Journal of Experimental Psychology: General
2004, Vol. 133, No. 4, 643–656

0096-3445/04/$12.00 DOI: 10.1037/0096-3445.133.4.643

Predicting One's Own Forgetting: The Role of Experience-Based and Theory-Based Processes

Asher Koriat
University of Haifa

Robert A. Bjork
University of California, Los Angeles

Limor Sheffer and Sarah K. Bar
University of Haifa

The authors examined the hypothesis that judgments of learning (JOL), if governed by processing fluency during encoding, should be insensitive to the anticipated retention interval. Indeed, neither item-by-item nor aggregate JOLs exhibited "forgetting" unless participants were asked to estimate recall rates for several different retention intervals, in which case their estimates mimicked closely actual recall rates. These results and others reported suggest that participants can access their knowledge about forgetting but only when theory-based predictions are made, and then only when the notion of forgetting is accentuated either by manipulating retention interval within individuals or by framing recall predictions in terms of forgetting rather than remembering. The authors interpret their findings in terms of the distinction between experience-based and theory-based JOLs.

In recent years, social and cognitive psychologists have given increasing emphasis to a possible distinction between two modes of thought that underlie judgments, decisions, and behavior (see Chaiken & Trope, 1999; Kahneman, 2003). The distinction has been variously described as nonanalytic versus analytic cognition (Jacoby & Brooks, 1984), associative versus rule-based systems (Sloman, 1996, 2002), impulsive versus reflective processes (Strack & Deutsch, in press), experiential versus rational systems (Epstein & Pacini, 1999), experience-based versus information-based (or theory-based) processes (Kelley & Jacoby, 1996; Koriat & Levy-Sadot, 1999), heuristic versus systematic processes (Chaiken, Liberman, & Eagly, 1989; Johnson, Hashtroudi, & Lindsay, 1993), and heuristic versus deliberate modes of thought (Kahneman, 2003). Stanovich and West (2000) have used the somewhat less committal terms System 1 versus System 2. Most researchers pointed out the overlap between their proposed contrasts and that between automatic and controlled processes (e.g., Posner & Snyder, 1975). Pooling across different theoretical proposals and borrowing the labels used by Stanovich and West, Kahneman (2003) provided the following summary characterization of the two modes of thought:

> The operations of System 1 are typically fast, automatic, effortless, associative, implicit (not available to introspection), and often emotionally charged; they are also governed by habit and are therefore difficult to control or modify. The operations of System 2 are slower, serial, effortful, more likely to be consciously monitored and deliberately controlled; they are also relatively flexible and potentially rule governed. (p. 698)

In the present study, we attempt to gain some insight into this distinction with regard to a specific question that emerges in the metacognitive monitoring of one's own knowledge during study—namely, how is one's memory likely to be affected by the anticipated retention interval? We adopt the standard view in metacognitive research that such judgments may be based either on heuristics that give rise to subjective feelings or on deliberate inference from beliefs and theories (see Kelley & Jacoby, 1996; Koriat & Levy-Sadot, 1999; Matvey, Dunlosky, & Guttentag, 2001). Our hypothesis is that heuristic-based judgments and theory-based judgments may yield divergent patterns of results as far as the effects of retention interval are concerned and that examination of these patterns can shed light on the conditions under which each of the two types of processes is likely to dominate and/or produce dissociations between subjective and objective measures of performance.

Experience-Based and Theory-Based Metacognitive Judgments During Study

Studies of online monitoring during study, that is, of participants' item-by-item judgments of learning (JOLs), have yielded two general findings. First, learners can estimate roughly the

Asher Koriat, Limor Sheffer, and Sarah K. Bar, Department of Psychology, University of Haifa, Haifa, Israel; Robert A. Bjork, Department of Psychology, University of California, Los Angeles.

Portions of this work were presented at the November 2003 Psychonomic Society meeting in Vancouver, British Columbia, Canada. The research was carried out at the Institute of Information Processing and Decision Making, University of Haifa, Haifa, Israel. It was supported by a grant from the German Federal Ministry of Education and Research (BMBF) within the framework of German-Israeli Project Cooperation (DIP). We thank Rinat Gil for her help in conducting the experiments.

Correspondence concerning this article should be addressed to Asher Koriat, Department of Psychology, University of Haifa, Haifa 31905, Israel. E-mail: akoriat@research.haifa.ac.il

percentage of items that they will recall and can also tell, with some accuracy, which items they will recall and which they will not (Dunlosky & Nelson, 1994). Second, under self-paced conditions, learners typically allocate more study time to items associated with relatively low JOLs than to items associated with relatively high JOLs (Mazzoni, Cornoldi, & Marchitelli, 1990; Nelson & Leonesio, 1988; but see Metcalfe & Kornell, 2003; Son & Metcalfe, 2000). These two findings suggest the operation of an adaptive process in which participants monitor online the study of different items relatively efficiently, regulating the allocation of learning resources according to the monitoring output (see Dunlosky & Hertzog, 1998).

The present study focuses on the extent to which recall predictions are sensitive to the expected interval between study and test. Consider a student who is preparing for an exam and has only one opportunity to go over the material. Given what is known about forgetting, she would be expected to invest more effort in studying when the exam is expected to take place after a 1-week interval as opposed to on the following day. For example, learners have been found to invest more study time when they expect a recall test than when they expect a recognition test (Mazzoni & Cornoldi, 1993). Similarly, when the amount of study time is controlled, students' predictions of their future recall would certainly be expected to decrease with an increasing study–test interval. As we shall see later, however, theoretical considerations as well as some empirical findings suggest otherwise.

As noted earlier, we draw on the dual-basis view of metacognitive judgments, which distinguishes between experience-based and theory-based metacognitive judgments (Koriat, 1997). Experience-based JOLs are assumed to rely on mnemonic cues that derive from the online processing of the items. These cues give rise to a sheer experience of knowing that can serve as a basis for the reported JOLs. Indeed, evidence has accumulated suggesting that JOLs are based on the ease with which studied items are processed during encoding (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Koriat, 1997). Other research supports the contention that JOLs are influenced by the ease and probability with which the to-be-remembered items are retrieved during learning (Benjamin & Bjork, 1996; Benjamin, Bjork, & Schwartz, 1998; Dunlosky & Nelson, 1992; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Matvey et al., 2001). Taken together, these results support the view that the fluency of perceiving or retrieving targets at study provides a basis for JOLs.

Turning next to theory-based judgments, there is little doubt that people make use of their a priori theories about memory in making JOLs. Theory-based JOLs rely on the deliberate application of metacognitive beliefs or theories about one's competence and skills (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Perfect, 2004) and about the way in which various factors can affect memory performance (see Dunlosky & Nelson, 1994; Mazzoni & Kirsch, 2002). For example, participants' JOLs appear to draw on the belief that generating a word is better for memory than reading it (Begg, Vinski, Frankovich, & Holgate, 1991; Matvey et al., 2001). The contribution of metacognitive beliefs has been spelled out most clearly by developmental psychologists (e.g., Flavell, 1971; see Koriat, 2002) in the context of children's memory functioning, but such beliefs clearly influence adults as well (see Koriat, 1997).

The Effects of Retention Interval on Recall Predictions

It is our conjecture that experience-based JOLs should be largely indifferent to retention interval. We arrive at this prediction because cues such as perceptual fluency or retrieval fluency during study do not seem likely to incorporate features associated with expected forgetting. That is, the study of an item should elicit the same degree of fluency independent of the expected delay of a subsequent recall test. Thus, to the extent that learners rely solely on processing fluency as a cue for JOLs, their recall predictions should be largely or entirely indifferent to retention interval. Given what is known about forgetting, this argument yields a strong prediction—namely, that there should be a dissociation between predicted and actual recall, a dissociation that should increase with increasing retention interval.

In contrast, theory-based recall predictions are expected to reveal systematic effects of retention interval. The phenomenon of forgetting is part and parcel of people's naive theories about memory. Thus, when learners base their metacognitive judgments on their metacognitive beliefs or theories, they can be expected to take into account what they know about the effects of retention interval on memory performance.

Experiments 1 and 2 focused on experience-based and theory-based recall predictions, respectively. In Experiment 1, the participants studied a list of paired associates, providing JOLs at the end of each study trial. Three retention intervals were used, and participants assigned to each interval were instructed to make their JOLs reflect performance at the scheduled time of testing. We expected the results to yield a dissociation between JOLs and actual recall such that actual recall performance should exhibit the typical decline with time, whereas JOLs should be indifferent to the expected retention interval. Experiment 2, in contrast, was designed to tap theory-based predictions: Participants were simply asked to predict the number of words that people would recall at different retention intervals. We expected recall predictions to yield a forgetting function similar to that observed for actual recall in Experiment 1.

Experiment 1: Online JOLs

Several scattered observations in the literature are consistent with the prediction that JOLs, under some conditions, may be insensitive to expected retention interval. For example, Maki and Swett (1987), in a study of memory for narrative text, found similar recall predictions when recall was expected to occur a week later versus when it was expected to occur immediately after study. Carroll, Nelson, and Kirwan (1997) had participants study paired associates to a recall criterion and then make JOLs on each item given either a 2-week or a 6-week retention interval. Although eventual recall performance was markedly lower for the 6-week interval, JOLs exhibited little difference between the two intervals. This pattern of insensitivity of JOLs to retention interval was replicated in a subsequent study that used meaningful textual material (Shaddock & Carroll, 1997).

Finally, in a recent study by Koriat and Bjork (in press, Experiment 4), participants studied a list of paired associates and made JOLs regarding cued recall 48 hours later. Across all pairs, participants' JOLs averaged 52% when actual recall averaged 21%! Such dramatic discrepancies are not found when the study–test interval is short (e.g., a few minutes): For example, Koriat, Sheffer, and Ma'ayan (2002) found that JOLs and recall (on the first study–test block of a list of paired associates) averaged 56% and 53%, respectively, across several experiments in which the study and test phases took place in the same session. Taken together, these observations suggest that JOLs might indeed—and counterintuitively—be insensitive to expected retention interval.

To test this hypothesis, we presented participants in Experiment 1 with a list of 60 paired associates and then tested them either immediately, 1 day, or 1 week after study. All participants were informed in advance about the format of the memory test (cued recall) and when it would take place.

In addition to making item-by-item JOLs, the participants were also asked to estimate at the end of the study phase how many of the 60 studied pairs they would recall successfully on the upcoming test. Previous studies have found that such aggregate estimates are generally well calibrated (or show underconfidence), even when item-by-item judgments yield overconfidence (Koriat et al., 2002; Mazzoni & Nelson, 1995). Given those findings, it would be surprising if aggregate judgments also exhibited overconfidence in the day and week conditions.[1]

Method

Participants. Sixty Hebrew-speaking University of Haifa students served as participants, with 20 students assigned randomly to each of the three retention-interval conditions.

Materials and apparatus. Two lists, each consisting of 60 Hebrew word pairs, were compiled. Half of the pairs in each list were constructed to represent a range of associative relatedness; those in the other half were unrelated. Pretesting verified that the two lists were comparable, and each list was assigned to half of the participants.[2] All of the experiments reported here were conducted in Hebrew. Hebrew-speaking participants were used and all materials and instructions were in Hebrew.

Procedure. The 60 word pairs were displayed one at a time on a computer screen, and participants were instructed to study the pairs so that later they would be able to recall the second word in each pair when the first was presented. They were given instructions about when the test phase was to take place: either immediately, the next day, or after a week. On each study trial, the word pair appeared at the center of the screen for 4 s, and when the pair disappeared, participants were asked to assess the chances that they would recall the second word in response to the first word even if they had to guess. The statement "probability to recall (0%–100%)?" appeared on the screen, with "tomorrow" and "in a week" added after "recall" for participants in the day and week conditions. When the study phase ended, participants were asked to make an aggregate estimate. The prompt, which appeared on the computer screen, was, "You were presented with 60 word pairs. How many of them do you think you will remember?" In the day and week conditions, the question ended with "tomorrow" and "in a week," respectively.

The test phase took place at the scheduled time. The stimulus words were presented one after the other, in a random order. Participants had 8 s to say the response aloud, after which a beep was sounded and the next stimulus word was presented. Participants' vocal JOL and recall responses were entered by the experimenter on a keyboard.

Results

Predicted versus actual recall. Mean percentages of actual and predicted recall (JOL) for both related and unrelated pairs are shown in Figure 1 as a function of retention interval. Overall, both predicted and actual recall were highly sensitive to cue–target

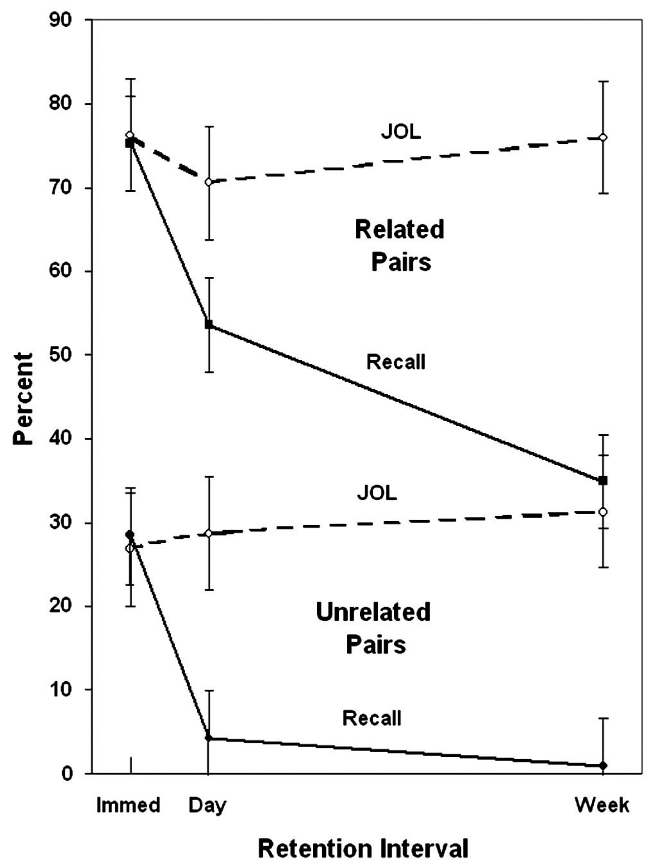

Figure 1. Predicted recall (judgments of learning [JOL], dotted lines) and actual recall (solid lines) as a function of retention interval for related and unrelated paired associates in Experiment 1. Immed = immediate. Error bars represent 95% confidence intervals.

relatedness but, as conjectured, predicted recall was not sensitive to anticipated retention interval for either type of word pair. A Retention Interval × Relatedness analysis of variance (ANOVA) yielded $F(1, 57) = 477.66$, $MSE = 129.09$, $p < .0001$, $\eta^2 = .89$, for relatedness, but $F < 1$ for retention interval and $F(2, 57) = 1.14$, $MSE = 129.09$, *ns*, for the interaction. Across both types of pairs, actual recall declined from 53% (standard error of the mean [*SEM*] = 2.33) to 18% (*SEM* = 1.08) over the 1-week interval, whereas item-by-item JOLs exhibited no drop (52% to 54%) over that interval. A two-way ANOVA, Measure (recall vs. JOL) × Retention Interval, yielded significant effects for measure, $F(1, 57) = 68.12$, $MSE = 154.20$, $p < .0001$, $\eta^2 = .54$, and for retention interval, $F(2, 57) = 13.06$, $MSE = 235.62$, $p < .0001$, $\eta^2 = .30$, but the interaction was also highly significant,

[1] There was a second part of Experiment 1 that is not reported in this article. It addressed an ancillary question: whether study–test experience gained by participants might enhance their monitoring accuracy when making recall predictions for a new list of paired associates. That question is not relevant to the present analysis.

[2] Two lists were used to test the question addressed in Part II; the assignment of the two lists to Part I and Part II was counterbalanced across participants.

$F(2, 57) = 21.16$, $MSE = 154.20$, $p < .0001$, $\eta^2 = .43$. One-way ANOVAs confirmed that retention interval exerted a significant effect on actual recall, $F(2, 57) = 38.94$, $MSE = 160.55$, $p < .0001$, $\eta^2 = .58$, but not on recall predictions, $F < 1$.

These results are in accord with the hypothesized indifference of JOLs to retention interval. Note that JOLs were very well calibrated in the immediate condition for both related and unrelated pairs (see Figure 1). Combining over related and unrelated pairs, predicted and actual recall levels averaged 52.3% ($SEM = 2.42$) and 52.6% ($SEM = 4.11$), respectively, $t(19) = 0.1$, *ns*. In the day and week conditions, however, predicted recall exceeded actual recall by 20.7% and 35.7%, respectively: $t(19) = 5.32$, $p < .0001$, and $t(19) = 8.26$, $p < .0001$, respectively.

We turn next to the aggregate judgments. The aggregate estimates (here and in all of the following experiments) were transformed into percentages. These estimates, 36.7%, 40.8%, and 42.8% in the immediate, day, and week conditions, respectively, were lower overall than the corresponding JOLs (51.6%, 49.6%, 53.6%), but they also evidenced complete indifference to retention interval. Thus, a Measure (JOLs vs. aggregate judgments) × Retention Interval ANOVA yielded significant effects for measure, $F(1, 57) = 35.59$, $MSE = 115.87$, $p < .0001$, $\eta^2 = .93$, but $F < 1$ for retention interval and $F(2, 57) = 1.07$, $MSE = 115.87$, $p < .35$, for the interaction. The effect of measure is consistent with previous findings indicating that aggregate judgments yield overall lower predictions than mean item-by-item JOLs (Koriat et al., 2002; Mazzoni & Nelson, 1995). Note that aggregate judgments, unlike JOLs, exhibited an underconfidence bias in the immediate condition (amounting to 15.9%), $t(19) = 4.43$, $p < .001$. In the day and week conditions, however, aggregate judgments were also overconfident: Predicted recall exceeded actual recall by 12.0% and 24.9%, respectively: $t(19) = 3.06$, $p < .01$, and $t(19) = 6.54$, $p < .0001$, respectively.

In sum, neither item-by-item recall predictions nor aggregate predictions evidenced a forgetting function typical of actual recall. As a result, recall predictions were markedly inflated when they concerned performance after a day or a week.

Sensitivity to associative relatedness. To rule out the possibility that JOLs are simply nonresponsive to factors that affect recall performance, it is important to stress the finding that JOLs were sensitive to cue–target relatedness. Consistent with previous results (Carroll et al., 1997; Connor, Dunlosky, & Hertzog, 1997; Dunlosky & Matvey, 2001; Koriat, 1997), participants' JOLs were highly sensitive to associative relatedness. In the immediate condition, the effects of relatedness on predicted and actual recall were very similar: Mean recall predictions for related and unrelated pairs were 76.3% ($SEM = 2.92$) and 26.8% ($SEM = 3.11$), respectively, when the respective means for actual recall were 75.3% ($SEM = 3.64$) and 28.6% ($SEM = 5.22$). A two-way ANOVA, Measure (JOL vs. recall) × Relatedness, on these means yielded $F(1, 19) = 200.13$, $MSE = 231.05$, $p < .0001$, $\eta^2 = .91$, for relatedness but $F < 1$ for both measure and the interaction.

The sensitivity of JOLs to associative relatedness, coupled with their insensitivity to retention interval, suggests that measures of calibration and resolution should vary differently as a function of retention interval. *Calibration* (or absolute accuracy) refers to the overall correspondence between mean JOLs and mean recall (see Lichtenstein, Fischhoff, & Phillips, 1982). *Resolution* (or relative accuracy), in contrast, refers to the accuracy of JOLs in monitoring the relative recallability of different items. Whereas calibration may affect a student's decision to continue studying for the exam or to stop, resolution may influence the allocation of study time between different parts of the material.

Calibration. Figure 2 depicts calibration curves for the three retention intervals, plotted according to the procedure described by Lichtenstein et al. (1982). Mean over- or underconfidence for each participant, computed as the weighted mean of the differences between the mean JOL and the percentage of correct recall for the 12 JOL categories (0, 1–10, 11–20, . . . 91–99, 100; see Lichtenstein et al., 1982), averaged −0.27 for the immediate condition, 20.71 for the day condition, and 35.70 for the week condition, $F(2, 57) = 21.16$, $MSE = 308.39$, $p < .0001$, $\eta^2 = .43$.

The calibration plot for the immediate condition is similar to that obtained in previous studies (e.g., Dunlosky & Nelson, 1992; Koriat et al., 2002). It exhibits the pattern of miscalibration that is also typical of retrospective confidence: a bias in the direction of underconfidence when JOLs are low and a bias in the direction of overconfidence when JOLs are high (see Erev, Wallsten, & Budescu, 1994). By and large, however, calibration for this condition is remarkably good. In contrast, the plots for the day and week conditions, although also evidencing a certain degree of underconfidence for low JOLs, demonstrate considerable overconfidence for high JOLs. Thus, considering only items assigned JOLs above 50%, mean actual recall averaged 76.13%, 50.25%, and 31.65% for the immediate, day, and week conditions, respectively, whereas the respective JOL means were practically identical: 80.70%, 80.65%, and 80.86%.

Resolution. Resolution, commonly indexed by JOL–recall gamma correlation (Nelson, 1984), yielded very different results. Gamma correlations, calculated for each participant across the 60

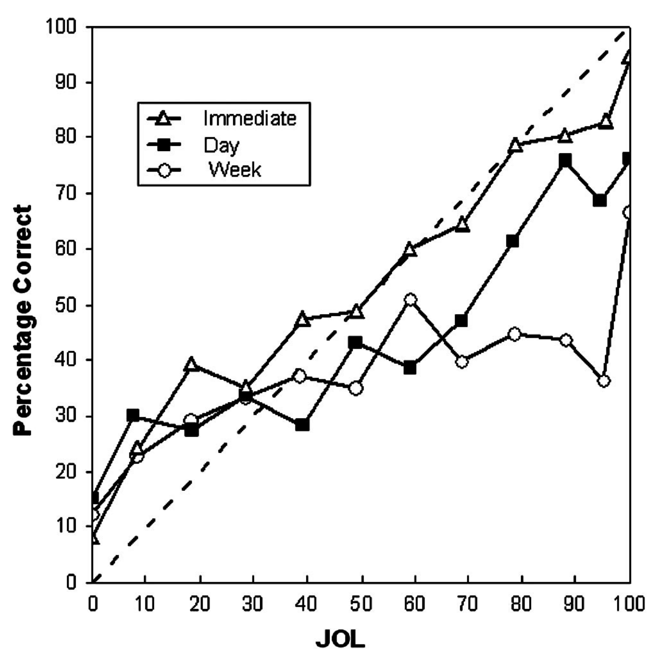

Figure 2. Calibration curves for the immediate, day, and week conditions in Experiment 1. The diagonal indicates perfect calibration. JOL = judgments of learning.

pairs, averaged .67, .74, and .70 for the immediate, day, and week conditions, respectively, $F(2, 57) = 1.38$, $MSE = 0.02$, *ns*. In addition, each of the means was significantly different from 0, $t(19) = 17.43$, $p < .0001$; $t(19) = 35.42$, $p < .0001$; and $t(19) = 24.16$, $p < .0001$, respectively. Thus, sensitivity to interitem differences in recallability was relatively good and was not affected by study–test interval.

Discussion

The results of Experiment 1 yielded a marked dissociation between the effects of retention interval on actual and predicted recall: Recall performance evidenced the typical decline with delay, whereas JOLs were completely insensitive to retention interval, even though participants were reminded of the expected test delay on each trial. The indifference of JOLs to retention interval is consistent with the idea that JOLs are generally based on cues pertaining to the online processing of the items during learning, cues that are insensitive to the effects of test delay.

The results for the aggregate judgments are particularly surprising. One might have expected that the requirement to make aggregate estimates would activate an analytic process that takes into account one's a priori knowledge and beliefs about forgetting. Apart, however, from the fact that aggregate JOLs were lower overall than item-by-item JOLs, they too exhibited complete indifference to retention interval.

Experiment 2: Eliciting Theory-Based Recall Predictions

The results for item-by-item JOLs and aggregate JOLs are at odds with the common observation that forgetting is a central part of everyone's naive beliefs about memory (Mazzoni & Kirsch, 2002). In fact, even young children understand that forgetting increases as the retention interval becomes longer (Lyon & Flavell, 1993; Macnamara, Baker, & Olson, 1976; Wellman & Johnson, 1979). For example, Lyon and Flavell (1993) observed that by 4 years of age, most children understand that forgetting would be more likely to occur after a longer retention interval than after a shorter one.

Presumably, beliefs about forgetting should affect recall predictions when such predictions are theory based rather than experience based. In Experiment 2, we attempted to elicit theory-based recall predictions by describing the experimental procedure of Experiment 1 to a new group of participants and asking them to estimate how many words participants would recall when tested after 10 min, 1 day, or 1 week. The question is whether such estimates, which we refer to tentatively as *theory based*, would yield a forgetting function similar to that observed for actual recall.

Method

Participants. Participants were 22 Hebrew-speaking University of Haifa undergraduates.

Procedure. The experiment took place at the beginning of a class meeting. A booklet that included all of the instructions and materials was distributed. The written instructions (translated from Hebrew) were as follows:

> In a previous experiment that we conducted, participants were presented with a list of 60 word pairs such as "table–apple" one after the other at a constant rate. Their task was to study these pairs so that when presented later with the first word ("table" in the example), they would be able to recall the second word ("apple" in the example).
>
> We would like you to estimate how many word pairs the participants were able to recall on average. Your estimate can range from 0 to 60 pairs. Write down your estimate at the appropriate space at the bottom of the next page.
>
> For your convenience we are enclosing the list of 60 word pairs that we presented to the participants. You are not required to study the word pairs, but only to estimate how many word pairs the participants recalled. Keep in mind that the participants had to recall the second word when presented with the first word.
>
> Please note: Three groups of participants took part in the experiment described. They all studied the same list under the same conditions. For Group A, however, the memory test took place ten minutes later, for Group B it took place one day later, and for Group C it took place one week later. Please give an estimate for each of the three groups.

One of the two lists of the 60 word pairs that were used in Experiment 1 appeared on the next page in two columns, followed by the question "How many word pairs were recalled on the average by each group (write a number between 0 and 60 in each space): After ten minutes ________? After a day ________? After a week ________?"

Results

The theory-based estimates, transformed into percentages, are plotted in Figure 3 along with actual recall scores from Experiment 1. They reveal two striking effects. First, in contrast to recall predictions in Experiment 1 (also plotted in Figure 3), theory-based judgments showed a strong monotonic effect of retention interval, $F(2, 42) = 53.83$, $MSE = 76.86$, $p < .0001$, $\eta^2 = .72$.

Second, there was an impressive correspondence between the theory-based estimates in Experiment 2 and the levels of actual recall in Experiment 1. A Measure (theory-based estimates vs. recall) × Retention Interval ANOVA yielded $F(2, 120) = 47.81$, $MSE = 212.87$, $p < .0001$, $\eta^2 = .44$, for retention interval, but $F < 1$ for both measure and the interaction.

Discussion

What critical aspect of Experiment 2 made participants apply their metacognitive knowledge in the theory-based condition? It is instructive to compare the theory task of Experiment 2 to the aggregate-judgment task in Experiment 1. These tasks differed in three respects, each of which may have contributed to the observed differences between them in the effects of retention interval. First, aggregate judgments were made after studying the list, which may have increased learners' reliance on online mnemonic cues in making recall predictions. Second, unlike aggregate judgments, which concerned one's own performance, the theory-based judgments referred to other people's memory performance, which may have caused greater detachment from one's own subjective experience and a greater tendency to apply one's a priori theory in making predictions. Note, however, that evidence suggests that predictions for others are generally based on processes similar to those underlying predictions for oneself (Kelley & Jacoby, 1996; Nickerson, 1999). Finally, in the theory-based task, retention interval was manipulated within individuals, perhaps increasing its salience and inducing participants to take their metacognitive

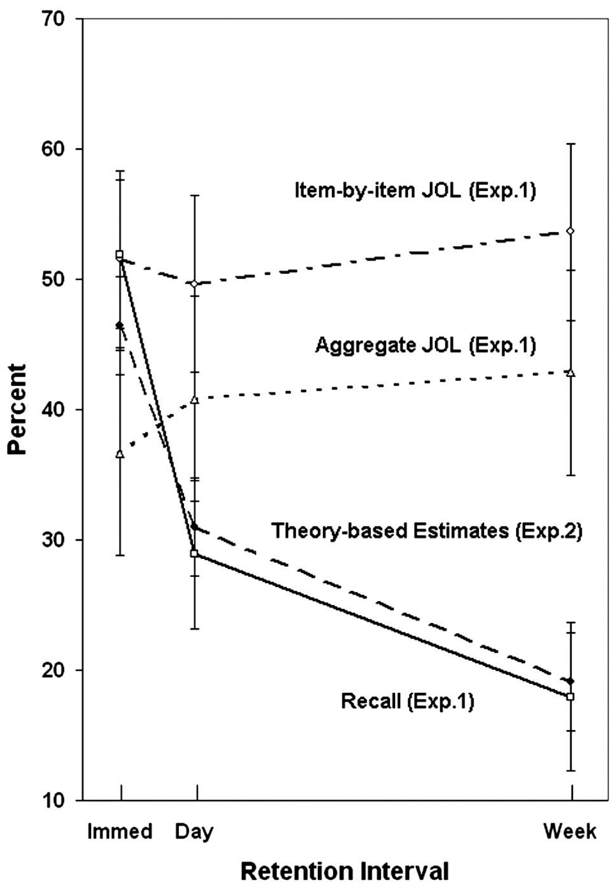

Figure 3. Theory-based judgments (Experiment 2) as well as actual recall, item-by-item judgments of learning (JOLs), and aggregate judgments (from Experiment 1 [Exp.1]) as a function of retention interval. Immed = immediate. The error bars for Experiment 2 (Exp.2) indicate the 95% confidence interval for within-participants comparisons (see Masson & Loftus, 2003), and the error bars for Experiment 1 indicate the 95% confidence interval for between-participants comparisons.

knowledge about forgetting into account in making recall predictions. The following experiments, then, were designed to help specify the critical conditions that allow participants to take retention interval into account in making recall predictions.

Experiment 3

In Experiment 3, we examined the possibility, just mentioned, that it is the within-person manipulation of retention interval that induced participants in Experiment 2 to draw on their metacognitive knowledge about forgetting in making recall predictions. Indeed, there is evidence that factors that affect degree of learning may also affect metacognitive judgments when manipulated within participants but not when manipulated between participants (Begg et al., 1989; Carroll & Nelson, 1993).

Thus, Experiment 3A was similar to Experiment 2 with the exception that retention interval was manipulated in a between-participant design. Experiment 3B, in contrast, was similar to Experiment 1 with the exception that retention interval was manipulated within participants by assigning different retention intervals to different items in the list.

Experiment 3A: Between-Participant Manipulation of Retention Interval in Making Judgments for Others

Method. Retention interval was manipulated in a between-participant design. The instructions and materials were the same as those in Experiment 2, but each participant was required to provide an estimate for only one particular retention interval (no mention was made of the other retention intervals tested). Participants were 14, 14, and 13 high school students in the immediate, day, and week conditions, respectively.

Results and discussion. The results yielded a flat function: The estimates—translated into percentages—averaged 46.90%, 55.48%, and 43.72% (*SEM*s = 5.13, 5.74, and 6.50, respectively) for the immediate, day, and week groups, respectively, $F(2, 38) = 1.19$, $MSE = 423.89$, *ns*. A two-way ANOVA comparing the results of Experiment 3A (between-participant condition) with those of Experiment 2 (within-participant condition, but ignoring the repeated-design feature of this condition) yielded $F(2, 101) = 4.78$, $MSE = 321.79$, $p < .01$, $\eta^2 = .09$, for the interaction.

These results suggest that indeed it is the within-individual manipulation of Experiment 2 that led participants to apply their knowledge in making recall predictions. When retention interval was manipulated between participants, as in this experiment, the global recall estimates for others were highly inflated for the day and week conditions, $t(32) = 5.05$, $p < .0001$, and $t(31) = 4.68$, $p < .0001$, respectively, similar to what was found in Experiment 1 for online JOLs about one's own recall.

Experiment 3B: JOLs With Retention Interval Differing for Different Items

Experiment 3B was similar to Experiment 1, where participants made item-by-item predictions. Retention interval varied between items, however.

Method. Twenty-four University of Haifa undergraduates participated in the experiment for course credit. The materials were the same as those used in Experiment 1. The procedure for the study phase was also the same except for the following changes. First, for each participant, the paired associates were randomly divided into three equal sets such that each set was assigned to one of three retention intervals: 10 min, 1 day, and 1 week. Second, the prompt for JOLs differed correspondingly so that it specified the expected retention interval: "Probability to recall in [ten minutes/one day/one week]: ________". No aggregate estimates were solicited and no recall test was actually administered.

Results and discussion. JOLs for the 10-min, 1-day, and 1-week conditions averaged 67.18%, 57.09%, and 48.84%, respectively (*SEM*s = 2.62, 2.72, and 3.16, respectively), $F(2, 46) = 43.72$, $MSE = 46.34$, $p < .0001$, $\eta^2 = .66$. Thus, JOLs declined monotonically with retention interval.

How successful were these recall predictions in mimicking the actual forgetting function observed in Experiment 1—that is, 52.60%, 28.88%, and 17.96% for the immediate, day, and week conditions, respectively? A two-way ANOVA, Measure (JOLs in Experiment 3B vs. actual recall in Experiment 1) × Retention Interval, yielded $F(1, 126) = 113.35$, $MSE = 174.42$, $p < .0001$, $\eta^2 = .47$, for measure; $F(2, 126) = 44.83$, $MSE = 174.42$, $p < .0001$, $\eta^2 = .42$, for retention interval; and $F(2, 126) = 4.75$, $MSE = 174.42$, $p < .05$, $\eta^2 = .07$, for the interaction. JOLs in the immediate, day, and week conditions exceeded actual recall by 14.58%, 28.21%, and 30.88%, respectively, $t(42) = 3.17$, $p < .01$; $t(42) = 8.04$, $p < .0001$; and $t(42) = 8.14$, $p < .0001$, respectively. Thus, not only was performance overestimated in comparison to the actual performance in Experiment 1 but also, and more important, the effects of retention interval were underestimated.

Even when retention interval is manipulated within participants, however, there may still be a tendency to underestimate its effects when online judgments about one's own performance are solicited versus when the judgments are purely theory based and concern others' performance (as in Experiment 2). Indeed, the effects of retention interval obtained in Experiment 3B were weaker than those observed in the theory condition (Experiment 2): An Experiment (2 vs. 3B) × Retention Interval ANOVA yielded $F(2, 88) = 3.87$, $MSE = 60.90$, $p < .05$, $\eta^2 = .08$. This pattern suggests that global estimates of others' recall are based primarily or solely on one's knowledge and beliefs, whereas the JOLs elicited in Experiment 3B disclosed a mixture of both mnemonic-based and theory-based processes.

Experiment 4

The comparison between the results of Experiments 2 and 3B suggests, however, that a greater sensitivity to the impact of forgetting can perhaps be achieved by having participants adopt an analytic attitude that is detached from their own subjective experience during learning. Would participants under such conditions resort to theory-based judgments, at least to some extent, even when retention interval is manipulated between participants?

We examined this question in Experiments 4A–4C using several modifications of the judgment task that were intended to make it more abstract and more remote from one's own subjective experience. In all three experiments, global predictions for others were solicited and retention interval was always manipulated in a between-participant design.

Experiment 4A: Describing the Experiment in the Abstract

In both Experiment 2 and Experiment 3A, participants received the entire list of word pairs to help them make accurate estimates of memory performance. The presentation of the actual study list may, however, have induced participants to base their predictions on the experience of attempting to master some of the items in the list. Would a between-participant manipulation of retention interval yield the expected effect of forgetting when the task is described in the abstract, without disclosing the specific items used? This question was examined in Experiment 4A.

Method. The procedure was similar to that of Experiment 3A except that the list of stimuli was not included. Thus, participants were given information about the task (paired associates) along with two examples of paired associates (one related and the other unrelated), as well as information about the test (cued recall) and about the length of the list (60 pairs). Participants were University of Haifa students who took part in the experiment at the beginning of a class meeting. They were randomly assigned to either an immediate condition ($n = 31$) or a 1-week condition ($n = 28$).

Results. The recall estimates, when transformed into percentages, averaged 23.71% ($SEM = 2.27$) and 29.29% ($SEM = 3.83$) for the two groups, respectively, $t(57) = 1.29$, *ns*. Thus, even when the memory task was described in the abstract, participants failed to take into account their metacognitive knowledge about forgetting.

Experiment 4B: A Precise Specification of Retention Interval

Experiment 4B was based on the conjecture that the specification of a precise retention interval might induce participants to adopt a more analytic attitude in estimating recall performance and would focus their attention on the possible contribution of retention interval. The two retention intervals used in this experiment were "ten minutes" and "six-and-a-half weeks"; each was assigned to a different group of participants.

Method. The experiment took place during a class meeting. The materials were the same as those used in Experiment 3A. Participants were randomly assigned to the 10-min ($n = 12$) and 6.5-week ($n = 12$) conditions.

Results. The estimates averaged 33.33% ($SEM = 5.42$) and 28.75% ($SEM = 4.40$) for the 10-min and 6.5-week retention intervals, respectively. A t test comparing these means yielded $t(22) = 0.69$, *ns*. Once again, the effect of retention interval was not significant.

Experiment 4C: A Year's Retention Interval

This experiment was an attempt to bring the matter to an absurdity: Would participants fail to take into account a year's retention interval in comparison to participants who are asked to predict performance under immediate testing?

Method. The experiment was conducted in an introductory course at the beginning of the class. The instructions were similar to those used in Experiment 4A. Participants were 90 Hebrew-speaking University of Haifa undergraduates who were assigned randomly to three retention intervals, immediate, week, and year, with 30 participants in each condition.

Results. The estimates reported averaged 35.17%, 40.44%, and 37.06% (*SEM*s = 3.84, 3.72, and 4.28, respectively) for the immediate, week, and year groups, respectively, $F < 1$. Amazingly, the estimates given for a 1-year interval were no lower than those given for immediate testing.

Discussion

The changes in procedure used in Experiments 4A–4C, each intended to induce a more analytic attitude that would lead participants to take into account the impact of retention interval, all proved ineffective. These results, together with those of Experiments 3A and 3B, suggest that participants do not spontaneously apply their theories about forgetting, even when their predictions would seem to be based on their theories and beliefs rather than on their experience studying the items.

Experiment 5

What, then, is special about the within-participant manipulation of retention interval? One hypothesis is that the presentation of different retention intervals activates a comparative mode of processing and it is this mode that induces people to consult their a priori beliefs and knowledge in making predictions. Indeed, it has been proposed that JOLs are comparative in nature—that people tend to focus on the relative recallability of different items within a list and, as a consequence, are less sensitive to factors that affect overall performance (see Begg et al., 1989; Koriat, 1997; Shaw & Craik, 1989). Consistent with this proposal, JOLs in Experiment 1 were found to be highly sensitive to interitem differences in associative relatedness. Thus, in Experiments 5A and 5B, we examined whether the activation of a comparative processing mode can produce sensitivity to retention interval even when each participant is exposed to only one retention interval.

Experiment 5A: Comparative Predictions for Recall and Recognition Testing

Experiment 5A was inspired by findings suggesting that learners are sensitive to the type of memory test expected. In a study by Thiede (1996, Experiment 2), for example, participants who anticipated a recall test spent more time studying a list of paired associates than did participants who anticipated a recognition test. Similar results were also reported by Mazzoni and Cornoldi (1993) when type of memory test expected was manipulated within participants. Thiede (1996, Experiment 3), also using a within-participant design, found participants' JOLs to be lower when a recall test was expected than when a recognition test was expected.

Given that the type of expected test affects JOLs even when manipulated between participants, the question is whether focusing participants' attention on the type of upcoming memory test might induce sensitivity to the effects of retention interval. Thus, in Experiment 5A, each participant was required to make two estimates, one when the expected test was a forced-choice recognition test and one when it was a cued-recall test. Expected retention interval, 10 min or 1 week, was manipulated between participants.

Method. Participants were 50 Hebrew-speaking college students. They were randomly assigned to two retention-interval conditions, 10 min and 1 week, with 29 and 21 participants, respectively, in each condition.

The instructions and materials were the same as those in Experiment 3A. In the written instructions, an experiment was described as follows:

> Students were asked to study a list of paired-associates and their memory for the list was then tested after [ten minutes/one week]. Two types of memory tests were used in that experiment. In the cued-recall test, the first word of a word pair appeared with a blank line next to it, and the student's task was to write down the second word on the blank line. In the recognition test, in contrast, the first word of each pair appeared along with two alternative response words, and the student's task was to choose the correct one out of the two.

After reading this description, participants were asked to estimate how many word pairs the students were able to remember on average in each of the two testing procedures that followed the assigned retention interval. The list of word pairs appeared on the next page followed by the question "How many word-pairs were remembered on the average following [ten minutes/one week] in each test (write a number between 0 and 60 in each space): Cued recall test ________. Recognition test ________."

Results. For the 10-min condition, the estimates averaged 30.40% (*SEM* = 3.80) for the cued-recall test and 50.06% (*SEM* = 4.52) for the recognition test. The respective estimates for the 1-week condition averaged 33.97% (*SEM* = 4.92) and 42.62% (*SEM* = 5.00). A two-way ANOVA, Test Format (cued recall vs. recognition) × Retention Interval (10 min vs. 1 week), on these means yielded a significant effect for test format, $F(1, 48) = 22.07$, $MSE = 222.63$, $p < .0001$, $\eta^2 = .31$, but $F < 1$ for retention interval and $F(1, 48) = 3.31$, $MSE = 222.63$, $p < .07$, $\eta^2 = .06$, for the interaction. The interaction suggested that the expected effect of retention interval was obtained for the recognition test. A *t* test comparing the two retention intervals for this test format was not, however, significant, $t(48) = 1.11$, $p < .25$. In sum, although the results disclosed systematic effects of test format, they failed to reveal an effect of retention interval.

Experiment 5B: Informing Participants About All Retention Intervals

In Experiment 5B, we sought to induce a comparative attitude by turning attention to variations in retention interval. We did so by informing participants about all of the retention intervals included in the study before soliciting an estimate for only one of them.

Method. The experiment was conducted in an introductory course at the beginning of the class. The materials were the same as those used in Experiment 4C. The instructions were also the same with the exception that prior to giving their estimate for a particular retention interval (10 min, 1 week, or 1 year), participants were told the following:

> Three groups of participants took part in the experiment described. They all studied the same list under the same conditions. For Group A, however, the memory test took place ten minutes later, for Group B it took place one week later, and for Group C it took place one year later.

Then, depending on their condition, participants were asked to estimate how many words would be recalled 10 min, 1 week, or 1 year later.

Participants were 73 Hebrew-speaking University of Haifa undergraduates who were assigned randomly to the three retention intervals, with 26, 25, and 22 participants in the immediate, week, and year groups, respectively.

Results. The reported estimates averaged 38.46%, 38.60%, and 21.59% (*SEM*s = 4.21, 5.17, and 4.15, respectively) for the immediate, week, and year groups, respectively, $F(2, 70) = 4.54$, $MSE = 485.80$, $p < .05$, $\eta^2 = .11$. A Scheffé post hoc analysis on these means established that the only significant difference ($p < .05$) was between the estimate for 1 year and the other two estimates.

Discussion. Experiment 5B was the first to yield some effect of retention interval in a between-participant design. The effect, however, was rather small. First, the estimates provided for the 1-week interval were no lower than those made for immediate testing. Second, recall estimates for the 1-year interval (21.59%) were most likely grossly overestimated because they were higher than what was actually demonstrated in Experiment 1 after a week (17.96%). Nevertheless, the results suggest that inducing a comparative attitude with regard to retention interval can make participants' estimates sensitive to the impact of forgetting.

A Reassessment: The Forgetting-Notion Hypothesis

The results reported so far call for a reassessment of our theoretical position. The initial motivation for the present project was the dual-basis view of metacognitive judgments. We expected mnemonic-based recall predictions to display indifference to retention interval, whereas theory-based predictions were expected to incorporate judges' beliefs about the impact of forgetting. The results indicated, however, that even when recall predictions were most likely based on the person's knowledge and beliefs rather than on the mnemonic feedback from study experience, they were insensitive to retention interval under a wide range of conditions. The only situation that produced a clear effect of retention interval was when each participant made predictions for several different intervals (Experiments 2 and 3B). A small effect was also found in Experiment 5B, in which participants were made aware that other retention intervals were also being tested. What is it about these situations that makes participants bring their knowledge to bear on their recall predictions?

It is our conjecture that when retention interval is varied within person, participants are induced to take into account the change that is expected to occur over different retention intervals, because the notion of forgetting implies a decline in memory performance over time. Unless the notion of forgetting is accentuated, recall predictions are largely insensitive to retention interval. Thus, the importance of a within-individual manipulation of retention interval is to activate the notion of forgetting.

This hypothesis implies that a stimulus situation must be consistent with the way in which beliefs are represented and activated for these beliefs to be brought to bear on that situation. Because beliefs about forgetting are represented in terms of memory changes that occur over time, these beliefs can only be activated when different intervals are presented. What is noteworthy, however, is that once the knowledge about such changes is activated, people seem to be able to make rather accurate predictions about the absolute level of recall at each point in time (Experiment 2), suggesting that they do possess some knowledge about the general level of memory performance at a given retention interval.

To strengthen the support for this hypothesis, we attempted to rule out an alternative hypothesis that also revolves around the notion that people react primarily to changes (Experiment 6). We then present a final experiment (Experiment 7) in which we attempted to frame the prediction task in terms of forgetting rather than in terms of remembering, with the expectation that perhaps the accentuation of the notion of forgetting would be sufficient to yield an effect of retention interval, even in a between-individual design.

Experiment 6

The idea that people are particularly responsive to change rather than to the absolute values of a variable has been expressed by researchers studying judgments of the well-being of others. Stated briefly, people overestimate the effects of objective life circumstances on subjective well-being. Individuals with paraplegia, for example, do not differ strongly from the average person in their reported happiness (Brickman, Coates, & Janoff-Bulman, 1978), but the subjective well-being of paraplegics and others with chronic health conditions is predicted by experimental participants to be much worse than that of the general public (see Ubel et al., 2001). Similarly, although the reported life satisfaction of students living in California was no better than that of students living in the Midwest, both student groups predict that Midwesterners are less satisfied with their lives than are Californians (Schkade & Kahneman, 1998).

Schkade and Kahneman (1998) explained the observed discrepancy between such predicted and actual judgments in terms of a focusing illusion: "When attention is drawn to the possibility of a change in any significant aspect of life, the perceived effect of this change on well-being is likely to be exaggerated" (p. 340). Perhaps, then, any manipulation that focuses attention on retention interval should increase sensitivity to the effects of that determinant.

There is, however, a subtle difference between the attention-focusing hypothesis and our hypothesis about forgetting. According to our hypothesis, exposing participants to different retention intervals does not merely focus participants' attention on retention interval but rather induces them to resort to theories and beliefs that specifically concern change. To test the attention-focusing hypothesis with regard to the effects of retention interval, we used manipulations that focus participants' attention on study–test interval to see whether such manipulations are sufficient to produce sensitivity to the impact of forgetting even in a between-participant manipulation.

Experiment 6A

A large number of studies have indicated that asking participants to imagine future events or to build scenarios that lead to future outcomes can sometimes help debias faulty predictions (see Koehler, 1991; Sanbonmatsu, Posavac, & Stasney, 1997). In one condition of Experiment 6A, participants were asked to imagine themselves being tested at the scheduled time and having to recall the target words in response to the cue words.

Method. The experiment took place at the beginning of a class meeting. Participants (University of Haifa undergraduates) were randomly assigned to one of the four conditions generated by crossing two levels of retention interval (10 min vs. 1 week) and attention instruction (attention focusing vs. control). Seventeen to 18 participants served in each of the four conditions.

The materials and instructions for the control groups were the same as those in Experiment 3A with the exception that participants made global predictions about themselves. They were told, "Suppose that you had to study a list of paired associates similar to that presented on the next two pages. How many words do you think you will recall if the test takes place [ten minutes/one week] after studying the list?" It was indicated that each pair would be presented for 4 s, and that the test would involve cued recall.

The instructions for the attention-focusing group were the same except that they included the following additional paragraph:

> In making your prediction, we would like you to imagine yourself sitting here [immediately after the presentation of the list/a week from now, that is, next (day of the week)] and going through the test. Imagine that you will be presented with the first word of each of the pairs and that you will be required to recall the corresponding second word. For how many stimulus words would you be able to recall the corresponding second word?

Results. The reported estimates averaged 43.15% (*SEM* = 5.11) and 49.90% (*SEM* = 6.02) for the 10-min and 1-week conditions, respectively, in the attention-focusing condition. The respective values for the control condition were 35.69% (*SEM* = 4.13) and 39.17% (*SEM* = 5.66). A Retention Interval × Attention Instructions ANOVA on these means yielded a near significant effect for attention instructions, $F(1, 66) = 3.04$, $MSE = 412.86$, $p < .08$, $\eta^2 = .04$. Estimates were somewhat higher for the attention-focusing condition ($M = 46.43$, $SEM = 3.86$) than for the control condition ($M = 37.48$, $SEM = 3.44$). Neither the effects of retention interval nor the interaction were significant, both Fs <1.

Experiment 6B

In Experiment 6B, we attempted to draw attention to retention interval by introducing a monetary incentive for making accurate predictions.

Method. The experiment took place during a class meeting, and the participants were University of Haifa undergraduates. The instructions were similar to those of Experiment 3A, but, in addition, it was indicated that participants would win NIS 15 (about \$3.50) if their estimates fell within five items of the correct answer. In calculating payoff, we used the recall results from Experiment 1 as a criterion. Two retention intervals were used, 10 min ($n = 36$) and 1 week ($n = 35$).

Results. The recall estimates averaged 38.43% ($SEM = 2.35$) and 35.14% ($SEM = 2.24$) for the 10-min and 1-week groups, respectively, $t(69) = 1.02$, *ns*. Thus, monetary incentives for accurate estimates were not successful in producing sensitivity to retention interval.

Actual recall in Experiment 1 averaged 53% and 18% for the 10-min and 1-week intervals, respectively. Fourteen participants (39%) in the 10-min group and 7 participants (20%) in the 1-week group provided estimates that fell within five items of the correct number, $t(69) = 1.75$, $p < .05$ (one sided).

A comparison of the results of Experiment 6B with those of Experiment 3A indicated that the estimates were overall lower in the incentive condition of Experiment 6B (36.81%) than in Experiment 3A (45.37%). Neither the effects of retention interval nor the interaction was significant, however: $F(1, 94) = 1.11$, $MSE = 245.53$, $p > .30$, and $F < 1$, respectively.

Discussion

The results of Experiments 6A and 6B failed to support the hypothesis that directing attention to the specified retention interval produces greater sensitivity to the effects of forgetting. These results suggest that the effect of introducing changes in retention interval is to activate beliefs about how such changes are likely to affect recall performance rather than simply to draw attention to retention interval.

Experiment 7

As a more direct test of the forgetting-notion hypothesis, Experiment 7 examined the idea that framing the prediction task in terms of forgetting rather than in terms of remembering would be effective in activating participants' beliefs about the decline in memory performance over time, even in a between-participant design. Thus, whereas in all of the previous studies participants were asked to estimate the number of words that students would be able to recall, in Experiment 7 they were asked to estimate the number of words that they would be likely to forget after different retention intervals.

Method

The procedure was similar to that of Experiment 4C, but the instructions read as follows:

> In a previous experiment that we conducted, students were presented with a list of 60 word pairs such as "table–chair" and "girl–eagle" one after the other. Each word pair was presented for 4 seconds. The students' task was to study these pairs so that when presented later with the first word, they would be able to recall the second word. The memory test took place [ten minutes/one week/one year] later.
>
> In the test, the first word was presented with a blank line next to it, for example:
>
> Table ________
>
> Girl ________
>
> and the students were asked to recall the second word and to write it down on the line.
>
> We would like you to estimate how many of the word pairs the students forgot after [ten minutes/one week/one year]. Your estimate can range from 0 to 60 pairs. An estimate of 0 pairs means that the students did not forget any of the word pairs, whereas an estimate of 60 pairs means that the students forgot all of the word pairs. Write down your estimate at the appropriate space at the bottom of the next page.

The experiment took place at the beginning of a class meeting. Participants were 80 Hebrew-speaking University of Haifa undergraduates who were assigned randomly to three retention intervals, 10 min, 1 week, and 1 year, with 28, 26, and 26 participants in each condition, respectively.

Results

The forgetting estimates averaged 52.62%, 66.86%, and 81.35% (*SEM*s = 4.42, 3.22, and 2.42, respectively) for the 10-min, 1-week, and 1-year groups, respectively, $F(2, 77) = 17.56$, $MSE = 316.87$, $p < .0001$, $\eta^2 = .31$. A Scheffé post hoc analysis on these means established that there was a significant difference in estimates between each pair of retention intervals ($p < .05$).

To compare these results with those of Experiment 4C, we transformed the reported forgetting estimates into recall estimates (by subtracting the means from 100%). Mean-derived recall estimates were 47.38%, 33.14%, and 18.65% for the 10-min, 1-week, and 1-year groups, respectively. The respective means in Experiment 4C were 35.17%, 40.44%, and 37.06%, respectively. A two-way ANOVA, Experiment (4C vs. 7) × Retention Interval, yielded $F(1, 164) = 2.19$, $MSE = 389.86$, *ns*, for experiment; $F(2, 164) = 6.75$, $MSE = 389.86$, $p < .005$, $\eta^2 = .08$, for retention interval; and $F(2, 164) = 8.78$, $MSE = 389.86$, $p < .001$, $\eta^2 = .10$, for the interaction.

Note, however, that the forgetting function observed in Experiment 7 falls short in capturing the actual recall function observed in Experiment 1 from the immediate (52.60%) and 1-week (18.0%) conditions. A two-way ANOVA comparing these means with the respective means in Experiment 7 (47.38% and 33.14%) yielded $F(1, 90) = 1.83$, $MSE = 312.93$, *ns*, for the actual predicted comparison; $F(1, 90) = 43.58$, $MSE = 312.93$, $p < .0001$, $\eta^2 = .33$, for retention interval; and $F(1, 90) = 7.59$, $MSE = 312.96$, $p < .01$, $\eta^2 = .08$, for the interaction.

Discussion

The results of Experiment 7 yielded a significant monotonic decline with retention interval in the estimated number of words forgotten. This effect is quite impressive in view of the failure to observe a similar effect even in Experiments 6A and 6B, which involved manipulations that focused participants' attention on retention interval. Activating the notion of forgetting apparently induces participants to consider the contribution of retention interval even when retention interval is manipulated between individuals.

General Discussion

In this study, we attempted to gain some insight into the distinction between experience-based and theory-based judgments by examining the impact of retention interval on predicted and actual recall. We addressed the question, To what extent and under what conditions does predicted recall mimic the forgetting curve shown by actual recall? The answer to this question, in addition to having applied implications, was expected to shed light on the monitoring processes that occur during study. In what follows, we present an overview of the findings and then explore explanations for these findings.

Overview of the Findings

1. Experiment 1 disclosed a sharp dissociation between predicted and actual memory performance. Whereas recall performance exhibited a clear forgetting function, both item-by-item JOLs and aggregate judgments yielded complete indifference to retention interval. The dissociation was such that JOLs for the immediate condition were well calibrated, whereas those for the 1-week condition were inflated by about 35%. This finding is surprising in view of the fact that participants were reminded of the study–test interval on each trial.

2. In Experiment 2, participants who were asked to estimate the recall performance of others at each of three retention intervals did so with remarkable accuracy. They clearly possessed not only the knowledge that forgetting happens but also some idea of the rate of forgetting of the materials in question.

3. Experiment 3A helped identify a critical variable that induces participants to apply their intuitive beliefs in making global recall predictions for others: the manipulation of retention interval within participants. When retention interval was manipulated between participants, there was no longer an effect of retention interval, which suggests that even theory-based recall predictions are insensitive to retention interval under these conditions.

4. Experiment 3B, in turn, indicated that online item-by-item JOLs were also sensitive to retention interval when the expected time of testing varied across items. The effect of retention interval was weaker, however, than the actual effect observed in Experiment 1 and also weaker than the global predictions made in Experiment 2. If sensitivity to expected retention interval is a kind of signature of theory-based judgments, this pattern of results suggests that participants' item-by-item judgments were influenced by mnemonic-based as well as theory-based processes.

5. Experiments 4A–4C, designed to induce analytic judgments based on calculated inferences (see Kahneman, 2003), demonstrated that insensitivity to the impact of forgetting is quite pervasive when retention interval is manipulated between participants. None of the manipulations used was effective in producing sensitivity to retention interval, including soliciting judgments for a year's retention interval (Experiment 4C).

6. Assuming that a within-person manipulation of retention interval induces participants to make comparative judgments, we designed Experiments 5A and 5B to activate a comparative mode of responding. The requirement to make two predictions, one for a recall test and one for a recognition test, yielded significant effects of test format but not of retention interval, which was manipulated between participants (Experiment 5A). A small but significant effect of retention interval was found, however, when participants were informed about all the retention intervals included in the study but were asked to provide a prediction for only one interval (Experiment 5B). These results suggest that alluding to variations in retention interval helps induce participants to apply their knowledge about forgetting.

7. Experiments 6A and 6B were motivated by the attention-focusing hypothesis (Schkade & Kahneman, 1998). Two manipulations that were intended to focus participants' attention on retention interval—imagination instructions and accuracy incentives—failed to produce sensitivity to retention interval, suggesting the effects observed for the within-person variation of retention interval (Experiments 2 and 3B) do not derive simply from the focusing of attention on retention interval per se but possibly because such variation activates participants' beliefs about forgetting as a process that occurs with time.

8. Finally, Experiment 7 yielded support for this idea by showing that the framing of the prediction question in terms of forgetting rather than in terms of remembering was sufficient to produce a significant effect of retention interval, even in a between-individual design.

How can this pattern of results be explained? In our view, the results on the whole can be conceptualized in terms of two general hypotheses regarding metacognitive judgments during study: the dual-basis hypothesis and the forgetting-notion hypothesis. We discuss each of these hypotheses in turn.

The Dual-Basis View of Metacognitive Judgments

The initial impetus for the research reported in this study was the dual-basis view of metacognitive judgments (Brown & Siegler, 1993; Jacoby & Kelley, 1987; Kelley & Jacoby, 1996; Koriat & Levy-Sadot, 1999; Strack, 1992). According to this view, metacognitive judgments may be based on one or both of two sources of information: (a) subjective experience and (b) domain-specific knowledge retrieved from memory. In the former case, various mnemonic cues contribute directly to produce an immediate feeling of knowing that can serve as the basis of judgments. Thus, for example, encoding and retrieval fluency may foster a feeling of competence that can serve as a basis for reported JOLs. Theory-based judgments, in contrast, rely on the deliberate use of specific beliefs and information to form an educated guess about one's own knowledge. Thus, JOLs may be based on such rules as "memory performance should be better on recognition than on recall memory tests."

The complete indifference of JOLs to retention interval in Experiment 1 is consistent with the assumptions that (a) online JOLs are based predominantly—perhaps exclusively—on the subjective experience associated with processing fluency and that (b) processing fluency during encoding is insensitive to the expected conditions of retrieval. These assumptions are reinforced by the results of Experiment 2, which yielded a strong effect of retention interval that corresponded closely to the forgetting function for actual recall. Thus, the contrast between the results of Experiments 1 and 2 highlights the importance of distinguishing between experience-based metacognitive judgments that rely on online mnemonic cues and theory-based judgments that rely on the explicit application of a priori beliefs.

The subsequent experiments, however, blurred the difference between the two types of processes to the extent of questioning the necessity of invoking the dual-basis view of metacognitive judgments in explaining the results. On the one hand, with the exception of Experiment 5B and Experiment 7, all of the experiments using a between-participant design failed to yield an effect of retention interval even though they most likely tapped theory-based judgments. On the other hand, Experiment 3B showed that even online JOLs exhibit sensitivity to retention interval when that interval is manipulated within person.

We believe, however, that the dual-basis view is important in explaining the results for two reasons. The first is conceptual: It is difficult to see how the fluency heuristic, as has been conceptualized in the literature (see Benjamin & Bjork, 1996), can incorporate the effects of expected time of test. Therefore, we tend to attribute the effects of retention interval (or of expected type of memory test, for that matter; see Thiede, 1996) to the application of one's beliefs rather than to reliance on fluency. The second reason is empirical: As noted earlier, Experiment 3B, which involved online JOLs with expected time of testing manipulated within individuals, yielded a significantly weaker effect of retention interval than was observed for global predictions for others in Experiment 2. We propose that JOLs elicited in Experiment 3B reflect a mixture of theory-based and mnemonic-based processes (see also Matvey et al., 2001).

Theory-Based Predictions: The Activation of the Notion of Forgetting

If the dual-basis view of JOLs is maintained, it must be supplemented with assumptions regarding the conditions that allow participants to bring their a priori knowledge about forgetting to bear on their recall predictions. The most surprising finding of this study is that across a wide range of conditions, people proved to be oblivious to the effects of forgetting. The finding that they did exhibit sensitivity to these effects under some circumscribed conditions (e.g., in Experiment 2) indicates that the information regarding the effects of retention interval was available to them and, when activated, they could use it to make rather accurate recall predictions.

What are the conditions that induce participants to apply their knowledge in making recall predictions? We found that a within-person variation of retention interval yields a clear forgetting function for JOLs, but why? Two hypotheses were entertained. The first was the attention-focusing hypothesis according to which the within-participant manipulation of retention interval acts to draw participants' attention to retention interval, much as the mention of California in Schkade and Kahneman's (1998) study might draw attention to climate. Inconsistent with this hypothesis, however, is the fact that two manipulations that were intended to draw attention to retention interval failed to yield any effect of retention interval in a between-individual manipulation.

The second hypothesis, the one we favor, concerns the conditions that induce people to apply their beliefs about the effects of retention interval when making theory-based recall predictions. The notion of forgetting as a decline in memory performance over time must be activated for people to consider the contribution of a specified retention interval to memory performance. This hypothesis implies that the pertinent beliefs in this case are organized hierarchically as far as their activation is concerned: Once the notion of forgetting is activated, people can then take into account what they know about the specified retention interval in making their recall predictions. It is indeed quite instructive that they do not do that spontaneously—even when the specified retention interval is a year (Experiment 4C)!

It should be stressed that such is probably not the case for the activations of other beliefs pertaining to memory performance. For example, as noted earlier, participants spent more time studying a list of paired associates when they anticipated a recall test than when they anticipated a recognition test, even though test format was manipulated between participants (Thiede, 1996). It is possible that participants take advantage of their beliefs about the effects of test format even when they are required to judge memory performance for only one test format.

In direct support of the forgetting-notion hypothesis, Experiment 7, which used a between-individual design, yielded a clear effect of retention interval when recall predictions were framed in terms of forgetting rather than in terms of remembering. Presumably, the mere mention of forgetting can activate people's knowledge about the decline in memory performance that is expected to occur with time, and this activation then affects their estimates for a given retention interval.

An alternative account of the results, however, must also be entertained. According to that account, people actually lack any knowledge about the proportion of items that can be recalled, and their estimates in the between-participant designs are based simply on a wild guess. People succeed in capturing the effects of retention interval in the within-participant designs because JOLs for the immediate retention interval are used as an anchor when they are subsequently making JOLs for a delayed retention interval. Thus, once people have provided an estimate for an immediate test, they use it as an anchor point and adjust their estimate downward when making JOLs for a delayed test.

This account, however, is inconsistent with findings indicating that JOLs are sensitive to a variety of factors even when these factors are manipulated between participants. These include semantic relatedness and presentation order (unrelated word pairs first vs. related word pairs first; Dunlosky & Matvey, 2001), type of encoding (generate vs. read; Mazzoni & Nelson, 1995), type of material (self-performed tasks vs. words; Cohen, Sandler, & Keglevich, 1991), specified criterion for memory performance (Carroll & Nelson, 1993), and type of word pairs (McDonald-Miszczak, Hubley, & Hultsch, 1996). Also, participants anticipating a recall test have been found to spend more time studying than do participants expecting a recognition test (Thiede, 1996). Finally, the finding of a monotonic effect of retention interval when the question was framed in terms of forgetting rather than in terms of remembering (Experiment 7) also challenges an explanation of the results strictly in terms of anchoring. Anchoring does appear to have an effect on JOLs (Scheck, Meeter, & Nelson, 2004), but as Scheck et al. reported, it is not sufficient to account for differences in the magnitude of JOLs in between-participant designs.

In sum, the present study disclosed sharp dissociations between predicted and actual memory performance with respect to the effects of expected retention interval. Examination of the conditions that did and did not affect participants' sensitivity to the effects of retention interval suggests that the dual-basis view of JOLs—that is, that such judgments draw on experience-based or theory-based knowledge—may need to be modified to reflect the way such bases interact. The present results also demonstrate that people do not spontaneously apply their knowledge about memory, even when they are required to make theory-based judgments. Understanding fully how and when theory-based knowledge is accessed and combined with experience-based subjective knowledge may require nothing less than a theory of how beliefs are organized and activated.

References

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28,* 610–632.

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition, 19,* 487–497.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Hillsdale, NJ: Erlbaum.

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.

Brickman, P., Coates, D., & Janoff-Bulman, R. (1978). Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology, 36,* 917–927.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100,* 511–534.

Carroll, M., & Nelson, T. O. (1993). Effect of overlearning on the feeling of knowing is more detectable in within-subject than in between-subject designs. *American Journal of Psychology, 106,* 227–235.

Carroll, M., Nelson, T. O., & 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.

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212–252). New York: Guilford Press.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Cohen, R. L., Sandler, S. P., & Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology, 45,* 523–538.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12,* 50–71.

Dunlosky, J. T., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 249–275). Mahwah, NJ: Erlbaum.

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic–extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1180–1191.

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.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83–87.

Epstein, S., & Pacini, R. (1999). Some basic issues regarding dual-process theories from the perspective of cognitive-experiential self-theory. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 462–482). New York: Guilford Press.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527.

Flavell, J. H. (1971). Stage-related properties of cognitive development. *Cognitive Psychology, 2,* 421–453.

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

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 18, pp. 1–47). New York: Academic Press.

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

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3–28.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58,* 697–720.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35,* 157–175.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110,* 499–519.

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. (2002). Metacognition research: An interim report. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 261–286). Cambridge, UK: Cambridge University Press.

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

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 483–502). New York: Guilford Press.

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.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Lyon, T. D., & Flavell, J. H. (1993). Young children's understanding of forgetting over time. *Child Development, 64,* 789–800.

Macnamara, J., Baker, E., & Olson, C. L. (1976). Four-year-olds' understanding of pretend, forget, and know: Evidence for propositional operations. *Child Development, 47,* 62–70.

Maki, R. H., & Swett, S. (1987). Metamemory for narrative text. *Memory & Cognition, 15,* 72–83.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29,* 222–233.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47–60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18,* 196–204.

Mazzoni, G., & Kirsch, I. (2002). Autobiographical memories and beliefs: A preliminary metacognitive model. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 121–145). Cambridge, UK: Cambridge University Press.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1263–1274.

McDonald-Miszczak, L., Hubley, A. M., & Hultsch, D. F. (1996). Age differences in recall and predicting recall of action events and words. *Journals of Gerontology: Series B. Psychological Sciences and Social Sciences, 51B,* P81–P90.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132,* 530–542.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 676–686.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125,* 737–759.

Perfect, T. J. (2004). The role of self-rated ability in the accuracy of confidence judgments in eyewitness memory and general knowledge. *Applied Cognitive Psychology, 18,* 157–168.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Sanbonmatsu, D. M., Posavac, S. S., & Stasney, R. (1997). The subjective beliefs underlying probability overestimation. *Journal of Experimental Social Psychology, 33,* 276–295.

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51,* 71–79.

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science, 9,* 340–346.

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

Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging, 4,* 133–135.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.

Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 379–396). New York: Cambridge University Press.

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.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23,* 645–665.

Strack, F. (1992). The different routes to social judgments: Experiential versus informational strategies. In I. I. Martin & A. Tesser (Eds.), *The construction of social judgment* (pp. 249–275). Hillsdale, NJ: Erlbaum.

Strack, F., & Deutsch, R. (in press). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review.*

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 49A,* 901–918.

Ubel, P. A., Loewenstein, G., Hershey, J., Baron, J., Mohr, T., Asch, D. A., & Jepson, C. (2001). Do nonpatients underestimate the quality of life associated with chronic health conditions because of a focusing illusion? *Medical Decision Making, 21,* 190–199.

Wellman, H. M., & Johnson, C. N. (1979). Understanding of mental processes: A developmental study of "remember" and "forget." *Child Development, 50,* 79–88.

Received April 20, 2004
Revision received August 5, 2004
Accepted August 5, 2004 ■