Journal of Experimental Psychology: General
2002, Vol. 131, No. 2, 147–162

0096-3445/02/$5.00 DOI: 10.1037//0096-3445.131.2.147

Comparing Objective and Subjective Learning Curves: Judgments of Learning Exhibit Increased Underconfidence With Practice

Asher Koriat, Limor Sheffer, and Hilit Ma'ayan
University of Haifa

When participants studied a list of paired associates for several study–test cycles, their judgments of learning (JOLs) exhibited relatively good calibration on the 1st cycle, with a slight overconfidence. However, a shift toward marked underconfidence occurred from the 2nd cycle on. This underconfidence-with-practice (UWP) effect was very robust across several experimental manipulations, such as feedback or no feedback regarding the correctness of the answer, self-paced versus fixed-rate presentation, different incentives for correct performance, magnitude and direction of associative relationships, and conditions producing different degrees of knowing. It was also observed both in item-by-item JOLs and in aggregate JOLs. The UWP effect also occurred for list learning and for the memory of action events. Several theoretical explanations for this counterintuitive effect are discussed.

In studying new material, learners typically monitor the extent to which they have mastered different segments of that material and may decide to go over some of these again to ensure comprehension or memory. Research on judgments of learning (JOLs) during study has indicated that these judgments are moderately accurate in predicting subsequent memory performance (e.g., Arbuckle & Cuddy, 1969; Dunlosky & Nelson, 1992; Koriat, 1997; Lovelace, 1984; Mazzoni & Nelson, 1995; Zechmeister & Shaughnessy, 1980). That research also indicates that learners generally allocate more study time to items associated with lower JOLs than to those associated with higher JOLs, suggesting that the allocation of study resources is based in part on the subjective judgments of degree of learning (but see Son & Metcalfe, 2000).

In this study, we compared the effects of repeated study–test cycles on predicted and actual memory performance. A typical learning situation often involves repeated practice. Students preparing for an exam, for example, typically go over the to-be-remembered material several times until they feel that they have attained the desired degree of mastery (Thiede & Dunlosky, 1999). There has been a vast amount of research both in the area of animal learning and in the area of human memory on the so-called learning curve, that is, on the improvement in performance that occurs with repeated practice. This research dates back to the work of Ebbinghaus (1885/1964) and Thorndike (1911). In contrast, there has been little research on the corresponding subjective learning curve, that is, the changes in the subjective sense of mastery (e.g., JOLs) that occur with repeated practice studying the same material. This lack of research is surprising in view of the work in metacognition that indicates that the subjective monitoring of one's own knowledge affects the strategic regulation of learning and remembering processes and, ultimately, memory performance itself (see Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Koriat & Goldsmith, 1996). Although there has been much research and theorizing about the on-line monitoring of learning that occurs in the course of a single study trial, little systematic work has been done on the monitoring that occurs across several trials. With regard to a single self-paced study trial, it has been proposed that learners continuously monitor the increase in encoding strength that occurs as more time is spent studying an item and cease study when a desired level of strength has been reached (e.g., Dunlosky & Hertzog, 1998). A similar process, perhaps, underlies the monitoring of degree of learning across several study trials. For example, in selecting items for restudy, participants generally rely on their subjective assessment of how much they have mastered different items in the list (Thiede & Dunlosky, 1999). Also, T. O. Nelson, Dunlosky, Graf, and Narens (1994) demonstrated that JOLs could be used to benefit multitrial learning. They found that computer-controlled allocation of restudy that was based on people's own JOLs was more effective for learning than an allocation based on normative performance. Thus, it is important to examine how such subjective assessments vary with repeated study.

A comparison of subjective and objective learning curves, that is, of the functions relating predicted and actual performance to repeated practice, should be important for an understanding of the effective management of learning. Assuming that (a) increased practice enhances subsequent memory performance, and that (b) learners decide whether to continue or stop practicing to-be-remembered items on the basis of their JOLs (T. O. Nelson & Leonesio, 1988), then any dissociation between the effects of

Asher Koriat, Limor Sheffer, and Hilit Ma'ayan, Institute of Information Processing and Decision Making, University of Haifa, Haifa, Israel.

A brief report of the results was presented at the 40th Annual Meeting of the Psychonomic Society, Los Angeles, November 1999. Support of this project by the German Federal Ministry of Education and Research (BMBF) within the framework of German–Israeli Project Cooperation (DIP) is gratefully acknowledged. The collection of some of the results reviewed in this project was supported by a grant to Asher Koriat and Robert Bjork from the United States–Israel Binational Science Foundation, Jerusalem, Israel. We are grateful to Ravit Levy-Sadot for her help in all stages of the research program and to Hadas Gutman and Michal Wind for conducting some of the experiments.

Correspondence concerning this article should be addressed to Asher Koriat, Institute of Information Processing and Decision Making, University of Haifa, Haifa 31905, Israel. E-mail: akoriat@research.haifa.ac.il

practice on predicted and actual performance should be detrimental to effective learning.

One kind of such dissociation has been observed by Koriat (1997) in a study of the basis of JOLs. Participants memorized a list of paired associates and, following the study of each pair, provided JOLs regarding their success in recalling, on a later test, the target word (response term) when presented with the cue word (stimulus term). The list was repeated for several study–test cycles (two in Experiment 1 and four in Experiment 2). A comparison of the effects of practice on JOLs and actual memory performance disclosed a pattern that we refer to as the underconfidence-with-practice (UWP) effect: With repeated presentation of the list, JOLs evidenced increased underconfidence, so that recall predictions became markedly lower than recall performance.

The results from Experiment 1 of Koriat (1997) can serve to illustrate the UWP effect. Participants studied a list of 50 paired associates, each presented for 5 s, and immediately after the disappearance of each pair, they made a recall prediction on a scale from 0% to 100%, expressing the chance that they would recall the test word in response to the cue word. The instructions included a detailed description of the conditions of study and test. A cued-recall test then followed. The study–test cycle was repeated a second time. As can be seen in Figure 1, participants were relatively well calibrated in Presentation 1 but became underconfident in Presentation 2: Whereas mean JOLs did not differ significantly from percentage recall on Presentation 1, JOLs were significantly lower than mean recall on Presentation 2, $t(15) = 6.27$, $p < .0001$. Thus, study–test practice impaired calibration, that is, reduced the correspondence between mean overall JOL and mean overall recall in the direction of increased underconfidence. The UWP effect depicted in Figure 1 is quite surprising for three reasons. First, the underconfidence evident on Presentation 2 is at odds with the general tendency for overconfidence that has been observed in a great many calibration studies involving retrospective confidence (see Keren, 1991; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994). In these studies, participants were presented with two-alternative, forced-choice questions (typically general information questions but also questions about a previously witnessed event; see, e.g., Granhag, 1997) and were asked to choose the correct answer and to assess the probability that it was correct. The typical result was that the average assessed probability exceeded the proportion of correct answers. This overconfidence effect has been observed across a wide range of conditions (for reviews, see Ayton & McClelland, 1997; Erev, Wallsten, & Budescu, 1994; Klayman, Soll, Gonzáles-Vallejo, & Barlas, 1999; McClelland & Bolger, 1994).

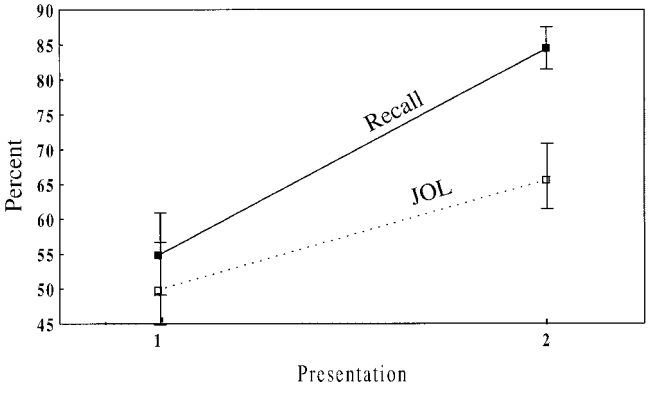

Figure 1. Mean judgment of learning (JOL) and recall for Presentations 1 and 2 in Experiment 1 of Koriat (1997). Error bars represent ± 1 *SE*.

Second, we might have expected calibration to improve with practice. In fact, on the first presentation of the list, participants did not have sufficient information about the items yet to come and about their own recall performance. Therefore, some improvement in calibration may have been expected after they had a chance to learn more about the task and about their own performance level. Instead, however, calibration deteriorated with practice.

Finally, the impairment in calibration with practice contrasts sharply with the observation that resolution actually improves steadily with practice. Calibration (or absolute accuracy; see T. O. Nelson & Dunlosky, 1991) refers to the correspondence between mean JOL and mean recall and reflects the extent to which recall predictions are realistic. Resolution (or relative accuracy) refers to the extent to which JOLs discriminate between recalled and unrecalled items and is commonly indexed by a within-subject gamma correlation between JOLs and recall (see T. O. Nelson, 1984). In Koriat's (1997) Experiment 1, mean gamma correlation actually increased from Presentation 1 (.58) to Presentation 2 (.67). Although this increase was not significant, $t(15) = 1.62$, $p < .13$, results indicating a significant improved resolution in a multitrial learning task have been reported by others (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; King, Zechmeister, & Shaughnessy, 1980; Leonesio & Nelson, 1990; Lovelace, 1984; Mazzoni, Cornoldi, & Marchitelli, 1990). A similar finding has also been noted by Rawson, Dunlosky, and Thiede (2000) for judgments of comprehension: Rereading texts improved the accuracy of comprehension ratings in predicting test performance compared with reading the texts only once. Thus, it is surprising that practice exerts opposite effects on resolution and calibration, improving sensitivity to interitem differences in recall while fostering underconfidence overall.

What is the explanation of the UWP effect? This effect is consistent with Koriat's (1997) cue-utilization model, according to which JOLs are based on a variety of cues that can be grouped into three classes—intrinsic, extrinsic, and mnemonic. Intrinsic cues refer to inherent characteristics of the study items that disclose their a priori difficulty (e.g., associative relatedness between paired associates). Extrinsic cues pertain to the conditions of learning (e.g., number of presentations) or to the encoding operations applied by the learner (e.g., level of processing). Finally, mnemonic cues are internal, subjective indicators that signal to the person the extent to which an item has been mastered (e.g., perceptual fluency and retrieval fluency; see Benjamin & Bjork, 1996).

The UWP effect is consistent with one proposition of the model, which states that in making JOLs, participants pay insufficient regard to the contribution of extrinsic factors relative to that of intrinsic factors (see also Carroll, Nelson, & Kirwan, 1997). The UWP effect accords with this prediction because it implies that the effect of list repetition (an extrinsic factor) is underweighted in the computation of JOLs. A second proposition of the model may explain the improvement in resolution that occurs with practice. According to that proposition, with repeated practice studying a

list of items, a shift occurs in the basis of JOLs from reliance on intrinsic cues toward greater reliance on internal, mnemonic cues. This shift is expected to improve resolution under the assumption that mnemonic cues closely reflect the cognitive processing of different list items. Indeed, the JOL–recall gamma correlation increased with practice, as noted earlier. In parallel, the correlation between JOLs and the judged a priori difficulty of the items (i.e., normative judgments by other participants) decreased systematically with practice, suggesting that the contribution of intrinsic cues to JOLs diminishes with repeated presentations.

The cue-utilization model, however, does not predict the specific pattern observed—increased UWP—and does not offer a process-type explanation of this pattern. In this article, we focus on the UWP effect in an attempt first to demonstrate its robustness and generality and second to provide experimental evidence that helps place some constraints over its explanation. Specifically, we pool results from several sources with the aim of (a) substantiating the existence of the UWP effect, (b) demonstrating its robustness across a variety of manipulations exerted in the context of paired-associate learning, (c) presenting new findings that suggest that the effect holds for other learning tasks as well, and (d) evaluating several theoretical accounts of this rather counterintuitive effect.

A comment is in order about the somewhat unusual format of this article. Much of the results reviewed in the first part of this article come from studies that have been published or will be published elsewhere but most of which were not designed specifically to examine the UWP effect or the conditions that affect it. Therefore, the pertinent statistical analyses are reported here for the first time.[1] Only in the later part of the article do we report new experiments that are designed to extend the study of the UWP effect to other memory tasks.

The Generality of the UWP Effect

We begin by examining the robustness and pervasiveness of the UWP effect in the context of paired-associates learning. We review evidence from previous studies that indicates that the effect is rather robust, surviving several manipulations.

The Effects of Feedback

One possible explanation for the UWP effect is that participants report unduly low JOLs on a second study of the list because they underestimate the correctness of the responses that they supplied on the preceding recall test. If so, feedback about the correctness of the answer produced during recall might reduce or eliminate the impairment in calibration with practice.

Such a feedback manipulation was included in Koriat's (1997) Experiment 2, mentioned earlier. In that experiment, a list of word pairs was presented for four study–test cycles. In a feedback condition, participants were informed on each test trial whether the response that they had just supplied was correct or wrong, whereas in the no-feedback condition no such feedback was provided. The results indicated a clear UWP effect for both the feedback and the no-feedback conditions: The interaction between measure (JOL vs. recall) and presentation was significant for both the no-feedback condition, $F(3, 33) = 7.43$, $MSE = 84.63$, $p < .001$, as well as the feedback condition, $F(3, 33) = 18.57$, $MSE = 43.81$, $p < .0001$. The triple interaction, Measure × Presentation × Condition, was not significant, $F(3, 66) = 1.23$, $MSE = 64.22$. For both conditions, mean recall exceeded mean JOL from the second presentation of the list on (for details, see Figure 4 in Koriat, 1997).[2] Thus, lack of feedback about the correctness of one's responses is not the source of the UWP effect.

Study Time Allocation

A second manipulation that may be expected to moderate the UWP effect concerns the allocation of study time. In typical experiments in which all items are presented for a fixed amount of time, some of the items are presented either for a longer or a shorter duration than what the learner feels is needed. Such is not the case in a self-paced procedure, in which learners are allowed to memorize each pair until they feel they have studied it long enough. In comparison with the self-paced procedure, the fixed-time procedure might yield a tendency for underconfidence if learners inaccurately judge that the amount of time allotted is insufficient to commit an item to memory (see Mazzoni et al., 1990). This possibility could be evaluated using the results from Koriat, Ma'ayan, and Levy-Sadot's (2002) Experiment 1. In that experiment, a list of word pairs was presented for two study–test cycles under one of three conditions, with 20 participants in each condition. Participants in the self-paced condition were allowed to memorize each pair until they felt they had studied it long enough. Participants in the other-paced and fixed conditions had no control over study time. Rather, they were yoked to one of the participants in the self-paced condition either by receiving the exact study time he or she allocated to each individual item (other paced) or by receiving the average study time he or she allocated to all items (fixed). The yoking was carried out separately for each of the two presentations. Mean study time per item was 5.37 s on Presentation 1 and 3.89 s on Presentation 2 for each of the three groups.

The results yielded a UWP effect that proved indifferent to the study time manipulation, because the effect was equally observed under all three conditions. Thus, the interaction between presentation and measure (JOL vs. recall) was significant for each of the three conditions, and when condition was added as a third factor, the triple interaction was not significant. Across all three conditions, JOLs and recall averaged 58.32% and 52.79%, respectively, on the first presentation, indicating some degree of overconfidence, $t(59) = 2.30$, $p < .05$. On the second presentation, in contrast, the respective means were 65.09% and 75.34%, indicating a substantial underconfidence, $t(59) = 4.98$, $p < .0001$. This pattern was observed for each of the three conditions: Mean over/underconfidence (see Lichtenstein et al., 1982) for the first and second presentations, respectively, averaged +8.46 and –9.40 for the self-paced condition, +3.17 and –12.90 for the other-paced condition, and +4.96 and −8.45 for the fixed condition. Thus, the UWP effect seems to be indifferent to whether study time is self-controlled or experimenter-controlled and to whether or not it is evenly distributed across items.

---

[1] A complete report of the as yet unpublished studies is available from Asher Koriat.

[2] Note that as a result of a sampling error, the feedback group's JOLs were consistently higher in that experiment than those of the no-feedback group even on the first presentation of the list.

The Effects of Incentives

A third manipulation that can potentially affect the magnitude of the UWP effect concerns the incentive given for memory performance. It has been observed that in self-paced learning, participants spend more study time on items whose recall is associated with a larger incentive than on those whose recall is associated with a smaller incentive, and memory performance improves accordingly (Dunlosky & Thiede, 1998; Experiment 2). In studies of retrospective confidence, manipulations that improve performance have also been found to improve calibration (see, e.g., Klayman et al., 1999). Thus, it is of interest to see whether incentives for recall can similarly improve calibration. Note, however, that to improve calibration with retrospective confidence judgments, an experimental manipulation must reduce mean confidence to match mean performance (see Koriat et al., 1980), whereas to improve calibration with JOLs, the manipulation must increase confidence (for presentations after the first).

One experiment in Koriat et al.'s study (2002; Experiment 3) included a manipulation of the incentive for correct recall. In one condition of that experiment (differential incentive), some of the pairs received a one-point bonus whereas others received a three-point bonus, and the list was presented for three study–test cycles. In a second condition (constant incentive), a two-point bonus was awarded for all items in the list. The number of points awarded to each item remained constant across the three presentations and was announced just before the presentation of each item. The differential incentive condition yielded significant effects of incentive on study time and JOL but not on recall. Thus, across the three presentations, JOL increased with incentive from 63.0% for one-point items to 66.9% for three-point items, whereas the respective means for recall were practically identical, 68.8% and 68.7%. This pattern implies that higher incentive may improve overall calibration. Therefore, it is of interest to see how incentive level may modulate the UWP effect. As can be seen in Figure 2, a UWP effect was nevertheless found for both incentive levels of the differential-incentive condition (top panel) as well as for the constant-incentive condition (bottom panel). Pooling data across all participants, a Presentation $\times$ Measure (JOL vs. recall) analysis of variance (ANOVA) yielded a significant effect for the interaction, $F(2, 62) = 51.79$, $MSE = 39.36$, $p < .0001$. On the first presentation, JOL and recall averaged 59.1% and 50.2%, respectively, exhibiting a significant overconfidence effect, $t(31) = 3.19$, $p < .005$. On the third presentation, in contrast, the respective means were 75.09% and 84.75%, indicating a marked underconfidence bias, $t(31) = 6.35$, $p < .0001$. Thus, although incentive appears to affect the overall level of over/underconfidence, the UWP effect was equally found for all incentive levels.

The Effects of Associative Relatedness

Another factor that may modulate the UWP effect is whether the members of the word pair are associatively related or unrelated. In Experiment 2 of Koriat (1997), in which feedback was manipulated, the list of 70 pairs consisted of 35 pairs (related) in which the stimulus word elicited the response word as a first associate in 5%–20% of the cases according to word association norms, whereas for the remaining 35 pairs (unrelated), the respective value was 0%. Would the UWP effect be found for both types of word pairs? An analysis across the two feedback conditions indicated that indeed the UWP effect was observed for both sets of items. For the related pairs, JOL and recall averaged 64.75% and 67.31%, respectively, on the first presentation, but 86.42% and 94.64%, respectively, on the fourth presentation. The Measure (JOL vs. recall) $\times$ Presentation interaction was significant, $F(3, 69) = 24.08$, $MSE = 51.41$, $p < .0001$. For the unrelated pairs, the respective means were 28.87% and 19.20% for the first presentation but 62.30% and 71.16% for the fourth presentation, $F(3, 69) = 7.27$, $MSE = 46.68$, $p < .0005$. Thus, despite the marked differences between related and unrelated pairs in predicted and actual recall, both types of pairs yielded increased UWP.

In studies of retrospective confidence, a correlation has generally been observed between probability correct and over/underconfidence such that the lower the proportion correct (i.e., the more difficult the items) the stronger the degree of overconfidence. In fact, for two-alternative choices (when probabilities are assessed on a scale ranging from 0.50 to 1.00) a shift from over- to underconfidence bias occurs close to a proportion correct of 0.75 (see Juslin, Winman, & Olsson, 2000 for a review). In light of these results, it is interesting that the UWP effect occurred for both the related and unrelated pairs despite the fact that the two sets of items differed markedly in recall performance. This result suggests that the UWP effect is independent of the hard–easy effect (Lichtenstein & Fischhoff, 1977; or the "difficulty effect," Griffin & Tversky, 1992), as we discuss later.

Items Producing an Illusion of Knowing

A fifth manipulation that may be pertinent to the UWP effect comes from an experiment by Koriat and Bjork (2001, 2002) concerned with the conditions that engender an illusion of knowing during study. Word pairs with a unidirectional association were included in the study list. For example, the likelihood of *kittens* eliciting *cats* in word association norms is 72%, whereas the likelihood of *cats* eliciting *kittens* is only 2% (Palermo & Jenkins, 1964). Some of the pairs were presented for study in the forward direction (e.g., with *kittens* as the cue word) and some in the backward direction (e.g. with *cats* as the cue word). As was expected, compared with forward and unrelated pairs, the backward pairs produced a strong illusion of knowing during the first presentation, eliciting much higher JOLs (73.52%) than recall (58.36%).

How did associative direction affect the UWP effect? The results indicated that although the backward pairs yielded a significant overconfidence on the first presentation, they too exhibited increased underconfidence from the first to the second presentation. Thus the UWP effect was observed for all three types of pairs: Whereas on the first presentation the signed difference between JOL and recall averaged –0.57, 15.16, and 10.55, respectively, for the forward, backward, and unrelated pairs, for the second presentation it averaged –14.95, –14.83, and –18.16, respectively.

It is important to examine the implications of these findings in light of the claim that the overconfidence effect that has been repeatedly found in retrospective judgments might be a by-product of biased or nonrepresentative sampling of general knowledge questions. The argument is that in most overconfidence studies, the items used included an overrepresentation of tricky or misleading items that were likely to yield high confidence but low probability

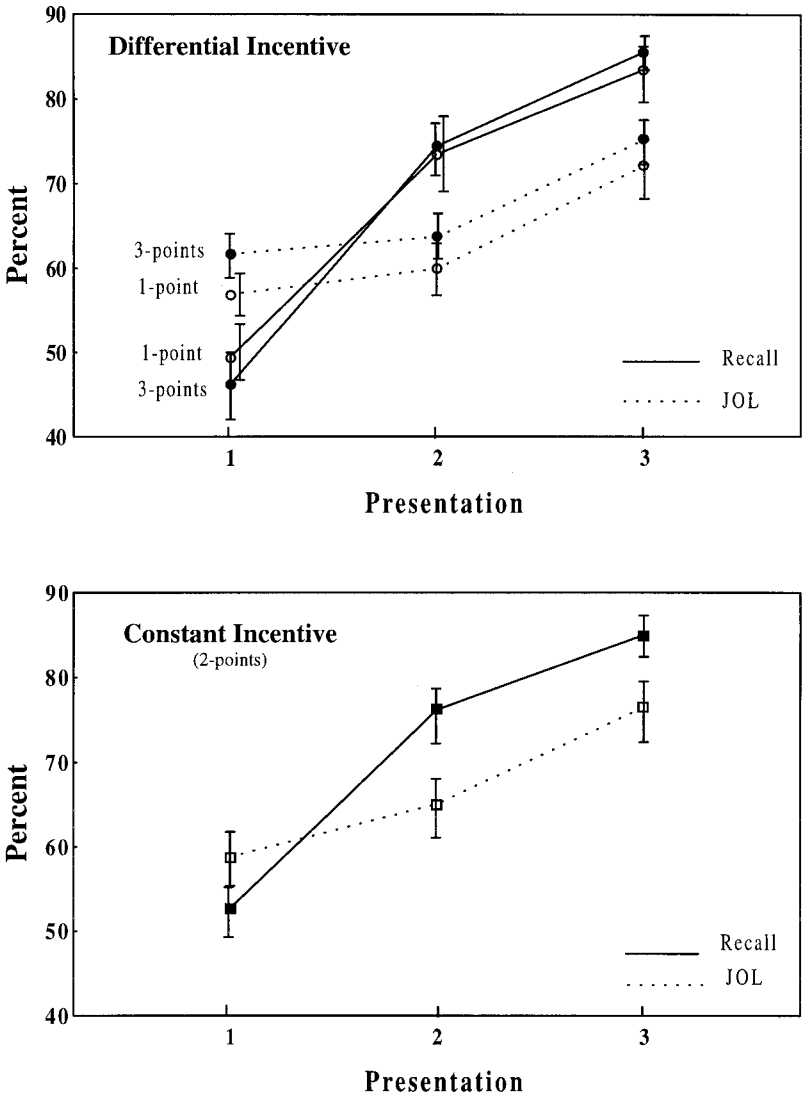

Figure 2. Mean judgment of learning (JOL) and recall as a function of presentation for each incentive level. Top: Results for 1-point and 3-point incentives in the differential-incentive condition. Bottom: Results for the constant (2-point) incentive condition. Error bars represent ± 1 *SE*.

of recall (e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991; see Juslin et al., 2000).

In light of this argument, it is interesting that the backward pairs in the study just mentioned produced a similar degree of underconfidence as the forward and unrelated pairs on the second study–test cycle. Backward pairs can be considered, in a sense, tricky or nonrepresentative (although it is unclear how representativeness could be defined in the context of a paired-associates task). Indeed, they produced marked overconfidence on the first study–test cycle in comparison with the forward items. The observation that they also displayed a UWP effect suggests that this effect may be independent of the kind of methodological biases that might contribute to the observed overconfidence in studies of retrospective judgments.

In summary, although as noted above the manipulations examined in this section were not specifically designed to test hypotheses about the source of the UWP effect, they are still quite informative because they help eliminate several potential accounts of this effect. Overall, the generality of the UWP effect across such varied conditions of paired-associates learning is impressive and testifies to the robustness of this effect.

A Grand Analysis: The UWP Effect Examined Across 11 Studies

In view of the remarkable robustness of the UWP effect, we thought it would be instructive to pool the data across the various experimental conditions in order to obtain a general picture of the changes in the calibration of JOLs that occur with practice. This analysis is important for two reasons. First, Mazzoni and Nelson (1995; Experiment 1) reported marked overconfidence for JOLs (see also W. Schneider, Visé, Lockl, & Nelson, 2000), and some of

our conditions also indicated overconfidence on the first presentation of the list. Therefore, we were interested in examining whether this overconfidence effect survived in the grand analysis. Second, it was important to examine the specific function relating recall and JOLs to presentation. In particular, it was instructive to see whether the increase in UWP continued after the second presentation of the list.

The grand analysis was based on 11 experimental conditions (labeled here "studies"; see Table 1) involving a total of 196 participants. Some of the studies actually represented different conditions in the same experiment (e.g., the feedback and no-feedback conditions in Experiment 2 of Koriat, 1997) and thus had a great deal in common, whereas others came from different experiments. What was common to all these studies was that in each of them (a) a list of paired associates was presented for more than one study–test cycle, and (b) JOLs were solicited immediately after the study of each pair on a 0–100 scale reflecting the assessed probability of recall. Table 1 summarizes some of the details of each study. As can be seen in this table, number of presentations varied between two and four across the 11 studies.

The Effects of Practice on JOL and Recall

Figure 3 presents mean JOL and recall as a function of presentation, based on all participants available for each presentation. A two-way, Measure (JOL vs. recall) × Presentation ANOVA was carried out on these means. The analysis yielded significant effects for presentation, $F(3, 401) = 1{,}090.74$, $MSE = 46.54$, $p < .0001$; for measure, $F(1, 195) = 52.82$, $MSE = 287.64$, $p < .0001$; and for the interaction, $F(3, 401) = 102.00$, $MSE = 53.32$, $p < .0001$.

As can be seen, mean JOL across all observations (65.0%) was lower overall than mean recall (71.7%). However, although both JOL and recall increased strongly with presentation, the function was steeper for recall than for JOL, as indicated by the interaction. In fact, in Presentation 1, mean JOL (55.8%) was higher than mean recall (52.6%), evidencing a significant overconfidence bias, $t(195) = 2.11$, $p < .05$. In contrast, in the presentations following the first, a clear underconfidence of about 10% was observed, with JOLs and recall averaging 69.4% and 81.0%, respectively. The underconfidence effect was significant for the second, $t(195) = 13.21$, $p < .0001$; the third, $t(119) = 10.08$, $p < .0001$; and the fourth presentations, $t(87) = 3.05$, $p < .0001$. Note, however, that the magnitude of underconfidence did not increase from the second presentation onward. In fact, it appeared to decrease somewhat, as indicated by a significant Presentation × Measure interaction in an ANOVA that included only Presentations 2–4, $F(2, 206) = 18.33$, $MSE = 18.37$, $p < .0001$. This interaction possibly derives from a ceiling effect on recall (18% of the participants exhibited perfect recall on the fourth presentation of the list).

In summary, three features should be noted in Figure 3. First, there was a small but significant overconfidence effect on the first presentation. Second, strong evidence was found for a UWP effect in comparing the results between the first and second presentations. Finally, there was no indication that underconfidence increased any further after the second presentation.

The Effects of Practice on Calibration

These observations were also brought to the fore by a calibration analysis. Figure 4 depicts the calibration curves for Presentation 1 and for the remaining presentations combined. These curves were plotted according to the procedure described by Lichtenstein et al. (1982). Mean over/underconfidence for each participant, computed as the weighted mean of the differences between the mean JOL and the percentage of correct recall for the 10 JOL categories

Table 1
A Brief Description of the 11 Studies Included in the Grand Analysis

Study	Source	Specification	No. of participants	No. of items	No. of presentations
1	Koriat (1997)	Experiment 1	16	50	2
2	Koriat (1997)	Experiment 2, feedback condition	12	70	4
3	Koriat (1997)	Experiment 2, no-feedback condition	12	70	4
4	Koriat, Ma'ayan, & Levy-Sadot (2002)	Experiment 1, self-paced condition	20	60	2
5	Koriat, Ma'ayan, & Levy-Sadot (2002)	Experiment 1, other-paced condition	20	60	2
6	Koriat, Ma'ayan, & Levy-Sadot (2002)	Experiment 1, fixed-time condition	20	60	2
7	Koriat, Ma'ayan, & Levy-Sadot (2002)	Experiment 2, self-paced condition	20	60	4
8	Koriat, Ma'ayan, & Levy-Sadot (2002)	Experiment 2, other-paced condition	20	60	4
9	Koriat & Bjork (2002)	Forward–backward experiment	24	48	4
10	Koriat, Ma'ayan, & Levy-Sadot (2002)	Experiment 3, self-paced, differential condition	16	60	3
11	Koriat, Ma'ayan, & Levy-Sadot (2002)	Experiment 3, self-paced, constant condition	16	60	3

(0–10, 11–20, . . . 91–100; see Lichtenstein et al., 1982), averaged 3.57 for Presentation 1 and –12.15 for Presentations 2–4 combined.

The calibration curve for Presentation 1 was very similar to that reported by Dunlosky and Nelson (1992) for JOLs elicited under similar conditions. This curve exhibited the typical pattern of so-called miscalibration observed for retrospective confidence (see Erev et al., 1994; Klayman et al., 1999)—a bias in the direction of underconfidence when JOL is low and a bias in the direction of overconfidence when JOL is high. The function for Presentations 2–4 closely paralleled that for Presentation 1 but exhibited marked underconfidence. However, the magnitude of the underconfidence bias decreased systematically with increasing JOLs, resulting in calibrated judgments only for very high JOLs.

The Effects of Practice on Resolution

Although the results pertaining to resolution were not directly pertinent, the finding that practice improves resolution should also constrain the explanation of the impairment in calibration that occurs with practice. Figure 5 presents mean within-subject gamma correlation between JOLs and recall as a function of presentation across all 11 studies combined. As can be seen in Figure 5, resolution increased systematically from Presentations 1 to 4, $F(3, 374) = 55.49$, $MSE = .042$, $p < .0001$.

A comparison of the resolution function to the calibration function, also plotted in Figure 5, is instructive. The calibration function was obtained by calculating over/underconfidence, as described earlier, for the 11 studies combined. One can see that the resolution function is monotonic, unlike the calibration function, which is almost stepwise. Thus, focusing only on the first two presentations, practice impairs calibration while improving resolution. That is, as a result of practice mean JOLs depart from mean recall in the direction of underconfidence, but the accuracy of JOLs in differentiating between recalled and not recalled items improves.

We were interested in examining the possibility that the changes that occur with practice in both resolution and calibration derive from the same source. Previous work comparing the accuracy of immediate and delayed JOLs (Dunlosky & Nelson, 1994) disclosed a curious finding: Delayed JOLs, which demonstrate a remarkably high relative accuracy (resolution), are typically associated with a polarized distribution of JOLs ratings such that participants tend to use the extreme values of the scale more frequently than the middle values. In contrast, in immediate JOLs, which produce lower JOL accuracy, participants tend to use middle values more often. Koriat and Goldsmith (1996) also noted that in confidence judgments, polarized distributions tend to be associated with better accuracy.

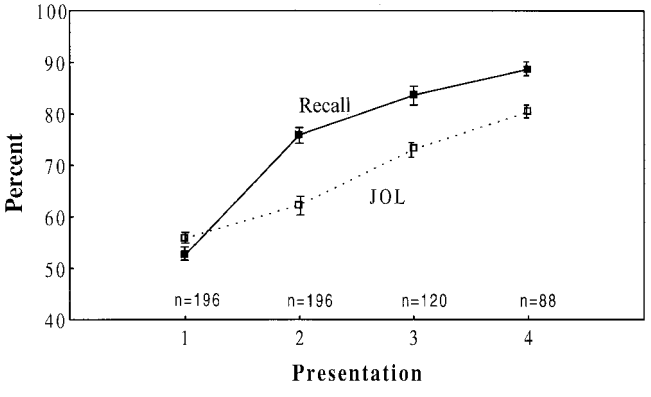

Figure 3. Mean judgment of learning (JOL) and recall as a function of presentation across all 11 studies combined. Error bars represent ± 1 *SE*.

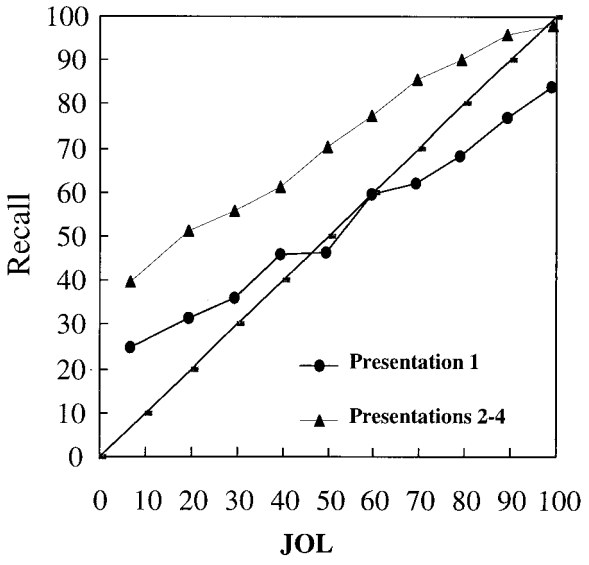

Figure 4. Calibration curves for the 11 studies combined, plotted separately for Presentation 1 and for Presentations 2–4 combined. The diagonal line indicates perfect calibration. JOL = judgment of learning.

Could the increased resolution with practice also reflect increased polarization of JOL ratings, and, if so, could such a change in JOL distribution explain the increased underconfidence? For example, assume that with increased practice participants realize that some items are quite easy whereas others are too difficult to be remembered. The resultant increase in JOL polarization may explain the UWP effect because the increased frequency of 0 values should constrain the increase in the means of JOLs across presentations.[3]

To examine this possibility, we collapsed the JOL values to form five classes: 0–20, 21–40, 41–60, 61–80, and 81–100. Figure 6 displays the frequency distribution of these classes across participants for each of four presentations. One can see that there is no increase in the use of the lower values with practice. Rather, the increase in usage of extreme JOL values is found only at the higher end of the JOL values. Thus, neither the enhanced resolution nor the increased underconfidence can be explained in terms of the change in JOL distributions across presentations.

The UWP Effect for Previously Recalled and Previously Not Recalled Items

Previous studies have shown that JOLs on one study block are highly correlated with the outcome of the memory test on the

[3] We are grateful to John Dunlosky for suggesting this possibility and the analysis that follows.

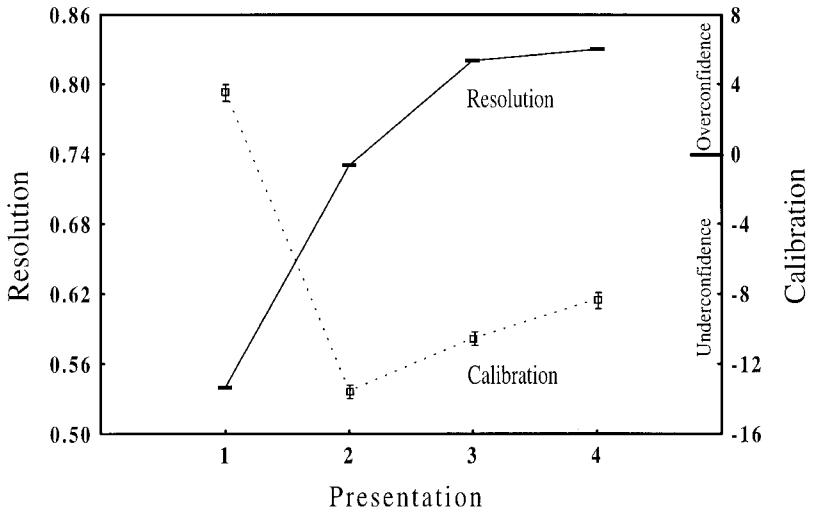

Figure 5. Resolution and calibration as a function of presentation for all 11 studies combined. Resolution was computed as the average within-subject JOL–recall gamma correlation, and calibration was computed as the weighted mean of the difference between JOL and recall across the 10 JOL categories. Error bars represent ± 1 *SE*. Note that for the resolution means, the *SE*s for the four presentations varied from .014 to .039. JOL = judgment of learning.

previous block (see King et al., 1980; Koriat, 1997; Lovelace, 1984; Mazzoni & Cornoldi, 1993). Is the UWP effect modulated by the outcome of previous recall attempts? Because JOLs on one presentation are much higher for items that were recalled on the previous presentation than for those that were not, it is possible that the underconfidence effect observed for Presentation 2 is confined to the items that participants failed to recall on Presentation 1. To examine this possibility, we analyzed the results for Presentation 2 across all 11 studies, comparing for each participant those items that he or she had recalled on Presentation 1 and those that he or she had failed to recall. Consistent with previous findings, higher JOLs were assigned to the previously recalled (82.40%) than to the previously not recalled items (43.80%), $t(189) = 36.59$, $p < .0001$. A similar difference was also found for recall, $t(189) = 30.38$, $p < .0001$, with respective means of 96.78% and 57.17%. It was interesting to find, however, that the two types of items exhibited an underconfidence bias of a similar magnitude on Presentation 2 (amounting to 14.5% and 13.8%, respectively), which was significant for the recalled items, $t(189) = 14.31$, $p < .0001$, as well as for the nonrecalled items, $t(189) = 9.08$, $p < .0001$. In fact, a two-way ANOVA, Recall (recalled vs. not recalled on Presentation 1) × Measure, yielded

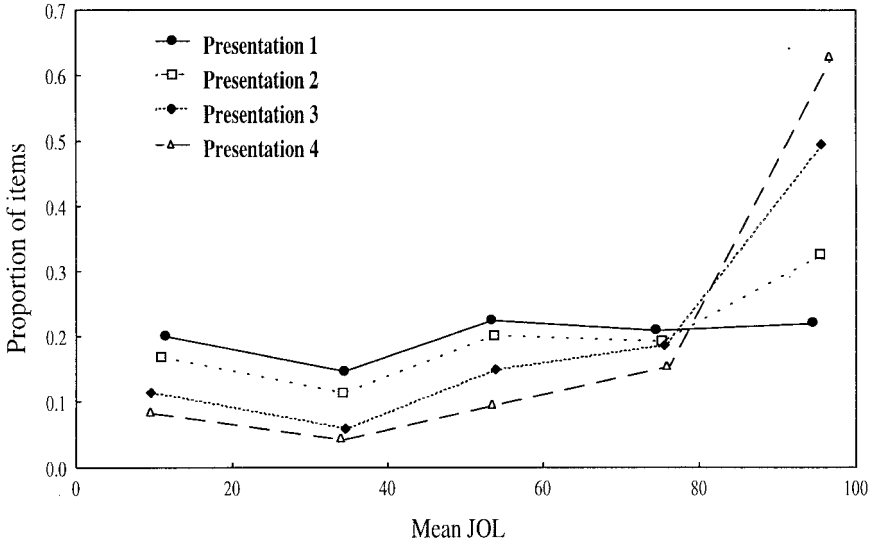

Figure 6. The mean across participants of the percentage of items that received a given judgment of learning (JOL) for each of the five JOL classes (see text) as a function of presentation.

$F < 1$ for the interaction. Thus, the UWP effect on Presentation 2 was not confined to the items that participants failed to recall on Presentation 1, nor was it stronger for such items than for items that were not recalled on that presentation.

Relating the UWP Effect to the Hard–Easy Effect

As noted earlier, a well-replicated effect in studies of retrospective confidence is that when participants are asked to indicate their confidence in their answer to a forced-choice question, the overconfidence effect is reduced as the difficulty of the questions decreases (see, e.g., Gigerenzer et al., 1991; Juslin et al., 2000; Suantak, Bolger, & Ferrell, 1996). In fact, easy items tend to produce a certain degree of underconfidence overall (e.g., Griffin & Tversky, 1992; Lichtenstein & Fischhoff, 1977; Yates, 1990).

Although there is little doubt about the empirical observations that support the hard–easy effect, there has been some debate over the interpretation of these observations. In particular, whereas some authors regard the hard–easy effect as a real phenomenon that deserves a substantive, psychological explanation, others claimed that the effect actually derives from several methodological problems. This debate has been reviewed recently (Juslin et al., 2000), and we shall not discuss it here. However, whatever is the explanation of the hard–easy effect, it is important to note that this effect can provide a framework for the analysis of the UWP effect. Briefly, it may be proposed that practice improves performance so that the effects of practice may be seen to parallel the shift from hard to easy items. Therefore, an attractive working hypothesis is that the processes responsible for the UWP effect are the same as those underlying the hard–easy effect.

To examine this possibility, we had first to test whether the hard–easy effect was obtained for JOLs as it was for retrospective confidence. We carried out two analyses that differed in the way in which item difficulty was operationalized (see, e.g., Juslin, 1993). In the first, item difficulty was defined on the basis of normative data, whereas in the second it was defined post hoc, on the basis of recall probability. These two analyses yielded inconsistent results.

In the first analysis, we took advantage of the fact that we had available an independent measure of item difficulty for the lists used in the 11 studies. In all, four different lists were used across these studies. For two Hebrew lists (one used in Study 1 and the other used in Studies 4–8 and 10–11), this measure was subjective memorability ratings provided by independent groups of participants.[4] For the other two lists (one in Hebrew in Studies 2 and 3 and one in English in Study 9), the measure was based on word association norms (see Footnote 4 for a description of these measures).

The items in each list were classified as hard and easy according to the independent measures just mentioned. Figure 7 (top panel) depicts mean JOL and recall as a function of presentation, plotted separately for easy and hard items. Focusing first on Presentation 1, it can be seen that there was a slight tendency for overconfidence, $F(1, 195) = 5.08$, $MSE = 441.90$, $p < .05$. This tendency was equally observed for the hard and easy items, as indicated by $F < 1$ for the Difficulty × Measure interaction.

This result suggested that the UWP effect cannot be explained in terms of a change in the difficulty of the items as a result of learning. Indeed, inspection of the effects of practice on hard and easy items (Figure 7, top panel) indicates that although easy items

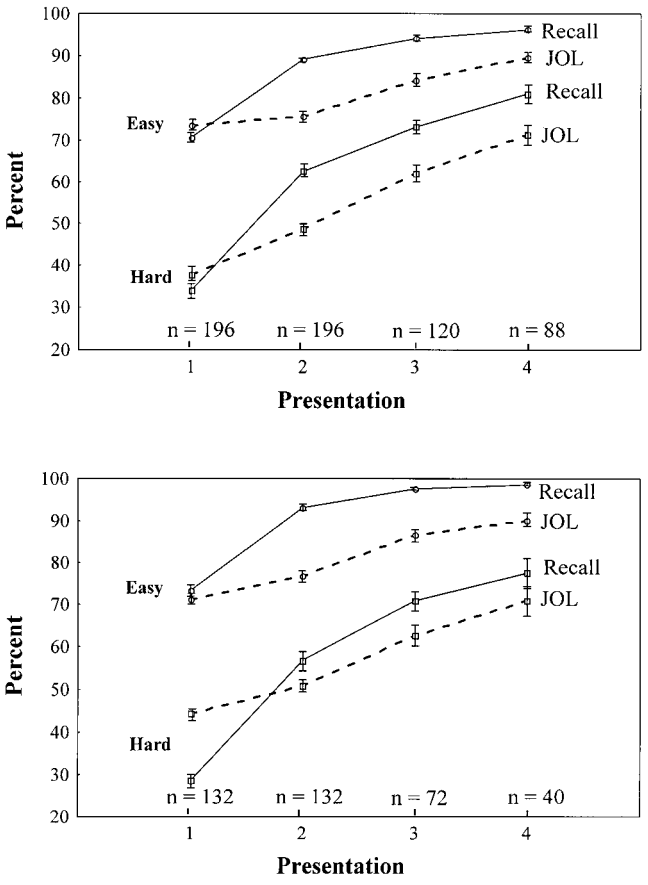

Figure 7. Mean judgment of learning (JOL) and recall as a function of presentation for hard and easy items. Top: Classification of items as hard or easy is based on independent measures. Bottom: Classification of items based on participants' actual memory performance. Error bars represent ± 1 *SE*.

yielded higher JOLs and recall overall than hard items, the two types of items exhibited very similar effects of practice. Because of the possibility of a ceiling effect for JOL and recall on the last presentations, we conducted a three-way ANOVA, Presentation × Measure × Difficulty, using only the data from Presentations 1 and 2 (for which we had data from all participants). The results yielded significant effects for presentation, $F(1, 195) = 912.83$, $MSE = 96.85$, $p < .0001$; for measure, $F(1, 195) = 19.64$, $MSE = 525.60$, $p < .0001$; and for difficulty, $F(1, 195) = 1{,}261.85$,

[4] Memorability ratings were obtained from different groups of participants than those that served in the experiments proper. For each list, a different group of participants was instructed to imagine that 100 people had been asked to learn a list of word pairs so that they could later recall the response word when shown the stimulus word. The participants were asked to estimate, for each pair, how many people would be likely to recall the correct response. The median memorability rating for each list was used to define the pairs as hard (below median) or easy (above median). This median was 45.00% for one list and 48.11% for the other. In the experiments in which item difficulty was defined in terms of associative norms, items were classified as easy when the stimulus elicited the response word with a probability of .05 or more and hard when associative strength was 0.

$MSE = 309.97$, $p < .0001$. More important, the Presentation × Measure interaction was significant, $F(1, 195) = 224.86$, $MSE = 126.41$, $p < .0001$, but neither the Difficulty × Measure interaction nor the triple interaction was significant, $F < 1$ and $F(1, 195) = 1.46$, $MSE = 40.88$, respectively. Thus, the hard and easy items did not differ either in the pattern of overconfidence exhibited on the first presentation or in the magnitude of the UWP effect.

In the second analysis, item difficulty was defined in terms of recall probability. Although this analysis is problematic because of the post hoc classification of the items (see Juslin, 1993), it is the one that is most common in confidence studies of the hard–easy effect. In this analysis, we included only the data from the seven studies in which the same list of Hebrew pairs was used (Studies 4–8, 10–11). For each presentation, the items were divided into hard and easy items in terms of the median of recall performance on that presentation. Figure 7 (bottom panel) presents the results obtained with this post hoc classification of items. Although the overall pattern is quite similar to that depicted at the top panel, the results on Presentation 1 suggest a hard–easy effect for JOLs. The hard items, but not the easy items, exhibited overconfidence, resulting in a significant Difficulty × Measure interaction, $F(1, 131) = 221.78$, $MSE = 49.01$, $p < .0001$. The overconfidence exhibited on the first presentation disappeared from the second presentation on, but the pattern for Presentations 2–4 indicated a reduced amount of underconfidence for the hard items. Indeed, a Presentation × Difficulty × Measure ANOVA for these presentations yielded a significant effect for measure, $F(1, 131) = 105.33$, $MSE = 237.82$, $p < .0001$, indicating an underconfidence bias, but the magnitude of underconfidence was smaller for the hard items, as indicated by a Measure × Difficulty interaction, $F(1, 131) = 35.05$, $MSE = 75.61$, $p < .0001$. On the whole, however, a clear UWP effect is evident for both types of items: A Presentation (4) × Measure (2) ANOVA yielded a significant interaction for both the easy items, $F(3, 241) = 32.04$, $MSE = 68.87$, $p < .0001$, and the hard items, $F(3, 241) = 122.91$, $MSE = 60.55$, $p < .0001$.

In conclusion, when item difficulty was defined on the basis of normative data, the results failed to yield any evidence for a hard–easy effect for JOLs. In contrast, when item difficulty was defined post hoc, in terms of actual recall performance, a hard–easy effect was obtained, which was most pronounced for the first presentation. This pattern of discrepant results paralleled that observed with regard to retrospective confidence judgments (Juslin, 1993). However, even when item difficulty was defined on the basis of recall probability, examination of the changes that occurred with practice indicated a clear UWP effect that appeared to be operating over and above the hard–easy effect. These results rejected the possibility that the effects of practice on the pattern of over/underconfidence were mediated by the effects of item difficulty.

The UWP Effect With Aggregate Judgments

In studies of retrospective confidence, a distinction has been drawn between two methods of eliciting probability judgments. In the item-by-item (or confidence) method, participants assess the probability that the answer to each single item is correct. In the aggregate (or frequency) method, in contrast, participants estimate the frequency of correct items across a series of items (Gigerenzer et al., 1991; Granhag, 1997; Griffin & Tversky, 1992; Juslin, 1993; Treadwell & Nelson, 1996; see also Slovic, Monahan, & MacGregor, 2000). A consistent finding that has been reported in comparing these two methods is that aggregate judgments, when transformed into percentages, are substantially lower than item-by-item judgments. In fact, whereas the item-by-item method typically yields overconfidence, aggregate judgments do not exhibit overconfidence and sometimes even yield underconfidence (Griffin & Tversky, 1992; S. L. Schneider, 1995; but see Keren, 1991). Several accounts for this discrepancy have been proposed. A question of interest in the present context was whether the UWP effect for JOLs was also found for the aggregate measure.

Mazzoni and Nelson (1995) speculated that people might be overconfident in item-by-item JOLs because each item may seem more recallable immediately after study than will subsequently be the case. However, people might actually be underconfident in their aggregate JOLs because after studying a long list, people are aware that they typically do not recall such a large number of items. Indeed, using list learning (rather than paired associates), they observed an aggregation effect for JOLs: Whereas item-by-item JOLs yielded overconfidence, aggregate judgments yielded underconfidence. Results obtained by Connor, Dunlosky, and Hertzog (1997) also suggested a trend toward underconfidence for aggregate JOLs in young adults, and W. Schneider et al. (2000) also replicated the aggregation effect for JOLs in children. Neither of these studies, however, examined the changes that occur in recall predictions with practice learning the same list.

In Experiment 2 of Koriat et al. (2002), which included a self-paced as well as an other-paced condition (see Studies 7 and 8 in Table 1), in addition to the item-by-item JOLs solicited at the end of each trial, we also obtained an aggregate prediction of recall. The following instruction was displayed on the computer screen at the end of each of the four study phases: "We showed you 60 word pairs. For how many of these do you think you will recall the correct response when presented with the cue word?" The results of the aggregate measure were not included in Koriat et al.'s report, and we present them here in some detail.

The aggregate estimates for each participant were transformed into percentage scores to allow comparison with mean item-by-item JOLs. Figure 8 presents the means of item-by-item JOLs, aggregate JOLs, and recall as a function of presentation for both the self-paced (top panel) and other-paced (bottom panel) conditions. Two patterns in this figure are noteworthy. First, the aggregate measure yielded lower overall predictions in both conditions than the item-by-item JOLs. Second, a UWP effect was observed not only for item-by-item JOLs, but also for the aggregate estimates. Because the results were very similar for the self- and other-paced conditions, we combined data across both conditions in the following analyses.

Let us first compare the results for the two recall predictions. A Presentation × Judgment ANOVA comparing the aggregate and item-by-item judgments yielded significant effects for presentation, $F(3, 117) = 69.83$, $MSE = 171.26$, $p < .0001$; for judgment type, $F(1, 39) = 43.12$, $MSE = 215.35$, $p < .0001$; and for the interaction, $F(3, 117) = 4.23$, $MSE = 33.31$, $p < .01$. The aggregate estimate was lower overall (57.54%) than the mean of the individual JOLs (68.32%), $t(39) = 6.57$, $p < .0001$. Note that this was also true on the first presentation, for which the respective

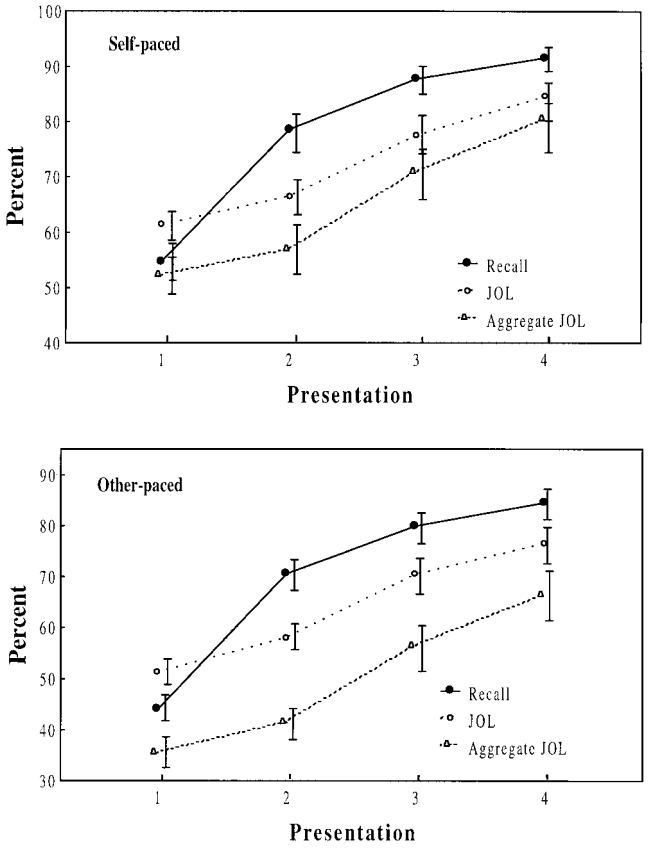

Figure 8. Mean item-by-item judgments of learning (JOLs), aggregate JOLs, and recall as a function of presentation for the self-paced (top) and other-paced (bottom) conditions. Error bars represent ± 1 *SE*.

means were 43.88% and 56.34%, $t(39) = 5.38$, $p < .0001$. This pattern was consistent with the results reported by Mazzoni and Nelson (1995) and by W. Schneider et al. (2000). The interaction apparently derived from the somewhat stronger effects of practice on aggregate than on item-by-item judgments.

Turning next to the calibration of aggregate judgments, a Presentation × Measure ANOVA comparing overall predicted recall (aggregate) and actual recall yielded significant effects for presentation, $F(3, 117) = 228.24$, $MSE = 78.05$, $p < .0001$; for measure, $F(1, 39) = 81.43$, $MSE = 265.06$, $p < .0001$; and for the interaction, $F(3, 117) = 18.67$, $MSE = 76.86$, $p < .0001$. Whereas on the first presentation of the list mean predicted recall (43.88%) did not differ significantly from mean actual recall (49.41%), on Presentations 2–4 the respective means averaged 62.11% and 82.16%, thus indicating a considerable underconfidence bias amounting to about 20%!

The underconfidence disclosed by aggregate judgments was quite impressive. As can be seen in Figure 8, after the third study phase, participants expected to recall fewer items (mean = 63.63% across the self-paced and other-paced conditions) than they actually recalled on the second test (74.57%), and after the fourth study phase they expected to recall fewer items (73.42%) than they did on the preceding, third test (83.84%)!

Note that item-by-item JOLs exhibited a certain degree of overconfidence on the first presentation, $t(39) = 2.41$, $p < .05$, whereas the aggregate judgments evidenced a near-significant trend toward underconfidence, $t(39) = 1.98$, $p < .06$ (see also Mazzoni & Nelson, 1995).

A methodological problem with the design of the experiment was that the aggregate estimates obtained at the end of each study phase may have been influenced by the item-by-item JOLs that the participant had reported earlier. However, in Mazzoni and Nelson's study (1995; Experiment 1) some of the participants provided only aggregate predictions, and these predictions were found to be similar to those made by other participants who had also made item-by-item JOLs. Nevertheless, it is still important to replicate the UWP effect for aggregate judgments under conditions in which item-by-item JOLs are not solicited.

In summary, the UWP effect was obtained even with aggregate judgments, with participants greatly underestimating their prospective recall on the presentations following the first. The finding that aggregate judgments were lower than mean item-by-item JOLs is consistent with the aggregation effect obtained in studies of retrospective confidence (e.g., Gigerenzer et al., 1991; Juslin, 1993; Griffin & Tversky, 1992) and may indicate that a common factor underlies the recall predictions made during study and the retrospective assessments of one's own performance.

Extending the UWP Effect to List Learning

The experiments reported so far have concentrated on the paired-associates task. Although this task has figured prominently in verbal learning research as a model of associative learning and cued recall (see, e.g., D. L. Nelson, McKinney, Gee, & Janczura, 1998), we were interested in examining whether the UWP effect generalized to other tasks as well. This examination is important not only for delimiting the boundaries of the UWP phenomenon, but also because it can help determine whether the explanation of this phenomenon should be sought in specific features of the paired-associates task. Does the UWP effect reflect a general tendency of learners to underestimate the benefits from practice and is it therefore characteristic of learning in general? In this and the following section, we report the results of two new experiments that addressed this question.

The first experiment examined the question of whether the increased UWP is also found in the context of a free-recall, list-learning paradigm. Participants were presented with a list of words and made JOLs at the end of each study trial. Their free recall of these words was then tested.

Method

Participants. Twenty Hebrew-speaking undergraduates at the University of Haifa were paid for participating in the experiment.

Apparatus and procedure. The experiment was conducted on a Silicon Graphics personal computer. A list of 40 common Hebrew words was used. Each word was presented on the computer screen for 5 s, and participants indicated their JOLs about each item as soon as it disappeared from the screen. JOLs were made on a 0%–100% scale reflecting the likelihood of recalling the word on a subsequent free-recall test. The study–test was repeated four times in total. The order of presentation of the words was randomly determined for each participant for each block.

Results

Mean recalls for the four presentations were 37.13%, 54.50%, 70.88%, and 76.75%, respectively. In comparison, the respective

means for JOLs were 55.48%, 57.76%, 64.45%, and 73.69%. The interaction between measure (JOL vs. recall) and presentation was significant, $F(3, 57) = 21.60$, $MSE = 55.75$, $p < .0001$. Thus, there was a tendency toward increased UWP. The pattern, however, was somewhat different from that observed for the paired-associates task: Whereas predicted recall was significantly higher than actual recall on the first cycle, $t(19) = 4.28$, $p < .001$, the tendency for underconfidence was significant only for the third cycle, $t(19) = 2.99$, $p < .01$. Thus, it seems that the UWP effect is obtained even for a free-recall task.

Extending the UWP Effect to Memory for Action

The second experiment concerned the monitoring of memory for self-performed tasks (SPTs; see review by Engelkamp, 1998). We were particularly interested in extending the investigation of the UWP effect to SPT memory because of the claim that people sometimes have severe difficulties in monitoring their own actions (e.g., Cohen, Sandler, & Keglevich, 1991; Koriat, Ben-Zur, & Druch, 1991).

Method

Participants. Twenty Hebrew-speaking undergraduates at the University of Haifa were paid for participating in the experiment.

Stimulus materials. A list of 30 so called Tumai words was constructed. They were one- to three-syllable pronounceable nonsense strings that were chosen so that they evoked little definite associations among Hebrew speakers. These words were randomly paired with 30 Hebrew phrases denoting minitasks, such as "touch your ear," and "stand up" (see Koriat, Pearlman-Avnion, & Ben-Zur, 1998). Some of the minitasks required the manipulation of an external object (e.g., "smell the flower"), whereas others involved mainly bodily actions (e.g., "lick your lips").

Apparatus and procedure. The apparatus was the same as in the previous experiment. Participants were told that the experiment concerned the memory for the meaning of words from an African language called Tumai, with each word denoting a particular action. They were instructed that they would have to study the meaning of these words by performing the actions that they denoted and indicate their JOLs at the end of each trial. When the task required the manipulation of an external object, they had to imagine the appropriate object and pantomime as if it were there. Participants were told that in the test phase they would see each Tumai word in turn and would be asked to recall the corresponding action by performing it.

The experiment involved four study–test cycles. In the study phase of each cycle, each Tumai word and its corresponding action phrase appeared at the center of the screen side by side for 8 s. Participants were instructed to study the meaning of each Tumai word by performing the action that it denoted. They were urged to use the entire 8 s for studying. The pair was then replaced by the statement "Probability to Recall:". Participants reported their estimate orally on a 0%–100% scale.

During the test phase, the 30 stimulus words were presented one after the other for up to 8 s each. Participants had to perform the corresponding action within the 8 s allotted and say the phrase aloud so that the experimenter could record it. One second thereafter a beep was sounded and the next stimulus word was presented.

The pairing of the action phrases with the Tumai words was random for each participant (but remained constant across presentations). The order of presentation of the items was randomly determined for each participant and for each study and test block.

Results

Mean recalls for the four presentations were 16.83%, 50.00%, 71.50%, and 83.33%, respectively, whereas the respective means for JOLs were 36.58%, 37.08%, 59.40%, and 73.09%. A Presentation × Measure ANOVA yielded $F(3, 57) = 36.52$, $MSE = 68.25$, $p < .0001$, for the interaction. Thus, a UWP effect was found for the memory for action events. The results for Presentation 1 indicated a significant overconfidence effect, $t(19) = 5.95$, $p < .0001$. All subsequent presentations yielded a significant underconfidence bias, $t(19) = 3.34$, $p < .005$; $t(19) = 4.28$, $p < .0005$; and $t(19) = 3.58$, $p < .005$, for Presentations 2–4, respectively.

In summary, the results presented in this section and the previous section suggested that the increased tendency to underestimate one's performance with practice may generalize to learning tasks other than the study of word pairs.

General Discussion

In this article, we explored a curious phenomenon concerning the monitoring of one's own knowledge during learning. The results suggest a dissociation between objective and subjective learning curves such that learners systematically underestimate the benefits from practice on memory performance.

Let us first summarize the main findings. First, across all experiments, there was increased underestimation of one's future recall performance on a repeated presentation of the study materials. Whereas on the first presentation there was a slight tendency toward overconfidence, this tendency changed to a marked underconfidence from the second presentation on (from the third presentation on for the free-recall task). This finding implies an underestimation of the effect of repetition. For example, the paired-associates data indicate that whereas recall increased on the average by about 23% from the first to the second presentation, JOLs increased by only 6%. With paired associates, there was no further increase in underconfidence beyond the second presentation, although participants continued to underestimate their recall even on the fourth presentation.

Second, the UWP effect was found to be very robust, surviving several experimental manipulations. Thus, a UWP effect was observed even when participants were given feedback about the correctness of their recall responses, suggesting that the effect does not derive from a tendency to underestimate the correctness of the recalled responses. Neither was the UWP effect sensitive to the mode of study time allocation: It was observed not only for a fixed-rate presentation, but also for a self-paced condition. The effect was also indifferent to the incentive for correct recall: Although a higher incentive increased initial JOLs, the UWP effect was equally obtained for items associated with high and low incentives. Similarly, although the presence of a backward (response-to-cue) association enhanced initial JOLs to the extent of generating an illusion of competence (i.e., unduly high JOLs; see Koriat & Bjork, 2001), a UWP effect was observed even for backward-associated pairs. An underconfidence bias on the second presentation of the list occurred equally both for the items for which recall had been successful on the previous test and for those in which recall failed. Finally, the effect was found for related as well as for unrelated word pairs and, in general, for easy and hard items. Altogether these results support the robustness of the UWP effect and help in eliminating several possible accounts of this effect.

Third, the impaired calibration with practice was found despite a clear improvement in resolution. Thus, consistent with previous reports (e.g., King et al., 1980; Lovelace, 1984), we found practice improves learners' relative accuracy, that is, their ability to discriminate between items that will be recalled and those that will not. At the same time, absolute accuracy deteriorated so that the discrepancy between mean predicted and actual recall widened. Note, however, that whereas the drop in confidence occurred between the first and second presentations, resolution improved monotonically with practice.

Fourth, the increase in UWP was also found for aggregate JOLs (prediction estimates of recall frequency), indicating that the UWP effect is not response specific. For example, on the second presentation, participants' aggregate JOLs underestimated the number of subsequently recalled words by 25%. Note that in discussing retrospective confidence, some researchers have argued that confidence judgments and frequency estimates are based on different types of evidence (e.g., Griffin & Tversky, 1992). The present results do not support such an argument with respect to JOLs, because apart from the fact that the aggregate judgments were overall lower than the item-by-item JOLs (as was also found to be the case with retrospective confidence), the effects of practice were very similar for the two types of judgments.

Finally, results from two experiments suggest that the UWP effect may generalize beyond the task of studying the association between words for future cued recall. Thus, it seems to generalize to the free recall of a word list and to the recall of action events. Of course, more work is needed before we can conclude that the UWP effect is characteristic of learning in general.

Taken together, the results reported in this article document an interesting phenomenon that deserves further investigation. This phenomenon has escaped notice so far primarily because much of the work on learning in general, and on verbal learning in particular, has traditionally confined itself to performance measures of learning (see Bjork, 1999). The recent upsurge of interest in metacognition, however, has brought to the fore the importance of considering the processes underlying the subjective monitoring of one's own knowledge. Not only are these processes of interest in their own right (e.g., the processes that lead to overconfidence and illusions of knowing), but they also affect actual memory performance (see Koriat, 2000; Koriat & Goldsmith, 1996; T. O. Nelson, 1996). For example, JOLs seem to affect the amount of time allocated to different items in a list as well as the selection of items for restudy (e.g., Dunlosky & Hertzog, 1998; Mazzoni & Cornoldi, 1993; T. O. Nelson & Leonesio, 1988; Thiede & Dunlosky, 1999). It is therefore surprising that although a vast amount of research has been done on the effects of repetition on performance measures of learning (the learning curve), little effort has been invested in the examination of the corresponding function relating subjective measures of learning to number of repetitions (the subjective learning curve). An important advantage of the methodology used in this study, that of obtaining measures of JOLs in the form of assessed probability (as is typically done in studies of subjective confidence), is that it allows objective and subjective learning curves to be directly compared. It is this type of comparison that disclosed the UWP effect.

Although much of the evidence for the UWP effect rests so far on one experimental paradigm—the study of paired associates for subsequent cued-recall testing—the results that we obtained with other paradigms hint at the possibility that the effect is more general. In discussing possible explanations of the UWP effect, however, we shall focus on the results from the paired-associates paradigm, for which this effect is clearly robust and pervasive.

Why then do people become underconfident in their subsequent memory performance when a study list is presented again? Paradoxically, it is precisely because the UWP effect turned out to be so robust that it is difficult to identify a lead toward a satisfactory explanation of this phenomenon. Therefore, our aim here is only to point out several possible directions that such explanations might take.

One explanation that can be immediately eliminated is that the UWP effect derives from the discrepancy between the time at which JOLs are solicited and the time at which recall is tested. If this were so, we should have found an overestimation rather than an underestimation of JOL predictions.

There are five other potential accounts of the UWP effect that we can offer. The first is that the UWP effect reflects an overshoot of the aggregate JOLs made (explicitly or implicitly) on a previous presentation. Perhaps the general impression that learners are left with after studying the entire list is best reflected in their aggregate (frequency) judgments rather than in the mean of item-by-item JOLs. Because the former are known to yield lower predictions (for reasons that are beyond the focus of this article), perhaps these predictions affect the item-by-item JOLs reported on the subsequent study presentation. One observation that argues against this explanation is that participants seem to also underestimate the effects of a repetition of items within the same list in the same way that they underestimate the effects of repetition of the entire list (see Koriat, 1997; Experiment 3).

A second account, which we have considered at some length, is that the UWP effect is essentially another manifestation of the hard–easy effect that has been documented in studies of retrospective confidence. Although there is still disagreement regarding the explanation of the hard–easy effect itself (Gigerenzer et al., 1991; Griffin & Tversky, 1992; Juslin et al., 2000; Lichtenstein & Fischhoff, 1977; Yates, 1990), a reduction of the UWP effect to the hard–easy effect would motivate a search for a common mechanism. The general idea is that the effects of practice may be conceived as involving a decline in the difficulty of the items, so that the effects of practice on the extent of over/underconfidence should parallel the discrepancy between hard and easy items.

This account, however, was not supported by the results. When the classification of items as hard or easy was based on criteria that were independent of participants' performance, no evidence for a hard–easy effect was found for JOLs. In contrast, when the classification was based on participants' actual recall performance, JOLs exhibited a hard–easy effect that was most pronounced on the first presentation. Even then, however, the UWP effect was found for both hard and easy items and appeared to occur over and above the hard–easy effect. Thus, the hard items, however defined, produced underconfidence from the second presentation on. These results suggest that the mechanism responsible for the UWP effect differs from that underlying the hard–easy effect.

A third potential account is the one originally proposed by Koriat (1997), namely, that the UWP effect represents a special case of the general tendency of learners to discount the effects of extrinsic factors in making JOLs. Extrinsic factors include the conditions of learning as well as the encoding operations that are

applied by the learner. The rationale for this hypothesis is that JOLs are comparative in nature. Therefore they are more sensitive to the relative memorability of different items within a list than to extrinsic factors that affect overall performance (see Begg et al., 1989). As Shaw and Craik (1989) argued, people "are largely unaware of memory effects associated with different mental processes but are somewhat sensitive to the effects associated with different materials" (p. 134).

This account gains some support from the finding that other extrinsic factors apart from list repetition also exert weaker effects on JOLs than on recall. These factors include, for example, depth of processing (Cutting, 1975; Shaw & Craik, 1989), elaborative versus maintenance rehearsal (Shaughnessy, 1981), interactive imagery instructions (Rabinowitz, Ackerman, Craik, & Hinchley, 1982), within-list item repetition, and stimulus duration (Koriat, 1997). Of particular interest is the finding of Carroll et al. (1997): They had participants study a list of related and unrelated paired associates to a criterion of two and eight correct recalls, respectively, resulting in better recall for the unrelated pairs. JOLs nevertheless displayed the opposite effect, being higher for the related pairs. This pattern suggests that the extrinsic factor of degree of learning is undervalued relative to that of the intrinsic factor of semantic relatedness. Note, however, that some of the studies reported in the literature failed to demonstrate a tendency to undervalue the effects of extrinsic factors (e.g., Begg, Vinski, Frankovich, & Holgate, 1991).

As already noted, however, although the UWP effect is consistent with the proposition that learners tend to undervalue the effects of extrinsic factors, it does not follow from it, because one can imagine a situation in which this tendency results in improved rather than in impaired calibration. As noted in the introduction, what is perplexing about the UWP effect is that it discloses an impairment in the subjective monitoring of one's own competence precisely where one would expect an improvement. Furthermore, the proposition that learners underestimate the contributions of extrinsic factors relative to those of intrinsic factors is, at best, a useful descriptive generalization. In order for this account to provide an explanation, it must be supplemented with a more complete specification of the cognitive mechanism responsible for the discounting of extrinsic factors in making JOLs.

The fourth and fifth accounts, finally, are presently the most promising and deserve investigation. The fourth account follows from the recent work by Runeson, Juslin, and Olsson (2000). That work suggests another way in which Koriat's (1997) cue-utilization model could be extended to account for the UWP effect. Runeson et al. studied the effects of practice on the discrimination of relative mass in observed collisions, contrasting predictions derived from constructivist (indirect) and Gibsonian (direct) approaches to perception. They proposed that early in training observers' performance accords better with the indirect view, reflecting the use of simple cues in a cognitive–inferential process. With practice, however, a shift occurs toward greater use of information in a direct–perceptual mode of apprehension. Capitalizing on the findings that cognitive or inferential tasks tend to be characterized either by overconfidence or by good calibration whereas sensory tasks tend to lead to underconfidence, Runeson et al. observed a shift toward greater underconfidence with practice performing the task. This finding was taken to suggest a shift that occurs with practice from a cognitive–inferential mode toward a more perceptual–intuitive mode in performing the task.

The distinction between the cognitive and perceptual modes of operation discussed by Runeson et al. bears some similarity to Koriat's (1997) distinction between theory-based and experience-based judgments (see also Koriat & Levy-Sadot, 1999). As noted earlier, Koriat (1997) proposed that with repeated practice studying a list of items, the basis of JOLs changes from reliance on the explicit application of rules toward increased reliance on mnemonic-driven subjective experience. This shift was seen to explain the improved resolution with practice, but Runeson et al's analysis suggests that it might also explain the increased UWP. Thus, it is possible that both the improved resolution and the impaired calibration that occur with practice (see Figure 5) result from the same process—increased reliance on subjective experience. At present, this possibility is highly speculative, but it certainly merits consideration.

The fifth account involves the distinction between the effects of study and test experience. Bjork and Bjork (1992) emphasized the fact that both study (or practice) experience and test (or retrieval) experience may increase the likelihood of future recall. In terms of their conceptual framework, study experience contributes to storage strength, whereas retrieval experience may help build retrieval strength. Furthermore, they postulated that the effects of retrieval strength are strongest when storage strength is weakest. Thus, retrieval experience may be particularly beneficial the more effortful retrieval is.

It may be proposed that learners are generally cognizant of the beneficial effects of study experience for future recall, and hence JOLs generally reflect these effects. In contrast, they are generally unaware that merely retrieving an item from memory can also facilitate its subsequent retrieval. In particular, participants "are apparently unaware that the more difficult or involved the process of retrieval, provided it succeeds, the greater its impact on subsequent recall" (Bjork, 1999, p. 451; see Benjamin, Bjork, & Schwartz, 1998). Thus, perhaps the UWP effect reflects the failure to take into account the beneficial effects of recall experience when making JOLs. If indeed recall experience is also most profitable when storage strength is low, this may also explain why the UWP effect is most pronounced between the first and second presentations. These hypotheses are currently under systematic investigation.

In summary, we have sketched several possible accounts of the UWP effect, but there are of course other potential accounts. It is very likely that more than one process contributes to this effect. The results presented in this review, although substantiating the generality and robustness of the UWP effect, also help place some constraints on its explanation. However, further research is clearly needed to clarify the puzzle of the UWP phenomenon.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81,* 126–131.

Ayton, P., & McClelland, A. G. R. (1997). How real is overconfidence? *Journal of Behavioral Decision Processes, 10,* 153–285.

Barnes, A. E., Nelson, T. O., Dunlosky, J., Mazzoni, G., & Narens, L. (1999). An integrative system of metamemory components involved in retrieval. In D. Gopher & A. Koriat (Eds.), *Attention and performance*

XVII—Cognitive regulation of performance: Interaction of theory and application (pp. 287–313). Cambridge, MA: MIT Press.

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. 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 metamemory index. *Journal of Experimental Psychology: General, 127*, 55–68.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII—Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy & S. M. Kosslyn (Eds.), *Essays in honor of William K. Estes* (Vol. 1, pp. 35–67). Hillsdale, NJ: Erlbaum.

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.

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.

Cutting, J. E. (1975). Orienting tasks affect recall performance more than subjective impressions of ability to recall. *Psychological Reports, 36*, 155–158.

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). Hillsdale, NJ: Erlbaum.

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.

Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica, 98*, 37–56.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (Rev. ed.). New York: Dover. (Original work published 1885)

Engelkamp, J. (1998). *Memory for actions*. East Sussex, England: Psychology 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.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Granhag, P. A. (1997). Realism in eyewitness confidence as a function of type of event witnessed and repeated recall. *Journal of Applied Psychology, 82*, 599–613.

Griffin, D. W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411–435.

Juslin, P. (1993). An explanation of the hard–easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5*, 55–71.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review, 107*, 384–396.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217–273.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology, 93*(2), 329–343.

Klayman, J., Soll, J. B., Gonzáles-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79*, 216–247.

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. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9*, 149–171.

Koriat, A., Ben-Zur, H., & Druch, A. (1991). The contextualization of memory for input and output events. *Psychological Research, 53*, 260–270.

Koriat, A., & Bjork, R. A. (2001). *Illusions of competence in monitoring one's knowledge during study: The foresight bias*. Manuscript submitted for publication.

Koriat, A., & Bjork, R. A. (2002). *Can illusions of competence be remedied? Effects of study and test experience on the foresight bias*. Manuscript in preparation.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

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., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Cognition, 6*, 107–118.

Koriat, A., Ma'ayan, H., & Levy-Sadot, R. (2002). *The intricate relationships between monitoring and control in metacognition*. Manuscript in preparation.

Koriat, A., Pearlman-Avnion, S., & Ben-Zur, H. (1998). The subjective organization of input and output events in memory. *Psychological Research, 61*, 295–307.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 464–470.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?: The calibration of probability judgments. *Organizational Behavior and Human Performance, 20*, 159–183.

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). Cambridge, England: Cambridge University Press.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 756–766.

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 and Cognition, 18*, 196–204.

Mazzoni, G., & Nelson, T. (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 & Cognition, 21*, 1263–1274.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probability: Theories and models 1980–94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester, UK: Wiley.

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review, 105*, 299–324.

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. (1996). Consciousness and metacognition. *American Psychologist, 51*, 102–116.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*, 267–270.

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

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.

Palermo, D. S., & Jenkins, J. J. (1964). *Word association norms: Grade school through college*. Minneapolis: University of Minnesota Press.

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I. M., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology, 37*, 688–695.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*, 1004–1010.

Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: Cue heuristics versus direct–perceptual competence. *Psychological Review, 107*, 525–555.

Schneider, S. L. (1995). Item difficulty, discrimination, and the confidence–frequency effect in a categorical judgment task. *Organizational Behavior and Human Decision Processes, 61*, 148–167.

Schneider, W., Vise, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring: Evidence from a judgment-of-learning (JOL) task. *Cognitive Development, 15*, 115–134.

Shaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning and Verbal Behavior, 20*, 216–230.

Shaw, R. J., & Craik, F. I. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging, 4*, 131–135.

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior, 24*, 271–296.

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

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67*, 201–221.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced learning. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25*, 1024–1037.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York: Macmillan.

Treadwell, J. R., & Nelson, T. O. (1996). Availability of information and the aggregation of confidence in prior decisions. *Organizational Behavior and Human Decision Processes, 68*, 13–27.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

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.

Received May 21, 2001
Revision received January 15, 2002
Accepted January 15, 2002 ■