Failing to Predict Future Changes in Memory: A Stability Bias Yields Long-Term Overconfidence

Nate Kornell
Williams College

Author Note

Nate Kornell, Department of Psychology, Williams College.

Thanks to Robert Bjork and Matt Hays for their comments on a draft of this chapter. Thanks also to Elizabeth Bjork, Lindsey Richland, and Sara Appleton-Knapp.

Correspondence concerning this article should be addressed to Nate Kornell, Department of Psychology, Williams College, Williamstown, MA, 01267. E-mail: nkornell@gmail.com.

This chapter is in press. It will be published in: *Successful remembering and successful forgetting: a Festschrift in honor of Robert A. Bjork.*

If you have ever experienced that panicky moment when you realize, while taking a shower in Los Angeles, that you are supposed to be in Seattle telling 75 business executives about their memories, you may have learned something about the instability of memory (Robert A. Bjork, personal communication, 11/15/2009). It is easy to forget travel plans that once seemed memorable, and it is easy to add embarrassing experiences to one's memory.

Human memory is anything but stable. We constantly forget old information and form new memories. Yet recent research has demonstrated a *stability bias* in human memory: People act as though their memories will remain stable in the future. They fail to predict future forgetting (Koriat, Bjork, Sheffer, & Bar, 2004) and future learning (Kornell & Bjork, 2009). In this chapter, I discuss the importance of assessing one's memory in everyday life, draw a distinction between predicting future remembering versus predicting future changes in remembering, and review evidence substantiating the stability bias. I then describe an experiment examining the cause of the stability bias. I asked participants (n = 430) to predict their ability to remember word pairs they would study once or four times and would be tested on in 5 minutes or one week. Participants predicted significant learning and forgetting, but vastly under-predicted both effects, demonstrating a stability bias. Asking participants to imagine the test situation had little or no effect. The results demonstrated long-term overconfidence: Relatively modest immediate overconfidence transformed into enormous overconfidence as the test delay increased.

The Importance of Metacognition

Metacognition—that is, our judgments and beliefs about memory—helps us regulate our cognitive processes. It is common, for instance, to withhold or modify statements based on uncertainty (Goldsmith, Koriat, & Weinberg-Eliezer, 2002). For example, "Duke's basketball coach is Mike Krzyzewski" conveys more confidence than "Duke's basketball coach is Mike Krzyzewski, or whatever." The phrase "or whatever" may not seem very sophisticated, but it signals that the speaker is unsure of the coach's name. It is another way of saying "I'm not sure" (which monkeys can also do; Kornell, 2009a; Smith & Washburn, 2005). In many cases, the act of assessing one's memories is almost as vital as the act of retrieval itself.

Memory assessments also help us manage our learning and memory, which is particularly important for students (Kornell & Bjork, 2007; Thiede, Anderson, & Therriault 2003). Students decide what to study and how much time to spend studying based on these assessments (Nelson, Dunlosky, Graf, & Narens, 1994; Son & Metcalfe, 2000). Assessing one's memory inaccurately can lead to ineffective study behaviors, such as studying too little, not studying the information most in need of attention, or studying inefficiently (Benjamin, Bjork, & Schwartz, 1998; Kornell & Metcalfe, 2006).

Beliefs about how memory works are also a kind of metacognition. Such beliefs affect all sorts of everyday situations; for example, people write shopping lists because they believe they will forget something otherwise. Students rely on metacognitive beliefs when they decide how to study (e.g., should I test myself, should I make flashcards, should I underline, etc.; see Karpicke, 2009; Kornell, 2009b; Kornell & Son, 2009).

Judgments Based on the Past Versus Judgments Based on the Future

Memory monitoring has been researched extensively in recent years (see Metcalfe & Shimamura, 1994; Dunlosky & Bjork, 2008). In the most common kind of memory-

monitoring experiment, participants study an item and then judge how likely they are to remember it on a later test (e.g., Dunlosky & Nelson, 1992). Such a judgment, which is called a *judgment of learning*, can be made based on cues associated with one's current memory; the stronger the memory, the more likely the item is to be recalled.

A judgment of learning is a prediction of one's ability to remember in the future. Predicting how one's ability to remember will *change* in the future has received much less research attention (see Koriat et al., 2004; Kornell & Bjork, 2009; Tiede & Leboe, 2009). Memories change, for example, as people forget over time. Memory can also improve in the future as a result of time spent studying.

Predicting future changes in one's memory offers potential advantages. For example, predicting future forgetting is a way to avoid overestimating one's ability to remember in the future. Predicting how much one will learn by studying in the future is probably valuable for students planning future study activities. Research reviewed in the next section suggests that predictions of future *changes* in remembering are less accurate than predictions of future remembering, however, because people act as though their memories will be stable in the future.

A Stability Bias in Human Memory

A wealth of evidence suggests that people cannot directly assess the strength of their memories. Instead, metamemory judgments are made based on inferences (Schwartz, Benjamin, & Bjork, 1997). These inferences are, in turn, based on cues. Koriat (1997) proposed that there are three categories of cues that underlie metamemory judgments. To illustrate, suppose a student is studying Chapter 7 of an introductory psychology textbook. Characteristics of the chapter itself, such as the lucidity of the writing, are *intrinsic* cues. Features of the students' interactions with the materials, such as the amount of time he or she spends studying, are *extrinsic* cues. The student's internal experiences with the material, such as the fluency with which he or she answers a practice quiz, are *mnemonic* cues.

People seem to rely heavily on intrinsic cues (e.g., Rhodes & Castel, 2008). They rely on mnemonic cues once they have developed such cues (e.g., Finn & Metcalfe, 2008). They tend to undervalue extrinsic cues (e.g., Koriat & Bjork, 2005; Carroll, Nelson, & Kirwan, 1997).

Future interactions with learning material mainly fall into the category of extrinsic cues. For example, the delay between a study trial and a test trial is an extrinsic cue; so is the number of times one will be allowed to study. The other two types of cues are less relevant to predictions of future learning: Mnemonic cues, which have to do with one's past interactions with the materials, are backward-looking, and intrinsic cues have to do with the learning material itself (e.g., a textbook chapter); neither will change in the future. This reasoning suggests that people will be stuck relying on extrinsic cues when making predictions of future learning. If people tend to undervalue extrinsic cues, they will tend to under-predict the ways their memories will change in the future. They should, for example, underestimate future forgetting.

Predicting Future Forgetting

Koriat et al. (2004) investigated people's sensitivity to the retention interval between a study session and a test. In a typical experiment, separate groups of participants were asked to predict how many word pairs they would remember on a test that would occur right away, in one day, or in one week. The groups' predictions were

essentially identical. Of course, their actual recall performance decreased dramatically as delay increased. In one experiment, predictions for a test that would occur in one *year* were basically the same as predictions for an immediate test. These results are evidence of a stability bias in human memory.

Koriat et al.'s (2004) participants surely knew that they were prone to forgetting, but they did not appear to apply that belief when making predictions. Koriat et al. distinguished between *theory-based* judgments, which are based on one's beliefs about memory, and *experience-based* judgments, which are based on the learning experience itself. Their participants did employ experience-based judgments—they predicted higher rates of recall for easier items—but they did not seem to apply theory-based judgments.

When Koriat et al. (2004) took steps to make the concept of forgetting salient, their participants began to apply their theory-based judgments. In within-participant experiments, where the same person was asked about multiple different intervals, the predictions became more accurate. The predictions also became accurate when participants were asked to predict how much they would forget rather than how much they would remember.

Predicting Future Learning

Along with gradual forgetting over time, one of the most obvious aspects of memory is that people learn by studying. If people under-predict future forgetting, do they also under-predict future learning? A set of studies by Kornell and Bjork (2009) suggests that the answer is yes. Participants were asked to predict how well they would do on a test that would occur after they had studied a word pair between one and four times. Although memory performance increased dramatically across trials, predicted performance increased very little. Unlike Koriat et al. (2004), Kornell and Bjork found that even predictions made on a within-participant basis revealed a strong stability bias. Again, these participants surely realized that they learned by studying (otherwise, why would the study?) but they did not appear to apply that belief.

Kornell and Bjork's (2009) participants made their predictions during the first study trial. Thus, a prediction for the first test did not require assessing future changes in memory—it was the same as a traditional judgment of learning. The test-one predictions were generally overconfident. This finding underscores the importance of distinguishing between judgments of learning and predictions of future learning; participants were simultaneously overconfident in their judgments of learning and underconfident in their future learning ability.

Possible Explanations of the Stability Bias

Why is there a stability bias? In the experiment presented below, I tested two hypotheses. One is that extrinsic cues are overshadowed by other cues (see Tiede & Leboe, 2009; Koriat, Sheffer, & Ma'ayan, 2002). In the experiments described above, the relatedness of the word pairs, an intrinsic cue, was highly salient. Its salience may have led people to focus on item difficulty and ignore extrinsic cues such as the number of study trials. If this explanation is correct, minimizing the salience of intrinsic cues should diminish the stability bias. In the present experiment, I minimized the influence of intrinsic cues in two ways: by using homogeneously difficult items and by presenting only one sample item per participant.

Another reason why people fall victim to the stability bias may be a failure of perspective taking. Seeing things from another person's perspective can be difficult for

adults, to say nothing of children or animals. It can even be difficult to take one's own perspective. For example, difficult questions often seem easy as soon as one knows the answer, a phenomenon known as the hindsight bias (Fischhoff, 1975). It is possible that in the experiments described above, participants made predictions from the perspective of the studier (which is what they were at the moment of the prediction) rather than from the perspective of the test taker (i.e., themselves in the future). If, for example, participants had imagined themselves a week in the future, taking the test, they might have been more sensitive to retention interval and less subject to the stability bias.

Experiment Overview

The central variable in the present experiment included three conditions. In the *baseline* condition, participants were told they would study a set of word pairs once and then take a test five minutes later; in the *extra study* condition they were told there would be four study trials; in the *extra delay* condition they were told the test would take place one week later.

These three conditions were compared in seven different ways, making a total of 21 between-participant conditions. The first of these seven sets of conditions was the three *actual recall* conditions, in which participants' memories were actually tested. In the other six sets of conditions (i.e., the other 18 conditions) participants predicted how they would do if their memories were tested. The six sets of conditions included the no perspective condition, the perspective condition, and four items-diminished conditions. In the *no perspective* condition, participants predicted how many items they would remember based on a description of the materials and procedure in an actual recall condition. The *perspective* condition was the same except that participants were asked, when making their prediction, to imagine themselves at the time of the test. The *items-diminished* conditions were the same as the perspective condition; the difference was that the influence of item-difficulty was diminished by presenting either homogeneously difficult pairs (which were all easy or all hard) or by presenting only one sample pair (which was either easy or hard).

The experiment was designed to answer six questions. 1) How would participants do in the actual recall conditions? 2) Were predictions accurate in the no perspective condition, which served as the baseline prediction condition? 3) Did imagining a test-taker's perspective increase prediction accuracy? 4) Did diminishing the influence of item difficulty increase prediction accuracy? 5) Overall, did participants suffer from a stability bias? 6) Were older adults less susceptible to the stability bias than younger adults?

This chapter does not present the entire methodology at once. Instead, the presentation of the experiment is organized around the six questions above. After the participants are described, the methods and results that answer each question are presented one at a time.

Participants

There were 430 participants. The number of participants in each condition is displayed in Table 1. Participants were recruited using Amazon's Mechanical Turk, a website that serves as a micro-task market, connecting employers to workers. Workers voluntarily sign up to do small tasks for pay (a typical job takes less than 5 minutes) and employers post jobs that they need done (see Kittur, Chi, & Suh, 2008).

Table 1
The number of participants in each experimental condition.

Condition	Perspective	Items Shown[a]	Difficulty	Extra Study	Baseline	Extra Delay	Total
			Actual Recall				
Actual	No	24	Mixed	8	8	8	24
			Predicted Recall				
No Perspective	No	24	Mixed	29	28	29	86
Perspective	Yes	24	Mixed	22	20	16	58
Items-Diminished	Yes	1 or 24	Easy or Difficult	86	96	80	262
			Items-Diminished (Uncombined)				
Easy 24	Yes	24	Easy	22	27	19	68
Difficult 24	Yes	24	Difficult	21	23	24	68
Easy 1	Yes	1	Easy	24	23	16	63
Difficult 1	Yes	1	Difficult	19	23	21	63

[a] All participants were informed that there would be 24 items; "items shown" refers to the number of sample items participants were shown.

Participants in the prediction conditions were paid 30 cents each for a task that generally took less than 3 minutes to complete. Participants in the actual memory conditions, which took longer to complete and involved two sessions, were paid $1.50 for completing the first session and $2.00 for completing the second session.

The sample was diverse with respect to age, location, and educational background. The mean age was 34 (median 31), with a standard deviation of 12 years and a range of 18-74 years. Eighty percent of the participants reported that they lived in the United States of America, 10% lived in India, 3% lived in Canada, 2% lived in the Philippines, and the 24 remaining participants came from 14 additional countries. (Participants were asked if they spoke English fluently; they were excluded from the dataset if they said no.) Education level ranged from some high school (3%) through high school graduate (36%), bachelors or associates degree (46%), and graduate degree (16%). Fifty-nine percent of the participants were female.

Actual Recall Performance

Method

As mentioned above, the central variable, which I refer to as future condition, included three conditions: study once and take a test five minutes later (baseline), study four times and take a test five minutes later (extra study), and study once and take a test in one week (extra delay).

In the actual recall conditions, participants completed a memory experiment without making predictions. After reading instructions describing the experiment, they studied 24 word pairs, which were split evenly into 12 easy pairs (e.g., Jelly-Bread, Usurp-Take) and 12 hard pairs (e.g., Figment-Satire, Criterion-Attitude). The pairs were presented one at a time for three seconds each. In the extra study condition, the same list of pairs was then presented three more times in a different random order each time.

After the study phase, all participants completed a five-minute task in which they were asked to recall the names of as many countries as they could. Following the country-naming distractor task, participants in the baseline and extra study conditions were tested on the 24 items; each cue word was presented, one at a time, and participants were asked to type in the target word. Participants in the extra delay condition were not tested during Session 1.

One week after Session 1, all participants were emailed and asked to participate in Session 2. Participants who did not complete the second session were excluded in all conditions. The test in Session 2 was the same as the test in Session 1.

Results and Discussion

The results are displayed in Figure 1. Compared to the baseline condition, recall accuracy was significantly lower in the extra delay condition, $t(14) = 5.52$, $p < .0001$, $d = 2.76$, and significantly higher in the extra study condition, $t(14) = 4.27$, $p < .0001$, $d = 2.14$. (All t-tests comparing the extra delay condition or the extra study condition to the baseline condition were one-tailed.) In other words, as expected, participants learned from additional studying and forgot over the course of a weeklong delay.

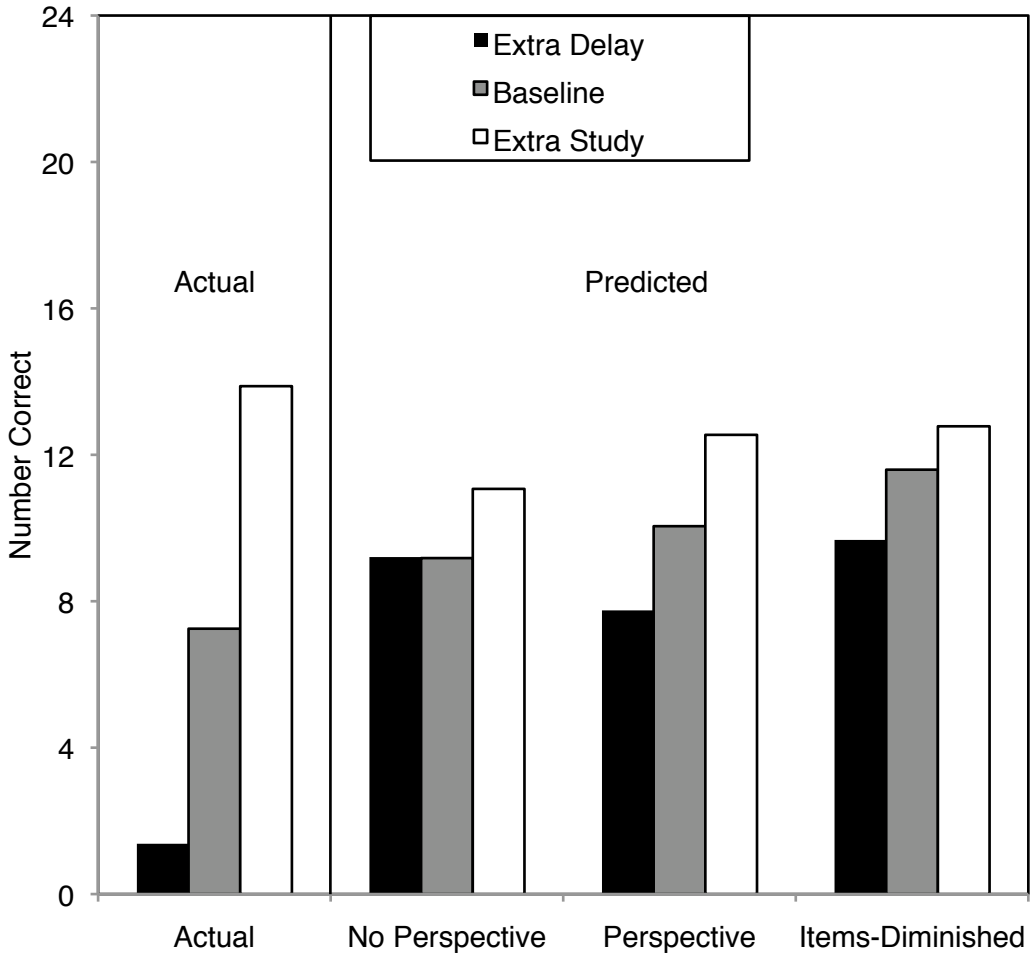

Figure 1. Number of items correct (out of 24) as a function of future condition. The three left-most bars are actual recall performance. The other nine bars are all predicted performance. The items-diminished conditions were created by collapsing the conditions in the Figure 2.

Were Predictions Accurate in the No Perspective Condition?

Method

As mentioned above, 18 groups made predictions. Each group read a description of one of the conditions in the actual memory experiment just described (i.e., the baseline, extra study, or extra delay condition). They were asked to predict how many of 24 items they would remember.

The no perspective condition served as the baseline prediction condition. It was similar to the prediction conditions in the research conducted by Koriat et al. (2004) and Kornell and Bjork (2009). Participants were asked to read a description of the actual memory experiment. In particular, they were told that 24 word pairs were studied for three seconds each; they were asked to read through the 24 mixed-difficulty word pairs used in the actual recall conditions; and the nature of the final cued-recall test was described. Separate groups made predictions for each of the three future conditions; depending on the participant's condition, they were told that the experiment involved one or four study trials and that the test took place after five minutes or one week. The information about the number of study trials and the test was then presented a second time in an alternative format, and then participants were asked to predict how many of the 24 items they would have recalled if they had participated in the experiment. The directions were designed to make the manipulation of study trials and test delay prominent and clear.

Results and Discussion

The results are displayed in Figure 1. Predicted recall in the extra delay condition was almost identical to predicted recall in the baseline condition. Predicted recall in the extra study condition was higher than baseline, but the difference was only marginally significant, $t(55) = 1.40$, $p = .08$, $d = .37$.

An ANOVA comparing predicted and actual accuracy confirmed that there was a significant interaction between future condition and recall task (predicted versus actual), $F(2, 104) = 8.71$, $p < .001$, $\eta_p^2 = .14$. That is, the predictions underestimated actual learning and forgetting, demonstrating a stability bias.

Did Imagining a Test-Taker's Perspective Increase Prediction Accuracy?

The predictions described above occurred at the time of study. The predictions would have been more accurate if they had occurred after participants had either completed all four study trials or waited out the weeklong delay—that is, at the time of the test (Kornell & Bjork, 2009; see also Carroll et al., 1997). Asking participants to imagine their perspective at the time of the test might also diminish or eliminate the stability bias.

Method

The perspective conditions were the same as the no perspective conditions with one exception: To encourage perspective taking, three sentences were added to the instructions immediately before participants were asked to make their prediction. The sentences in the extra delay condition, for example, read: "Now we want you to imagine that you are participating in the experiment. Imagine that you are about to start the test. You studied the word pairs once each 1 week ago."

Results and Discussion

The results are displayed in Figure 1. Compared to the baseline condition, predictions were significantly higher in the extra study condition, $t(40) = 1.97$, $p < .05$, d

= .61, and lower in the extra delay condition, $t(34) = 1.86$, $p < .05$, $d = .63$. In other words, participants predicted significant learning and forgetting when they were encouraged to take a test-taker's perspective.

Compared to actual recall, however, predicted recall still significantly underestimated both learning and forgetting. A Future Condition X Recall Measure (predicted versus actual) ANOVA produced a significant interaction, $F(2, 76) = 6.00$, $p < .01$, $\eta_p^2 = .14$.

I also compared predictions in the perspective condition to predictions in the no perspective condition. A 3x2 ANOVA examining the Future Condition x Perspective interaction did not yield a significant effect, $F(2, 138) = 1.25$, $p = .29$. That is, the magnitude of the stability bias was not significantly different in the perspective condition compared to the no perspective condition. Inspecting Figure 1 suggests that perspective taking primarily affected the extra delay condition, but even when the extra delay condition was compared to baseline (i.e., the extra study condition was excluded from the analysis), predictions were not significantly more sensitive to delay in the perspective condition than they were in the no perspective condition, $F(1, 89) = 1.62$, $p = .21$.

Koriat et al. (2004) also compared a condition in which participants were encouraged to take a test-taker's perspective to a condition in which they were not (see Experiment 6a). They found that the perspective instruction did not have a significant effect on predictions, consistent with the present results. They also found that predictions were not sensitive to retention interval in either case. In the present experiment, by contrast, predictions were sensitive to retention interval, and number of study trials, when they were made from the test-taker's perspective.

In summary, encouraging participants to take a test-taker's perspective did not eliminate the stability bias. The stability bias remained robust. Nevertheless, participants did predict significant amounts of both learning and forgetting. No previous between-participant experiment has shown significant effects of number of learning trials (Kornell & Bjork, 2009) or retention interval (Koriat et al., 2004) on predictions of future remembering—including the no perspective condition of the present experiment. These results seem to suggest that a failure of perspective taking contributes to the stability bias. This conclusion is tentative, however, because the apparent effect of perspective taking on the stability bias was not significant.

Did Diminishing the Influence of Item Difficulty Increase Prediction Accuracy?

The items people study often have a powerful influence on metacognitive judgments (Koriat, 1997). Highly salient item characteristics, such as the relatedness of word pairs, might overshadow extrinsic cues that would otherwise influence predictions, such as delay before a test. The last four sets of conditions were designed to reduce item difficulty's influence on predictions. (To foreshadow, these four conditions were collapsed into the items-diminished condition in later analyses.)

Method

Except for the sample stimuli, the last four conditions were all identical to the perspective condition. In one of the four sets of conditions, all 24 sample-items were relatively easy; in another they were all relatively difficult. I expected item difficulty to become less salient, and thus impact predictions less, when it was homogenous. In the other two sets of conditions, participants were told that the experiment involved learning 24 word pairs, but they were only shown one sample pair, which was either easy or hard.

Displaying only one item was intended prevent any comparison between items, further diminishing item difficulty's influence on predictions. Thus there were four sets of conditions: 24 easy, 24 hard, 1 easy, and 1 hard.

Results and Discussion

Number and difficulty of items. Before examining the stability bias, it is worth comparing the four sets of conditions to each other. The results are displayed in Figure 2. There was a main effect of future condition, $F(2, 250) = 5.00$, $p < .01$, $\eta_p^2 = .04$. (More on this effect below.) Predictions were higher for easy than hard items, even on a between-participants basis, $F(1, 250) = 20.46$, $p < .0001$, $\eta_p^2 = .08$. Predictions were also higher when one item was presented than when 24 items were presented, $F(1, 250) = 16.33$, $p < .0001$, $\eta_p^2 = .06$. (It is possible that seeing all 24 items made participants consider the effect of list length and that led them to predict lower performance.)

There were no significant interactions between item difficulty, number of items, and future condition (all Fs < 1). Thus, the four sets of conditions designed to minimize item differences were collapsed in further analyses. The collapsed conditions will be referred to as the items-diminished condition.

Stability bias. Figure 1 displays the collapsed data for the items-diminished condition. Predicted recall was lower in the extra delay condition than it was in the baseline condition, $t(174) = 2.09$, $p < .01$, $d = .32$. The difference between the baseline condition and the extra study condition was marginally significant, $t(180) = 1.42$, $p = .08$, $d = .21$.

Predicted recall underestimated actual learning and forgetting. A Future Condition x 2 Recall Measure (predicted versus actual) ANOVA produced a significant interaction, $F(2, 280) = 5.14$, $p < .01$, $\eta_p^2 = .04$.

The items-diminished conditions were compared to the perspective conditions (in which 24 mixed-difficulty sample pairs were displayed). The perspective versus items-diminished variable did not interact with future condition ($F < 1$). If anything, contrary to the hypothesis, the stability bias was larger in the items-diminished condition.

These findings are consistent with Experiment 4a by Koriat et al. (2004). In that experiment, instead of asking participants to predict based on all of the items they were going to study, Koriat et al. showed only two sample items, one easy and one hard. They, too, found that the stability bias persisted.

In summary, minimizing item differences did not diminish the stability bias. The hypothesis that intrinsic cues overshadow extrinsic cues was not supported. Although the stability bias was robust, there was a significant effect of test delay and a marginally significant effect of number of study trials.

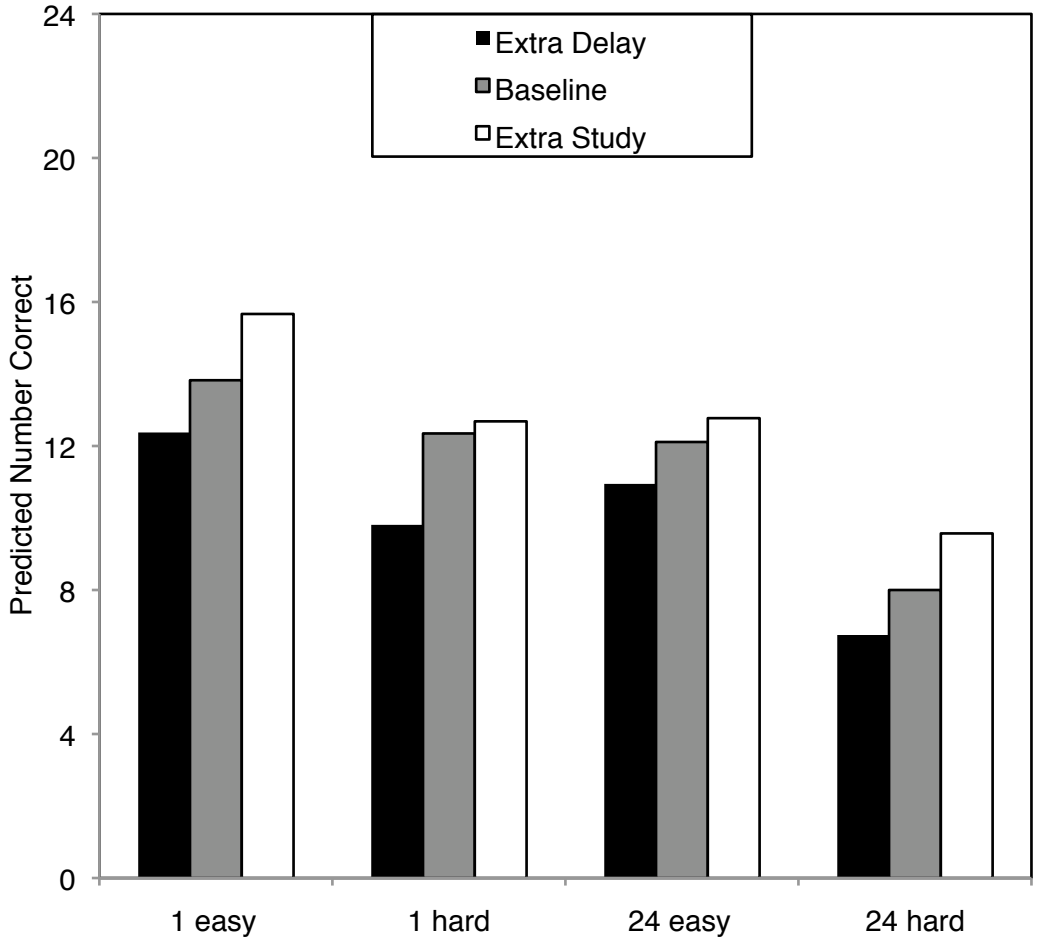

Figure 2. Predicted number of items correct (out of 24). The sample items were either all easy or all hard, and the number of sample items presented was either 1 or 24. In all conditions participants were told they would study and be tested on 24 items. These four sets of conditions were collapsed to create the items-diminished condition (see Figure 1).

Overall, Did Participants Suffer from a Stability Bias?

In a final set of analyses, all of the prediction conditions in Figure 1 were compared. Prediction condition (no perspective, perspective, items-diminished) did not significantly influence the effect of future condition (i.e., there was not a significant interaction, $F < 1$). Future condition did have significant effects, however. Predicted recall was higher in the extra study condition than the baseline condition, $t(279) = 2.30$, $p < .05$, $d = .27$. Predicted recall was lower in the extra delay condition than the baseline condition, $t(267) = 2.35$, $p < .01$, $d = .29$.

Although predicted recall was sensitive to the number of trials and test delay, the predictions still showed clear evidence of a stability bias. Future condition influenced actual recall much more than it influenced predicted recall, as a significant Future Condition X Recall Measure (actual, predicted) ANOVA demonstrated, $F(2, 424) = 5.95$, $p < .01$, $\eta_p^2 = .03$. Overall, predicted recall was overconfident compared to actual recall, $F(1, 424) = 9.08$, $p < .01$, $\eta_p^2 = .02$. As expected, there was a main effect of future condition, $F(2, 424) = 16.12$, $p < .0001$, $\eta_p^2 = .07$.

Were Older Adults Less Susceptible to the Stability Bias?

As people age, they may become more sensitive to their own forgetting and, more generally, to how their memories change over time. If so, the stability bias might become less pronounced as people age. Participants were median split into a relatively young group (range 18-31; mean = 25) and a relatively old group (range 32-74; mean = 44), and collapsed across all predictions conditions. There were 203 participants per age group. As Figure 3 shows, the older group showed less stability bias than the younger group. However, the effect (i.e., the interaction between age group and future condition) was not significant, $F(2, 400) = .83$, $p = .44$.

The young participants predicted higher recall levels than the older participants, $F(1, 400) = 5.19$, $p < .05$, $\eta_p^2 = .01$. Actual recall performance was almost identical between the age groups, however, and age group did not interact with future condition ($Fs < 1$). In short, the stability bias appeared to be less pronounced in older participants, but the effect was not significant.

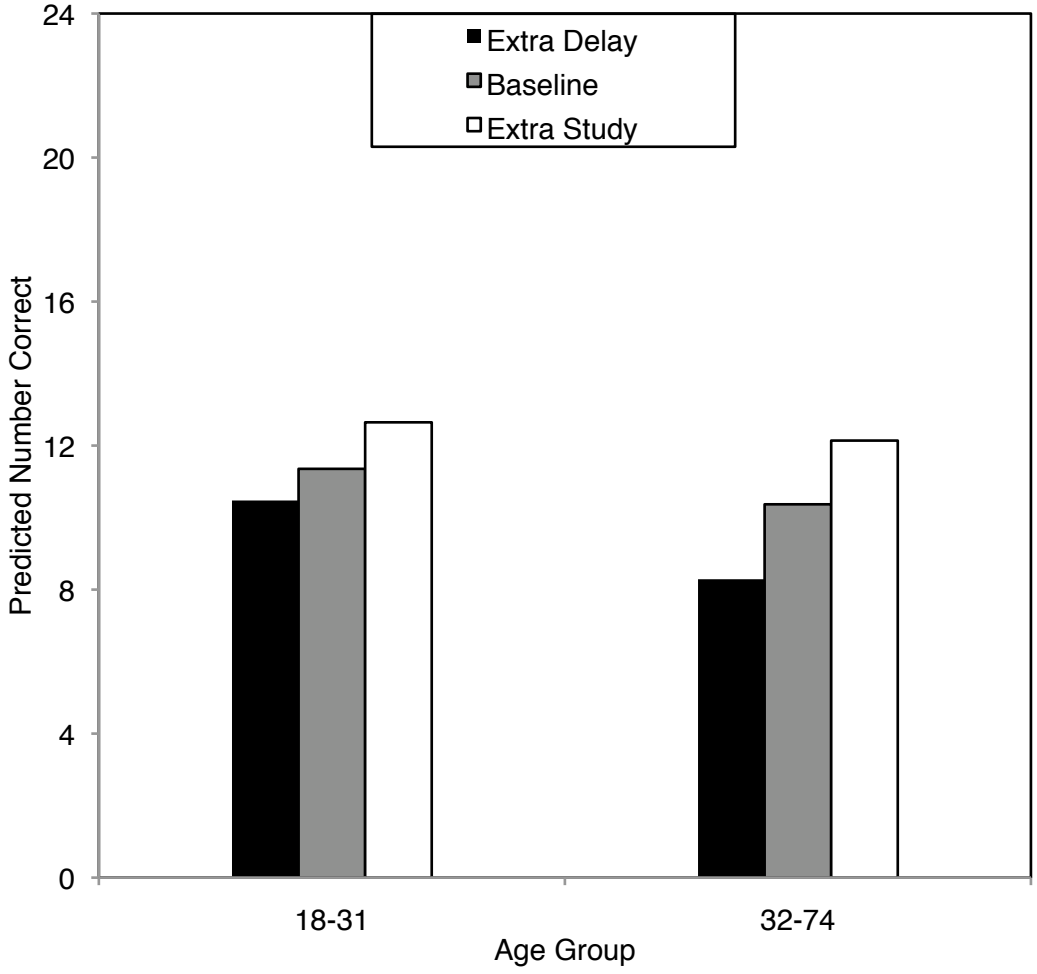

Figure 3. Predicted number of items correct (out of 24), collapsed across all prediction conditions, for participants who were relatively young (mean age = 25) and relatively old (mean age = 44).

General Discussion

The experiment reported here demonstrated a stability bias: Participants underestimated their learning ability, replicating Kornell and Bjork (2009), and they underestimated their propensity to forget, replicating Koriat et al. (2004). Encouraging participants to take a test-taker's perspective did not significantly increase prediction accuracy, nor did diminishing the salience of item differences.

Despite the robust stability bias, participants were not completely insensitive to learning and forgetting. Compared to the baseline condition, predictions were significantly higher in the extra study condition and significantly lower in the extra delay condition. In previous between-participant experiments, predictions have not been significantly affected by delay (Koriat et al., 2004) or number of study trials (Kornell & Bjork, 2009). This finding suggests that imagining a test-takers perspective may have had some effect in the present experiment. The effect was not statistically significant, however.

Long-Term Overconfidence

People are frequently overconfident in their memories (Metcalfe, 1998; Fischhoff, Slovic, & Lichtenstein, 1977) and the current experiment was no exception. Examining Figure 1, though, makes clear what the largest source of the overconfidence was: forgetting. When all prediction conditions were collapsed, the average predictions in the extra study, baseline, and extra delay conditions were, 12.4, 10.9, and 9.3, respectively. The actual accuracy scores were 13.9, 7.3, and 1.4. I computed an overconfidence score by subtracting predicted accuracy from actual accuracy and dividing by the standard deviation of actual accuracy. Extra study participants were underconfident by .5 standard deviations. Baseline participants were overconfident by 1.3 standard deviations—a fairly large amount of overconfidence by metacognition standards. The delayed group, however, was overconfident by 8.7 standard deviations.

These findings suggest it is important to distinguish between short-term overconfidence and long-term overconfidence. The phrase long-term overconfidence refers to a feeling of confidence that one will be able to retrieve a memory in the relatively long-term future. (It should be distinguished from feelings of confidence at the time of retrieval.) As Figure 4 illustrates, when actual forgetting outstrips predicted forgetting, small amounts of short-term overconfidence can grow into large amounts of long-term overconfidence.

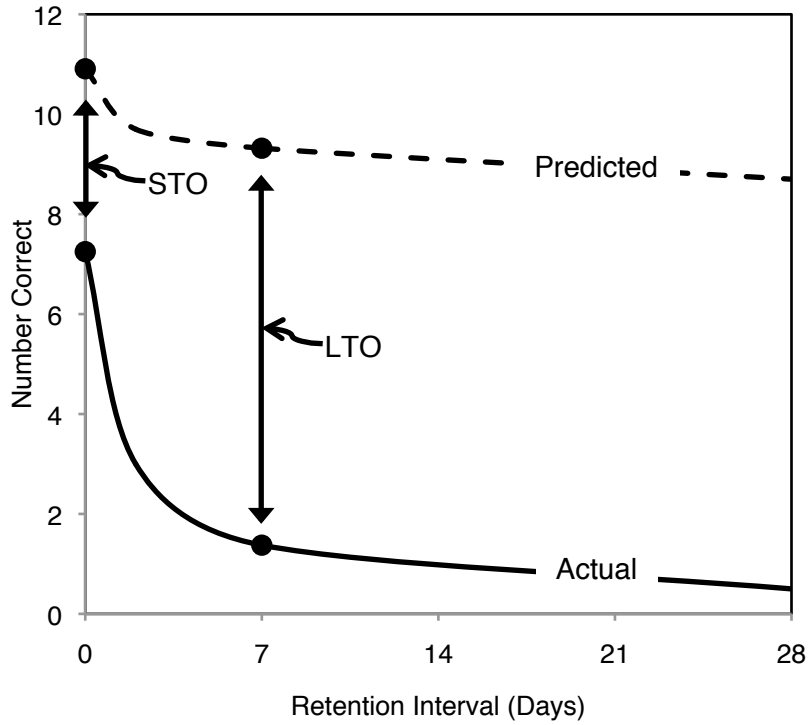

Figure 4. Hypothetical changes in overconfidence as a function of time. The solid line represents actual recall; the dashed line represents predicted recall. The solid circles are actual data points computed by averaging across conditions in the present experiment. The two vertical lines represent short-term overconfidence (STO) and long-term overconfidence (LTO).

In real life, we often make confidence judgments about our ability to retrieve information in the relatively distant future. For example, if a student reads a textbook three weeks before a cumulative exam, and then makes a judgment that he or she will remember the information on the exam, that judgment may be subject to long-term overconfidence. The same is true when we judge our ability to remember new information at an undetermined future time, such as a new acquaintance's name or an interesting fact from a news article.[1] In these situations, people may be highly overconfident about their ability to remember in the future. More broadly, long-term overconfidence may be a more ecologically valid measure than short-term overconfidence.

Most metacognition experiments (e.g., experiments on judgments of learning) take place in one-hour sessions. These experiments measure short-term overconfidence. The metacognition literature may systematically underestimate long-term overconfidence.

Reinterpreting Judgments of Learning Based on the Stability Bias

In the experiment reported here, participants were asked to make predictions for a delayed test. Other researchers, including Koriat et al. (2004), have asked for similar predictions. For example, Roediger and Karpicke (2006, Experiment 2) tested their participants either immediately after they studied or a week later. Consistent with the stability bias, the predictions were not affected by retention interval. The participants were asked to learn passages, and those given four chances to read the passage (SSSS) predicted they would do better than those who read the passage once and then took three tests (STTT). (Predictions were intermediate in a third SSST condition.) In the delayed test condition, these predictions were essentially backwards, because free-recall accuracy was highest in the STTT condition and lowest in the SSSS condition.

Roediger and Karpicke's (2006) findings can be interpreted as meaning that people underestimate the value of tests. But they look different in light of the stability bias, which causes people to make judgments based on the current state of their memories. Relative to the current state of their memories, Roediger and Karpicke's participants made accurate predictions, because their immediate test performance, like their predictions, was highest in the SSSS condition and lowest in the STTT condition. Thus, perhaps these participants' mistake was not misunderstanding the benefits of testing. Their mistake was more basic: They judged their future remembering based on their current memory state. Of course, the outcome was the same: Participants rated testing, the most effective long-term strategy, as least effective.

In general, in situations where memory changes over time, the stability bias will tend to make people's predictions inaccurate. This hypothesis may help explain Roediger and Karpicke's (2006) findings as well as other research involving predictions of long-term learning (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Carroll et al., 1997; Karpicke & Roediger, 2008; Shaddock & Carroll, 1997).

Conclusion

How is it possible that people act as though their memories will not change in the future? We all know we learn by studying and we forget over time. And in some circumstances we act on that knowledge; for example, everyone knows it is important to record speaking engagements in a calendar—otherwise, one might forget to fly to Seattle

to give a lecture about memory. Yet unless special measures are taken to make learning or forgetting salient, people consistently demonstrate a stability bias.

The situation is reminiscent of another failure to anticipate future events: the planning fallacy.[2] As anyone who has had their kitchen remodeled knows, people generally underestimate how long it will take to complete complex tasks (Kruger & Evans, 2004). The stability bias and the planning fallacy have a number of features in common. They concern predictions about future events. They seem to be associated with a failure to take past events (e.g., forgotten names, late papers) into consideration when making predictions about the future. Perhaps most striking, knowledge does not eliminate the planning fallacy: Even psychologists who have personally experienced the planning fallacy every time they have written an article or prepared a lecture—even an article or lecture *about* the planning fallacy—cannot seem to overcome it.[3] Similarly, knowledge does not seem to eliminate the stability bias. We all know about forgetting and learning, but we often act as though we don't.

References

Agarwal, P., Karpicke, J., Kang, S., Roediger, H., & McDermott, K. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861-876.

Benjamin, A. S, Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55-68.

Carroll, M., Nelson, T., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Psychologica, 95*, 239-253.

Dunlosky, J., & Bjork, R.A. (Eds.). (2008). *A handbook of metamemory and memory*. Hillsdale, NJ: Psychology Press.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed JOL effect. *Memory & Cognition, 20*, 374-380.

Finn, B., & Metcalfe, J. (2008). Judgments of Learning are influenced by Memory for Past Test. *Journal of Memory and Language, 58*, 19-34.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effects of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 288–299.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 552-564.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General, 131*, 73-95.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: an eternal Golden Braid.* Basic Books, New York.

Karpicke, J. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469-486.

Kittur, A., Chi, E., & Suh, B. (2008). Crowdsourcing User Studies With Mechanical Turk. *CHI 2008: Proceedings of the ACM Conference on Human-factors in Computing Systems*. New York: ACM Press.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966-968.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349-370.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 187–194.

Koriat, A., Bjork, R. A., Sheffer, L, & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643-656.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131,* 147-162.

Kornell, N. (2009a). Metacognition in humans and animals. *Current Directions in Psychological Science, 18*, 11-15.

Kornell, N. (2009b). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297-1317.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219-224.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449-468.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 32,* 609-622.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*, 493-501.

Kruger, J., & Evans, M. (2004). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology, 40*, 586-598.

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality & Social Psychology Review, 2*, 100-110.

Metcalfe, J., & Shimamura, A.P. (Eds.). (1994). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5,* 207–213.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615-625.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249-255.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential basis of metamemory. *Current Directions in Psychological Science*, *6*, 132–137.

Shaddock, A., & Carroll, M. (1997). Influences on metamemory judgements. *Australian Journal of Psychology, 49*, 21-27.

Smith, J. D., & Washburn, D. A. (2005). Uncertainty monitoring and metacognition by animals. *Current Directions in Psychological Science, 14*, 19–24.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 204–221.

Thiede, K. W. & Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66-73.

Tiede, H., & Leboe, J. (2009). Metamemory judgments and the benefits of repeated study: Improving recall predictions through the activation of appropriate

knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 822-828.

Footnotes

[1] In real life, when people predict whether they will be able to remember something in the future, the exact time of the future retrieval is often, perhaps usually, uncertain. It may be that we don't take into account when we will be tested because we usually don't know when we will be tested. Similarly, perhaps we don't take into account how much time we will spend studying in the future because often we do not know that either. This reasoning might explain one of the most puzzling questions about the stability bias: When predicting future remembering, why do people ignore their own beliefs about how memory works? Under natural circumstances, perhaps people do not apply their beliefs because future study trials and retention intervals are unknown quantities.

[2] Thanks to Matt Rhodes for pointing this out.

[3] This situation evokes Hofstadter's Law: It always takes longer than you expect, even when you take into account Hofstadter's Law (Hofstadter, 1979). Thanks to Jim Kornell for pointing out this similarity.