ = .55. No difference was observed for targets that were allocated to the Study and Read conditions, $t$ (39) = 0.65, $p$ = .52, $d_z$ = 0.10, with Bayesian evidence for the null, $BF_{10}$ = 0.21.

We also computed the Goodman-Kruskal gamma correlation for each participant to examine the relationship between motivation/surprise rankings and final recognition performance. We did not include the Generate trials in this analysis, since these trials were affected by the experimental manipulation (Guess versus Study facts). Particularly with the motivation rankings, the Guess facts received higher rankings than the Study facts. This limits the range of scores that the participants will have used within both the Guess and Study condition, which in turn reduces the capacity to detect a correlation between motivation rankings and final recognition performance for Generate trials. We therefore looked only at the facts that were presented on Read trials, which did not incorporate an experimental manipulation. Gamma correlations could not be calculated for two participants (one from each group), because they correctly recognised all of the Read targets. For the remaining participants, one-sampled $t$-tests showed that the mean gamma correlations were significantly greater than zero for both the Motivation ($M$ = 0.32, $SEM$ = 0.08), $t$ (17) = 3.96, $p$ = .001 $d$ = 0.93, and Surprise ($M$ = 0.26, $SEM$ = 0.11), $t$ (19) = 2.44, $p$ = .02 $d$ = 0.54, group.

**Discussion**

The results of Experiment 1 complement the previous literature, in that we demonstrated that errorful generation attempts improve subsequent recognition memory, compared with pure study trials that do not involve errorful generation. Participants were asked to guess two facts that were presented alongside a photograph of an unfamiliar person (Generate trials). These facts were subsequently more likely to be recognised than facts that were presented on trials in which no

guesses were made (Read trials). Thus, we demonstrated that the previous errorful generation effects that have been observed with novel word pairs (Potts et al., 2019; Potts & Shanks, 2014; Seabrooke et al., 2019) generalise to unfamiliar face-fact materials. Errorful generation attempts also produced a specific improvement to the guessed facts, rather than conferring a more general benefit to all facts that were presented on those trials (memory was worse for studied facts than guessed facts presented on Generate trials). Furthermore, the participants ranked their motivation to learn the guessed facts more highly than the studied facts that were presented alongside those guessed facts. Surprise rankings, by contrast, did not differ between the guessed and studied facts. Finally, in the absence of any errorful generation manipulation on Read trials, both motivation and surprise rankings were positively related to final recognition memory. This suggests that both motivation and surprise are related to recognition memory, but that errorful generation attempts only affect motivation. It is possible, then, that guesses serve to increase motivation to learn the correct answers, which then improves processing of those answers.

## Experiment 2

Experiment 2 primarily aimed to provide an extension of the key effects that were observed in Experiment 1. In Experiment 1, motivation to learn the facts was assessed using rankings. This procedure provided a useful initial test of the role of motivation, because the participants had to give different rankings to each fact. This encouraged them to think carefully about each fact, rather than giving the same answer to all facts on each trial. Rankings do not, however, allow the participants to express their overall levels of motivation. For instance, the participants had no way of expressing that they felt similar levels of motivation to learn all facts on any given trial. They also had no way to say if they had different overall levels of motivation to learn the facts on Generate and Read trials. In Experiment 2, we therefore asked the participants to freely *rate*, rather than rank, their motivation to learn the facts during the encoding phase. We did not measure surprise in Experiment 2 since the data from Experiment 1 provided evidence for the null hypothesis in this respect.

The participants also performed very well on the two-alternative forced-choice recognition test in Experiment 1. Indeed, it could be argued that performance approached ceiling levels of performance, particularly in the Guess condition. It should be noted that, if performance was at ceiling in the Guess condition, it did not prevent us from detecting an effect of errorful generation. It is still possible, however, that a ceiling effect led us to underestimate the size of our recognition effects in Experiment 1. In Experiment 2, we therefore took two approaches to reduce overall recognition performance. We first inserted a delay between the encoding and test phases. Thus, the participants completed an unrelated distractor task for approximately 20 minutes after the encoding phase and before the test phase. The final test also took the form of a four-choice multiple choice test (for similar procedures, see Potts & Shanks, 2014 and Seabrooke et al., 2019, Experiment 5). We expected this test to be more difficult than the two-choice test used in Experiment 1 and should, therefore, reduce overall performance levels.

Finally, the motivation condition in Experiment 1 was underpowered to detect medium-sized effects (power = 67% at $d_z = 0.5$). We increased the sample size in Experiment 2 to address this.

**Method**

**Design.** A within-subjects design with encoding condition (Guess vs. Study vs. Read facts) as a single variable was employed. During the encoding phase, the primary measures were the accuracy of the participants' guesses on Generate trials, and motivation ratings for the Guess, Study and Read facts. During the recognition test, the primary dependent measure was the percentage of Guess, Study and Read facts that were recognised correctly.

**Participants.** Thirty-two University of Plymouth students (24 females, aged between 18 and 31 years, *M* = 20.31 years, *SEM* = 0.40 years) completed the experiment for course credit. This sample size was chosen because it provided good power to detect recognition (Guess > Study and Guess > Read) and motivation (Guess > Study) effects at the sizes seen in Experiment 1 (90% at $d_z$ =

0.53, 92% at 0.55 and over 99.5% at 1.44, respectively). Power calculations were performed using the *pwr* package (Champely, 2017) in R (R Core Team, 2018).

**Apparatus and materials.** We used a new set of photographs in Experiment 2 to allow greater generalisation of our findings. Thus, twenty-seven face stimuli (14 females, 13 males) were selected from the "smiling_front" archive from DeBruine and Jones' (2017) database of adult face stimuli. These stimuli were reduced in size to 405x405 pixels. These stimuli also have the benefit of being from an open-source database, which should make them readily accessible for future experiments. The targets/foils were 108 facts from each of the four categories (occupations, hobbies, foods and names). In each category, 24 facts served as targets, another 72 facts served as foils, and the remaining 12 facts served as targets or foils for the practice trials. The facts were randomly allocated to serve as targets and foils for each participant. All other aspects of the apparatus and materials were as in Experiment 1.

**Procedure.** Participants completed the same encoding phase as the Motivation Group from Experiment 1 (see Figure 1), except that the participants were asked to rate (rather than rank) their motivation of the four facts. Thus, before the correct facts had been revealed, the four categories were highlighted (in blue) in a random order and the participants were asked to rate their motivation to learn the highlighted fact by choosing a number between one ("Not at all motivated") and five ("Very motivated"). Once motivation ratings had been collected for all four facts, the four facts were presented together with the face stimulus for 12 seconds. All other aspects of the encoding phase were the same as the Motivation group in Experiment 1. After the encoding phase, the participants completed a filler task that involved rating pairs of green dot-filled rectangles for similarity. On average, this unrelated task took 24.70 minutes (*SEM* = 0.22 minutes) to complete.

The distractor task was followed by a multiple-choice test that, as in Experiment 1, assessed target recognition memory. Four practice trials were administered to begin with. These trials followed the format of the main multiple-choice test (described below), but used the faces and

targets (one from each category) from the practice encoding trials (as well as randomly selected foils from the same categories as the targets). On each test trial, the participants were presented with one of the face stimuli from the encoding phase, plus four possible facts from the same category (e.g., "Hobby"). One of those facts was the correct fact for that face stimulus, and the rest were novel foils that were not presented during the encoding phase. The foils were presented only once each on test. The location of the target and the foils were randomly determined on each trial. All other aspects of the test were as in Experiment 1.

**Results**

Upon publication of the manuscript, the trial-level raw data will be publicly archived at https://osf.io/5ne72/.

**Encoding phase.** During the encoding phase, the participants spent an average of 8141ms (*SEM* = 370ms) guessing each fact on Generate trials. In total, four facts were correctly guessed on Generate trials. As in Experiment 1, these trials were not removed for the analysis of motivation ratings, because the correct answers were only revealed after the motivation ratings had been taken. Figure 3a shows the mean motivation rankings given to Guess, Study and Read facts. A one-way ANOVA revealed a main effect of encoding condition, $F$ (2, 62) = 15.26, $p < .001$, $\eta_g^2 = .09$. Pairwise comparisons showed that the participants gave higher motivation rankings to Guess facts than both Study facts, $t$ (31) = 4.29, $p < .001$, $d_z = 0.76$, and Read facts, $t$ (31) = 3.97, $p < .001$, $d_z = 0.70$. Motivation ratings for the Study and Read facts did not differ, $t$ (31) = 0.37, $p = .71$, $d_z = 0.07$, $BF_{10} = 0.20$.

**Final multiple-choice test performance.** Figure 3b shows the mean percent correct for each encoding condition in the final recognition test. As in Experiment 1, all facts that were presented on trials in which a participant generated a correct guess at encoding were removed from the test dataset. A one-way ANOVA revealed a main effect of encoding condition, $F$ (2, 62) = 6.86, $p = .002$, $\eta_g^2 = 0.05$. Pairwise comparisons revealed that the Guess facts were better recognised than both the

Study facts, $t$ (31) = 2.70, $p$ = .01, $d_z$ = 0.48, and the Read facts, $t$ (31) = 3.37, $p$ = .002, $d_z$ = 0.60. Recognition of the Study and Read facts did not differ, $t$ (31) = 0.32, $p$ = .75, $d_z$ = 0.06, $BF_{10}$ = 0.20. These results are consistent with those of Experiment 1.

We also computed the Goodman-Kruskal gamma correlation for each participant to examine the relationship between motivation ratings and final recognition performance. As in Experiment 1, the gamma correlations were restricted to the Read trials. Gamma correlations could not be calculated for two participants, because they gave the same motivation rating for all Read facts. In contrast to Experiment 1, the gamma correlations ($M$ = 0.01, $SEM$ = 0.09) for the remaining participants did not significantly differ from zero, $t$ (29) = 0.11, $p$ = .91, $d$ = 0.02, $BF_{10}$ = 0.20. We suspected that this null result might have arisen because the participants used a small range of ratings during the encoding phase, which would have limited the capacity to detect a significant correlation between motivation ratings and recognition memory. This contrasts with Experiment 1, where the participants had to use the full range of scores because they gave rankings rather than ratings. To test this idea, we directly compared the standard deviation of rankings/ratings given on Read trials in each experiment. In Experiment 1, the average standard deviation of rankings (across motivation and surprise groups) was 1.14. In Experiment 2, the comparable rating score was 0.81, which suggests that the participants did indeed use a smaller range of motivation scores in Experiment 2 than in Experiment 1. A Welch two sample $t$-test confirmed that this difference was significant, $t$ (31) = 5.15, $p$ < .001, $d$ = 1.37. We note that the participants used a significantly smaller range of scores in Experiment 2, even though they were able to use a *larger* range of scores than in Experiment 1 (ratings in Experiment 2 were on a five-point scale, whereas ranking scores in Experiment 1 varied from one to four).

**Discussion**

Experiment 2 extended the pattern of results observed in Experiment 1. When participants freely rated their motivation to learn each fact during the encoding phase, they gave higher ratings

to the facts that they had previously guessed (Guess facts) than facts that they had not guessed (Study and Read facts). Consistent with Experiment 1, the participants recognised more Guess facts than Study and Read facts on the multiple-choice test. This latter result shows that the errorful generation benefit seen in Experiment 1 is robust, and can survive a 20-minute retention interval and a change in test format.

## General Discussion

Two experiments examined the effect of errorful generation on recognition memory in a novel paradigm that paired novel cues with multiple targets. There were several noteworthy findings. First, both experiments revealed a benefit of errorful generation on recognition memory. That is, the participants recognised more facts that they had incorrectly guessed (Guess facts) at encoding than facts that were presented on Read trials, in which those facts were simply presented for study without any guessing period. This finding is consistent with previous work with novel word pairs (Potts et al., 2019; Potts & Shanks, 2014; Seabrooke et al., 2019), and it reiterates that errorful generation can improve target recognition (relative to an equivalent period of time spent studying) for cues and targets do not have pre-existing semantic associations.

The two experiments also showed that errorful generation attempts produce a specific benefit to the guessed item. That is, errorful generation attempts did not improve subsequent recognition of Study facts that were presented alongside the Guess facts (relative to Read facts). The experiments therefore provide two well-controlled laboratory demonstrations of an errorful generation effect that cannot be explained by changes in overall attention at the trial level. That is, our errorful generation effect cannot result from guessing increasing attention to all information on that particular trial, because every Generate trial included both Guess and Study targets, and yet the Guess targets were better recognised than the Study targets. This can be contrasted with previous errorful generation experiments, where the effect was assessed for single targets that had been either guessed or not guessed on any given trial (Potts et al., 2019; Potts & Shanks, 2014; Seabrooke

et al., 2019). The effect also cannot be explained by appealing to differences in attention to the *cue*, because the same face stimuli served as cues for both Guess and Study targets. Thus, the results provide unique evidence to suggest that errorful generation attempts produce a selective enhancement of processing of the guessed item, rather than improving recognition for all subsequent feedback more generally. This pattern was apparent in both experiments, suggesting that our effects are robust and can outlive at least a 20-minute retention interval.

The current experiments also allow us to say something about the possible effects of errorful generation with respect to Guess facts on *other* material (Study facts) also presented on Generate trials. We might have expected that, if errorful generation improves subsequent processing of the corrective feedback, it might hinder processing of other information that is presented alongside such feedback (the Study targets) – perhaps via a process of reallocation of attention. This might be especially likely given recent evidence to suggest that errorful generation increases participants' curiosity (Potts et al., 2019) and motivation (in the present experiments) to learn the true answers. However, we found no evidence to support this claim. Although Read facts were less well recognised than Guess facts, the participants showed similar recognition performance for Study and Read targets.

In both experiments, we also found that errorful generation attempts affected participants' self-reported motivation to learn the correct answers. When the participants were asked to report their motivation to learn each fact during the encoding phase, the guessed facts were ranked (Experiment 1) and rated (Experiment 2) more highly than the non-guessed facts. This result is akin to Potts et al.'s (2018) recent demonstration that guessing the definition of unfamiliar words can increase participants' curiosity to learn those definitions (see also Gruber, Gelman, & Ranganath, 2014; Kang et al., 2009). More broadly, our findings also accord with studies demonstrating that participants choose to restudy information that they previously guessed more than information that they have simply studied (Yang et al., 2017). Together, the data are consistent with the hypothesis

that generating guesses increases motivation to learn the correct answers to the questions posed, which then facilitates encoding of those answers when they are revealed.

Unlike the motivation data, we saw no evidence in Experiment 1 to suggest that errorful generation increases participants' self-reported surprise when corrective feedback is presented. It is possible that more sensitive assays of surprise (e.g., eye-tracking; see Brod et al., 2018) would reveal evidence of an effect of errorful generation on surprise. It is also possible that our face-fact materials limited our capacity to detect differences in surprise. When participants were asked to guess a fact (e.g., favourite food) for a novel face photograph on Generate trials, they had no real basis on which to make a guess (because the face-fact pairs were novel and arbitrarily paired). This is the nature of the *errorful generation* paradigm, in which participants are asked to generate guesses (or study) when there is virtually no chance that those guesses will be correct. Hence, it is possible that our use of an errorful generation paradigm, in which participants study novel associations, limited our capacity to detect differences in surprise. The participants might not have had any great expectation that their guesses would be correct on Generate trials, and this could have reduced the potential to detect differences in surprise for guessed and non-guessed facts. It is possible that we would have obtained greater differences in surprise with an unsuccessful retrieval paradigm, where participants generate guesses for familiar materials that they have prior knowledge of (although see Zawadzka & Hanczakowski, 2018). While this would be an interesting avenue for future research, it would not, of course, shed light on the mechanisms that underlie errorful generation effects (which was the purpose of the present experiments). We also note that the absence of a difference in surprise ratings for Guess and Study facts was seen in the context of a robust errorful generation effect. Hence, errorful generation led to a clear improvement in subsequent target recognition, even though participants' surprise ratings for Guess and Study targets did not differ.

We are not the first to use face-fact pairs when exploring the effect of generating errors on learning and memory. Kessels and De Haan (2003) presented younger and older adults with

unfamiliar faces and either presented a name for the participants to study, or asked the participants to guess the name before it was revealed. Both younger and older adults showed *worse* recall of the names that they had previously guessed than those that they had simply studied. McGillivray and Castel (2010), on the other hand, found that both older and younger adults showed *better* recall of face-age pairs that they had guessed than face-age pairs they had studied. Notably, this benefit of generating guesses was only seen when the face cues were consistent with the target ages. McGillivray and Castel therefore suggested that generating guesses might only be beneficial when the cues are informative; a face provides useful information about a person's age, but does not give any information about their name. The current results suggest that the picture may be more complex. In our experiments, the cues (unfamiliar faces) did not provide any information about the targets (facts), and yet errorful generation improved subsequent target recognition. The question then is, why did we see a benefit of errorful generation with non-informative cues, while Kessels and De Haan (2003) and McGillivray and Castel (2010) did not? One notable difference concerns the final test procedures. Both Kessels and De Haan (2003) and McGillivray and Castel (2010) administered cued recall tests, whereas we assessed target recognition performance. This distinction has been shown to be crucial in a related set of studies that used simple word pairs. In particular, generating errors aids cued recall for related word pairs such as *pond-frog* (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012), but only boosts subsequent target recognition (not cued recall) for unrelated word pairs (Potts et al., 2019; Potts & Shanks, 2014; Seabrooke et al., 2019). Taken together, both the studies that used face-fact pairs and those that used simple word pairs are consistent with the idea that generating errors only aids subsequent cued recall when the cues provide information that is useful for informing the participants' guesses, or when the participants' guesses are likely to provide information that is useful for retrieving the correct answer.

To conclude, the current experiments extended previous work by demonstrating a clear benefit of errorful generation in a novel paradigm that involved pairing cues (faces) with multiple targets (facts). Errorful generation had a selective effect, in that its benefit on subsequent

recognition memory did not extend to other targets that were presented alongside guessed targets. Furthermore, participants rated their motivation to learn facts more highly when they had previously guessed those facts, and those facts rated most highly for motivation were also those that produced the best recognition memory. Surprise ratings, by contrast, did not significantly differ for guessed and studied facts. The findings therefore add to a growing body of literature suggesting that errorful generation improves subsequent recognition, possibly by increasing motivation to attend to and process corrective feedback.

## References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*, 659–701. https://doi.org/10.3102/0034654316689306

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, *32*, 53–68. https://doi.org/10.1016/0028-3932(94)90068-X

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*, 158–175. https://doi.org/10.3758/s13428-011-0123-7

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction*, *55*, 22–31. https://doi.org/10.1016/j.learninstruc.2018.01.013

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

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*, 69–84. https://doi.org/10.1007/s11409-006-6894-z

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*, 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Rahman, S., & Perkins, K. (2018). The effects of prequestions on classroom learning. *Journal of Experimental Psychology: Applied*, *24*, 34–42. https://doi.org/10.1037/xap0000145

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition*, *6*, 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

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

Champely, S. (2017). pwr: Basic functions for power analysis. *R Package Version 1.2-1.* Retrieved from https://cran.r-project.org/package=pwr

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*, 841–850. https://doi.org/10.1037/xlm0000073

Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: How does testing impair new learning? *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*, 1741–1754. https://doi.org/10.1037/xlm0000126

DeBruine, L., & Jones, B. (2017). Face Research Lab London Set (Version 3). figshare. https://doi.org/10.6084/m9.figshare.5047666.v3

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

Finn, B., & Roediger, H. L. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*, 1665–1681. https://doi.org/10.1037/a0032377

Forlano, G., & Hoffman, M. N. H. (1937). Guessing and telling methods in learning words of a foreign language. *Journal of Educational Psychology*, *28*, 632–636. https://doi.org/10.1037/h0056518

Geller, J., Carpenter, S. K., Lamm, M. H., Rahman, S., Armstrong, P. I., & Coffman, C. R. (2017). Prequestions do not enhance the benefits of retrieval in a STEM classroom. *Cognitive Research: Principles and Implications*, *2*, 42. https://doi.org/10.1186/s41235-017-0078-z

Griffiths, L., & Higham, P. A. (2017). Beyond hypercorrection: remembering corrective feedback for

low-confidence errors. *Memory*, *0*, 1–18. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.neuron.2014.08.060

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. https://doi.org/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. https://doi.org/10.3758/s13421-011-0167-z

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology*, *70*, 626–635. https://doi.org/dx.doi.org/10.1037/0022-0663.70.4.626

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., Mcclure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, *20*, 963–973. https://doi.org/10.1111/j.1467-9280.2009.02402.x

Kang, S. H. K., Mcdermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*, 471–479. https://doi.org/10.1080/09658210802647009

Kessels, R. P. C., & De Haan, E. H. F. (2003). Mnemonic strategies in older people: A comparison of errorless and errorful learning. *Age and Ageing*, *32*, 529–533. https://doi.org/10.1093/ageing/afg068

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. https://doi.org/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. https://doi.org/10.1037/a0033699

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219–224. https://doi.org/10.3758/BF03194055

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

Kornell, N., Jacobs Klein, P., & Rawson, K. A. (2015). Retrieval attempts enhance learning , but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 283–294. https://doi.org/10.1037/a0037850

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation*, *65*, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513. https://doi.org/10.1080/09541440701326154

McGillivray, S., & Castel, A. D. (2010). Memory for age-face associations in younger and older adults:

The role of generation and schematic support. *Psychology and Aging*, *25*, 822–832. https://doi.org/10.1037/a0021044

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). Package 'BayesFactor.' Retrieved from http://bayesfactorpcl.r-forge.r-project.org/

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*, 1023–1041. https://doi.org/10.1037/xlm0000637

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667. https://doi.org/10.1017/CBO9781107415324.004

Pressley, M., Tanenbaum, R., McDaniel, M. A., & Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology*, *15*, 27–35. https://doi.org/10.1016/0361-476X(90)90003-J

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447. https://doi.org/10.1016/j.jml.2009.01.004

R Core Team. (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from https://www.r-project.org/

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, *21*, 64–99. https://doi.org/10.1101/gr.110528.110

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. https://doi.org/10.1037/a0016496

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. https://doi.org/10.1111/j.1467-8721.2008.00612.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463. https://doi.org/10.1037/a0037559

Seabrooke, T., Hollins, T. J., Kent, C., Wills, A. J., & Mitchell, C. J. (2019). Learning from failure: Errorful generation improves memory for items, not associations. *Journal of Memory and Language*, *104*, 70–82. https://doi.org/10.1016/j.jml.2018.10.001

Skinner, B. F. (1958). Teaching machines. *Science*, *128*, 969–977. https://doi.org/10.1109/TE.1959.4322064

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

Squires, E. J., Hunkin, N. M., & Parkin, A. J. (1997). Errorless learning of novel associations in amnesia. *Neuropsychologia*, *35*, 1103–1111. https://doi.org/10.1016/S0028-3932(97)00039-0

Tanaka, S., Miyatani, M., & Iwaki, N. (2019). Response format, not semantic activation, influences the failed retrieval effect. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00599

Terrace, H. S. (1963). Discrimination learning with and without "errors." *Journal of the Experimental Analysis of Behavior*, *6*, 1–27. https://doi.org/10.1901/jeab.1963.6-1

Toftness, A. R., Carpenter, S. K., Lauber, S., & Mickes, L. (2018). The limited effects of prequestions on learning from authentic lecture videos. *Journal of Applied Research in Memory and Cognition*, *7*, 370–378. https://doi.org/10.1016/j.jarmac.2018.06.003

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for

enhancing memory? *Psychonomic Bulletin & Review*, *19*, 899–905. https://doi.org/10.3758/s13423-012-0276-0

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. https://doi.org/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. https://doi.org/10.1037/xlm0000363

Zawadzka, K., & Hanczakowski, M. (2018). Two routes to memory benefits of guessing. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000676

**Figure 1**

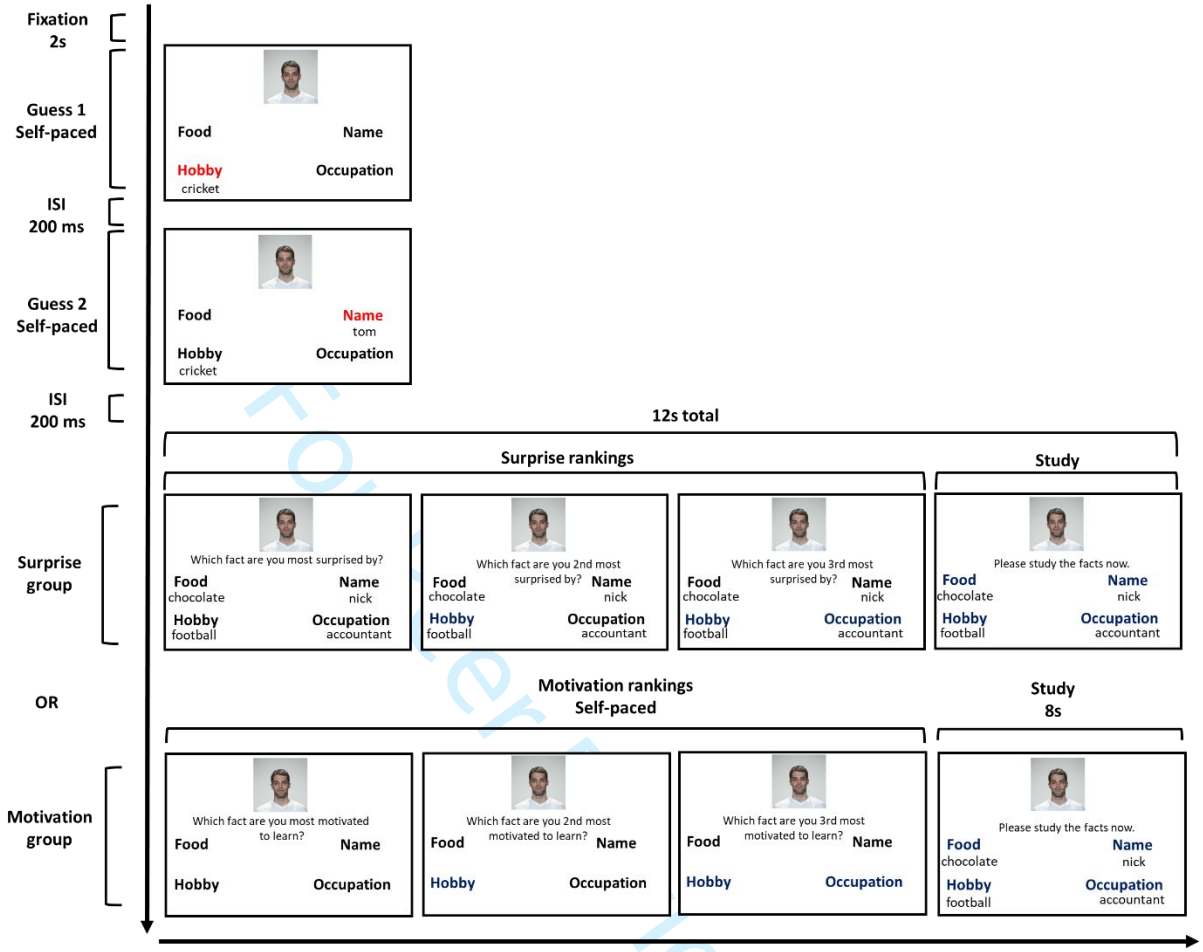


**Figure 1.** Schematic of an example Generate trial during the encoding phase for the Motivation and Surprise groups in Experiment 1. Participants were presented with a face and four category options, and were asked to first guess two facts. For the Surprise group, the four facts were then revealed and the participants had 12 seconds to rank their surprise and study all of the facts. For the Motivation group, the participants ranked their motivation to learn the four facts before they were presented for study for eight seconds. During the fixation period, a fixation cross (+) was presented for two seconds. During the inter-stimulus interval (ISI) periods, the stimuli disappeared for 200ms before reappearing. The Read trials were the same as the Generate trials, except that the participants did not guess any of the facts. Note that the photograph shown is from the open source database used in Experiment 2.

**Figure 2**

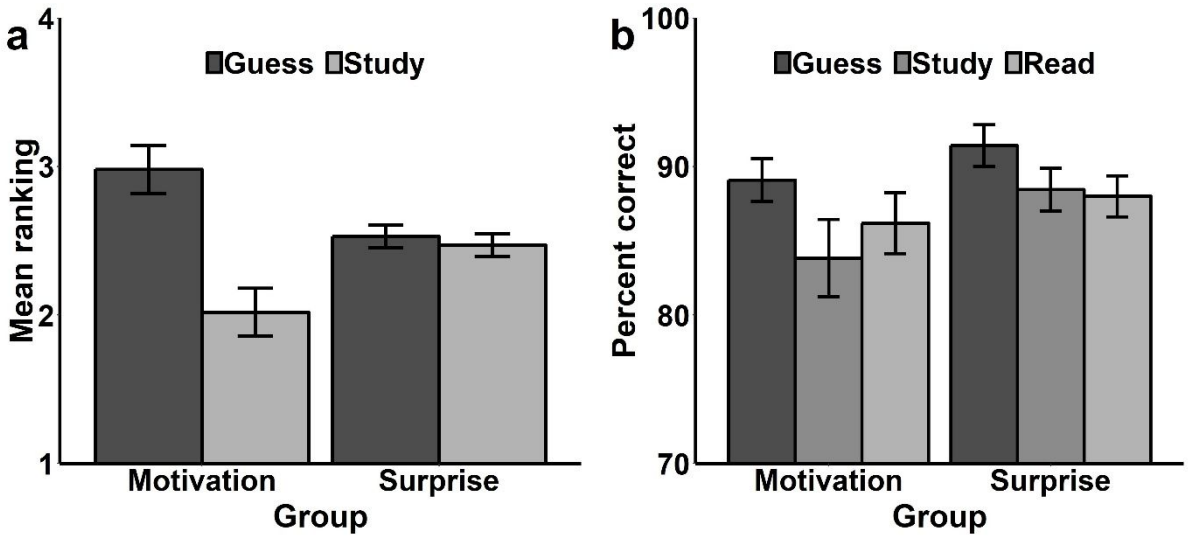


**Figure 2.** Results of Experiment 1. **(a)** Mean motivation and surprise rankings given to Guess and Study targets on Generate trials during the encoding phase. Rankings of four represent the highest level of motivation/surprise; rankings of one represent the lowest level of motivation/surprise. **(b)** Mean percent correct per encoding condition on the final recognition test. Error bars represent difference-adjusted within-subject 95% confidence intervals (Baguley, 2012).

**Figure 3**

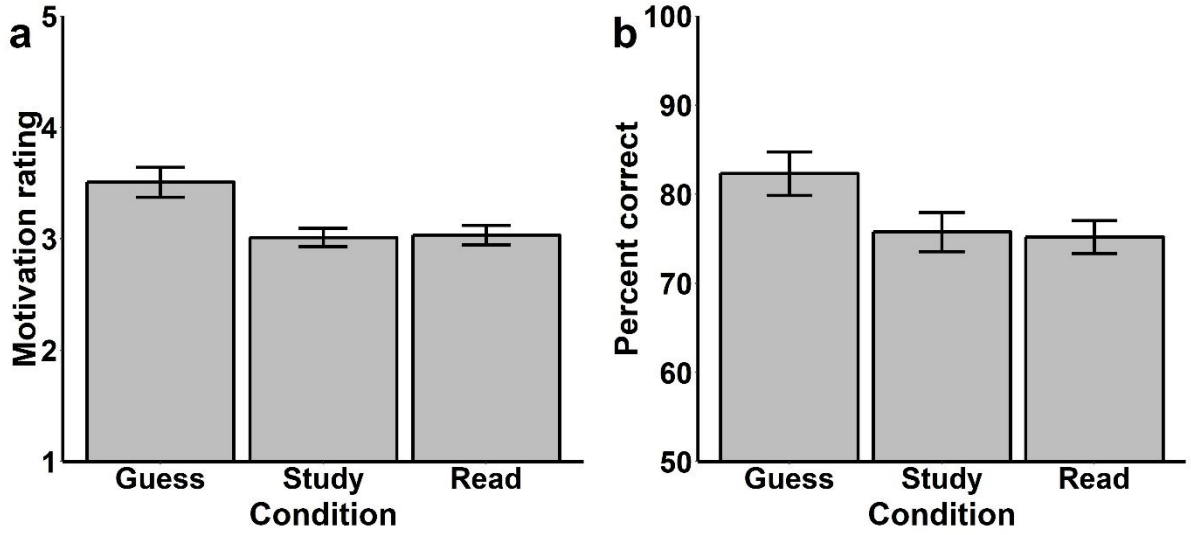


**Figure 3.** Results of Experiment 2. **(a)** Mean motivation ratings given for Guess, Study and Read facts during the encoding phase. Ratings of one and five represent "Not at all motivated" and "Very motivated", respectively. **(b)** Mean percent correct on the final multiple-choice test. Error bars represent difference-adjusted within-subject 95% confidence intervals (Baguley, 2012).

## Supplementary materials

**Mean rankings (Experiment 1) and ratings (Experiment 2) given to each category on Read trials.**

| | Experiment 1 | | Experiment 2 |
|---|---|---|---|
| | **Motivation rankings** | **Surprise rankings** | **Motivation ratings** |
| **Name** | 1.47 [1.25, 1.68] | 1.57 [1.42, 1.71] | 2.70 [2.56, 2.85] |
| **Food** | 2.21 [2.05, 2.37] | 2.47 [2.35, 2.59] | 2.98 [2.87, 3.08] |
| **Job** | 3.49 [3.24, 3.74] | 3.04 [2.88, 3.19] | 3.26 [3.08, 3.43] |
| **Hobby** | 2.83 [2.67, 2.99] | 2.92 [2.80, 3.05] | 3.20 [3.07, 3.32] |

*Note*. Numbers in parentheses denote difference-adjusted within-subject 95% confidence intervals (Baguley, 2012).

**Time taken to provide motivation rankings (Experiment 1) and ratings (Experiment 2).**

The participants had unlimited time to give motivation rankings in Experiment 1 and ratings in Experiment 2. On average, the participants took 5256ms (*SEM* = 92ms) to complete the ranking task in Experiment 1. In Experiment 2, the participants took, on average, 6576ms (*SEM* = 110ms) to complete the rating task.