Mem Cogn (2017) 45:1270–1280
DOI 10.3758/s13421-017-0740-1

Performance bias: Why judgments of learning are not affected by learning

Nate Kornell[1] · **Hannah Hausman**[2]

Published online: 24 July 2017
© Psychonomic Society, Inc. 2017

Abstract Past research has shown a performance bias: People expect their future performance level on a task to match their current performance level, even when there are good reasons to expect future performance to differ from current performance. One explanation of this bias is that judgments are controlled by what learners can observe, and while current performance is usually observable, changes in performance (i.e., learning or forgetting) are not. This explanation makes a prediction that we tested here: If learning becomes observable, it should begin to affect judgments. In three experiments, after practicing a skill, participants estimated how they performed in the past and how they expected to perform in the future. In Experiments 1 and 2, participants knew they had been improving, as shown by their responses, yet they did not predict that they would improve in the future. This finding was particularly striking because (a) they did improve in the future and (b) as Experiment 3 showed, they did hold the conscious belief that past improvement predicted future improvement. In short, when learning and performance are both observable, judgments of learning seem to be guided by performance and not learning.

Keywords Skill acquisition · Metacognition · Judgment · Memory

✉ Nate Kornell
nkornell@gmail.com

[1] Department of Psychology, Williams College, 880 Main St, Williamstown, MA 01267, USA

[2] Department of Psychology, Colorado State University, Fort Collins, CO, USA

There is a difference between performance and learning, just as there is a difference between position and speed. For instance, when a person takes a test, performance is a function of their level of knowledge. Learning is how much their knowledge changes as a result of taking the test. Performance and learning sometimes diverge. This phenomenon can be seen in research on desirable difficulties: By making studying more difficult, it is possible to decrease performance during study and, at the same time, increase learning (E. L. Bjork & Bjork, 2011; R. A. Bjork, 1994).

In this article, we investigated the influences of learning and performance on judgments of learning (JOLs), which are predictions of how well one will do when tested in the future. First, a note on our use of the term *performance*. We define *objective performance* as how accurately and quickly one responds while studying or being tested. Objective performance can differ from *subjective performance*, which is the learner's internal feeling of how accurately they are doing a task and how easy or difficult it is. There are times when these two kinds of performance are similar, such as when someone is quizzing themselves using flashcards; objective and subjective estimates of accuracy will tend to be similar, and response time will be correlated with subjective difficulty. At other times they can diverge—for example, when one reads two pages of text, one simple and one complex. Subjective performance will be high and low, respectively, while objective performance might not differ, given that there is no accuracy to be measured and reading speed might not differ much. Moreover, objective and subjective performance can differ when subjective performance is wrong, such as when someone is overconfident. In this article, at times we specify that we are referring to objective or subjective performance, and at other times we use the unmodified term performance to refer to situations where either or both types of performance are relevant.

The primary goal of studying is learning, not performance. As people study, it behooves them to make accurate JOLs, and accurate JOLs entail being sensitive to how much one is learning. Research suggests, however, that a person's current level of subjective performance has an oversized influence on JOLs. The amount the person is learning, by contrast, seems to have little influence on JOLs. Again, desirable difficulties provide evidence for this claim: As one example of desirable difficulty, spacing or interleaving one's practice increases long-term learning as compared to massing practice (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), but massing increases subjective performance during practice (Rohrer & Taylor, 2007; Taylor & Rohrer, 2010). If a JOL is made while one is studying, it is possible to ask which guides JOLs, learning or performance, because the two are at odds: There is more learning in the spaced condition and better subjective performance in the massed condition. If JOLs were guided by learning, they would be higher following spaced or interleaved practice; instead, JOLs seem to be guided by subjective performance, given that people consistently rate massing as being more effective than spacing or interleaving (Dunlosky & Nelson, 1994; Kornell, 2009; Kornell & Bjork, 2008; Zechmeister & Shaughnessy, 1980; Zulkiply, McLean, Burt, & Bath, 2012).

Another example of the strong influence of current performance on JOLs comes from research on the memory for past test heuristic (Finn & Metcalfe, 2007, 2008). In these studies, participants are tested on an item and then restudy the item on subsequent trials. When they are asked to make a JOL after some number of subsequent trials, their JOL seems to be determined based on their performance the last time they took a test on that item (i.e., it is correlated with a previous experience of objective performance) more than by the subsequent study trials. Given that people see tests as more diagnostic of their knowledge than restudy (Kornell & Bjork, 2007; Kornell & Son, 2009), it seems that participants are making JOLs on the basis of their performance the last time they were tested and failing to be influenced by the learning that has happened on subsequent trials. (For the purpose of this article, we will consider a judgment based on memory for a past test to be a judgment of current subjective performance because we assume that they are judging how well they know the information now based on objective performance that occurred on a previous trial.)

In short, JOLs seem to be guided by what we will call a performance bias: People expect their future level of performance to match their current level of performance.[1] When making judgments about their memories, this means they expect their future knowledge to match their current knowledge. The performance bias is a more general form of another bias in the metacognitive literature, known as the stability bias: JOLs tend to be the same regardless of what participants expect to happen in the future. For example, JOLs are the same regardless of whether people are judging their ability recall items in 10 minutes or a week, but recall is not (Koriat, Bjork, Sheffer, & Bar, 2004). Similarly, JOLs are the same regardless of how many times people are told they will be allowed to study in the future, but memory accuracy is higher when people they are allowed to study more (Kornell & Bjork, 2009; Kornell, Rhodes, Castel, & Tauber, 2011). It is worth noting that participants in these studies did not lack the explicit belief that they forget over time and that they learn by studying. Indeed, when learning or forgetting was manipulated on a within-participant basis, making the manipulation more salient, participants made (more) accurate JOLs. In the between-participants studies, therefore, it appears that participants' error was not inaccurate beliefs but rather a failure to apply their beliefs when making judgments. (To foreshadow, we observed a similar phenomenon in the studies reported here.)

Thus, it seems that three disparate sets of findings can be explained by the performance bias: the memory for past test heuristic, the stability bias, and metacognitive errors under conditions of desirable difficulty. These sets of findings are different in an important way. In the case of desirable difficulty and memory for past test, performance bias results from looking backward and failing to be sensitive to the amount of learning that has already occurred; in the case of the stability bias it results from looking forward and failing to appreciate the importance of what will happen in the future. Given that performance bias is a single (albeit simple) mechanism that explains these three phenomena, it appears to be a fairly general metacognitive heuristic.

The cause of performance bias

In this study, we investigated what causes the performance bias. One important factor may be that, as E. L. Bjork and Bjork (2011) point out, performance is easier to observe than learning:

> Performance is what we can observe and measure during instruction or training. Learning—that is, the more or less permanent change in knowledge or understanding that is the target of instruction—is something we must try to infer, and current performance can be a highly unreliable index of whether learning has occurred. (p. 57)

Desirable difficulties, such as spaced practice, are an example of this unreliability: Subjective performance is not only

[1] The performance bias is not to be confused with the *performance heuristic*, a term introduced by Critcher and Rosenzweig (2014). We describe their research in the General Discussion, but, in short, both terms are about how well people expect to do in the future; but whereas we investigated the influence of past improvement on these expectations, Critcher and Rosenzweig investigated the influence of overall past performance on these expectations.

highly unreliable as a guide to learning, it is downright backward (e.g., Kornell & Bjork, 2008).

The subjective experiences one has while learning have a relatively strong effect on JOLs (Jacoby & Kelley, 1987; Koriat, 1997). One kind of subjective performance is the ease or difficulty or retrieving information from memory on a test; people give higher JOLs after answering a question correctly and quickly than after answering slowly or not at all (Benjamin, Bjork, & Schwartz, 1998; Dunlosky & Nelson, 1992; Kelley & Lindsay, 1993). We also consider the ease, or fluency, of perceptual processing as a kind of subjective performance, in the sense that people often think that they are doing well when they find the information they are learning easy to process (Besken, 2016; Rhodes & Castel, 2008, 2009; Undorf, Zimdahl, & Bernstein, 2017). While cues like retrieval fluency and processing fluency influence JOLs, the amount one is learning does not necessarily affect one's experience at the time of the JOL.

The present experiments examined two possible explanations of why the performance bias occurs. One, which we call the observability hypothesis, relies on the idea that people are only influenced by things they can observe. The performance bias comes about, according to the observability hypothesis, because it is possible to observe current performance, but it is frequently impossible to observe learning while it is happening (E. L. Bjork & Bjork, 2011; Soderstrom & Bjork, 2015). The observability hypothesis makes a testable prediction: The performance bias should decrease or disappear if learning becomes observable. In other words, when people know they have been improving by practicing, they should predict that they will continue to improve if they continue to practice, because both current performance and learning are observable.

A second explanation of performance bias, which we call the disregard hypothesis, says that the problem is deeper than observability. According to the disregard hypothesis, the performance bias would come about even if learning were observable, because when we make JOLs we are heavily influenced by performance but we disregard learning. The disregard hypothesis makes its own prediction: The performance bias should persist if learning becomes observable.

As an analogy, consider someone who has just failed to stop at a stop sign that is hidden behind a tree branch. According to the observability hypothesis, our driver is competent and can be counted on to stop at stop signs; his problem is that he could not see the sign. The disregard hypothesis says our driver is incompetent, and would have not have stopped even if the sign had been visible. In the studies reported here, we attempted to make learning observable—or in the analogy, make the stop sign visible—and see whether our or not this led our participants to make competent responses about their expected future performance.

Previous research in which learning was observable

Existing evidence is consistent with the disregard hypothesis. We review this evidence next. (To foreshadow, this evidence is also consistent with other explanations, as we explain later.) Koriat, Sheffer, and Ma'ayan (2002) reviewed data from 10 experiments in which participants learned word pairs through at least three study–test cycles. On each study trial, participants made a JOL for the subsequent test. At the time when participants made their JOLs regarding Test 3, their past learning was, at least potentially, observable: Actual recall had improved from Test 1 to Test 2. Consistent with the performance bias, participants did not expect to improve much from Test 2 to Test 3 (even though their actual recall continued to improve).

Research by Kornell and Bjork (2009) also supports the disregard hypothesis. In their Experiment 12, participants were tested on word pairs and then given feedback. In the key condition, for our purposes, participants went through the same set of pairs four times, and each time they were asked to predict how they would do their fourth time through the list. At the time of their third judgment they could have observed their improvement from Test 1 to Test 2, and at the time of their fourth judgment they could have observed their improvement from Test 1 to Test 3. Consistent with the disregard hypothesis, participants did not seem to expect to keep improving; even though they were predicting their response accuracy on the fourth test, they consistently predicted they would do as well in the future as they had on their previous test, whether it was Test 1, 2, or 3. (One might wonder whether they had noticed their own improvement, but Experiment 11 in the same article suggests that they had.)

The studies just reviewed are consistent with the disregard hypothesis, but they can also be explained other ways, one of which is that participants did not actually observe their own learning. A stronger test of this hypothesis would need to ensure, and verify, that learning was observable. In the studies just described, observing improvement was not easy; participants improved across only two or three test trials, and tracking their improvement might have been difficult given that they were studying many different items and only improving on some of them (Koriat et al., 2002; Kornell & Bjork, 2009). It is not clear whether, or how much, participants thought they had been improving. Furthermore, for any given item, the only way to improve is to answer the item correctly, but doing so prevents any future improvement from occurring on that particular item; in other words, at the level of individual items, once a participant had improved, there was no room for them to expect to keep improving.

Our study went beyond previous research by making it very easy for participants to observe, and respond to, their improvement. To this end, we allowed participants to practice a skill for 50 or 60 trials, rather than two or three, before

making a judgment, and thus build up a longer history of improvement. We asked them to judge how well they would do on this skill across a set of trials, so that they had room for improvement. We also made past improvement highly salient: Immediately before they were asked how much they would improve in the future, participants answered questions that made their own past improvement explicit. Our study differs from previous studies in two additional ways: We examined procedural skill learning rather than verbal learning, and in Experiment 1 we manipulated improvement, with random assignment to an improvement condition or a control condition. In short, the present studies represent a strong test of the disregard hypothesis: Past improvement was made so obvious that the observability hypothesis clearly predicts JOLs should predict future improvement; only by disregarding past improvement could participants' JOLs fail to predict future improvement.

The present experiments

We pitted the observability and disregard hypotheses against each other by allowing participants to observe their learning as they practiced a new skill and then asking them to predict how they would do in the future if they kept practicing. These predictions took the form of an aggregate JOL, because each participant made a single prediction, instead of making separate predictions for each individual item. In Experiment 1 we manipulated the amount participants thought they were improving across trials in a numerical estimation task (estimating the number of stars on a screen). In one condition participants received honest feedback, whereas in the other they received specious feedback designed to make them think they were improving more than they actually were. In Experiments 2 and 3, which did not involve deception, participants improved through successive trials of learning the positions of letters in the alphabet (e.g., A = 1). After completing 50 trials (Experiment 1) or 60 trials (Experiments 2 and 3), participants made three judgments. They estimated their accuracy in the first and second half of the trials they just completed. Then they predicted their accuracy on future trials. Experiments 1 and 2 asked participants to predict the number of those future trials they would get correct. In Experiment 3, to make learning even more salient, we framed the prediction of future performance in terms of how much better or worse they would do on additional trials relative to their prior performance.

Assuming that we succeeded in making learning observable (which the results show we did), there are two opposing predictions. The observability hypothesis predicts that participants should have expected to do better on future trials than they were currently doing. The disregard hypothesis predicts that the performance bias should persist, and therefore that predictions of future performance should have been about the same as estimates of current performance.

Experiment 1

On each trial in Experiment 1, participants were shown an image containing 11 to 20 stars for 2 seconds; they were asked to estimate how many stars were in the image. The feedback was honest in the honest feedback condition, but in the specious feedback condition, feedback was manipulated slightly in a way that made participants seem to be improving in the task. After 50 trials, participants were asked how well they had done on the first 25 trials and second 25 trials, and how they would do on the next 25 trials.

Method

Participants One hundred nine participants were recruited using Amazon's Mechanical Turk and were paid $2.00 for completing Experiment 1. At the end of the experiment we asked if they noticed anything unusual about the experiment. Five participants (four in the specious feedback condition and one in the honest feedback condition) wrote that they believed the feedback was not always truthful. We excluded these participants because we assume they did not believe the feedback they were given.

The 104 remaining participants were included in the data analysis. Fifty-three were randomly assigned to the honest feedback condition (32 females, 21 males; median age = 34 years, range: 23–74), and 51 were randomly assigned to the specious feedback condition (27 females, 24 males; median age = 32 years, range: 18–64). All participants reported living in the United States and being fluent English speakers, except for three who did not answer the country question and one who did not answer the language question.

Materials The materials were 50 images of between 11 and 20 stars (asterisks) arranged irregularly. There were five different arrangements for each number of stars. Figure 1 shows an example of an item with 17 stars.

Procedure There were two phases: a learning phase and a JOL phase. During the learning phase, the 50 star images were shown in a random order. Each one was shown for 2 seconds, which meant that participants typically did not have enough time to count the stars. Then, with the stars no longer on the screen, participants had 4 seconds to type their guess for how many stars had just been displayed. Feedback was then given for 2 seconds.

Participants were assigned to either the honest or specious feedback condition. In the honest feedback condition,

Fig. 1 An example of the stars stimuli used in Experiment 1

participants were given truthful feedback about the number of stars that had been shown, and whether or not they had been correct. In the specious feedback condition, participants were given false feedback designed to make it seem as though their performance was improving throughout the experiment. We could have examined participants whose performance actually improved versus those whose performance did not improve, but that would have led to subject selection effects (e.g., perhaps improvers are generally optimistic about the future regardless of whether they have been improving recently). Instead, we decided to lie so that we could randomly manipulate how frequently participants were told their responses were correct. By indicating that they had done a little worse at the start and a little better at the end than they had, we attempted to create the impression that participants were improving.

To prevent participants from noticing the lies, our paradigm gave false feedback as infrequently as possible. The paradigm worked as follows. We considered the 50 trials as five blocks of 10 trials, though this was not apparent from the participants' perspective. In the first block of 10 trials, we gave truthful feedback most of the time, but said that three correct answers were actually incorrect. When we erroneously told participants they were incorrect, we also told them the correct answer was some number of stars that was randomly chosen to be between two fewer and two more stars than the correct answer they had actually given. For example, if a participant correctly said that 17 stars were shown, we might have told them that they were wrong, and that the correct answer was 15, 16, 18, or 19. In the second block of 10 trials, we only said one correct answer was incorrect. In the third block of 10 trials, we gave truthful feedback. Finally, in the last two blocks of 10 trials, we made performance appear better than it actually was. We told participants they were correct when they were actually wrong on one and three trials in the fourth and fifth block, respectively. Of the eight lies we told, four were negative and four were positive, so they cancelled each other out, and the number of trials a participant was told they had gotten correct was the same, across all 50 trials, as the number they had actually gotten correct (e.g., a participant who responded correctly 24 times end up being told they had been correct 24 times). For this reason, we expected to observe no difference between conditions in terms of how many times participants were told they had been correct overall.

The JOL phase followed the learning phase. During the JOL phase, participants estimated their previous performance and predicted their future performance. (We use the term JOL for all of these judgments even though a judgment of past performance is technically a confidence judgment, not a JOL.) When estimating previous performance, participants were told they just completed 50 trials and then asked two questions: How many of the first 25 of the past 50 trials do you think you got correct? How many of the second 25 of the past 50 trials do you think you got correct? When predicting future performance, they were asked how many they thought they would get right if they completed 25 more trials. Whether a participant was asked to estimate past performance first or predict future performance first was counterbalanced across participants.

Results

In the honest feedback condition, accuracy was similar on the first 25 trials ($M = .44$, $SD = .22$) and second 25 trials ($M = .46$, $SD = .23$). In the data analyses presented here, accuracy in the specious feedback group refers to accuracy according to the false feedback, not to actual proportion correct. In the specious feedback condition, we manipulated accuracy to be lower on the first 25 trials ($M = .22$, $SD = .17$) than on the second 25 trials ($M = .57$, $SD = .21$).

JOLs are presented in Fig. 2. We first analyzed judgments of past performance by computing the difference between JOLs for the second 25 trials and the first 25 trials. Consistent with the accuracy data, participants in the specious feedback condition judged that they had improved more than participants in the honest feedback condition, $t(102) = 4.59$, $p < .001$, $d = .88$. Thus, participants were able to observe their own improvement.

Next, we analyzed predictions of future performance. Half of the participants estimated previous performance first, and half of the participants predicted future performance first. To examine the effect of question order, we conducted a 2 (learning condition: honest vs. specious feedback) × 2 (question order: previous-first vs. next-first) mixed-design analysis of variance. There was no main effect of question order on predicted improvement, $F(1,100) = .03$, $p = .85$, $\eta_p^2 = .0004$, and no significant interaction between condition and question order, $F(1,100) = .81$, $p = .37$, $\eta_p^2 = .008$. Therefore, we collapsed our data across question order.

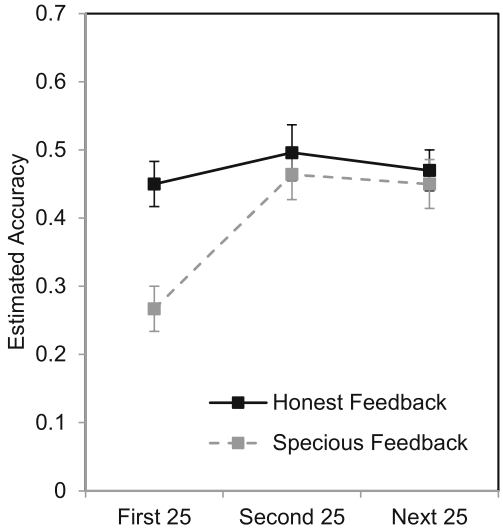

Fig. 2 Estimated proportion correct in Experiment 1

To answer our primary question, we computed an improvement score for each participant by calculating the difference between their second 25 estimate and their next 25 estimate. The observability hypothesis predicted that participants in the specious feedback condition would expect to keep improving and thus predict better performance in the future. This hypothesis was not supported: Predicted improvement scores in the specious and honest feedback conditions were not significantly different, $t(102) = .40$, $p = .69$, $d = .08$ (see Fig. 2). There was actually a slight decline in predicted performance for both groups, but it was not significant in the honest feedback condition ($M = -.03$, $SD = .18$), $t(52) = -1.02$, $p = .31$, $d = .28$, or the specious feedback condition ($M = -.01$, $SD = .12$), $t(50) = -.78$, $p = .44$, $d = .22$.

Discussion

The results showed that participants in the specious feedback condition thought they had learned more than those in the honest feedback condition. In other words, learning was observable in Experiment 1. Yet predicted future performance did not differ between the specious and honest feedback conditions. In fact, the specious feedback condition did not expect to improve at all. These findings are inconsistent with the observability hypothesis. The bias seems deeper than that: Even when learning was observable, people seemed to estimate their future performance based on their current performance (i.e., how they did on the most recent set of trials), which is consistent with disregard hypothesis.

The fact that we lied to participants is a potential limitation of Experiment 1. It leaves open the possibility that participants did not actually think they were getting better at the task. Participants who noticed our deception were excluded from the analyses, but still, one could argue there is a problem: Even if the participants all believed the feedback, it is possible that they chalked up their improvement to the vagaries of luck rather than to actual learning. We addressed this possibility in Experiments 2 and 3 by using a task where the learning was real. Another virtue of the task used in Experiments 2 and 3 was that performance never came near perfection and thus it was always possible to improve on future trials.

Experiment 2

Experiment 2 was a conceptual replication of Experiment 1. In this case, there was no deception and participants actually improved, instead of just being told they had improved. The question, again, was whether they would expect to continue to improve in the future. Participants learned the numerical position of letters in the alphabet (e.g., G = 7). As in Experiment 1, participants completed the learning phase and then predicted their performance on future trials. Unlike in Experiment 1, participants went on to complete these additional trials. In Experiment 2, therefore, predictions of future performance were compared to actual future performance, whereas in Experiment 1 they had been compared to the predictions of future performance in the control condition. (This meant there was no need for a control condition in Experiment 2.)

Method

Participants Ninety-three participants were recruited using Amazon's Mechanical Turk and were paid $2.00 for completing Experiment 2. None of these participants said they had already memorized the numerical position of the letters of the alphabet before the experiment started. Six participants were excluded for reaching 100% accuracy during the learning phase, since they could not improve further.

We analyzed the data from the remaining 87 participants (55 females, 32 males; median age = 32 years, range: 18–61). All participants reported living in the United States and speaking English, except for three who did not report whether they were fluent English speakers.

Materials and procedure There were three phases: a learning phase, a JOL phase, and a test phase. During the learning phase and test phase, participants were shown a letter on the screen and had 4 seconds to enter its numerical position in the alphabet. For example, *H* is the eighth letter in the alphabet. We used only letters *G* (7) through *U* (21), to prevent participants from counting up to or down to a given position of the letter in the alphabet. The correct answer was then shown for 2 seconds. There was no false feedback.

In the learning phase, the list of 15 letters was tested four times, for a total of 60 trials. The letters were presented in a

random order each time through the list. The JOL phase was next. Similar to Experiment 1, participants were told they had just completed 60 trials and were asked to estimate how they did on the first 30 trials and the second 30 trials. They also predicted how they would do if they did 30 more trials. Again, the order of the question about past performance and the question about future performance was counterbalanced across participants. Then came the final test phase, in which the list of 15 letters was tested two additional times, for a total of 30 more trials. As in the learning phase, the letters were tested in a random order each time through the list.

Results and discussion

Actual accuracy improved during the learning phase from the first 30 trials to the second 30 trials (see Fig. 3). JOLs increased at a similar rate. We computed improvement scores by calculating the difference between the second 30 and first 30 trials for both actual accuracy and JOLs. There was no significant difference between the actual and estimated improvement scores, $t(86) = 1.55$, $p = .12$, $d = .17$. Thus, our participants learned and they accurately observed their own learning.

Once again, our primary question was whether predictions of future performance would be sensitive to observable learning. For each participant, we calculated a predicted improvement score by subtracting their second 30 estimate from their next 30 prediction. Then we calculated an actual improvement score by subtracting second 30 accuracy from next 30 accuracy. As in Experiment 1, half of the participants estimated past performance and then predicted future performance, while the other half predicted future performance first.

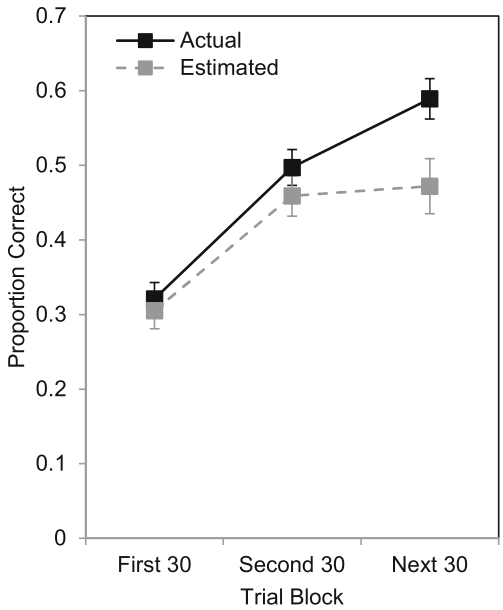

Fig. 3 Actual and estimated proportion correct in Experiment 2

Neither predicted or actual improvement scores were significantly affected by question order. A 2 (predicted vs. actual improvement) × 2 (previous-first vs. next-first) mixed-design analysis of variance revealed no significant main effect of question order, $F(1, 85) = 3.53$, $p = .06$, $\eta_p^2 = .03$, and no significant interaction effect, $F(1, 85) = .014$, $p = .90$, $\eta_p^2 = .0002$. Therefore, we collapsed our data across question order and compared predicted and actual improvement scores.

Actual improvement from second 30 to next 30 was significantly larger than predicted improvement, $t(86) = 2.34$, $p = .02$, $d = .27$. In fact, consistent with Experiment 1, predicted improvement scores were not significantly different from zero, $t(86) = .43$, $p = .67$, $d = .09$ (see Fig. 3).

In short, with a new task that led to actual learning, participants were able to accurately estimate how much they learned during practice, but they did not expect to improve in the future. Taken together, Experiments 1 and 2 suggest that judgments of learning are sensitive to current performance but not observable learning. These findings substantiate the disregard hypothesis but not the observability hypothesis.

Experiment 3

Given that our participants knew they had been improving, why did they not expect to keep improving? Experiment 3 contrasted two possibilities. One possibility is that participants believed they had reached their maximum performance level after 60 trials and could not improve in the future. The other possibility is that participants believed that past improvement portended future improvement, but this belief did not affect their predictions of future performance.

To sort out which of these explanations is correct, we borrowed a strategy that has been successful in prior research (e.g., Koriat et al., 2004): We made improvement very salient at the time the JOLs were made. Koriat et al.'s participants predicted the same level of performance on a memory test regardless of the retention interval (e.g., 1 day vs. 1 week). They became sensitive to retention interval when Koriat et al. made forgetting salient, though. They did this either by manipulating retention interval as a within-participants variable or by having participants predict the number of items they would forget (instead of the number of items they would remember). Oddly, predictions of future learning seem to be even more resistant to remediation; Kornell and Bjork (2009) found that JOLs were insensitive to the amount of learning that would occur in the future, even when future learning was manipulated within-participants design (which should have made it salient). Thus, it is not clear whether making future learning highly salient should affect JOLs in Experiment 3.

Experiments 2 and 3 were nearly identical, but to make learning salient, we changed the framing of the JOL question

to be about learning instead of absolute performance. Previous research suggests that the framing of the JOL question can affect JOLs (Finn, 2008; Koriat et al., 2004; Tauber & Rhodes, 2012). In Experiment 2 we asked, "How many of the next 30 trials do you think you will get correct?" In Experiment 3 we asked, "How well do you think you will do on the next 30 trials in comparison to the previous 30 trials?" If our participants believed that that past improvement heralds future improvement, we expected them to apply those beliefs in answering this question.

Method

Participants Fifty-one participants were recruited using Amazon's Mechanical Turk and were paid $2.00 for completing Experiment 3. The number of participants was smaller than in the previous studies because, based on prior research (e.g., Koriat et al., 2004), we expected a large average increase in JOLs rather than the small increases in Experiments 1 and 2. None of these participants reported knowing the numerical position of the letters of the alphabet before the experiment started. One participant was excluded for reaching 100% accuracy on at least one block of letter trials during the learning phase because he could not improve further.We analyzed the data from 50 participants (27 females, 23 males; median age = 31 years, range: 20–54). All but one participant reported that English was their first language and all participants reported living in the United States, except for two who did not report a country.

Materials and procedure Experiment 3 was very similar to Experiment 2. The only difference was the JOLs phase in the middle of the experiment. Unlike Experiment 2, all participants were asked to estimate past performance first (first 30 and second 30) and then predict future performance. The other change was the phrasing of the question that asked participants to predict future performance. In Experiment 3 participants were asked, "In comparison the last 30 trials you completed, how do you think you will do on the next 30 trials?" A drop-down menu appeared and participant could choose *Same 1 worse*, *2 worse*, *3 worse*, etc., or *1 better*, *2 better*, *3 better*, etc. This phrasing made level of improvement very salient; some might even say it created a demand characteristic that impelled participants to report that they would do better.

Results and discussion

As in Experiment 2, participants improved throughout the training phase and their estimates of their performance on the first 30 and second 30 trials were largely consistent with their actual accuracy (see Fig. 4). Participants' JOLs

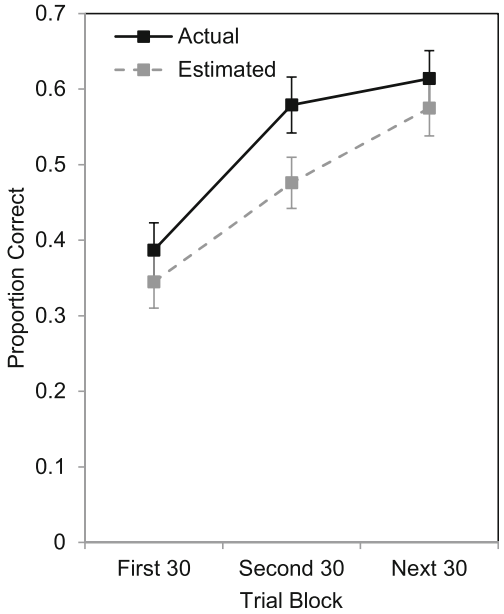

Fig. 4 Actual and estimated proportion correct in Experiment 3

significantly increased from the first 30 trials to the second 30 trials, $t(50) = 6.34$, $p < .001$, $d = 1.79$, indicating they were aware of their learning. However, actual improvement between the first 30 and second 30 was significantly larger than estimated improvement, $t(49) = 2.80$, $p = .007$, $d = .50$.

We were primarily interested in participants' predictions about their performance on the next 30 trials. In the first two experiments, we calculated predicted improvement from second 30 and next 30 estimates, but doing so was unnecessary in Experiment 3 because we asked participants to predict their improvement directly. On average, participants predicted that they would improve by nearly three trials on the next 30 from the second 30 ($M = 2.96$, $SD = 2.86$), which translates into an improvement in accuracy of 0.098, a value significantly greater than zero, $t(50) = 7.33$, $p < .0001$, $d = 1.02$. We calculated actual improvement by subtracting second 30 accuracy from next 30 accuracy. Predicted improvement was greater than actual improvement, $t(50) = 3.44$, $p = .001$, $d = .48$.

That our participants expected to improve relative to their past performance is consistent with previous research examining what has been referred to as a performance heuristic (Critcher & Rosenzweig, 2014). The performance heuristic is the tendency of people who have performed well in the past to expect to improve a large amount in the future. Critcher and Rosenzweig's studies were similar to Experiment 3 in the sense that they made the concept of future improvement highly salient. However, their data analysis asked a different question than ours. Ours asked whether participants use past improvement to predict future improvement and did not consider overall levels of performance. Critcher and Rosenzweig asked whether participants think that high levels of past performance predict future improvement and did not examine past

improvement. Although Critcher and Rosenzweig's question was not important for our hypothesis, the data we collected in Experiment 3 allowed us to examine it. For the sake of completeness, we did so. Specifically, we computed a correlation between performance during the first 60 trials and expected improvement. Although Critcher and Rosenzweig's performance heuristic would predict a positive correlation, we did not find one ($r = -.01$). This finding is not a failed replication because their paradigms (which involved playing darts and solving anagrams) differed from ours in multiple ways, but it might suggest a limitation to the generalizability of their findings.

In summary, Experiment 3 suggests that when observable improvement is made highly salient, participants do believe past improvement will tend to continue in the future. Given that Experiment 2 was almost identical to Experiment 3, it is safe to assume participants in Experiment 2 held similar beliefs. It seems clear, therefore, that mistaken beliefs cannot explain participants' failure to predict future learning in Experiment 2. Instead, apparently, these participants failed to apply their beliefs.

General discussion

This article began with a question: Why are judgments of future performance so closely tied to current performance? According to the observability hypothesis, people are sensitive to whatever they can observe, and it is usually easy to observe current performance, but difficult to observe learning or improvement. Our data did not support this hypothesis. Experiments 1 and 2 showed that participants' JOLs were largely controlled by their current performance even when they had clearly observed (and reported on) their own improvement. One could criticize these experiments by agreeing that participants knew they had improved in the past, but hypothesizing that, perhaps because of features of the task, they did not believe they would continue to improve in the future. This criticism would be inconsistent with Experiment 3, which showed that our participants did believe past improvement portended future improvement; the problem, apparently, was that our participants did not apply this belief in Experiments 1 or 2.

These findings are consistent with the disregard hypothesis: When making judgments about future performance, participants are controlled by their current performance even if they know they have been improving up until now. The disregard hypothesis, in turn, is consistent with—and an explanation for—the broader performance bias, whereby people's judgments of learning are controlled by their current performance.

The finding that our participants predicted no improvement at all in Experiments 1 and 2 is particularly striking given how salient the concept of improvement was in both experiments. Our paradigm made it easy to perceive one's own improvement while doing the trials. Moreover, at the time of the JOL, improvement became even more explicit: In short order, our participants estimated their performance in the first half and second half of the preceding trials—and could hardly have failed to notice that the numbers they entered showed improvement—and then predicted their future performance (or they made the prediction first, but the results were the same in either order). In short, the performance bias must have been powerful indeed for participants to fail to predict future improvement, given how obvious past improvement was in these studies.

As mentioned in the introduction, the performance bias is helpful in explaining the memory for past test heuristic (Finn & Metcalfe, 2007, 2008) the stability bias (Koriat et al., 2004; Kornell et al., 2011), and the metacognitive error of rating desirable difficulty as harmful to learning (e.g., E. L. Bjork & Bjork, 2011). The stability bias and the studies presented here have something else in common: Participants based their judgments on current performance and ignored their beliefs.

One question that remains is why our participants disregarded their improvement, and their beliefs, when predicting future performance. One possibility has to do with the role of experience. Metacognition theories make a distinction between cues that can be experienced and cues that cannot be, and predict that the former will influence judgments much more than the latter (R. A. Bjork, Dunlosky, & Kornell, 2013; Jacoby & Kelley, 1987; Koriat, 1997; Koriat et al., 2004; Kornell et al., 2011). In the present experiments, improvement could not be experienced on any particular trial; it had to be estimated by comparing across trials. Thus, the cause of the performance bias might be that even when learning or improvement can be observed, they are not typically experienced on any given trial, and that is why they do not influence judgments.

However, recent research suggests that experience-based cues do not always dominate metacognitive judgments. Researchers have begun to test the relative importance of beliefs versus processing fluency for judgments of learning. Mueller, Tauber, and Dunlosky (2013) carried out a series of studies showing that participants gave higher JOLs for related than for unrelated words. Mueller, Dunlosky, Tauber, and Rhodes (2014) found that JOLs were higher for large-font words than small-font words. The novelty of these studies is that the authors claim, based on their evidence, that it was beliefs (about relatedness and font size) that controlled participants' responses, not processing fluency. These claims are controversial, however (e.g., Besken, 2016; Frank & Kuhlmann, 2016; Undorf & Erdfelder, 2015), and research in this area has stimulated a healthy debate about the relative importance of beliefs and processing fluency. The studies presented here cannot settle this debate, because we measured

neither perceptual fluency nor relatedness, and the relative balance between beliefs and fluency surely varies from one metacognitive cue to the next. We would simply point out that in the present studies, participants clearly had beliefs about the relationship between past improvement and future improvement (as shown by Experiment 3), and just as clearly, those beliefs had no effect on their estimates of future performance in Experiments 1 or 2. Consistent with other findings, beliefs only influenced JOLs when they were made highly salient (e.g., Koriat et al., 2004).

In the present experiments, current performance affected metacognitive judgments, but we assume it could also affect metacognitive control. For example, performance bias could influence how much effort students put into studying for exams or practicing skills. A person whose goal is to reach a certain level of performance, or knowledge, will be better off if she sees it as possible to improve her performance through practice. If, instead, the performance bias makes her underestimate how much better she can become in the future, she will presumably be more likely to lose hope and give up. For example, children, those inveterate learners, seem to struggle with the performance bias; the girl who tries a new skill twice, decides it is impossible, and gives up, is often the same girl who, after some parental encouragement, is successful and happy after 10 tries—as long as she tries 10 times.

Author note Scholar Award 220020371 from the James S. McDonnell foundation to N. Kornell supported this research. H. Hausman is supported by the National Science Foundation Graduate Research Fellowship Grant No. DGE-1321845. Bridgid Finn provided valuable comments on this manuscript.

References

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*(1), 55–68.

Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(9), 1417–1433. doi:10.1037/xlm0000246

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge: MIT Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York: Worth.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64,* 417–44. doi:10.1146/annurev-psych-113011-143823

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. doi:10.1037/0033-2909.132.3.354

Critcher, C. R., & Rosenzweig, E. L. (2014). The performance heuristic: A misguided reliance on past success when predicting prospects for improvement. *Journal of Experimental Psychology: General, 143*(2), 480–485. doi:10.1037/a0034129

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.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545–565.

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36,* 813–821.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology Learning Memory and Cognition, 33*(1), 238–44. doi:10.1037/0278-7393.33.1.238

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*(1), 19–34. doi:10.1016/j.jml.2007.03.006

Frank, D. J., & Kuhlmann, B. G. (2016). More than just beliefs: Experience and beliefs jointly contribute to volume effects on metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10). doi:10.1037/xlm0000332

Jacoby, L. L., & Kelley, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin, 13*(3), 314–336. doi:10.1177/0146167287133003

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. doi: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. doi:10.1037//0096-3445.126.4.349

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*(4), 643–656. doi:10.1037/0096-3445.133.4.643

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*(2), 147–162. doi:10.1037//0096-3445.131.2.147

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*(9), 1297–1317. doi:10.1002/acp.1537

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219–224. doi:10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585–592. doi:10.1111/j.1467-9280.2008.02127.x

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*(4), 449–468. doi:10.1037/a0017350

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease of processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*(6), 787–794. doi:10.1177/0956797611407929

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*(5), 493–501. doi:10.1080/09658210902832915

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language, 70,* 1–12. doi:10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*(2), 378–84. doi:10.3758/s13423-012-0343-6

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615–25. doi:10.1037/a0013684

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*(3), 550–4. doi:10.3758/PBR.16.3.550

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35,* 481–498. doi:10.1007/s11251-007-9015-8

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10,* 176–199. doi:10.1177/1745691615569000

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging, 27*(2), 474–83. doi:10.1037/a0025246

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*(6), 837–848. doi:10.1002/acp.1598

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition, 43*(4), 647–658. doi:10.3758/s13421-014-0479-x

Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language, 92,* 293–304. doi:10.1016/j.jml.2016.07.003

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15,* 41–44.

Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction, 22*(3), 215–221. doi:10.1016/j.learninstruc.2011.11.002