Running head: REPEATED TESTING AND SCHOLASTIC APTITUDE

Repeated Testing Sessions and Scholastic Aptitude in College Students' Metacognitive Accuracy

William L. Kelemen

California State University, Long Beach

Robert G. Winningham

Western Oregon University

Charles A. Weaver, III

Baylor University

Correspondence:

William L. Kelemen

Department of Psychology

1250 Bellflower Blvd

Long Beach, CA 90840-0901

E-mail: wkelemen@csulb.edu

Phone: (562) 985-5030

FAX: (562) 985-8004

Abstract

We performed three experiments to examine the effects of repeated study-judgment-test sessions on metacognitive monitoring, and to see if better students (those with higher SAT scores) outperform low-SAT students. In all experiments, mean metacognitive accuracy (bias scores and Gamma correlations) did improve across sessions. In addition, students with high SAT scores recalled more items, were less overconfident, and adjusted their predictions more effectively. Thus, high-SAT students may be able to adjust their metacognitive monitoring effectively without feedback, but low-SAT students appear unlikely to do so. Educators may need to devise more explicit techniques to help low-SAT students improve their metacognitive monitoring during the course of semester.

Repeated Testing Sessions and Scholastic Aptitude in College Students' Metacognitive Accuracy

The best college students effectively manage an assortment of requirements in a number of courses during a typical semester. Modes of instruction, course-related activities, and instructors' testing styles may vary from course to course even within a single major. Many students greet the first test in a course with apprehension, however, because they do not know exactly what to expect. Instructors may even counsel students who score poorly on the first test that "You should do better next time, now that you know what to expect." Anecdotally, good students often seem better able to assess their preparation for their first exam, or at least to adjust to subsequent exams effectively. The present research was designed to test whether students' ability to predict their memory performance improves with repeated testing, and whether good students are indeed more effective at monitoring and adjusting their predictions compared with poor students.

Being able to predict performance accurately can impact test performance. For example, Thiede (1999) asked students to provide judgments of learning (JOLs) during study of vocabulary items followed by a recall test. Participants then selected items for restudy, and the process repeated for a total of five trials. Both monitoring accuracy and degree of self regulation were significant predictors of test performance across trials. Likewise, Nelson, Dunlosky, Graf, and Narens (1994) also showed the importance of using JOLs effectively in subsequent test performance; recall was better when items receiving low JOLs were restudied compared with items receiving high JOLs. In both of these cases, JOLs were used to influence future study and test performance improved. Recent work by Metcalfe and her colleague suggests good learners identify a zone of "proximal learning," where additional study is most likely to produce the

greatest gains (Metcalfe, 2002; Metcalfe & Kornell, 2003, 2005; Son & Metcalfe, 2000). Thus, there is solid evidence that one's degree of monitoring accuracy can influence academic success.

A separate, but related, question is whether or not good students (based on some index of learning ability) show superior monitoring accuracy compared with poor students. Such a finding should emerge if good students succeed at least in part because they have superior monitoring ability. Surprisingly, there is little evidence to support this idea. Lovelace (1984) found that JOL accuracy was unrelated to paired-associate recall. Cull and Zechmeister (1994) similarly found no differences in metacognitive strategies related to recall. Additional null results have been obtained using ease-of-learning judgments (Kearney & Zechmeister, 1989; Underwood, 1966), JOLs for texts (Pressley & Ghatala, 1990), and retrospective confidence judgments for general knowledge questions (Lichtenstein & Fischhoff, 1977). With few exceptions (e.g., Maki & Berry, 1984; Shaughnessy, 1979), differences in learning ability do not typically influence metacognitive accuracy.

Many of the studies just described assessed learning ability by computing a median split of participants based on the number of items recalled. However, this technique may be problematic because recall scores are used both to form the independent variable (learning ability group) and also to assess the dependent variable (monitoring accuracy). The present study used college students' scores on the Scholastic Aptitude Test (SAT, combined verbal and quantitative scores) to provide an independent, standardized estimate of academic ability. Demonstrating a relationship between SAT scores and monitoring accuracy might be useful in devising effective interventions for students of different aptitudes.

Repeated Testing and Metacognition

Repeated testing in the classroom (e.g., administering weekly quizzes) has been shown to be an effective aid to course performance. In one study (Tuckman, 1996), regular quizzes over a

5-week interval significantly improved test scores compared with weekly homework assignments. In a follow-up study, Tuckman confirmed this finding over the course of an entire semester, and showed that low-GPA students tended to benefit the most from repeated testing. In addition to classroom studies, there is a growing body of laboratory evidence showing the effectiveness of repeated testing on long-term memory test accuracy (e.g., Roediger & Karpicke, 2006). In regard to metacognitive monitoring, repeated study-test cycles of the same stimuli tend to produce higher levels of recall without commensurate increases in JOL magnitude, resulting in marked underconfidence (i.e., the "underconfidence with practice effect," Koriat, Sheffer, and Ma'ayan, 2002, but see Scheck & Nelson 2005 for boundary conditions). Unfortunately, most of the laboratory experiments have examined repeated testing of the same stimulus materials, which would be relevant to a student preparing for a single exam, but does not speak to the question of how students use the outcome of one exam to prepare for subsequent ones on different material.

Some studies on training have examined repeated testing using different stimulus materials and shown increased metacognition accuracy. Lichtenstein and Fischhoff (Experiment 1, 1980) found that participants who completed 11 sessions (each containing 200 trials and followed by explicit feedback) showed increased calibration in their postdiction ratings, primarily in the early sessions. In Experiment 2, they obtained similar improvements using only three sessions of training and feedback. In subsequent research, Zechmeister, Rusch, and Markell (1986) showed that just one session of training was enough to improve postdiction calibration. , Furthermore, low-achieving students (as measured by classroom performance) initially were more overconfident than high-achieving students, and as a result, the low achievers benefited more from training. In sum, it appears that formal training can influence monitoring accuracy, and the effectiveness of this training may differ for low- versus high-achieving students.

In both studies above, the instructions/training sessions were multifaceted, and the relative contributions of various components were not evaluated. Other research suggests that self-generated feedback alone can affect subsequent JOLs. Hertzog, Dixon, and Hultsch (1990) asked participants to predict future recall of categorized lists and texts. Participants predicted either how many total words they would recall, or how many ideas they would recall, respectively. Two important findings emerged. First, the correlation between JOLs and recall increased across the three lists (*r*s = .24, .52, and .62 for word recall, and *r*s = .44, .54, and .58 for texts). As a group, participants' predictions of their overall performance improved with practice alone. In addition, structural equation modeling suggested that this improvement was due primarily to the effects of past recall performance. Together, these findings demonstrated that participants can use the outcome of previous recall attempts to predict future recall with increasing accuracy, even without explicit feedback. Koriat (1997) obtained an increase in item-by-item predictive accuracy using two different lists of paired associates. In Experiment 1, he found a small but reliable increase in mean *G* from List 1 (*G* = .59) to List 2 (*G* = .65). Koriat attributed this increase to general experience with the task. Unfortunately, no further experiments on practice were conducted using different items on each trial, and subsequent work on repeated testing has focused primarily on repeated trials using the same stimuli.

These results suggest that practice alone can improve participants' estimates of their overall recall (absolute metacognitive accuracy; Hertzog et al., 1995), and that this improvement may extend to item-by-item JOLs (relative metacognitive accuracy; Koriat, 1997). It is difficult to draw firm conclusions, however, because the effect of repeated study-JOL-test sessions on monitoring accuracy was not the primary focus these experiments. Moreover, previous studies included only two or three practice trials completed during the same testing session, and so it is impossible to determine (a) if further improvement would have been obtained using additional

sessions, and (b) if the improvements in JOL accuracy would extend to testing sessions completed at different times.

Goals and Hypotheses

The present study examined the effects of repeated study-JOL-test sessions on metacognitive accuracy over a five-week period. Participants completed five experimental sessions that involved learning foreign vocabulary items, making JOLs, and completing a recall test. Different vocabulary items were used each week, so these procedures were akin to the experiences of a student in a foreign language course having weekly vocabulary quizzes on new items each week. Two main issues were considered. First, we examined the effects of five study-JOL-test sessions on two aspects of predictive accuracy (i.e., estimates of overall recall and item-by-item JOL accuracy). Previous research suggests that monitoring accuracy may improve, but no studies have examined this issue directly, using multiple testing sessions over a period of weeks. Second, we examined the relationship between academic aptitude (SAT scores) and metacognitive monitoring. We predicted high-SAT students would produce more accurate JOLs overall. If so, the benefits of practice might differ between groups, with low-SAT students perhaps gaining more over time than high-SAT students (cf. Zechmeister et al., 1986).

Experiment 1

Method

Participants and materials. Eighty-four undergraduates enrolled in Introductory Psychology volunteered to participate and received course credit. The stimuli were 100 pairs of Swahili - English translations (e.g., *WINGU - cloud*) drawn from Nelson and Dunlosky's (1994) norms. Five lists of 20 different vocabulary items were constructed to be of similar normative difficulty. The stimuli were projected on a large screen, and all participants were tested together in a lecture hall.

Design and procedure. Participants were informed that they would be learning foreign vocabulary and making judgments about their future memory performance. They also agreed to allow access to their SAT scores from their university records following completion of the study. Experimental sessions were held once a week, for five consecutive weeks. Before beginning each of the five sessions, participants were reminded that they would be asked to study 20 Swahili - English vocabulary items, provide a confidence rating about future recall for each item, and complete a memory test on the items. The to-be-learned vocabulary items appeared sequentially at a rate of 6-s/item. Immediately after studying an item, the Swahili cue word appeared below the following question, "How confident are you that in about 10 minutes you will be able to recall the English translation of the Swahili word below?" Participants circled one of the following six ratings on a separate judgment sheet: 0% confident (labeled "will not recall"), 20%, 40%, 60%, 80%, or 100% confident (labeled "will recall"). In order to discourage participants from rehearsing previous items during the study and rating phase, the Swahili cues did not appear on the participants' judgment sheets. Instead, the Swahili cues shown on the screen during the rating phase were numbered consecutively from 1-20, and participants circled their confidence rating listed next to the corresponding number. Participants were allowed five seconds to make each rating. A two-second warning on the screen preceded presentation of the next vocabulary item. After studying and rating all 20 items, participants completed an unrelated filler activity for five minutes. Finally, a cued-recall test was distributed. The 20 Swahili cues were listed in a random order, and participants were asked to recall the English translation. Participants were allowed as much time as necessary to complete the memory test.

Results

A total of 67 participants completed all five experimental sessions. In order to minimize missing observations across sessions, data from participants who failed to complete all five

sessions were excluded from statistical analyses. All tests of statistical significance were conducted at $p < .05$.

JOL magnitude. Mean JOL ratings for each session were computed. For ease of comparison, these values were converted into proportions and are listed in Table 1. Repeated-measures analysis of variance (ANOVA) revealed a significant decrease in JOL magnitude (i.e., confidence) over experimental sessions, $F(4, 264) = 17.69$, $MSE = .54$. A post-hoc analysis was conducted to determine which sessions differed reliably. Differences between all pairs of means were compared to a critical difference, $q(4, 264) = .03$, computed using the Fisher-Hayter multiple-comparison test (Hayter, 1986).[1] Significant differences between sessions are noted in Table 1. Two findings emerged: (a) participants showed higher mean confidence in Session 1 compared to all the other sessions, and (b) mean confidence in Session 2 was higher than in Sessions 4 and 5.

Cued recall performance. Mean proportion correct on the memory test for each session was modest and consistent (means ranged from .31 to .35; see Table 1). No significant difference in recall across sessions was observed $F(4, 264) = 1.36$, $MSE = 0.01$.

Metacognitive accuracy. JOL accuracy can be assessed in several ways. Relative metacognitive accuracy is best measured by Goodman-Kruskal Gamma (*G*) correlations (Nelson, 1984, 1996). *G* is an ordinal measure of association that ranges from -1.0 to +1.0, with 0 indicating a complete lack predictive accuracy. *G* assesses item-by-item predictive accuracy, i.e., whether an item that received a high JOL is more likely to be recalled compared to a different item that received a lower JOL.

Mean *G*s for each experimental session were computed. Some *G*s could not be computed for two participants due to a lack of variability in recall. In Sessions 1 and 4, one participant failed to correctly recall any stimuli; in Session 5, one participant correctly recalled all 20

stimuli. Data from these participants were excluded from the following ANOVA and post-hoc test. A significant change in *G* was detected over sessions, $F(4, 256) = 4.47$, $MSE = 0.10$. A post-hoc analysis was performed to compare all pairwise combinations of mean *G*s. Differences in mean *G* exceeding the critical difference, $q(4, 256) = .14$, are noted in Table 1. Relative metacognitive accuracy improved with practice: *G* was reliably higher in Sessions 3-5 compared to Session 1.

In addition to *G*, another way to assess metacognitive accuracy is to compare participants' mean confidence for all items to their mean recall. This procedure provides a measure of absolute metacognitive accuracy called bias. A bias score greater than 0 indicates overconfidence, and a score less than 0 indicates underconfidence. Mean bias scores for each experimental session are listed in Table 1. All bias scores were positive, indicating general overconfidence in all sessions. ANOVA revealed a significant decrease in overconfidence over sessions, $F(4, 264) = 11.09$, $MSE = 1.41$. All pairwise combinations of means were compared to a critical difference, $q(4, 264) = .08$ using the Fisher-Hayter test. Participants were more overconfident in Session 1 than in Sessions 3-5.

SAT scores and performance. In addition to the changes across sessions analyzed above, we also examined whether individual differences reflected in students' SAT scores were related to their performance. Pearson product-moment correlation coefficients (*r*s) were computed between SAT scores and the four dependent measures (confidence, recall, *G*, and bias) for each session. SAT scores were unavailable for seven participants; data from these individuals were excluded from subsequent analyses. The mean SAT score for the remaining participants was 1110 ($SD = 150$). SAT scores and confidence were not correlated in any of the five sessions (mean $r = -.01$). However, the correlation between SAT scores and recall was reliable, albeit modest, in four of five sessions (mean $r = .24$). No correlation was observed between SAT

scores and *G* (mean *r* = -.04). Finally, SAT scores and bias scores were negatively correlated; this relationship was statistically reliable in three of five sessions (mean *r* = -.27). In sum, students with higher SATs tended to recall more on the memory test, and these students were less overconfident.

These correlational results led us to analyze individual differences in performance more systematically. To determine whether good students performed better than poor students, we rank-ordered the 60 participants according to their SAT scores and then compared students with low SATs (less than 1000, *n* = 14) to students with high SATs (greater than 1200, *n* = 16). These groups were roughly the bottom and top quartiles, respectively, of our sample. The mean SAT score for the low SAT group was 912 (*SD* = 42); mean SAT in the high SAT group was 1303 (*SD* = 56). Summary statistics for confidence, recall, *G*, and bias for these two groups of students are shown in Table 2.

Separate mixed-design ANOVAs were conducted on each of the four dependent measures. Session (1-5) was a within-subjects variable and group (low versus high SAT) was included as a between-subjects variable. Because no significant interaction between session and group was detected in any of the ANOVAs, only main effects are discussed. For mean confidence, a significant decrease over sessions was obtained, $F(4, 112) = 8.62$, $MSE = .51$, but there was no significant difference in confidence between SAT groups. No significant main effects were obtained for recall, although Table 2 shows a clear trend toward higher recall in the high SAT group, $F(1, 28) = 3.23$, $MSE = .11$, $p = .08$. No significant main effects were observed for *G*. Thus, no significant differences between high SAT students and low SAT students were detected in confidence, recall, and *G*.

Significant differences in bias were detected between groups and across sessions, $F(1, 28) = 7.76$, $MSE = 6.24$, and $F(4, 112) = 4.34$, $MSE = 1.25$, respectively. In both groups,

overconfidence declined in later sessions, but, low SAT students remained more overconfident than high SAT students all sessions. In order to examine this difference in bias more closely, we plotted calibration curves for both groups (see Figure 1). These curves show actual recall as a function of predicted recall for both SAT groups. If participants were perfectly calibrated, they would recall none of the items given a JOL = 0, about 20% of items given a JOL = 20, and so forth. Examination of Figure 1 confirms that high SAT students were less overconfident than low SAT students across the entire rating scale.

Discussion

Experiment 1 examined changes in confidence, recall, and metacognitive accuracy over five sessions of foreign vocabulary learning. Recall did not differ between sessions, suggesting that our attempt to equate the difficulty of stimuli across sessions was successful. In contrast, a reliable decrease in confidence was observed over sessions. Because participants were always overconfident, the decrease in mean confidence produced smaller (i.e., more accurate) bias scores in later sessions. By Session 5, the difference between participants' mean JOLs and mean recall was significantly reduced (.04) compared to Session 1 (.14). In other words, absolute metacognitive accuracy increased as participants became more familiar with the task.

In addition to the reduction in overconfidence, item-by-item metacognitive accuracy (*G*) also improved in Sessions 3-5. This finding was somewhat surprising, because previous studies (e.g., Kelemen, Frost, & Weaver, 2000; Thompson & Mason, 1996) have compared metacognitive accuracy over two sessions and found no increase in mean *G*. The present improvements in *G* and bias emerged only when performance in Session 1 was compared to performance in Sessions 3-5. Thus, students may be able to improve their memory monitoring performance given enough practice in a particular task.

An alternative explanation of these results is that, because the five lists of stimuli were comprised of different items, the range of item difficulty may have varied between sessions even though mean difficulty did not. Specifically, the variability of difficulty could have been greater in Sessions 3-5 compared with Session 1. For example, imagine Session 1 contained 20 items of roughly equal difficulty, whereas Session 5 contained 10 very easy items and 10 very difficult items. Mean recall might be the same for both sessions, but *G* likely would be higher in Session 5 because it contained two subgroups of items that could be discriminated from each other quite easily. This alternative, however, does not explain why absolute metacognitive accuracy (bias scores) also improved over sessions. Nevertheless, Experiment 2 was designed to eliminate this potential confound by varying the amount of practice between groups of participants, but comparing the same groups of items.

We found evidence that individual differences in absolute metacognitive accuracy (bias) were related to SAT scores. Figure 1 shows that high-SAT students showed less overconfidence than low-SAT students, across the entire rating scale. In addition, a reliable difference in bias scores was found between high- and low-SAT groups, with the former group predicting their mean level of performance on a test better than the latter group. Table 2 shows that low-SAT students failed to reduce their confidence adequately to match future recall. Even after four sessions of practice, low-SAT students were no better (bias = .11) than unpracticed high-SAT students in Session 1 (bias = .09). These results contrast with Shaughnessy et al.'s (1986) finding that poorer students benefited more from training; in our study, practice alone (in the absence of training) produced benefits only for the high-SAT students.

Experiment 2

Experiment 2 was designed to further investigate the improvement in memory monitoring over sessions. In order to avoid potential confounds created by comparing performance across

different sets of items, we formed three experimental conditions. The procedures of Condition A were similar to those of Experiment 1: participants studied lists of 20 Swahili - English pairs, made a JOL for each item immediately after study, and then received a cued-recall test. Participants completed five experimental sessions, each containing a unique set of 20 vocabulary items. In Condition B, participants also studied the lists of items and received memory tests, but they only provided JOLs for items in Session 5. The manipulation between Conditions A and B was whether participants made JOLs during Sessions 1-4. In Condition C, participants studied items, provided JOLs, and received a memory test, but they completed only Session 5. These three conditions are summarized in Table 3.

This design allowed us to test several hypotheses. First, we wanted to determine the source of memory monitoring improvement in Experiment 1: did participants' monitoring improve because of practice, or was the increase an artifact of stimulus variability between sessions? If memory monitoring genuinely improves because of practice, then metacognitive accuracy in Session 5 should be reliably higher in Condition A compared with Condition C. The same set of items was used in each condition, but in Condition A participants had four sessions of practice, whereas participants in Condition C were naïve to the procedures.

If the improvement was genuine, then a second question concerns whether making JOLs is necessary, or whether merely being exposed to the stimuli, study procedures and cued-recall tests would improve memory monitoring. This issue can addressed by comparing results from Conditions 1 and 2 in Session 5. In the former case, participants made JOLs during all five experimental sessions; in the latter, participants provided JOLs only in Session 5. If practice making JOLs is necessary, then metacognitive accuracy should be higher in Condition A.

We also obtained an additional type of prediction from our participants. In Experiment 1, mean confidence was computed by averaging the 20 JOLs elicited from each participant. Bias

scores then were calculated from these mean values, showing general overconfidence. Mazzoni and Nelson (1995), however, found an "aggregation effect" for JOLs: participants were less overconfident when making a single judgment about a collection of items than when their individual JOLs for each item were averaged. Perhaps the initial overconfidence in Experiment 1 reflected the averaging procedure we used, rather than a genuine metacognitive effect. To address this possibility, participants in Experiment 2 provided JOLs for each item, and then they made a single "aggregate JOL" after studying all the items. In this way, confidence and bias scores could be computed using both measures.

Method

Participants and materials. A total of 135 undergraduates enrolled in Introductory Psychology volunteered for Experiment 2 (49 in Condition A, 47 in Condition B, and 39 in Condition C). All students received course credit for participation. The same five lists of Swahili - English items from Experiment 1 were used. Participants in each condition were tested as a group, but the three groups were tested separately.

Design and procedure. Three experimental conditions were devised. In Condition A, participants completed one session per week, for five weeks. In each session, participants studied a list of 20 Swahili - English vocabulary items and provided a JOL immediately after studying each item. In order to increase the modest proportion of items recalled from Experiment 1, study time was increased to eight seconds for the Swahili - English pairs, and participants were allowed six seconds to provide each JOL. A two-second warning preceded the presentation of each vocabulary item. After studying and rating all 20 items, participants were shown the following prompt, "Out of the 20 pairs you just studied, how many will you get right on a test in about 10 minutes?" Participants entered a number ranging from 0-20 at the bottom of

their rating sheets. This judgment was their aggregate JOL. Participants then completed a filler activity for 10 minutes and completed a cued-recall test, as in Experiment 1.

Condition B was similar, except that participants did not provide JOLs during Sessions 1-4. Instead, the Swahili word alone appeared for six seconds below the prompt "Continue to think about the item below" immediately after study. The stimuli and all other procedures were the same as Condition A for Sessions 1-4. In Session 5, participants were informed that in addition to studying the 20 items, they would now be asked to predict their future memory performance. Thus, the stimuli and procedures in Session 5 were the same in Conditions A and B.

Only one experimental session was included in Condition C. Participants were instructed to study the items and make judgments about their future memory performance. The stimuli and procedures were the same as those used in Session 5 of previous conditions.

Results

A total of 35 participants completed all five sessions in Condition A, 28 completed all five sessions in Condition B, and 39 participants completed Condition C (which included only one session). We included all available data from participants in the following statistical analyses, which resulted in small fluctuations in sample sizes across sessions for Conditions B and C.

Changes in JOLs over sessions. Participants in Condition A provided JOLs during each of the five sessions. Two types of JOLs were elicited from participants: (a) a judgment for each vocabulary item (item-by-item JOLs), and (b) one judgment about how many of the 20 words would likely by recalled (an aggregate JOL). One participant failed to provide an aggregate JOL in Session 2, and another participant failed to provide one in Session 4. Both types of JOLs were converted into proportions to permit direct comparison.

We first considered item-by-item JOLs. Data from 35 participants in Condition A who completed all five sessions were analyzed. Confidence appeared to decrease after the initial session: mean item-by-item JOLs were .44, .37, .38, .35, and .37, for Sessions 1-5, respectively (the standard error of the mean, *SEM*, for each value was less than .03). Consistent with Experiment 1, a repeated-measures ANOVA revealed a significant change, $F(4, 136) = 10.65$, $MSE = .01$. A critical difference between means was calculated using the Fisher-Hayter post-hoc test, $q(4, 136) = .04$. This post-hoc test confirmed that confidence was higher in Session 1 than in all subsequent sessions.

A similar pattern of results emerged using aggregate JOLs. Confidence again decreased after the first session (mean aggregate JOLs for Sessions 1-5 = .41, .35, .38, .36, .35; all *SEM*s < .03). A repeated-measures ANOVA on data from 33 participants who provided aggregate JOLs for all five sessions showed a significant change, $F(4, 128) = 3.51$, $MSE = .01$. A critical difference between means was computed, $q(4, 128) = .05$. Aggregate JOLs in Session 1 were reliably higher compared to those in Sessions 2, 4 and 5. In short, the magnitude of both mean item-by-item and aggregate JOLs decreased after the initial session.

Item-by-item versus aggregate JOLs. The mean values for item-by-item JOLs and aggregate JOLs reported above were similar. A 2 X 5 (type of JOL by session) ANOVA was conducted to test for differences between confidence measures. Thirty-three participants provided both types of JOLs for all five sessions. A main effect of session was observed, $F(4, 128) = 9.12$, $MSE = .01$, confirming the results of the two previous one-way ANOVAs. However, no main effect was observed for type of JOL (item-by-item versus aggregate), and there was no significant interaction. Both measures produced similar mean levels of confidence.

JOL magnitude between conditions. The first two columns of Table 4 contain mean JOLs from Session 5 between conditions. These values represent the effects of practice on

confidence. For item-by-item JOLs, a one-way between-groups ANOVA was conducted using 104 participants who provided judgments for all items from Session 5. A significant effect of condition was obtained, $F(2, 101) = 8.73$, $MSE = .03$. Participants with previous JOL experience (Condition A) were less confident than participants with no experience in making JOLs (Conditions B and C), $q(2,101) = .05$.

A similar pattern emerged for aggregate JOLs. A significant effect of condition was found, $F(2, 100) = 9.36$, $MSE = .03$, and a critical difference between means was computed, $q(2, 100) = .05$. The magnitude of aggregate JOLs differed reliably across all three conditions. Participants with JOL experience were the least confident, and participants in Condition B were the most confident.

Cued recall performance. Participants in Conditions A and B learned vocabulary items and completed cued-recall tests during Sessions 1-5. Mean recall in Condition A was .34, .41, .43, .39, and .42 for Sessions 1-5, respectively. Corresponding recall values for Condition B were .42, .52, .50, .53, and .42 (SEMs for both groups were all less than .03). A 2 X 5 (condition by session) ANOVA was conducted. Recall was reliably higher in Condition B compared to Condition A, $F(1, 60) = 7.84$, $MSE = .12$. This may reflect a slight difference in procedures between conditions during Sessions 1-4. Participants in Condition A studied the cue-target pair for eight seconds and then made a cue-alone JOL for six seconds; participants in Condition B studied the cue-target pair for eight seconds and then continued to study the cue alone for six seconds. The extra study time in Condition B may have increased cued-recall performance. A significant effect of session was noted, $F(4, 240) = 4.99$, $MSE = .01$, and there was a significant interaction between condition and session, $F(4, 240) = 3.48$, $MSE = .01$.

More important were potential differences in recall between conditions in Session 5 (see third column in Table 4). In the final session, all three conditions received eight seconds of

study time and six seconds to make a JOL. ANOVA showed no significant differences in recall between conditions, although there was a trend for practiced participants to recall more than unpracticed participants. Thus, when study time between groups was equal (Session 5), no differences in recall were observed.

Relative metacognitive accuracy (G). Gamma correlations were calculated between JOLs and recall for 35 participants in Condition A. Mean values across sessions were .53, .61, .65, .60, .51 (SEMs ranged from .04-.08). In contrast to Experiment 1, no significant change in metacognitive accuracy was found across sessions, $F(4, 136) = 1.07$, $MSE = .11$.

The analysis of major interest, however, was the comparison of *G*s across Conditions A-C in Session 5. The mean values are listed in Table 4. A significant difference in *G* between conditions was detected, $F(2, 99) = 4.26$, $MSE = .18$. A critical difference was calculated post-hoc, $q(2, 99) = .12$, which confirmed that mean *G* was higher in Conditions A and B (*G*s = .51 and .52, respectively) than in Condition C ($G = .26$). These results suggest that practice with the experimental procedures, but not necessarily practice making JOLs, underlies the increase in relative memory monitoring accuracy over sessions.

Absolute metacognitive accuracy. Bias scores (i.e., the signed difference between confidence and accuracy) were computed for each participant in Condition A using two measures of confidence: (a) the mean of item-by-item JOLs, and (b) aggregate JOLs. Mean bias scores for Sessions 1-5 using item-by-item JOLs were .10, -.03, -.05, -.03, and -.05, respectively. Using aggregate JOLs, mean bias scores were .06, -.06, -.05. -.02. and -.06 (all *SEMs* were less than .03). ANOVA revealed a significant effect of session on bias, $F(4,136) = 11.96$, $MSE = .02$ for item-by-item JOLs, and $F(4,128) = 8.69$, $MSE = .02$ for aggregate JOLs. A critical difference between means of .08 was calculated for both measures of bias. These post-hoc test

indicated a significant difference between Session 1 and Sessions 2-5. As in Experiment 1, participants were most overconfident in Session 1.

We also examined differences in bias between conditions in Session 5 (see Table 4). ANOVA revealed a significant difference between conditions, $F(2,101) = 7.56$, $MSE = .06$ for item-by-item JOLs, and $F(2,101) = 9.82$, $MSE = .03$ for aggregate JOLs. Critical differences between the means were calculated, $q(2,101) = .06$ for item-by-item JOLs and $q(2,101) = .05$ for aggregate JOLs. Significant differences between means are noted in Table 4. In general, participants with JOL experience (Condition A) were slightly underconfident, compared to the overconfidence demonstrated by participants without JOL experience (Conditions B and C). This pattern was consistent whether bias was measured using aggregate JOLs or mean item-by-item JOLs.

SAT scores and performance. SAT scores were unavailable for 19 of the 135 participants. For 12 of these students, scores from the American College Test (ACT) were obtained and converted to equivalent SAT scores using a conversion table provided by the Institutional Research and Testing department at Baylor University. Data from the remaining seven students were omitted from subsequent analyses because no college admission scores were available. The mean SAT score for participants was 1101 ($SD = 133$).

Partial correlation coefficients (controlling for the effect of condition) were computed between SAT scores and all dependent measures for Session 5. SAT scores were not significantly correlated with confidence as measured by item-by-item JOLs ($r = .01$) nor aggregate JOLs ($r = .14$). However, SAT scores were significantly correlated with recall ($r = .36$): higher-SAT students tended to recall more vocabulary items. Significant negative correlations also were obtained between SAT and bias: $r = -.29$ for mean item-by-item JOL bias and $r = -.25$ for aggregate JOL bias. These results indicate that higher-SAT students tended to be

less overconfident than lower-SAT students. Finally, a statistically significant, negative correlation was observed between SAT and *G* (r = -.21). Surprisingly, higher SAT students tended to produce lower *G*s than lower SAT students.

To explore these issues further, participants were ranked according to their SAT scores. We compared the performance of low-SAT students (SAT = 1010 or below; n = 33) to high-SAT students (SAT = 1200 or above; n = 31). These two groups corresponded to approximately the bottom and top quartiles of our entire sample (N = 128), respectively. The mean score for low-SAT students was 940 (SD = 70) and the mean for high-SAT students was 1271 (SD = 76).

Twenty-four participants from each SAT group completed all sessions, including Session 5. Mean confidence, recall, and metacognitive accuracy by group are shown in Table 5. ANOVAs were performed on these data to test for differences in performance between high and low SAT students.[2] Significant differences between groups emerged for recall (high SAT mean recall = .48 and low SAT mean recall = .28), $F(1, 46) = 17.07$, MSE = .03. High SAT students also showed less overconfidence as measured by item-by-item JOLs, $F(1, 46) = 9.42$, MSE = .07 and aggregate JOLs, $F(1, 46) = 11.33$, MSE = .03. No significant differences between groups emerged for confidence or *G*s. These findings replicate the pattern of differences between high and low SAT groups obtained in Experiment 1.

Discussion

Four issues were addressed in Experiment 2. First, we examined whether the improvement in monitoring accuracy obtained in Experiment 1 was genuine or an artifact of stimulus variability. Reliable differences in the predicted direction for *G* and bias scores emerged between conditions in Session 5. These findings suggest that the improvements over sessions in Experiment 1 were not stimulus artifacts.

Given that the improvement in metacognition over sessions was real, a second issue concerned whether practice with making JOLs was necessary, or whether students would improve simply by completing previous study and recall trials. Practice with the study and recall tasks alone (Condition B) was sufficient to increase *G* compared to performance by unpracticed participants (Condition C). However, practice making JOLs themselves (Condition A) did not further increase *G*. In sum, *G* improved if participants had studied and recalled similar items in the past, regardless of whether or not they practiced making JOLs.

A slightly different pattern emerged when absolute metacognitive accuracy (bias) was examined. As before, participants were least overconfident when they had practice making JOLs (Condition A). However, experience with the study and recall procedures alone (Condition B) did not reduce this overconfidence. In fact, participants were most overconfident in Condition B. This finding may be related to the fact that participants in Condition B were instructed to continue studying each item (rather than make a JOL) during Sessions 1-4. These procedures produced higher levels of recall in the first four sessions compared to Condition A, and may have produced overconfidence in Session 5. One way to address this issue would be to elicit JOLs in Condition A and stimulus ratings on some unrelated dimension (e.g., pleasantness) in Condition B. We followed these procedures in Experiment 3 to control this potential confounding.

Third, we examined whether the general overconfidence observed in Experiment 1 was due to our measurement procedures. Previous research (Mazzoni & Nelson, 1995) has found that the magnitude of overconfidence is greater using mean item-by-item JOLs compared to a single rating based on the aggregate of items. We compared confidence as measured by mean item-by-item JOLs and aggregate JOLs, and there were no differences in bias using the two methods. The initial overconfidence in Experiments 1 and 2 was genuine, and it declined across sessions.

Finally, we replicated findings from Experiment 1 regarding scholastic aptitude and performance. Low-SAT students recalled fewer items and produced less accurate estimates of overall memory performance. Bias scores in Session 5 were substantially larger for low-SAT students (bias = .19 for mean item-by-item JOL bias and .14 for aggregate JOL bias), compared to high SAT students (bias = -.03 for both measures of confidence). For *G*, a small (unexpected) negative correlation was observed between SAT scores and *G*, but no reliable difference emerged in the ANOVA between low and high SAT groups.

Experiment 3

In Experiment 2, participants who completed four sessions of practice, but did not make JOLs (Condition B), were overconfident in Session 5 when JOLs were required. This may have occurred because participants were allowed to study items longer in Sessions 1-4 than in Session 5. The purpose of Experiment 3 was to control for this potential confound. We formed three experimental conditions in Experiment 3. In Condition A, participants completed five sessions of study-JOL-test procedures as before. In Condition B, participants studied each item and provided a pleasantness rating (instead of a JOL), followed by a test, during Sessions 1-4. During Session 5, participants followed the same study-JOL-test procedures as Condition A. Finally, participants in Condition C completed only Session 5, as before. These procedures are summarized in Table 3.

Method

Participants and materials. A total of 118 college undergraduates volunteered for Experiment 3 (54 in Condition A, 40 in Condition B, and 24 in Condition C). All students received course credit for participation. The same 100 Swahili - English items from previous experiments were used, but five new lists were created. Participants in each condition were tested as a group, but the three groups were tested separately.

Design and procedure. Three experimental conditions were devised. The procedures for Condition A were identical to Experiment 2. In Condition B, participants provided ratings of item pleasantness, ranging from 1 ("very unpleasant") to 6 ("very pleasant"), during Sessions 1-4. Participants were instructed to provide JOLs instead of pleasantness ratings in Session 5. Thus, the stimuli and procedures during Session 5 were the same in Conditions A and B. In Condition C, participants completed only Session 5 to provide an unpracticed comparison group. To maintain consistency across sessions and conditions, participants always provided aggregate JOLs. Participants in Conditions A and B were tested twice per week (on Monday and Friday), because testing did not begin until late in the semester. SAT scores were not obtained.

Results

A total of 40 participants in Condition A and 21 participants in Condition B completed all 5 sessions. Twenty-four participants completed Condition C, which included only one session. Five participants reported participating in a previous version of this experiment on a post-test questionnaire during Session 5. Data from these participants and all others who did not complete Session 5 were omitted from statistical analyses. As a result, the final sample sizes for Conditions A-C were 36, 20, and 24, respectively.

JOL magnitude and recall. Confidence was assessed using mean item-by-item JOLs and aggregate JOLs in Session 5 (see Table 6). One participant did not provide an aggregate JOL during Session 5. Although there was a trend toward higher JOL magnitude for participants in Condition C, no significant differences were observed between groups using either measure, $F(2, 77) = 1.62$, $MSE = .02$, for item-by-item JOLs and $F(2, 76) = 1.23$, $MSE = .03$, for aggregate JOLs. No significant differences in recall were found between groups, $F(2, 77) = .82$, $MSE = .04$.

Metacognitive accuracy. As in Experiments 1 and 2, two measures of metacognitive accuracy (*G* and bias) were calculated. Mean values for each measure are reported in Table 6. A

significant effect of condition on *G* was obtained, $F(2, 77) = 4.73$, $MSE = .10$. Post-hoc tests revealed that participants who received practice (Conditions A and B) showed higher *G* than unpracticed participants (Condition C), $q(2, 77) = .10$. This finding was consistent with previous results showing an improvement due to practice with the study and test procedures. Making JOLs during Sessions 1-4 (versus pleasantness ratings) did not improve *G* during Session 5. In fact, *G* was higher in Condition B than Condition A in the final session.

Bias was calculated using mean item-by-item JOLs and aggregate JOLs (see Table 6). Unpracticed participants in Condition C tended to be more overconfident than participants in the other conditions. For mean item-by-item bias, a significant effect of condition emerged, $F(2, 77) = 3.27$, $MSE = .04$. A significant difference was also found using aggregate JOLs, $F(2, 76) = 4.64$, $MSE = .03$. Post-hoc tests showed that for both measures of bias, unpracticed participants were more overconfident than practiced participants, $q(2, 77) = .06$ and $q(2, 76) = .06$. In contrast to Experiment 2, no significant difference was observed between Conditions A and B.

Discussion. One purpose of Experiment 3 was to control for a difference in study time in Experiment 2 that may have produced overconfidence. Specifically, participants in Condition B of Experiment 2 were allowed to continue studying items during the time that participants in Condition A made JOLs. In order to keep the amount of study time consistent in Experiment 3, participants in Condition B provided ratings of item pleasantness during Sessions 1-4 and JOLs during Session 5. Using this procedure, no differences in confidence or bias were found between Conditions A and B. These results suggest that the previous overconfidence was indeed due to the difference in study time. Experiment 3 also confirmed previous main findings that unpracticed participants (Condition C) were most overconfident and showed the lowest *G*s.

General Discussion

We examined the effects of multiple study-JOL-test sessions on students' metacognitive performance for foreign vocabulary items over a five-week period. Participants completed five experimental sessions, and different vocabulary items were included in each session. In all three experiments, the predictive accuracy of JOLs increased with practice. Improvements in memory monitoring were observed using mean *G* scores, which assessed item-by-item JOL accuracy, as well as bias scores, which assessed global under- and overconfidence. Most of the increase in metacognitive accuracy occurred in the first three sessions (see Table 1), but performance remained high in subsequent sessions. In contrast to past research (e.g., Zechmeister et al., 1986) these increases were achieved without any special training or explicit feedback.

Experiments 2 and 3 were designed to determine whether making JOLs during practice sessions was necessary to increase monitoring accuracy. In both experiments, practice making JOLs was not required to increase mean *G*. For bias, overconfidence was observed in Experiment 2 when participants did not have previous JOL experience, suggesting that JOL practice may affect absolute monitoring accuracy. When participants were asked to provide pleasantness ratings during the first four sessions of Experiment 3, however, bias in Session 5 was no greater than when participants made JOLs during Sessions 1-4. Thus, when the procedures between conditions were more closely matched, no differences in bias emerged. Together, these two experiments showed that exposure to the study and recall procedures alone was sufficient to improve both *G* and bias scores; making JOLs during previous sessions did not increase monitoring accuracy further.

Why did participants' memory monitoring improve? One possibility is that participants used self-generated feedback from previous test trials to adjust their future JOLs. For example, Hertzog et al. (1990) found that the correlation between predictions of overall recall and test performance increased over three different lists. Hertzog et al. argued that this improvement in

predictive accuracy stemmed from self-generated feedback about recall on the previous trial. They showed that the correlation between previous recall and subsequent JOL was sometimes larger than the correlation between JOL and subsequent recall. Using memory performance on the previous list as a basis for future recall would have been relatively easy for their participants, because the prediction and memory tasks were conducted sequentially during a single testing session.

Our results suggest that participants were able to generate feedback about their own performance and then adjust their JOLs up to one week later. Participants adjusted both their estimates of overall performance (aggregate JOLs) and their judgments of individual items (item-by-item JOLs) to reflect future mean recall. Impressively, participants were able to estimate how many items they had answered correctly on a test (e.g., 7 out of 20, or 35%), and then adjust (a) their subsequent aggregate JOL (e.g., predict 7 out of 20 next week), and (b) the magnitude of all 20 JOLs a week later such that the average rating was appropriate (e.g., the mean of all JOLs next week would be about 35%). This latter adjustment seems especially remarkable, especially given that participants received no feedback or instructions to adjust their future judgments.

In regard to the increase in mean item-by-item metacognitive accuracy (*G*), Koriat (1997) has proposed that participants may use one or more classes of cues when making JOLs. In the present research, JOLs in the initial session may have been based on intrinsic cues (e.g., perceived relatedness between the items, etc.), but in later sessions extrinsic cues (e.g., conditions of learning including study time, retention interval, etc.) might have become available and thus enhanced predictive accuracy. Experience with the recall test may have provided especially important information for participants. Vye, Schwartz, Bransford, Barron, and Zech (1998) reported on a program to increase students' metacognitive awareness and argued that knowing what to expect on a test provides critical cues for future study. They claimed, "The

better individuals can imagine (model) the situation in which they must use their knowledge, the easier it is for them to assess their level of preparation. *Increasing experience with particular situations* [italics added]…increases the accuracy with which one can anticipate the kinds of knowledge and skills necessary to perform adequately." (p. 309). In future research, it might be possible to tease apart the relative contributions of knowledge about study versus test procedures by comparing groups with practice in each of these components. Regardless of the precise theoretical mechanism, the metacognitive benefits of administering multiple realistic practice tests (or quizzes) to students in the classroom can be inferred from these laboratory data.

Learning Ability and Metacognition

A final aim of the present study was to determine whether individual differences in confidence, recall and metacognition were related to learning ability (as indexed by SAT scores). No relationship between confidence and SAT scores was detected. A modest correlation between SAT and recall, however, was detected in Experiment 1. In Experiment 2, this relationship appeared even stronger, and a reliable difference in recall emerged between high-SAT (1200 or more) and low-SAT (1010 or less) students. Low-SAT students remembered less than their high-SAT counterparts, but they were just as confident in their future memory performance.

It follows from these findings that low-SAT students were more significantly overconfident than high SAT students. This relative overconfidence was observed across the rating scale (see Figure 1). In addition, low-SAT students failed to adequately reduce their overconfidence in later sessions. Even during the final session, low-SAT students were more overconfident than unpracticed high-SAT students. High-SAT students, conversely, did reduce their overconfidence across sessions, and even appeared slightly underconfident by Session 5 in both experiments. Not only were low-SAT students worse at predicting their overall level of

recall, but they also failed to adjust their future predictions to reflect past recall performance. Future research is required to determine whether these students were less adept at constructing self-generated feedback, or whether they were less sensitive to their own feedback concerning previous recall.

No consistent relationship was found between *G* and SAT scores, consistent with previous studies comparing learning ability and item-by-item monitoring accuracy (e.g., Lovelace, 1984). However, the reliability of *G* in testing for individual differences has been found lacking in past research (e.g., Kelemen et al., 2000). It is unrealistic to expect *G* to correlate with learning ability, or any other psychological variable, until a procedure for obtaining reliable differences is devised.

In sum, the present increases in monitoring accuracy emerged over 5 weeks without specific instructions or training, and even without practice making JOLs themselves. These improvements were consistent across sessions, using lists of different items. Unfortunately, low-SAT students did not benefit from the five sessions. Educators may need to devise more explicit techniques to help low-SAT students improve their metacognitive monitoring during the course of semester.

References

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory & Cognition, 22,* 249-257.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging, 5,* 215-227.

Kearney, E. M., & Zechmeister, E. B. (1989). Judgments of item difficulty by good and poor associative learners. *American Journal of Psychology, 102,* 365-383.

Kelemen, W. L., Frost, P. J., & Weaver, C. A., III. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition.*

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

Lichtenstein, S. & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159-183.

Lichtenstein, S. & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26,* 149-171.

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

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10,* 663-679.

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

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131,* 349-363.

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

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, 52, 463-477.

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). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance of an individual item. *Applied Cognitive Psychology, 10,* 257-260.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2,* 325-335.

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.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25,* 19-33.

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

Scheck, P. & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134,* 124-128.

Shaughnessy, J. J. (1979). Confidence-judgment accuracy as a predictor of test performance. *Journal of Research in Personality, 13,* 505-514.

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

Thompson, W. B., & Mason, S. E. (1996). Instability of individual differences in the association between confidence judgments and memory performance. *Memory & Cognition, 24,* 226-234.

Tuckman, B. W. (1996). The relative effectiveness of incentive motivation and prescribed learning strategy in improving college students' course performance. *The Journal of Experimental Education, 64,* 197-210.

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology, 71,* 673-679.

Vye, N. J., Schwartz, D. L., Bransford, J. D., Barron, B. J., & Zech, L. (1998). SMART environments that support monitoring, reflection, and revision. In D. L. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 305-346). Mahwah, NJ: Erlbaum.

Zechmeister, E. B., Rusch, K. M., & Markell, K. A. (1986). Training college students to assess accurately what they know and don't know. *Human Learning: Journal of Practical Research & Applications, 5,* 3-19.

Author Notes

William L. Kelemen, Department of Psychology, California State University, Long Beach; Robert G. Winningham, Western Oregon University; Charles A. Weaver, III Department of Psychology and Neuroscience, Baylor University.

Portions of Experiments 1 and 2 were presented at the 40th annual meeting of the Psychonomic Society in November 1999. The authors thank Rita Massey for retrieving students' SAT scores.

Correspondence concerning this study should be addressed to William L. Kelemen, Department of Psychology, 1250 Bellflower Blvd., Long Beach, CA 90840-0901. *E-mail*: wkelemen@csulb.edu.

Footnotes

[1] For computational procedures and rationale for using the Fisher-Hayter test, see Kirk (pp. 148-150, 1995).

[2] Before computing the ANOVAs, a Chi-square test for independence was performed on condition and SAT group. No reliable effect was observed, χ^2 (2) = 3.69, $p > .15$, so ANOVAs were computed using data summed across all three conditions.

Table 1

Mean JOL magnitude, recall, and metacognitive accuracy across Sessions 1-5 in Experiment 1.

Session	Confidence*	Recall	Gamma*	Bias*
One	0.45 a	0.31	0.46 a.	0.14 a
Two	0.41 b	0.33	0.58 a, b	0.08 a, b
Three	0.38 b, c	0.35	0.68 b	0.04 b
Four	0.36 c	0.33	0.60 b	0.02 b
Five	0.37 c	0.32	0.60 b	0.04 b

Note. Main entries are mean values. Standard errors of the mean were less than .03 for all confidence, recall, and bias scores; standard errors of the mean were less than .05 for *G*s. Entries sharing a common subscript within a column were not significantly different in post-hoc analyses.

* A statistically significant effect was observed for that dependent variable, $p < .05$

Table 2

JOL magnitude, recall, and metacognitive accuracy as a function of SAT scores in Experiment 1.

	Session				
Measure	One	Two	Three	Four	Five
Confidence					
High SAT	0.44 (.03)	0.41 (.03)	0.37 (.04)	0.35 (.04)	0.33 (.04)
Low SAT	0.45 (.05)	0.42 (.04)	0.38 (.05)	0.36 (.05)	0.38 (.05)
Recall					
High SAT	0.36 (.04)	0.38 (.05)	0.36 (.05)	0.39 (.05)	0.40 (.06)
Low SAT	0.29 (.03)	0.29 (.03)	0.29 (.04)	0.28 (.03)	0.28 (.05)
Gamma					
High SAT	0.39 (.08)	0.48 (.08)	0.63 (.10)	0.62 (.09)	0.58 (.10)
Low SAT	0.39 (.13)	0.61 (.09)	0.59 (.09)	0.64 (.07)	0.49 (.14)
Bias*					
High SAT	0.08 (.04)	0.03 (.04)	0.02 (.04)	-0.04 (.03)	-0.07 (.04)
Low SAT	0.16 (.04)	0.14 (.04)	0.09 (.04)	0.09 (.04)	0.11 (.04)

Note. Main entries are mean values; standard errors of the mean are in parentheses. In the "High SAT" group, SAT scores were greater than 1200; in the "Low SAT" group, scores were less than 1000.

* A statistically significant difference between SAT groups was observed, $p < .05$

Table 3

Types of judgments provided by participants in each condition of Experiments 1-3.

	Session				
Experiment	One	Two	Three	Four	Five
Experiment 1	item-by-item JOL	item-by-item JOL	item-by-item JOL	item-by-item JOL	item-by-item JOL
Experiment 2					
Condition A	item-by-item JOL aggregate JOL	item-by-item JOL aggregate JOL	item-by-item JOL aggregate JOL	item-by-item JOL aggregate JOL	item-by-item JOL aggregate JOL
Condition B	none	none	none	none	item-by-item JOL aggregate JOL
Condition C	N/A	N/A	N/A	N/A	item-by-item JOL aggregate JOL
Experiment 3					
Condition A	item-by-item JOL aggregate JOL	item-by-item JOL aggregate JOL	item-by-item JOL aggregate JOL	item-by-item JOL aggregate JOL	item-by-item JOL aggregate JOL
Condition B	item pleasantness aggregate JOL	item pleasantness aggregate JOL	item pleasantness aggregate JOL	item pleasantness aggregate JOL	item-by-item JOL aggregate JOL
Condition C	N/A	N/A	N/A	N/A	item-by-item JOL aggregate JOL

Table 4

JOL magnitude, recall, and metacognitive accuracy for Session 5 in Experiment 2 by condition.

Task Experience	Item-by-item JOLs*	Aggregate JOLs*	Recall	Gamma*	Bias (item-by-item)*	Bias (aggregate)*
Condition A (Four Sessions of JOLs and Recall)	0.37 (.02) a	0.35 (.02) a	0.42 (.03)	0.51 (.08) a	-0.05 (.03) a	-0.06 (.03) a
Condition B (Four Sessions of Recall Only)	0.50 (.03) b	0.54 (.04) b	0.42 (.04)	0.52 (.08) a	0.08 (.03) b	0.10 (.03) b
Condition C (No Previous Experience)	0.52 (.03) b	0.44 (.03) c	0.35 (.03)	0.26 (.06) b	0.17 (.05) c	0.09 (.03) b

Note. Main entries are mean values; standard errors of the mean are in parentheses. Entries sharing a common subscript within a column were not reliably different in post-hoc analyses.

* A statistically significant effect was observed, $p < .05$

Table 5

JOL magnitude, recall, and metacognitive accuracy in Session 5 of Experiment 2 for high- versus low-SAT students.

Group	Item-by-item JOLs	Agg-JOL	Recall*	Gamma	Bias (item-by-item)*	Bias (aggregate)*
High SAT	0.45 (.03)	0.45 (.03)	0.48 (.04)	0.30 (.10)	-0.03 (.03)	-0.03 (.03)
Low SAT	0.47 (.04)	0.42 (.04)	0.28 (.03)	0.41 (.09)	0.19 (.05)	0.14 (.04)

Note. Main entries are mean values; standard errors of the mean are in parentheses. In the "High SAT" group, SAT scores were 1200 or greater; in the "Low SAT" group, scores were 1010 or less.

* A statistically significant difference between SAT groups was observed, $p < .05$

Table 6

JOL magnitude, recall, and metacognitive accuracy in Session 5 of Experiment 3 as a function of previous experience.

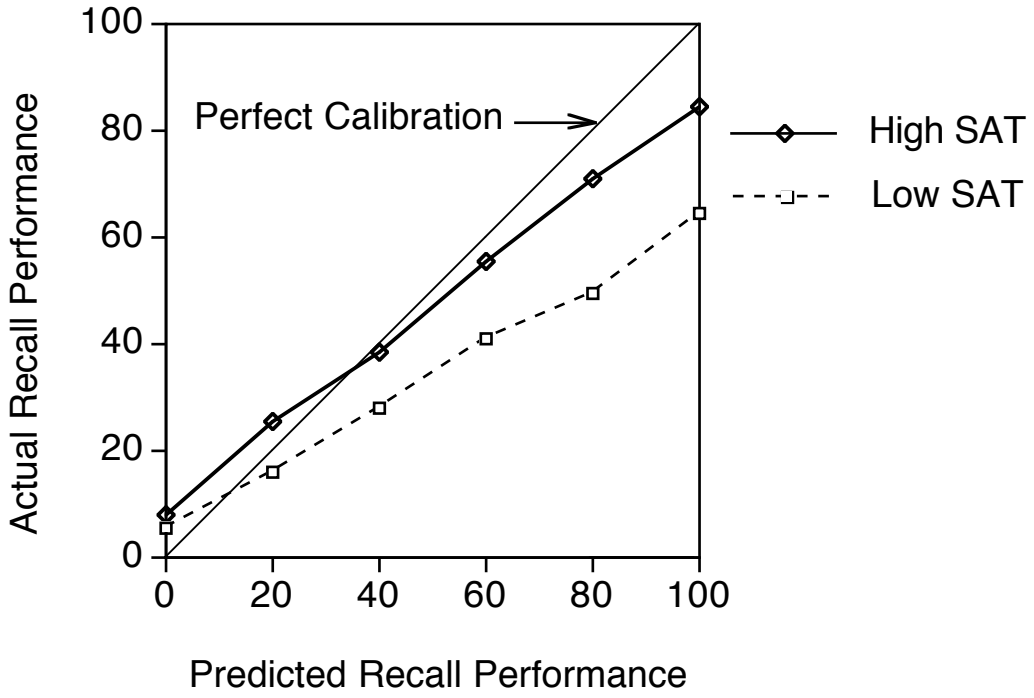
Task Experience	Item-by-item JOLs	Aggregate JOLs	Recall	Gamma*	Bias (item-by-item)*	Bias (aggregate)*
Condition A (Four Sessions of JOLs and Recall)	0.42 (.02)	0.40 (.03)	0.40 (.03)	0.53 (.06) a	0.02 (.03) a	0.0 (.02) a
Condition B (Four Sessions of Ratings and Recall)	0.44 (.04)	0.40 (.04)	0.39 (.05)	0.65 (.06) b	0.05 (.04) a	-0.01 (.03) a
Condition C (No Previous Experience)	0.49 (.03)	0.46 (.04)	0.34 (.04)	0.36 (.06) c	0.15 (.05) b	0.13 (.05) b

Note. Main entries are mean values; standard errors of the mean are in parentheses. Entries sharing a common subscript within a column were not reliably different in post-hoc analyses.

* A statistically reliable effect was observed, $p < .05$

Figure Caption

Figure 1. Mean recall as a function of predicted recall and SAT scores in Experiment 1.


100
80
60
40
20
0
Actual Recall Performance
Perfect Calibration
High SAT
Low SAT
0
20
40
60
80
100
Predicted Recall Performance