Memory & Cognition (2019) 47:365–382
https://doi.org/10.3758/s13421-018-0872-y

Memory and availability-biased metacognitive illusions for flags of varying familiarity

Adam B. Blake[1] · Alan D. Castel[1]

Published online: 23 October 2018

Abstract

Research on everyday attention suggests that frequent interaction with objects often does not benefit memory or metamemory for them. Across three experiments, participants gave confidence judgments and completed eight-alternative forced-choice tests of the US, Canadian, and Mexican flags. In Experiment 1, environmental availability was correlated with confidence for the US flag, despite similar recognition performance at a saturated time point in the US (July 4th) and a neutral time point (August 6th). In Experiment 2, participants that were asked to verbally describe the flags before judging and remembering them were less accurate and more overconfident than were controls. Experiment 3 utilized a draw-study paradigm wherein participants who first drew the flag had reliably more accurate recognition and confidence scores than those who only studied it. These findings illuminate a persistent metacognitive bias, demonstrate a powerful learning intervention, and extend theories of errorful learning by highlighting the role of attention.

Keywords Recognition · Everyday attention · Metamemory · Flags · Errorful learning

Overall, visual memory tends to be accurate in humans, such that these memories are stored distinct and protected from interference; even when hundreds of photos intervene between the first and second appearance of a photo, recognition accuracy is high (Nickerson, 1965). Other research has shown an immense capacity for visual detail in long-term memory, with high accuracy for more than 2,000 images (Brady, Konkle, Alvarez, & Oliva, 2008). However, in a classic study, people were shown to have difficulty recognizing the correct locations of features on a penny (Nickerson & Adams, 1979). Similarly, people often fail to recall the location of previously seen fire extinguishers, despite the fact that fire extinguishers are in high-visibility locations (Castel, Vendetti, & Holyoak, 2012). Explicit memory is poor for items that people interact with daily, such as the keypads of calculators and telephones (Rinck, 1999), computer keyboards (Snyder, Ashitaka, Shimada, Ulrich, & Logan, 2014), the layout of frequently used elevator buttons (Vendetti, Castel, & Holyoak, 2013), and aspects of road signs (Martin & Jones, 1998), among other items (Castel, Nazarian, & Blake, 2015).

✉ Adam B. Blake
adamblake@ucla.edu

[1] Department of Psychology, University of California, Los Angeles, 1285 Franz Hall Box 951563, Los Angeles, CA 90095, USA

Poor memory for common objects may be due to a form of attentional saturation, which could later result in "inattentional amnesia" (Wolfe, 1999). For common objects, it becomes unimportant to remember their explicit details due to the frequent presence of those objects in the environment. An extreme case of this is the letter *g*. The lowercase letter *g* is commonly written with a "looptail" (like in the current font) or an open tail (like in print handwriting). Despite massive visual experience, participants show poor awareness of these two forms and have difficulty drawing them (Wong, Wadee, Ellenblum, & McCloskey, 2018).

It can be argued that such inattention is an efficient mental adaptation, and that changing the context of encoding that information may lead people to remember it better. That is, it may be that under intentional learning conditions (e.g., Marmie & Healy, 2004), people are better able to memorize information, even information associated with objects previously seen many times. However, in naturalistic settings, there is likely no intent to encode the details of various logos and symbols, which leads to an interesting dissociation: Increased exposure increases familiarity and confidence, but does not reliably affect memory. Despite frequent exposure to simple and often visually pleasing symbols, what we *think* is memorable may not reflect processes in memory and attention that underlie what is actually memorable (Castel et al., 2015).

The familiarity of highly available items may lead people to think they have a good memory for the items. In many cases, this is a good diagnostic cue for memory (e.g., multiple

presentations of an item will lead to better memory at an immediate test than a single presentation would; Ebbinghaus, 1913). However, in the case of very frequently seen items, familiarity may impair attention to their details: The items saturate the environment so thoroughly that the benefit of having a strong memory for them is minimal—if needed, a representation can be found very quickly. In a study regarding memory and confidence in the Apple company logo, participants gave judgments of their confidence in their memory for the logo before and after drawing and choosing the logo from a set of alternatives (Blake, Nazarian, & Castel, 2015). Unlike the prior work with the penny (Nickerson & Adams, 1979), Blake et al. (2015) examined a logo that is prominently advertised, that people attend to frequently, and that was designed to be recognizable. Only one participant was able to draw it with all of the correct features, and roughly half of the participants in the study were unable to pick the correct logo from a set of alternative versions. Participants in the study were also asked to give metacognitive judgments of their performance; in this case, they indicated their confidence in their choices. Participants were overconfident in their memory for the logo when judgments were made prior to both the drawing and recognition tasks (see Iancu & Iancu, 2017, for a replication). The discrepancies between metamemory evaluations and memory performance indicate that participants were relying on inappropriate strategies or information when assessing their memory. These findings resonate well with work suggesting that judgments of performance are inferred through subjective experiences rather than objective performance (Werth & Strack, 2014).

What are the subjective experiences that may be inflating confidence for highly available items? A common influence on metacognitive judgments is the ease of processing information at encoding (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Koriat & Ma'ayan, 2005). Generally speaking, easily "learned" information is judged as easy to remember. In particular, when participants are able to rapidly generate an image of a to-be-studied item, they will give higher likelihood judgments of later recall even though this fluency is not well-correlated with recall (Hertzog et al., 2003). Logos, flags, and other brands are designed to be easy to encode and recognize. Advertisers often strive to create minimalistic, simple logos, which are processed more fluently than overly detailed or complex logos (Janiszewski & Meyvis, 2001).

The design and inherent processing fluency of these highly fluent and available stimuli likely lead to one's overconfidence in memory for them. The present research utilizes the processing fluency inherent to national flags. Many of the design considerations for national flags regard the speed of recognizing the flag. For example, the number of points on the iconic maple leaf of the Canadian flag was decided following wind-tunnel tests of identification and blurriness (Matheson, 1980). Similarly, the design for the flag of the United States of America (US) is based on a naval design where the white stripes were placed on a red background, presumably because a red border is easier to distinguish against a bright sky (Williams, 2012). Additionally, national flags tend to have specific verbalizable rules that are often taught to schoolchildren. For example, the US flag has 13 alternating red and white stripes and a blue field of 50 white stars in the upper left corner. An Italian colleague informed us anecdotally that children in Italy are taught that the green part of their flag should touch the pole. Having both a verbal code and a mental image for a particular object might enhance memory for that object because details are encoded in different ways, resulting a stronger memory trace (Paivio, 1986). However, the presence of both a verbal and visual code may impair metacognition because the verbal code specifically highlights aspects of the object to attend to (e.g., the number of stars that should be present). If those aspects are not critical for identifying the correct flag, they may hinder mnemonic performance and be overconfident prior to the choice and after making the choice.

Another source of metacognitive bias related to overconfidence for common items is retrieval fluency: When it is easy to retrieve information from memory, that information is judged as better learned than information that takes longer to bring to mind (Kelley & Lindsay, 1993; Koriat & Bjork, 2006; Koriat & Ma'ayan, 2005; Schwarz et al., 1991). For frequently seen items such as logos, it is presumably a relatively fluent experience to generate a vague mental image of the logo. Further, this ease of generation may lead to high confidence in the memory that prevents critical inspection of the mental image: If confidence is at ceiling, then there can be no perceived ambiguity in the memory. Indeed, in the study with the Apple logo, participants showed apparent ceiling effects in confidence judgments elicited prior to each memory task (Blake et al., 2015).

Finally, the well-documented availability heuristic suggests that people often use cues such as relative frequency and recency to guide their judgments (Tversky & Kahneman, 1973). In this research, Tversky and Kahneman (1973) showed that judgments of ecological frequency—how often something occurs in the natural world—for items in a category correlate with the number of examples in a category that participants can bring to mind. For instance, a person's estimate of the class's grade-point average might be higher if that person has more friends with higher grades than not. This bias has been explained as an effect of the ease of retrieval for instances rather than the number of instances recalled (Schwarz et al., 1991). Primarily, these availability effects are related to judgments of frequency and probability, but availability may have downstream consequences for metacognitive judgments. We hypothesize that when participants make judgments about highly available items, the judgments are partially influenced by the ease of recalling *encounters* with the items rather than critically assessing their memories for detail. That is, it may be

that when an item is ubiquitous in nature, participants substitute a judgment of the frequency of their recent interactions with the item instead of their memory for it.

The present study approaches this topic from a novel perspective by examining the effects of environmental saturation on metacognition and memory the flag of the United States of America (US). Importantly, relevance and availability of the US flag are highly variable over the course of a year. On the 4th of July and surrounding weeks, the flag is featured prominently in many public venues, leading to a saturated state of availability in memory. Our first aim in this study was to assess how metacognitive judgments parallel the relative availability of the flag. In the case of lexical materials, participants have better recall and faster response times for words congruent to nearby holidays (e.g., "haunted" in October) than for words not associated with nearby holidays (Coane & Balota, 2009). Though we do not anticipate better memory for the US flag, as it is a member of the highly available items discussed here, it is expected that a more flag-saturated environment will lead participants to be more confident in their ability to recall the flag.

To rectify errors in metacognition and memory that may arise from these biases, Experiments 2 and 3 direct participants to attend to more relevant, diagnostic cues for memory. Using availability as a cue is useful in many contexts. However, availability is not always a good diagnostic cue for memory, as has been shown in numerous cases in which frequent interaction with an item has not resulted in better recall (Castel et al., 2015). Retrieval fluency is a more diagnostic cue when used in relevant contexts—namely, when an attempt is made to recall an item in a manner similar to the test context. For example, when participants were given multiple test events during a study phase, they gave more accurate judgments of how they would perform at a final test (slightly underconfident) than did participants that had no test events at study (grossly overconfident; Roediger & Karpicke, 2006). An explanation for this improved accuracy following delayed recall, or test, is that it encourages individuals to attempt to recall a prior memory rather than make a judgment predicated on the current task difficulty (Nelson & Dunlosky, 1991, 1992)—although there is an important difference when people experience multiple tests, as is often done with typical testing effect experiments (Roediger & Karpicke, 2006). In the present task, if participants are making their judgments of confidence based on the availability of encounters with the US flag instead of retrieving a prior memory, then prompting them to think more specifically about their memories for the items and express the details of them, either verbally or visually, is expected to improve metacognitive judgments due to retrieval dynamics rather than by simply monitoring one's memory.

Finally, in the case of highly available images, participants maintain overconfidence following free recall and recognition tests, although their overconfidence is attenuated by experiencing the recall episodes (Blake et al., 2015; Iancu & Iancu, 2017). The final aim of the current study was to correct postrecall overconfidence. Metacognitive biases are relatively resilient, and people do not always fully update their knowledge with experience (Mueller, Dunlosky, & Tauber, 2015). However, overconfidence may be remedied by improving memory through the use of a learning paradigm that specifically highlights errors in memory and corrects those errors. Related research has shown that people are sometimes less overconfident when asked to retrieve specific details of a process before giving their confidence judgment (Keil, 2003; Rozenblit & Keil, 2002)

Generating errors during learning has positive effects on memory when coupled with immediate corrective feedback (Kang et al., 2011; Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009; Yang, Potts, & Shanks, 2017; but see also Cyr & Anderson, 2015). Notably, this research on *errorful learning* focuses on the learning of new information, specifically for novel word associations though it has been extended to more semantically rich information such as trivia questions (Kornell, 2014). However, testing with feedback has also been shown to improve memory for prior-learned information (Fenesi, Sana, & Kim, 2014; see Rowland, 2014, for a meta-analysis). This benefit was limited to practice questions that tested basic retention of facts, which is relevant to memory for visual materials such as flags and logos, where nearly all features are low level. In light of this research, it is expected that introducing a recall event prior to study will lead to error generation that will complement study, improve later recognition of the studied item, and consequently reduce overconfidence (by improving memory to match confidence).

Experiment 1

In Experiment 1, we examined the effects of relative availability of flags on memory and metamemory for those flags. We see flags frequently, and they likely have a high personal relevance for some people. Comparing memory for flags of different countries with that of one's own country provides a foundation for understanding how personal relevance and availability heuristics may affect mnemonic phenomena. Further, national flags have different levels of frequency and extrinsic relevance throughout the year, a point that is integral to this experiment.

If participants make memory judgments based on availability, it is expected that a priori confidence in their ability to correctly recognize the flag will be miscalibrated for available objects like their country's national flag. Further, it follows that at time points during which the US flag is more available, overconfidence will increase compared with more neutral time points. In particular, the US flag is much more prominent and visually available during the weeks surrounding

Independence Day in the US (July 4), which may lead participants to think they can recall its details better—or at least that they "should" be able to, due to the flag's increased cultural significance (and, perhaps, participants' increased patriotism) during that holiday. Additionally, it is expected that participants will be less overconfident in their memory for the Canadian (CA) and Mexican (MX) flags, which are not prominently featured on a daily basis in most parts of the US. However, the CA flag is relatively simplistic in its design compared with the more complex MX flag, and presumably this results in higher encoding and retrieval fluencies. This may foster a sense of confidence in the CA flag compared with the MX flag.

Method

Participants and design Data were collected from 86 participants recruited through Amazon Mechanical Turk, who were paid $6/hr. Participation was limited to people in the US and to workers who had not already participated in any pilot studies involving these or similar materials. Data collection for this experiment occurred at two time points in 2016: July 4 (n = 43, 23 females, M_{age} = 34.14, SD_{age} = 11.68), and August 6 (n = 43, 24 females, M_{age} = 31.58, SD_{age} = 9.10). This variable was manipulated in a naturalistic, quasi-experimental manner where participants were not randomly assigned to a collection date. Participants participated only at one of the time points, during which their recognition and metamemory performance (prerecognition confidence and postrecognition confidence) for three flags was recorded.

Materials A set of eight flag stimuli was constructed for each of the US, MX, and CA flags. Each set included the correct flag along with seven alternatives created by manipulating key features of the flag. Only three prominent features of each flag were systematically varied for each of the alternatives. These alternatives were informed by pilot studies to yield highly competitive lures. The correct features and the corresponding alternate features for each flag are detailed in Table 1. For each flag the emblem was modified, the layout was altered, and the proportion of the flag taken by the main emblem was changed to create the alternate features. A flag was created for each combination of correct and incorrect features, yielding eight flags per country (see the Appendix for the exact materials used).

Procedure Participants started the experiment by entering their Amazon Mechanical Turk worker IDs. On the following screen they were instructed that they would be viewing pictures of common objects and would answer questions about them, but, importantly, they should not look around or navigate away from the page when answering. It was emphasized that their data would be unusable if they were to "cheat," so to speak.

The order and sequence of the flags was fully counterbalanced such that each flag appeared in each position and following each of the other flags across participants, which yielded a total of six counterbalanced orders. Participants were randomly assigned to one of the counterbalanced orders. For each flag, participants were first prompted to rate their confidence in the upcoming flag:

> Imagine that you are shown a set of flags of the United States of America. In the set, one flag is the correct flag, and the others are similar versions. How confident are you that you could pick out the correct version, on a scale from 0 (*not confident at all*) to 100 (*completely confident*)?

Then, they were told that they would be shown a set of eight flags as a grid of choices on a neutral gray background (see the Appendix for an example of each grid). To select a choice, a participant would click on the flag that they felt depicted the correct flag for that particular country. A yellow rectangle would show which flag was selected until the participant chose to submit the response. For each participant, the position of each flag in the grid was randomized. Once the response was submitted, participants were again asked their confidence on a 0 to 100 scale, this time regarding whether or not they chose the correct flag. This sequence would repeat until the participant had responded to each of the three flags.

After all of the prompts and flag sets, participants were asked to answer how many stars were on the US flag, how many stripes were on the US flag, their awareness of relevant holidays (Canada Day on July 1; Independence Day on July 4), and a number of demographic questions.

Results and discussion

All participants in the study were able to accurately report the number of stars and stripes on the US flag, and they were all aware of Independence Day. Only five people were aware that July 1 is a holiday in Canada, which is relatively unsurprising given that participation was limited to the US Figure 1 shows the average prerecognition confidence, recognition accuracy, and postrecognition confidence as a function of flag shown and environmental saturation. The pattern of results shows relatively similar recall for each of the flags that does not differ across time points. Additionally, the rate of decrease from prerecognition to postrecognition judgment appears to be stable. However, there is a clear spike in confidence judgments for the US flag at the 4th of July (high environmental saturation), as expected. To test these apparent effects, an analysis of variance (ANOVA) was run for the recognition and confidence judgments.

Table 1 Altered features for each of the flag stimuli

Flag	Feature	Correct	Incorrect	% correct
CA	1	11-pointed maple leaf	15-pointed maple leaf	79.9
	2	Correctly-sized leaf	Reduced-size leaf	84.3
	3	Straight leaf stem	Naturally bent leaf stem	60.5
US	1	Five-pointed star	Six-pointed star	80.0
	2	50 stars	41 stars	78.1
	3	Field spans seven lines	Field spans six lines	61.9
MX	1	Mexican flag emblem	Color-shifted US presidential seal	61.1
	2	Emblem facing left	Emblem facing right	75.6
	3	Green–White–Red	Red–White–Green	50.5

Note. See Appendix for the stimuli constructed from these descriptions. CA = Canada; US = United States; MX = Mexico; % correct = percentage of people across relevant experiments that chose a flag with correct version of that feature

In the following analyses the conditions are collapsed across the counterbalanced orders. All of the analyses were initially conducted with counterbalancing terms, but there were neither significant effects on recognition (all Fs < 1) nor confidence, main effect of flag order not significant at $F(5, 81) = 1.43$, $\eta_p^2 = .016$, $p = .21$ (all other Fs < 1). It should be noted that these counterbalancing tests were somewhat underpowered, with only approximately 14 participants in each counterbalanced order. To address this, Experiment 2 used a similar design with a greater number of participants.

Recognition The percentage of correct choices was analyzed in a 3 (flag: CA, US, MX) × 2 (environmental saturation: low at August 6, high at July 4) mixed-factor ANOVA. These data indicated no main effect of environmental saturation on recognition, $F(1, 84) = 0.13$, $\eta_p^2 = .002$, $p = .72$, consistent with prior work demonstrating that the high availability of items does not enhance memory for those items (Castel et al., 2015). Further, there was no interaction between the type of flag shown and environmental saturation, $F(2, 168) = 0.06$, $\eta_p^2 = .001$, $p = .94$. Although it was expected that participants might perform more accurately for the US (very familiar, $M = 38.37$, $SD = 48.91$) and CA (less familiar yet very simplistic; $M = 30.23$, $SD = 46.20$) flags than for the MX flag (less familiar and complex; $M = 23.26$, $SD = 42.49$), there were no significant differences, $F(2, 168) = 2.67$, $\eta_p^2 = .031$, $p = .07$. However, the lack of differences here is not particularly surprising: The flag stimuli alternatives were crafted by the researchers to only have three possible alterations, but due to the nature of flags, these alterations cannot be considered equivalent across flags. Thus, it is possible that there are differences in recognition based on both memory for the flag and

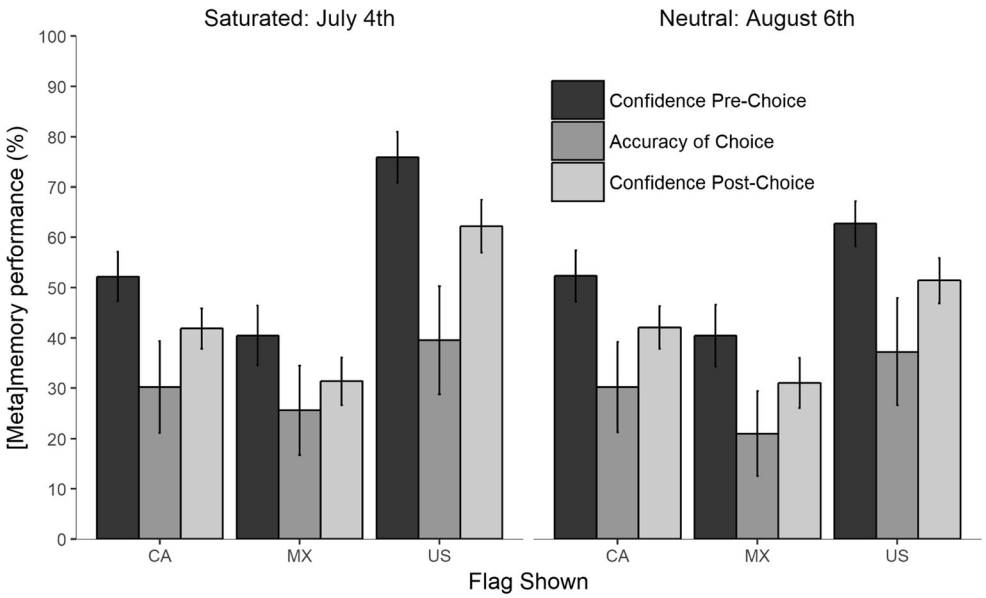


Fig. 1 Metamemory (confidence) and memory (recognition) performance for each of the CA (Canadian), US (United States), and MX (Mexican) flags at the saturated and neutral time points. Error bars indicate 95% confidence intervals

differences in the difficulty of the recognition task across flags, and these differences are cancelling out one another.

Despite a lack of differences among the recognition performance for the flags, it is clear that the percentage of correct choices was very low. Referring back to the percentages in Table 1, there are higher percentages of correct recognition for each individual feature across flags compared with the full representation. It appears that many participants are able to pick out one or two of the correct features, but do not know, or have trouble binding, all three correctly. This highlights the thresholding inherent to recognition memory. It is possible to have some knowledge of the target, but a correct response requires that the participant know enough to distinguish between *all* of the tested foils and to have the patience and ability to adequately compare alternatives. This partial knowledge is likely to increase confidence in memory in the case where the participant is unaware of the remaining features being tested, and possibly even when participants know that a feature is being tested but do not know the correct answer; that is, the partial knowledge may inflate confidence even though doubt over a single binary feature reduces the probability of recognition to 50%.

Confidence Participants' metamemory for the flags was compared using a 3 (flag: CA, US, MX) × 2 (environmental saturation: August 6, July 4) × 2 (prerecognition vs. postrecognition) mixed-factor ANOVA. First, considering the three-way interaction, there were no significant differences in the rate of change in confidence from prerecognition to postrecognition judgments, $F(2, 168) = 0.13$, $\eta_p^2 = .002$, $p = .88$. Similarly, there were no significant changes in prerecognition and postrecognition confidence as a function of flag, $F(2, 168) = 0.66$, $\eta_p^2 = .008$, $p = .52$, or date, $F(1, 84) = 0.04$, $\eta_p^2 < .001$, $p = .84$. However, there was a marked decrease in confidence judgments from those given prior to the recognition task ($M = 54.02$, $SD = 31.26$) to those given after the recognition task ($M = 43.32$, $SD = 30.99$), $F(1, 84) = 35.63$, $\eta_p^2 = .298$, $p < .001$.

We next considered the more critical effects regarding the flags shown and the environmental saturation. For the post hoc tests reported here, the Holm–Bonferroni method was used with independent-samples *t* tests to maintain an alpha level of .05. There was no significant main effect of environmental saturation, $F(1, 84) = 0.70$, $\eta_p^2 = .008$, $p = .41$, on confidence judgments, yet there was a significant difference among the flags, $F(2, 168) = 47.34$, $\eta_p^2 = .360$, $p < .001$. Pairwise comparisons of the average confidence in memory for each flag indicated that participants gave higher confidence judgments for the US flag ($M = 63.06$, $SD = 25.53$) than for the CA flag ($M = 47.11$, $SD = 28.14$), $t(168) = 3.89$, $d = 0.59$, $p_{adj.} < .001$, and higher confidence judgments for the CA flag than for the MX flag ($M = 35.83$, $SD = 26.98$), $t(168) = 2.68$, $d = 0.41$, $p_{adj.} = .03$. Importantly, confidence judgments for the flags interacted with environmental saturation, $F(2, 168) = 3.04$, $\eta_p^2 = .035$, $p = .05$. Pairwise comparisons showed that this interaction was likely driven by the US flag: Average confidence in the US flag at the 4th of July ($M = 69.07$, $SD = 27.20$) was higher than at August 6 ($M = 57.06$, $SD = 22.49$), $t(84) = 2.23$, $d = 0.48$, $p_{adj.} = .03$. There were no significant differences between the time points for either the CA or MX flags ($ts < 1$, $p_{adj.} = 1$). Thus, environmental saturation was associated with elevated confidence, but no difference in accuracy.

Looking at the broader comparisons of recognition, participants were highly overconfident in their ability to choose the correct flag both before (~54%) and after (43%) the recognition task (hit rate ~30%). There is a decline between the two confidence judgments, yet participants remain overconfident. One explanation for this overconfidence might be that participants are retrieving improperly stored representations of the flag, and thus the confidence is high because they feel they chose a matching representation. Alternatively, these findings may suggest that participants are not consulting their memory at all for the flags before making their judgments; instead, they use the most salient heuristics when making their confidence judgments. This is particularly evident in the case of the US flag, where changes in availability of the flag predict changes in participants' confidence regarding it. That is, on the 4th of July when the US flag is relatively saturated in the environment, participants give higher ratings of their confidence for that flag compared with the CA and MX flags, which are not found in great abundance during either time point.

Experiment 2

Given the findings from Experiment 1, indicating relative overconfidence and poor memory for the CA, US, and MX flags, Experiment 2 sought to debias participants' metacognition by prompting them to consider their memory before making confidence judgments. In Experiment 1, it is possible that participants were making their preconfidence judgments regarding the flags based solely on nondiagnostic factors such as availability (Tversky & Kahneman, 1974) or rote knowledge (e.g., memorized rules) about the flags. Considering that the recognition task is visual in nature, it follows that consulting a mental image of each flag would result in a more accurate metacognitive prediction. By asking participants to describe the flags before making judgments about them, we expected participants to bring to mind the mental images of the flags, creating more diagnostic cues to factor into their judgments. Similarly, when participants describe how to complete a task it lowers their overconfidence in understanding the process (Rozenblit & Keil, 2002). These descriptions are expected to improve metacognitive performance by decreasing the overconfidence seen in Experiment 1.

Method

Participants and design Data were collected from 214 participants (114 females, M_{age} = 36.83, SD_{age} = 12.48). Participants were recruited through Amazon Mechanical Turk and paid $6/hr. Participation was limited to people in the US and to workers who had not already participated in Experiment 1 or any pilot studies involving these or similar materials.

In this experiment, participants were primed with either neutral (workspace-related) or targeted (flag-related) prompts for descriptions. Their recognition memory and metacognitive judgments prior to and after recognition were recorded for the CA, US, and MX flags. Considering the relatively small effect sizes in Experiment 1, more participants were recruited in this experiment to ensure an appropriate power level. Assuming a small effect size (η_p^2 = .01) with a moderate correlation between the within-subjects measures, a sample size of approximately 95 participants per between-subjects priming condition was deemed to be sufficient. We posted 214 openings on Amazon Mechanical Turk (107 per condition) to account for attrition and cheating, but did not need to exclude any.

Materials The flag materials for the US, CA, and MX alternatives in this task were the same as those used in Experiment 1 and can be found in the Appendix.

Two types of prompts were created for the experiment. Targeted prompts instructed participants to briefly describe each of the US, CA, or MX flag in their own words. For neutral prompts, participants were asked to briefly describe their computer keyboard, the wall behind them, or the chair that they were sitting in. The orders of the flag-targeted prompts and neutral prompts were each counterbalanced such that they appeared equally often in each position and sequence.

Procedure Participants were randomly assigned in equal numbers to either answer targeted or neutral prompts at their own pace. The procedure in this experiment was nearly identical to Experiment 1, with two differences: The data were collected at only one time point, and prior to each flag sequence (preconfidence judgment, recognition task, postconfidence judgment), participants answered the prompt assigned to the upcoming flag. The orders of flags were again counterbalanced across participants such that each flag appeared in each position and sequence equally often, yielding six orders, as in Experiment 1.

Results and discussion

In Figure 2, summaries of the metacognitive and recognition performances are displayed as a function of the flag shown and the priming prompt type. Compared with Experiment 1, the confidence scores look very similar in that the US averages are higher than CA, and CA are higher than MX. Further, there appears to be a general overconfidence for the US flag that is not seen in the CA or MX flags. As in Experiment 1, there were no effects of counterbalancing on the outcome measures (all Fs < 1), and the analyses are collapsed across the counterbalancing conditions.

Recognition A mixed-subjects ANOVA tested the effects of the flags shown (CA, US, MX) and priming prompt (targeted, neutral) on recognition. The ANOVA revealed effects of both the prompts and the flags, but these effects were not qualified by an interaction, $F(1, 422) = 2.40$, $\eta_p^2 = .011$, $p = .09$.

Participants in the neutral priming condition (M = 54.29, SD = 49.90) performed better on the recognition task than participants in the targeted condition (M = 43.25, SD = 49.62), $F(1, 211) = 6.49$, $\eta_p^2 = .03$, $d = 0.18$, $p = .01$. This finding was somewhat unexpected in that thinking of an object does not usually impair memory for the object. However, research in retrieval-induced forgetting has shown that the act of retrieving some subset of information can reduce memory for other related information (Anderson, Bjork, & Bjork, 1994). It is possible that the act of verbally retrieving some of the details of the flag simultaneously selected against other nonsalient details that would become important at test, thus reducing performance. Similarly, research with the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990) has shown that identification of previously seen faces is impaired when immediately preceded by a verbal prediction task (Meissner & Brigham, 2001). In the current experiment, the verbal descriptions may have oriented participants to nondiscriminative characteristics of the flag foils or caused poor reconstruction of the flag memory.

There was also a significant main effect of flag shown, $F(1, 422) = 10.54$, $\eta_p^2 = .048$, $p < .001$. To examine this main effect, independent samples t tests were conducted using the Holm–Bonferroni method to maintain an alpha level of .05. The CA flag (M = 58.41, SD = 49.40) was not recognized more often than the US flag)M = 49.77, SD = 50.12), $t(213) = 1.93$, $d = 0.13$, $p_{adj.} = .07$, but was recognized more often than the MX flag (M = 37.85, SD = 48.62), $t(213) = 4.57$, $d = 0.31$, $p_{adj.} < .001$. The US flag was also correctly recognized more often than the MX flag, $t(213) = 2.64$, $d = 0.18$, $p_{adj.} = .03$. This pattern of results was what was expected in Experiment 1, although the main effect of flag shown was only trending ($p = .07$) in that instance. The same caveats regarding the construction of the materials and possible differences in relative difficulty of recognition still apply, as this experiment uses the same materials. However, it may be that the previous experiment simply had less power to find the recognition effect with these materials, and the larger sample size for this experiment addressed this issue.

Confidence Participants' confidence scores were compared using a mixed-subjects ANOVA using the flags (CA, US,

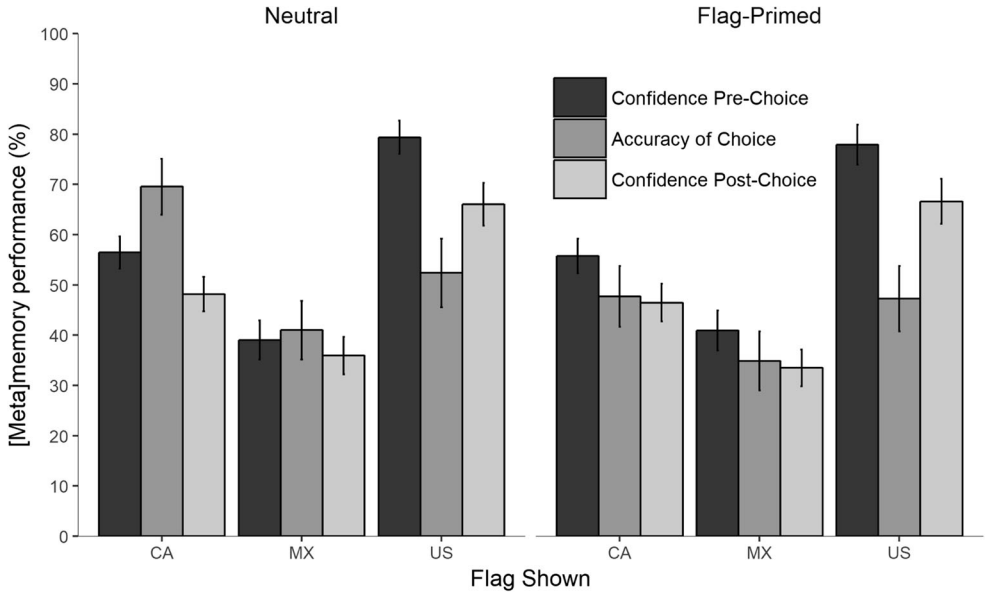


Fig. 2 Metamemory (confidence) and memory (recognition) performance for each of the CA (Canadian), US (United States), and MX (Mexican) flags in the neutral and targeted priming conditions. Error bars indicate 95% confidence intervals

MX), priming condition (targeted, neutral) and judgment timepoints (prechoice, postchoice) as independent variables. The three-way interaction was not significant, $F(2, 424) = 0.99$, $\eta_p^2 = .005$, $p = .37$. Further, there was no significant interaction between priming prompt and judgment time point, $F(1, 212) = 0.33$, $\eta_p^2 = .002$, $p = .57$, no interaction between priming prompt and flag, $F(2, 424) = 0.02$, $\eta_p^2 < .001$, $p = .98$, nor a main effect of priming prompt, $F(1, 212) = 0.09$, $\eta_p^2 < .001$, $p = .77$. The lack of priming effects on confidence is intriguing and defies the a priori expectations for this experiment. It was hypothesized that the largely inaccurate metacognitive judgments for Experiment 1 were caused by the use of heuristics instead of memory appraisal through retrieval. However, even when participants actively recalled the flag, they were unable to make an appropriate judgment of their memory for it.

Despite the lack of priming effects on confidence, there were significant effects on confidence depending on the flag shown, $F(2, 424) = 98.05$, $\eta_p^2 = .316$, $p < .001$. Dependent-samples t tests were conducted using the Holm–Bonferroni method to maintain an alpha level of .05. Participants were more confident in their memory for the US flag ($M = 72.45$, $SD = 25.45$) compared with the MX flag ($M = 37.32$, $SD = 26.60$), $t(212) = 13.93$, $d = 0.95$, $p_{adj.} < .001$, and compared with the CA flag ($M = 51.0$, $SD = 26.90$), $t(212) = 8.22$, $d = 0.56$, $p_{adj.} < .001$. Participants were also more confident in their memory for the CA flag compared to the MX flag, $t(213) = 5.70$, $d = 0.39$, $p_{adj.} < .001$. This pattern is consistent with Experiment 1, where overall confidence for the US flag was greater than for the CA flag was greater than for the MX flag.

Lastly, participants were more confident in their judgments prior to each choice ($M = 58.23$, $SD = 31.60$) than those after ($M = 49.42$, $SD = 32.84$) the recognition choice was made, $F(1, 212) = 87.28$, $\eta_p^2 = .292$, $d = .639$, $p < .001$. This effect was consistent across the flags, as shown in Table 2. In each case, the postrecognition confidence judgment was much better calibrated to actual memory performance. This change in confidence was greater for some of the flags than the others, $F(2, 424) = 4.99$, $\eta_p^2 = .023$, $p = .01$.

To determine where these differences arose, difference scores were computed for each of the flags by subtracting prechoice confidence from postchoice confidence. The change in the US ratings ($M = -12.30$, $SD = 25.00$) was significantly greater than change in the MX ratings ($M = -5.27$, $SD = 22.5$), $t(210) = -3.20$, $d = -0.22$, $p = .002$, but not the change in the CA ratings ($M = -8.83$, $SD = 22.80$), $t(210) = -1.50$, $d = -0.10$, $p = .10$. The change in ratings for the MX and CA flags was not significantly different, $t(210) = 1.6$, $d = 0.11$, $p = .10$.

The reduced variance from prerecognition to postrecognition in the MX flag data belies a "confidence in one's own confidence"; that is, participants understand that

Table 2 Descriptive statistics and t tests for the confidence ratings in Experiment 2

Flag shown	Prechoice		Postchoice		*df*	*t*	*d*	*p*
	M	*SD*	*M*	*SD*				
US	78.62	25.10	66.29	31.29	213	7.20	0.49	<.001
CA	56.12	28.55	47.29	29.88	213	5.65	0.39	<.001
MX	39.96	28.32	34.69	29.42	213	3.43	0.24	<.001

Note. US = United States; CA = Canada; MX = Mexico

they have a poor recollection of the MX flag and are able to make a relatively accurate judgment for it that does not change over time. On the other hand, the very available US flag and relatively simplistic CA flag are associated with increased levels of overconfidence prior to the recognition task. This discrepancy in overconfidence suggests that participants are attending to different cues when making their judgments for nonavailable and unfamiliar items compared with familiar items.

Experiment 3

In Experiment 2, participants who typed out the features of the flag before attempting to identify it (priming condition) performed more poorly than participants who completed a control task. Possibilities for this result include the effects of verbal overshadowing (Meissner & Brigham, 2001) or retrieval-induced forgetting (Kornell et al., 2009). Further, the act of describing the flag using verbal codes involves different mental faculties than simply picturing the flag in one's mind (e.g., Paivio, Rogers, & Smythe, 1968). It may be that the verbal code for such a well-known flag invokes different details than a more visual code. Given that the flag is relatively simple in its visual design—as compared with, for example, a human face—it may be more helpful to attempt to draw the flag as a method of retrieving the memory. Drawing also has been shown to have strong memorial effects, not unlike the related production effect (MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010), possibly because drawing involves the integration of semantic, visual, and motor memories (Wammes, Meade, & Fernandes, 2016) and can aid in producing recollection based memories with more intact source memory (Wammes, Meade, & Fernandes, 2018).

It is expected that when participants are forced to attempt to draw out the features of the flag, they will be more likely to focus on features they had not considered critical before. For example, participants from the US are very likely know that the US flag has 50 stars and 13 stripes, but may not have considered the arrangement of the stars, the shape of the blue field, and the number of stripes below the blue field. When attempting to draw the flag, these details must be considered in order to create a picture. Generating errors surrounding such details and then providing corrective feedback should benefit memory for those details (Fenesi et al., 2014; Kang et al., 2011; Richland et al., 2009). This research on feedback suggests that if participants were to see the correct flag after they had trouble attempting to draw it, they will develop stronger memories for the flag because they are cued to attend to crucially overlooked features. However, if participants were to only study the flag, it is unlikely that they would consider such features, as they have not attended to them in countless past viewings of the flag.

Extending these predictions to metamemory judgments, past research shows that participants are frequently overconfident prior to any attempts to retrieve common objects from memory (Blake et al., 2015; Iancu & Iancu, 2017). It is unlikely that simply studying the flag will alter that confidence, as it is already a well-known and commonly seen object. However, attempting to recall the flag and experiencing the disfluency of retrieval for unknown details may force participants to temper their metacognitive judgments (Miller & Geraci, 2014; Rozenblit & Keil, 2002). The more salient cue of retrieval strength should force participants to realize which features they do not have committed to memory and lower their confidence.

Method

Participants Data were collected from 52 participants (35 females, M_{age} = 19.86, SD_{age} = 1.43) through the Psychology Department subject pool at University of California, Los Angeles. Participants received course credits for participating in the study. Given that drawing alone has been shown to have very large effects (ds > 1) on recall (Wammes et al., 2018), a sample of 26 participants per condition is justified for this research to achieve an appropriate power level of .80 for a large effect (d = 0.8).

Materials and procedure Participants were randomly assigned in equal numbers to either a study-only condition or a draw-then-study condition. All participants were seated at a desk with a computer that displayed the questions and images in the experiment.

To begin the experiment, all participants were asked to rate how confident they were that they could correctly choose the US flag from a group of alternatives on a scale from 0 (*not at all*) to 100 (*extremely*). Then, in the draw-then-study condition, participants were given a sheet of paper and colored pencils and told to draw the US flag on the provided sheet of paper. After 40 s, the paper was removed and they were shown a correct image of the US flag on the computer screen to study for 40 s. Participants in the study-only condition were shown the correct image for 80 s and not asked to draw anything. All participants then made another rating of their confidence that they could choose the correct US flag from a set of similar alternatives.

Prior to the recognition phase of the experiment, all participants completed other laboratory experiments for approximately 20 min. These experiments were primarily word-learning experiments and did not have relevant visual stimuli.

After the intervening experiments were completed, participants again rated their confidence in their ability to choose the correct US flag. Then, they were shown the US flag

alternatives (see the Appendix) in a grid and made their choice of the correct flag, as in Experiments 1 and 2. Finally, they were asked to rate their confidence that they chose the correct US flag.

Results and discussion

Recognition This experiment utilized a drawing task to promote error generation and facilitate learning. An independent-samples *t* test showed that a larger percentage of participants in the draw-then-study condition (M = 76.92, SD = 42.97) chose the correct US flag than in the study-only condition (M = 38.46, SD = 49.61), $t(50) = 2.99$, $d = 0.83$, $p = .004$. These data indicate that drawing the flag benefited participants' subsequent learning of the flag at study. This is particularly interesting in that participants in the study-only condition had twice as long to study the flag (80 s) compared to the draw-then-study participants (40 s draw, 40 s study). This is consistent with research on errorful learning showing that generating an answer is more beneficial than an equivalent amount of study time, even if the answer is incorrect (Kornell et al., 2009). Attempting to retrieve the flag served as a powerful learning event that produced learning beyond intentional study of the flag. We suggest that the benefit in this experiment is derived from the *productive failures* made during the drawing phase. These failures likely serve to direct participants' attention to study the features of the flag that they were unsure of when attempting to draw it. This explanation fits with other data showing that immediate feedback improves memory for fact-based information (Fenesi et al., 2014), even when errors are made at test (Kang et al., 2011). Lastly, these data complement recent research on the benefits of drawing information rather than just studying or restudying it (Wammes et al., 2016, 2018).

Confidence In the last two experiments, metacognitive judgments have been poorly calibrated in that participants overestimated their performance on the recognition task. Additionally, the recognition test has acted as something of a metacognitive intervention where judgments made posttest have been lower, indicating a lower overconfidence. Figure 3 shows the confidence at each judgment time in this experiment (prestudy, poststudy, prechoice, postchoice), with separate lines indicating the condition (draw, study) that is being summarized. The general pattern suggests that participants were very confident in their ability to recognize the US flag and that this confidence seems to be unaffected by the condition they were in. This lack of difference would be in sharp contrast to Rozenblit and Keil (2002) where participants were less overconfident when they had attempted to recall a process than when they had not. Further, there appears to be drop in the final judgment which would indicate that the test is acting as a debiasing event in this experiment.

A 4 (judgment timepoint: prestudy, poststudy, prechoice, postchoice) × 2 (condition: study only, draw study) mixed-subjects ANOVA was used to analyze these apparent effects. The interaction between judgment time point and condition was indeed nonsignificant, $F(3, 150) = 0.73$, $\eta_p^2 = .014$, $p = .54$. Similarly, participants who were asked to draw the flag during the study phase (M = 86.58, SD = 24.32) were no more confident that participants who only studied the flag (M = 78.32, SD = 32.08), $F(1, 50) = 0.04$, $\eta_p^2 = .001$, $p = .83$. However, there was a main effect of judgment time point, $F(3, 150) = 12.17$, $\eta_p^2 = .196$, $p < .001$.

Multiple paired-samples *t* tests were run to elucidate the nature of the time point main effect. The Holm–Bonferroni method was used to maintain an alpha level of .05. From prestudy (M = 88.98, SD = 15.04) to poststudy (M = 93.14, SD = 12.65), participants became more confident, $t(51) = -2.85$, $d = -0.40$, $p_{adj.} = .02$. This is likely because the participants have just seen a perfect rendition of the flag, which is a common object, and thus the judgment is made in a very fluent retrieval context compared with prestudy. In the interval during which participants completed other lab tasks, confidence dropped from poststudy to prechoice (M = 90.37, SD = 14.46), $t(51) = 2.89$, $d = 0.40$, $p_{adj.} = .02$. Presumably, the intervening tasks degraded participants' retrieval fluency of the flag, making it harder to pull the details to mind, and reducing their confidence in turn. The values at prestudy, when the flag has not been seen in the lab yet, and at prechoice, when the intervening tasks had just completed, were similar; participants went back to baseline after the intervening tasks, $t(51) = -0.76$, $d = -0.11$, $p_{adj.} = .45$. Lastly, from prechoice (M = 90.37, SD = 14.46) to postchoice (M = 82.08, SD = 15.89), participants again were debiased somewhat by the recognition task, which is "harder than [they] thought," as some participants reported, $t(51) = 2.914$, $d = -0.40$, $p_{adj.} = .02$.

It was expected that participants would be less confident in their memory for the flag after the study opportunity in the draw-then-study condition. The act of drawing likely highlighted many participants' missing or false memories for details of the flag, as was the case in other related studies (Blake et al., 2015; Iancu & Iancu, 2017). The salience of these errors generally decreases confidence in memory, and some research has shown that participants are unaware of the benefits of error generation on learning (Yang et al., 2017). However, in this experiment, the opposite was true: Participants in the drawing condition did not show a decrease in confidence following the study phase, and all participants were more confident at the poststudy

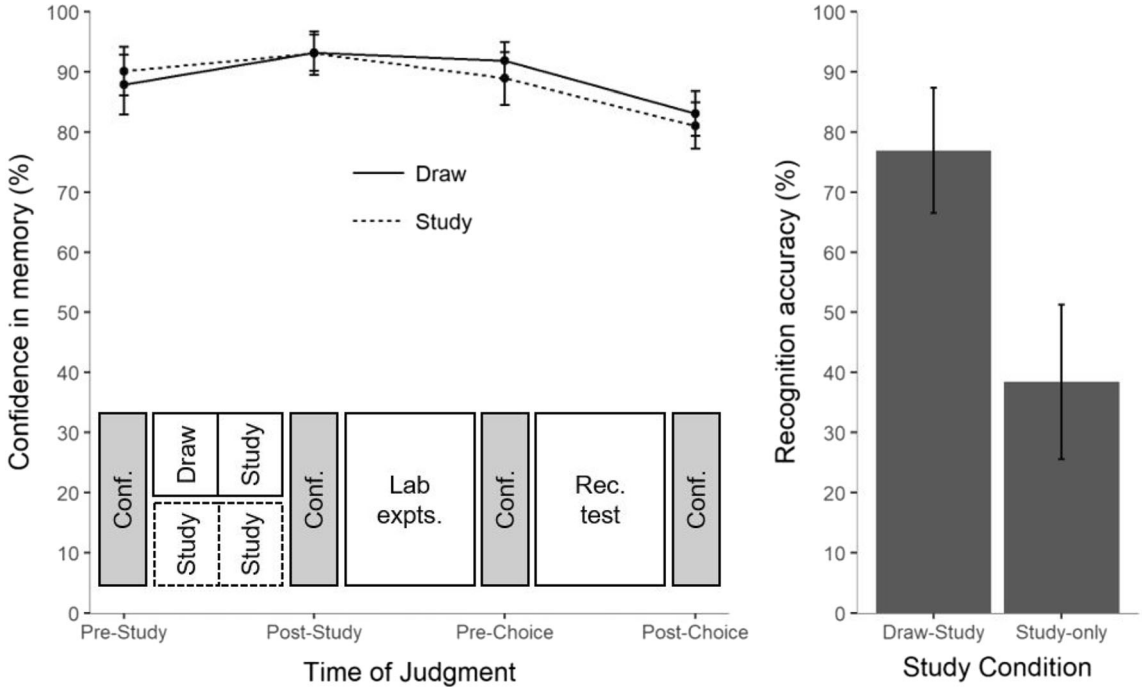


Fig. 3 Confidence in memory for the flag of the United States at each of the four metacognitive-judgment time points (left panel) compared with the recognition accuracy as a percentage of correct responses (right panel). A study outline diagram is overlaid in the left panel to clarify when each measure was taken. Error bars indicate 95% confidence intervals

judgment. A likely explanation is that the feedback portion of the draw-then-study condition was enough for participants to evaluate and rectify problems with their memory for the flag and thus reinflate their confidence. Indeed, poststudy confidence appears to be exhibiting a ceiling effect.

Finally, participants in the draw-then-study condition showed less of a deviation of their metacognitive judgments from their recognition scores than did the study-only group. For each participant, the average of all metacognitive judgments was computed and then subtracted from the recognition score. An independent-samples t test showed that participants in the drawing condition (M = −12.07, SD = 41.43) gave less extreme judgments than those in the study-only condition (M = −49.83, SD = 50.39), $t(50) = 2.95$, $d = 0.82$, $p = .005$. A difficulty with interpreting these results in confidence is that the recognition performance is changing, as well. Because calibration is dependent on both performance and metacognitive ability, it is unclear if the changes in confidence are due to improved calibration or if the discrepancy is entirely driven by the increased ability to recognize the US flag. One method of addressing this might be to equate performance across contexts as a method of isolating changes in confidence and calibration (Bodner & Lindsay, 2003); however, that may be difficult with these types of materials. The lack of differences between the confidence patterns in each group possibly suggests that participants were performing similar mental tasks. Nonetheless, participants in the drawing group did show less deviation between metacognitive judgments and memory due to the enhancement in memory.

General discussion

In this study, memory and metamemory for flags was examined across three experiments. A clear thread in these three studies is that participants remained relatively overconfident, especially prior to test. In each case, the testing event served to debias metacognitive judgments significantly: Once faced with the test, participants were made aware that their memory was not as accurate as they thought. The only apparent changes in metacognition prior to the recognition task were seen in Experiment 1, where participants in August showed less overconfidence than in July (presumably as a function of flag availability), and in Experiment 3, where the study event *increased* overconfidence.

Interpreting this time-of-year effect in light of past research on metacognitive biases, the results of these experiments support a theory of cue utilization (Koriat, 1997), but also show how availability can bias use of certain familiar cues. The US

flag is a highly salient national symbol: Students learn about it early in school, it is part of history, and Americans likely feel that we should know it well. Additionally, as discussed in the introduction, national flags are often constructed to be relatively simple and easy to encode, which may lead to biased metacognition due to the processing fluency. In these studies, participants showed particular overconfidence in the US flag that correlated with natural changes in the environmental saturation of the flag. The US flag tends to be more available in the days surrounding July 4 than a baseline comparison at August 6, as the flag appears in Independence-Day-themed advertisements, social media posts, and even apparel and lawn decorations. Participants tested at the saturated time point showed a stronger tendency for overconfidence than did those at the neutral time point, suggesting that environmental saturation and availability may play a large role in how people make their judgments about these types of frequently seen items. Though these findings may support an availability-biased explanation, the findings could also be considered in terms of significance of the flag or associative strength of a flag-based holiday. Stronger tests of this may include examining this issue in people who have had greater exposure to flags over a lifetime, with strong emotional allegiances (potentially military veterans), or at times when flags are prominent for a variety of other countries (such as during the Olympics), and measuring degree of allegiance to the flag.

In this study, all of the participants were tested in the US, and most were US citizens. It is unclear whether participants in other countries would show such systematic overconfidence for their own flag compared with other countries' flags. We hypothesize that the very simplistic nature of many countries' flags—such as Germany, France, and Italy, which only include three bands of color—will yield an accurately high confidence in memory for citizens of those countries. On the other hand, participants in Canada and Mexico may perform similarly to these US participants due to the more feature-rich nature of their flags. Further, these effects may be more complex when considering globalization and national identity, which not only have a complex effect on one another (Ariely, 2012) but also may change how people see and remember the flags of their country and others.

This present research differs somewhat from other research on prospective judgments of learning and cue utilization in metacognition. Particularly, when participants are asked to retrieve information from memory, this is usually a very salient diagnostic cue of that memory, which results in very accurate judgments (Nelson & Dunlosky, 1991). In these studies, participants were also asked to judge their memory for the US flag and others when the flag was not present in memory, and yet were unable to produce accurate confidence judgments in their memory. One explanation for the inaccuracy may be that participants were simply unable to comprehend the difficulty of the recognition test despite it being described to them. That is, confidence may have been inaccurate because they were unfamiliar with the task in general or because they had an inaccurate idea of the nature of the test. We discount this notion with the counterargument that the systematic bias for the US flag was present in Experiments 1 and 2, where the order of the flags was counterbalanced; counterbalancing the flags' orders ensures that two thirds of the ratings for the US flag came after having experienced the same task for the CA flag, the MX flag, or both. There were similarly no effects of the counterbalanced orders on ratings in either experiment. Additionally, the increase in overconfidence across time point in Experiment 1 suggests that this overconfidence is related, in part, to the relative availability of recent interactions or sightings with the flag. We suggest that participants are not only considering that they should know the flag due to its ease-of-encoding, cultural significance but also because of the ease of recalling encounters with it.

In Experiment 2, we sought to bridge the discrepancy and resolve the overconfidence issue by forcing participants to verbally describe the items before making judgments or recognition attempts. If participants were making their judgments more on availability rules than on assessments of retrieval fluency, then their metacognitive judgments would likely improve by making the diagnostic retrieval cues more salient. This description task had no effect on metacognitive judgments, suggesting that participants already attempt to bring the item to mind when making their judgment. Further, participants exhibited a verbal overshadowing effect (Schooler & Engstler-Schooler, 1990) where their recognition performance was weakened by the description task, and incidentally resulted in more overconfidence in that condition as a result of the attenuation. A possible explanation for specifically how their performance was weakened is that by retrieving specific components of the flag via their verbal descriptions, participants simultaneously suppressed memory traces for the nonretrieved components (retrieval-induced forgetting; Anderson et al., 1994). Retrieval-induced forgetting is generally studied in the context of multiple study trials, and it may be argued that there is not enough time for substantial suppression effects to occur in the window of time from description to recognition in this study. Perhaps in this situation it would be more apt to label this effect a retrieval-induced attentional neglect, where participants are hyperfocusing on the features they feel to be important based on their verbal descriptions.

Finally, Experiment 3 demonstrated the power of a study-then-draw learning paradigm that greatly enhanced recognition memory for the US flag. By asking participants to attempt to draw the US flag, rather than report a verbal code for it, as in Experiment 2, memory for the flag was improved beyond that of study alone. This improvement in memory complements findings in extant literature regarding generation (Slamecka & Graf, 1978) and production (MacLeod et al., 2010) effects on memory. More recently, drawing has additionally been shown to enhance to-be-remembered information via motoric, elaborative, and pictorial mechanisms (Fernandes, Wammes, & Meade, 2018). Because memory for the flag was enhanced, participants in the drawing group showed less overconfidence than those in the study-only group. However, the reduction in overconfidence appears to be strongly driven by the improvement in recognition performance, and it is unclear if there is any difference in how participants are making their confidence judgments. Though this drawing effect was shown for enhancing recall of verbal materials, the current study effectively extends the effect to restudy of visual materials. Further, we suggest that the act of drawing made errors in participants' memories for the flag salient, allowing the feedback to have a stronger effect like in other errorful-learning research (e.g., Kornell et al., 2009; Richland et al., 2009).

Failure to recall all the details of a flag is somewhat opaquer than errors in recalling a list of words. When participants are asked to recall as many words as possible from a list, it is very clear when the recalled words number fewer than the studied words. In the case of visual materials, unless the image is drawn participants may not be able to fully assess how many and which details are missing. It is possible that participants are partially relying on semantic rather than visual memory for the flags, and that matches with semantic knowledge inflate confidence. For example, the semantic knowledge that the CA flag has a maple leaf in the center increases confidence, but participants do not readily employ mental strategies to *reduce* confidence and are unable to appropriately assess how detailed that maple leaf should be until they put pen to paper. In essence, the act of drawing the mental representation forces the learner to produce their failures, a process that invokes retrieval dynamics and diagnostic monitoring that result in better memory for the items than additional study alone does.

A limitation of this paradigm for obtaining confidence judgments at each time point is the lack of ability for the upper bound of the scale to grow. Participants were very, very confident on average when making their initial judgments of confidence. In Experiment 3, judgments rose following the study phase in both conditions, but it is likely that there were ceiling effects that masked the true amount of change. Subsequent comparisons to these truncated judgments may also not show the true change in confidence for participants. Future studies should employ methods of training participants on similar tasks to improve their understanding of the scale points, or perhaps allow for relative judgments to be made (i.e., *more* or *less* than the previous judgment) instead of absolute values.

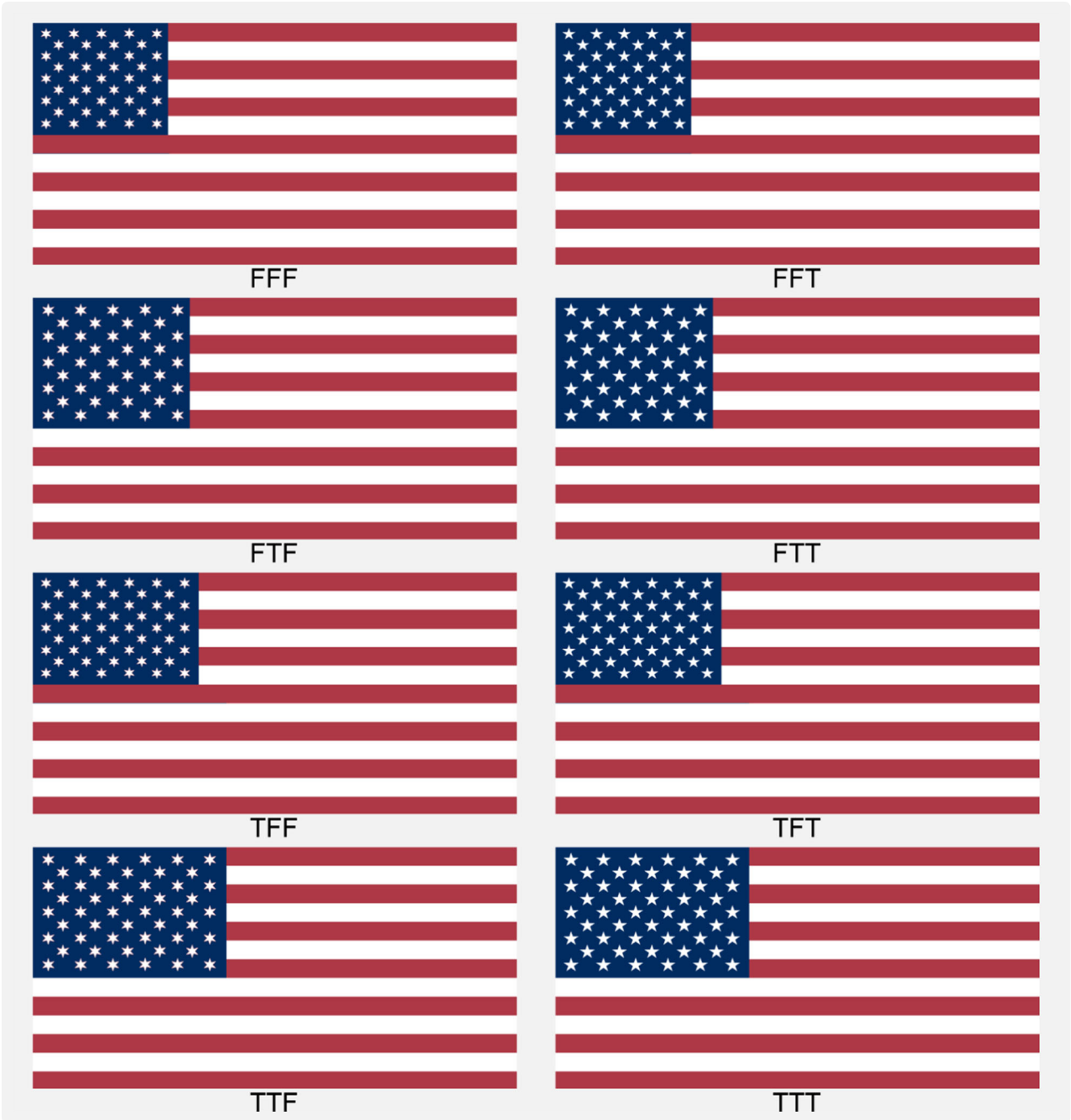
In sum, this collection of studies shows the improper utilization of cues in metacognitive judgments about national flags, which are highly available, frequently seen objects that we often feel we *should* be able to remember. Further, there is a nontrivial relationship between overconfidence and environmental availability of items (Experiment 1) as well as recent mnemonic activity (Experiment 2). It is clear that these types of items are special in their ability to bias metacognitive faculties across a number of domains and everyday settings (Castel et al., 2015), but this study also shows that this bias is not insurmountable. We demonstrate a powerful learning tool (Experiment 3) that can be useful in a variety of contexts: the draw-then-study method for invoking productive failure. Though this tool improves recognition memory for visual materials, it is unclear if this learning is reflected in metacognitive judgments. This work extends theories of errorful learning and generation to visual materials and highlights the role of productive failures in focusing attention to previously overlooked features. We suggest further research to address the long-term effects of this method, and the application of this method to rectify other everyday memory failures, some of which can be very important, such as the location of the nearest fire extinguisher (Castel et al., 2012). As metamemory both matures and changes across the life span (Blake & Castel, 2015), it is also of interest to examine the effects of the variables considered here in children and older adults. Flags are especially interesting to study across the life span, as national attitudes and even the representation of the flag shift over time.

Author note We thank Matthew Rhodes, Mary Hargis, Catherine Middlebrooks, and the CogFog research group for their helpful insights on this research, and also Tyson Kerr for his input in designing the study implementation, Kelly Masuda for copy editing, and Aashna Oberoi for help with data collection and coding. Portions of this work were presented at the 56th (Chicago, IL, 2015) and 58th (Vancouver, B.C., Canada, 2018) annual meetings of the Psychonomic Society.

Appendix

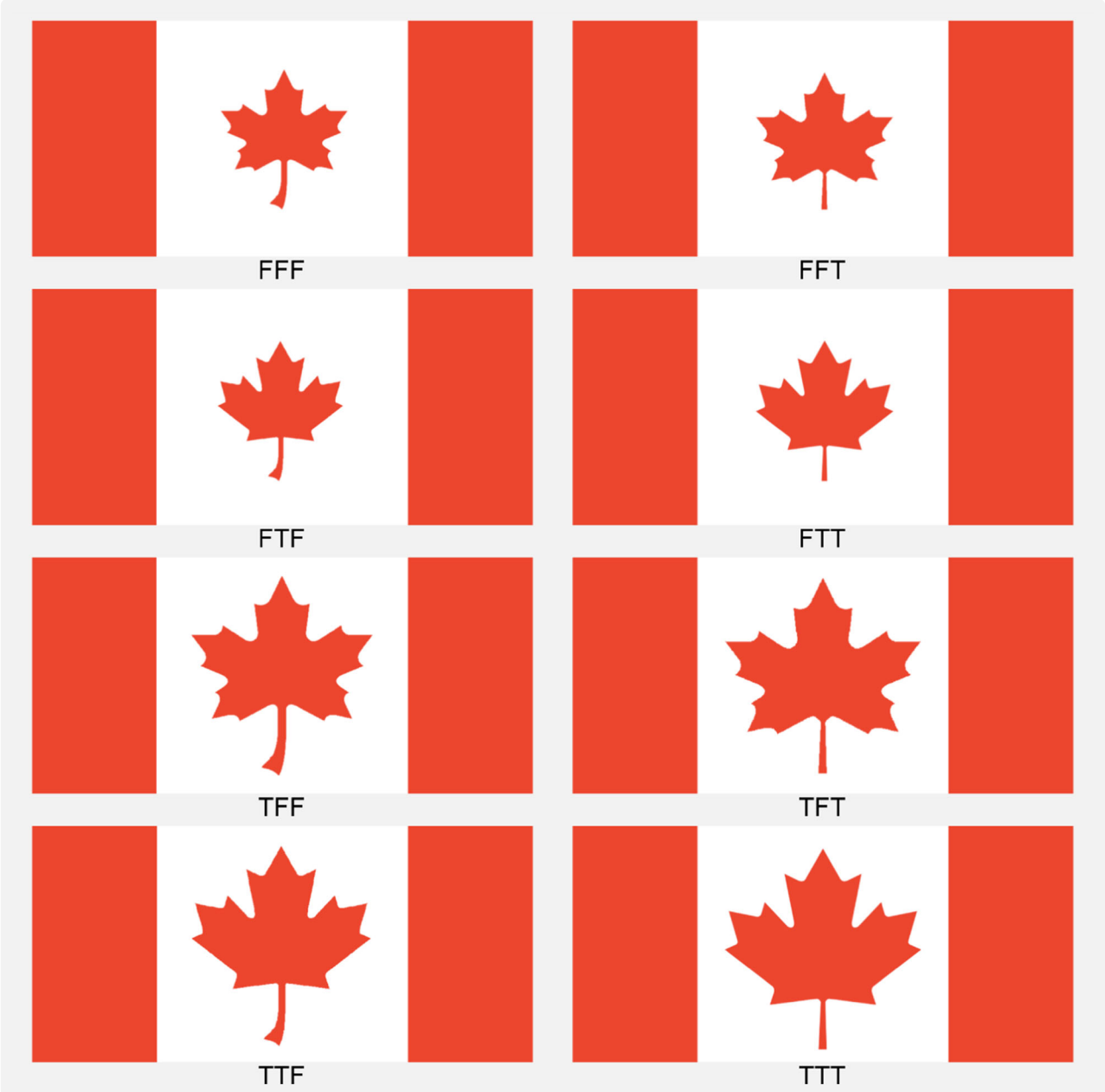
Flag stimuli used in the experiments. A label is given below each flag indicating which features are correct or incorrect (see Table 1). For example, *FTF* is short for *false–true–false* and indicates that the first and third features are incorrect, but the second is correct.

United States of America



Mexico

Canada



References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063

Ariely, G. (2012). Globalisation and the decline of national identity? An exploration across sixty-three countries. *Nations and Nationalism*, *18*(3), 461–482. https://doi.org/10.1111/j.1469-8129.2011.00532.x

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Blake, A. B., & Castel, A. D. (2015). Metamemory. In S. K. Whitbourne (Ed.), The encyclopedia of adulthood and aging (pp. 1–5). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118521373.wbeaa043

Blake, A. B., Nazarian, M., & Castel, A. D. (2015). The Apple of the mind's eye: Everyday attention, metamemory, and reconstructive memory for the Apple logo. *The Quarterly Journal of Experimental Psychology*, *68*(5), 858–865. https://doi.org/10.1080/17470218.2014.1002798

Bodner, G. E., & Lindsay, D. S. (2003). Remembering and knowing in context. *Journal of Memory and Language*, *48*(3), 563–580. https://doi.org/10.1016/S0749-596X(02)00502-8

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Castel, A. D., Nazarian, M., & Blake, A. B. (2015). Attention and incidental memory in everyday settings. In J. M. Fawcett, E. F. Risko, & A. Kingstone (Eds.), The handbook of attention (pp. 463–483). Cambridge, MA: MIT Press.

Castel, A. D., Vendetti, M., & Holyoak, K. J. (2012). Fire drill: Inattentional blindness and amnesia for the location of fire extinguishers. *Attention, Perception, & Psychophysics*, *74*(7), 1391–1396. https://doi.org/10.3758/s13414-012-0355-3

Coane, J. H., & Balota, D. A. (2009). Priming the holiday spirit: Persistent activation due to extraexperimental experiences. *Psychonomic Bulletin & Review*, *16*(6), 1124–1128. https://doi.org/10.3758/PBR.16.6.1124

Cyr, A.-A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 841–850. https://doi.org/10.1037/xlm0000073

Ebbinghaus, H. (1913). Retention as a function of the number of repetitions. In H. A. Ruger & C. E. Bussenius (Trans.), Memory: A contribution to experimental psychology (pp. 52–61). New York, NY: Teachers College Press. https://doi.org/10.1037/10011-006

Fenesi, B., Sana, F., & Kim, J. A. (2014). Evaluating the effectiveness of combining the use of corrective feedback and high-level practice questions. *Teaching of Psychology*, *41*(2), 135–143. https://doi.org/10.1177/0098628314530344

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science*, *27*(5), 302–308. https://doi.org/10.1177/0963721418755385

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22–34. https://doi.org/10.1037/0278-7393.29.1.22

Iancu, I., & Iancu, B. (2017). Recall and recognition on minimalism. A replication of the case study on the Apple logo. *Kome*, *5*(2), 57–70. https://doi.org/10.17646/KOME.2017.24

Janiszewski, C., & Meyvis, T. (2001). Effects of brand logo complexity, repetition, and spacing on processing fluency and judgment. *Journal of Consumer Research*, *28*(1), 18–32. https://doi.org/10.1086/321945

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103*(1), 48–59. https://doi.org/10.1037/a0021977

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368–373. https://doi.org/10.1016/S1364-6613(03)00158-X

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*(1), 1–24. https://doi.org/10.1006/jmla.1993.1001

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037//0096-3445.126.4.349

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133–1145. https://doi.org/10.1037/0278-7393.32.5.1133

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*(4), 478–492. https://doi.org/10.1016/j.jml.2005.01.001

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*(1), 106–114. https://doi.org/10.1037/a0033699

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(3), 671–685. https://doi.org/10.1037/a0018785

Marmie, W. R., & Healy, A. F. (2004). Memory for common objects: brief intentional study is sufficient to overcome poor recall of US coin features. *Applied Cognitive Psychology*, *18*(4), 445–453. https://doi.org/10.1002/acp.994

Martin, M., & Jones, G. V. (1998). Generalizing everyday memory: Signs and handedness. *Memory & Cognition*, *26*(2), 193–200. https://doi.org/10.3758/BF03201132

Matheson, J. R. (1980). Canada's flag: A search for a country. Boston, MA: G. K. Hall & Co.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, *15*(6), 603–616. https://doi.org/10.1002/acp.728

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition*, *29*, 131–140. https://doi.org/10.1016/j.concog.2014.08.008

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2015). Why is knowledge updating after task experience incomplete? Contributions of encoding experience, scaling artifact, and inferential deficit. *Memory & Cognition*, *43*(2), 180–192. https://doi.org/10.3758/s13421-014-0474-2

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, *2*(4), 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Dunlosky, J. (1992). How shall we explain the delayed-judgment-of-learning effect? *Psychological Science*, *3*(5), 317–318. https://doi.org/10.1111/j.1467-9280.1992.tb00681.x

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology*, *19*(2), 155–160. https://doi.org/10.1037/h0082899

Nickerson, R. S., & Adams, M. J. (1979). Long-term memory for a common object. *Cognitive Psychology*, *11*(3), 287–307. https://doi.org/10.1016/0010-0285(79)90013-6

Paivio, A. (1986). Mental representations: A dual coding approach. New York, NY: Oxford University Press.

Paivio, A., Rogers, T. B., & Smythe, P. C. (1968). Why are pictures easier to recall than words? *Psychonomic Science*, *11*(4), 137–138. https://doi.org/10.3758/BF03331011

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

Rinck, M. (1999). Memory for everyday objects: Where are the digits on numerical keypads? *Applied Cognitive Psychology*, *13*(4), 329–350. https://doi.org/10.1002/(SICI)1099-0720(199908)13:4<329::AID-ACP583>3.0.CO;2-3

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1016/S0364-0213(02)00078-2

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*(1), 36–71. https://doi.org/10.1016/0010-0285(90)90003-M

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*(2), 195–202. https://doi.org/10.1037//0022-3514.61.2.195

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Snyder, K. M., Ashitaka, Y., Shimada, H., Ulrich, J. E., & Logan, G. D. (2014). What skilled typists don't know about the QWERTY keyboard. *Attention, Perception, & Psychophysics*, *76*(1), 162–171. https://doi.org/10.3758/s13414-013-0548-4

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vendetti, M., Castel, A. D., & Holyoak, K. J. (2013). The floor effect: Impoverished spatial memory for elevator buttons. *Attention, Perception & Psychophysics*, *75*(4), 636–643. https://doi.org/10.3758/s13414-013-0448-7

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752–1776. https://doi.org/10.1080/17470218.2015.1094494

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2018). Creating a recollection-based memory through drawing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(5), 734–751. https://doi.org/10.1037/xlm0000445

Werth, L., & Strack, F. (2014). An inferential approach to the knew-it-all-along phenomenon. *Memory*, *11*(4/5), 411–419. https://doi.org/10.1080/09658210244000586

Williams, E. P., Jr. (2012). Did Francis Hopkinson design two flags? *The Quarterly Newsletter of the North American Vexillogical Association*, *216*, 7–9. Retrieved from http://www.flagguys.com/pdf/NAVANews_2012_no216.pdf

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), Fleeting memories (pp. 71–94). Cambridge, MA: MIT Press.

Wong, K., Wadee, F., Ellenblum, G., & McCloskey, M. (2018). The devil's in the g-tails: Deficient letter-shape knowledge and awareness despite massive visual experience. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/xhp0000532

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*(7), 1073–1092. https://doi.org/10.1037/xlm0000363