Metacognition Learning (2011) 6:303–314
DOI 10.1007/s11409-011-9083-7

Training metacognition in the classroom: the influence of incentives and feedback on exam predictions

Tyler M. Miller · Lisa Geraci

Received: 23 April 2010 / Accepted: 3 August 2011 /
Published online: 12 August 2011

Abstract In two semester-long studies, we examined whether college students could improve their ability to accurately predict their own exam performance across multiple exams. We tested whether providing concrete feedback and incentives (i.e., extra credit) for accuracy would improve predictions by improving students' metacognition, or awareness of their own knowledge. Students' predictions were almost always higher than the grade they earned and this was particularly true for low-performing students. Experiment 1 demonstrated that providing incentives but minimal feedback failed to show improvement in students' metacognition or performance. However, Experiment 2 showed that when feedback was made more concrete, metacognition improved for low performing students although exam scores did not improve across exams, suggesting that feedback and incentives influenced metacognitive monitoring but not control.

Keywords Metacognition · Prediction · Overconfidence · Calibration

In an ideal world each one of us, when asked about the quality or efficiency with which we can accomplish goals, could provide a correct answer. As it is though, no one is immune to flawed self-assessment; doctors, nurses, business managers, and other trusted professionals routinely commit errors of self-assessment, which is one aspect of metacognition (Dunning et al. 2004). *Metacognition* refers to a person's "knowledge and cognition about cognitive phenomena" (Flavell 1979, p. 906). Three aspects of metacognition that have been researched extensively are metacognitive knowledge, metacognitive monitoring, and metacognitive control. In this paper, we focus on metacognitive monitoring, which is an individual's ability to assess the state of their cognitive activity, and metacognitive control, which is an individual's ability to regulate cognitive activity (Dunlosky & Metcalfe 2009). To illustrate these two aspects of metacognition, consider the following example: A student is studying for an anatomy exam and she asks herself how well she remembers the bones of the hand. When she answers this question about the current state of her learning, she has

T. M. Miller (✉) · L. Geraci
Department of Psychology, Texas A&M University, College Station, TX 77843, USA
e-mail: milltyl@tamu.edu

made a metacognitive monitoring judgment. If her current learning state is not satisfactory and she uses that information to decide to spend more time studying she has used metacognitive control.

Individuals can demonstrate their metacognitive monitoring ability by making a prediction about how well they believe they remember a particular fact or a large number of facts. When people make metacognitive monitoring judgments about particular facts on an item-by-item basis, their accuracy is measured by computing a correlation coefficient, this type of measurement is referred to as a relative accuracy, or resolution. In contrast, when people make monitoring judgments about a large number of items, accuracy is measured by the degree to which the prediction corresponds to the actual level of performance; this type of measurement is referred to as absolute accuracy, or calibration (Dunlosky & Metcalfe 2009). Results from research on metacognitive monitoring accuracy show that people aren't always accurate when they make monitoring judgments, and often their inaccuracy is systematic—biased in the direction of overconfidence. Countless studies reveal that individuals overestimate the state of their knowledge or their ability to perform tasks (e.g., Ehrlinger et al. 2008), they believe that they are "better-than-average" (Alicke 1985). Some have argued that there are benefits to overestimation (see Gramzow et al. 2008), like that it encourages good future behavior, but having accurate metacognitive abilities is also important for educational reasons.

Having accurate metacognitive monitoring abilities is important in educational settings for a variety of reasons. Perhaps the most important reason is that research shows accurate metacognition is associated with better academic performance. In one such study to show this relationship, researchers assessed the monitoring ability of incoming college freshman students and compared monitoring ability to their grade point average (GPA) and the end of the semester (Everson & Tobias 1998). In the assessment, students were first asked to identify words they knew and did not know from a word list and then were asked to take an objective test on the same words. Monitoring ability was measured by the difference between the student's estimates and their test performance. Results indicated that students with the highest metacognitive monitoring ability, or the smallest differences between estimates and performance, were also the ones with the highest GPA (Everson & Tobias).

Experimental research also shows that more accurate metacognitive monitoring leads to better academic performance. For example, Thiede et al. (2003) manipulated monitoring accuracy by asking participants to generate keywords about expository texts immediately after reading, after a 5-min delay or not at all. Afterwards, all participants took a comprehension test and then were allowed to self-select and reread texts of their choice. Following the rereading session, participants took another comprehension test. Participants who generated keywords after a delay had better monitoring accuracy and were better able to regulate their study (by choosing and rereading texts appropriately). This combination of superior monitoring and regulation of study led to improved test performance.

Researchers have attempted to improve metacognitive accuracy with the goal of improving performance, both in the laboratory and in the classroom. However, the findings are mixed. For example, in one lab study, repeated practice making performance predictions did improve calibration (Kelemen et al. 2007). Participants were instructed to study different Swahili-English word pairs on five occasions and to indicate the likelihood that they would remember the English words. Results showed that by session 5, participant's predictions significantly improved relative to previous sessions. An interesting finding was that it was only the high achieving students (as measured by participant's SAT scores) who were able to increase their monitoring accuracy, or calibration. Cued-recall performance did not change across test sessions.

In classroom studies, calibration has proven more difficult to modify. For example, in one such classroom study, student participants were asked to predict exam scores on each of three mid-term exams and one final comprehensive exam (Nietfeld et al. 2005). Students made item-by-item performance predictions and predictions for the entire exam. After each exam, students were encouraged to review their predictions, although no feedback or formal monitoring training was provided. Review of the accuracy of predictions was considered to require self-directed feedback because students could assess how well calibrated they were in their predictions. Results showed that performance predictions for the entire exam were more accurate than item-by-item predictions but that both types of monitoring actually decreased from exam 1 to 2. Based on the pattern of data, the authors concluded that self-directed feedback, or simply asking students to review the accuracy of their predictions, was not a sufficient intervention to improve students' metacognitive calibration.

Others have tried to improve calibration in the classroom by providing practice and specific types of training on the value of accurate self-assessment. For example, in one study, students made predictions and postdictions on each of three different exams (Hacker et al. 2000). Students were encouraged to make accurate self-assessments and were informed about the value of accurate self-assessments (e.g., how accurate self-assessment could improve use of feedback, time-management, and goal-setting). They also completed practice exams prior to each exam to obtain more accurate feedback on the state of their knowledge and, after each exam, they were advised to reflect on the accuracy of their predictions and develop a plan to improve their monitoring accuracy. Under these conditions, students' predictions improved across exams whereas postdictions remained stable and consistently more accurate than predictions. When students were split into high and low performance groups based on the percentage of total items answered correctly, results showed that the prediction improvement was carried by the high-performing students. Notably, even though prediction accuracy improved for the high-performing group, overall exam performance did not.

The authors offered reasons why high- but not low-performing students were able to improve their prediction accuracy. First, they suggested that students' use of feedback may vary according to the extent to which they externalize negative outcomes. When poor students receive negative feedback about the accuracy of an exam prediction, they would ideally use the feedback to adjust their performance predictions. Alternatively, they may externalize the negative outcome by attributing it to an external factor such as inferior course instruction or a poorly constructed exam. Another reason the authors offered about why low performers showed no improvement in prediction accuracy was that the incentive used (i.e., motivation to graduate) may have only been effective for high-performing students.

In a subsequent study, Hacker et al. (2008) examined student's attributions for monitoring inaccuracy. Participants in the study made performance predictions for an upcoming exam. After they were told how inaccurate their predictions were, they completed a short attributional style questionnaire that included task-centered questions (e.g., "The instruction wasn't really helpful in preparing us for the test"), student-centered testing questions("I usually get really anxious while taking tests"), student-centered studying questions ("I didn't study as much as I should have"), and social-centered questions ("My interactions with other students in class influenced my judgments"). Participants answered each of the 20 questions on a 5-point Likert scale with the degree to which they believed the question explained the discrepancy between their performance prediction and their actual performance. Results indicated that low-performing students attributed the discrepancy between their prediction and actual performance to external

factors (e.g., "The instruction wasn't really helpful in preparing us for the test.") significantly more so than high-performing students.

Thus, the evidence is somewhat mixed regarding whether calibration can be improved, and research in the classroom suggests that it is difficult to improve students' self-assessments, especially for poorer students. The reasons why poor students in these previous classroom studies have not been found to improve their self-assessments are unknown. One possibility is that low performing students did not improve their performance predictions because the nature of the feedback was simply too general for them to use. For example, in the Hacker et al. (2000) study, participants were instructed to reflect on the accuracy of their judgments after receiving their calibration scores, but poor students may not be able to make use of this type of feedback or instruction.

The purpose of the current studies was to attempt to improve metacognitive accuracy and exam performance for low and high performing students by providing tangible incentives and concrete feedback. Given previous studies have shown that having accurate metacognitive monitoring abilities is important for educational reasons (i.e., improved GPA), and research has shown that attempting to improve students' calibration in the classroom with minimal feedback or incentives is a difficult endeavor. We reasoned that a stronger intervention, with concrete and specific feedback regarding how students could bring their predictions in line with their performance and immediate and tangible incentives would allow both high- and low-performing students to improve their metacognition accuracy. In both experiments participants were asked to make predictions regarding the outcome of 4 different mid-term exams (they predicted whether they would receive an A, B, C etc. on each exam). We examined the accuracy of these predictions how well predictions matched grades, whether grade prediction accuracy improved for all exams 1–4, and whether prediction accuracy was related to class performance. Note that in previous work, improvements from the first exam to the second are not always evaluated for methodological reasons (see Hacker et al. 2008), even though one might expect the biggest improvements early in the course. We predicted that giving students practice and concrete feedback predicting their own grades would lead them to become more proficient at self-monitoring and possibly better students. In addition, to determine whether students were using the feedback appropriately, we asked students to complete a questionnaire at the end of the course regarding their general strategies for incorporating the feedback they received.

In Experiment 1 students were asked to make a global prediction about their exam score on 4 exams immediately prior to taking the exam. Students were given extra credit incentives for prediction accuracy and minimal feedback regarding their accuracy across repeated exams. In Experiment 2 we used the same extra credit incentives for accurate predictions and also provided more explicit, concrete feedback to students regarding their prediction accuracy. Additionally, in Experiment 2, we examined student's self-reported use of the feedback. In previous studies using feedback, there was no way to know how much students were attending to the feedback.

Experiment 1

The purpose of Experiment 1 was to examine whether providing incentives in conjunction with minimal feedback would increase metacognitive accuracy and exam performance across multiple exams. The purpose of this study was to replicate previous work and to serve as a baseline comparison for Experiment 2, which tested the influence of a stronger feedback manipulation.

Method

Participants One hundred thirty students in a cognitive psychology course taught by the one of the authors at Texas A&M University participated in the study.

Design and procedure Immediately before each exam, participants recorded a letter grade prediction on the exam cover sheet (e.g., "B+"). If a student predicted a "B+" that score was converted to a numeric value of 88% based on the midpoint of the B+ range in the standard grading scale. Similarly, "D" predictions, with their corresponding range of 64–66% were given a numeric value of 65% as it is the midpoint of the range. The instructor used the standard grading scale for all exams. For example, scores between 100% and 90% received an "A," between 89% and 80% received a "B," between 79% and 70% received a "C," between 69% and 60% received a "D," and students scoring 59% or less earned an "F."

Students' predictions as well as their actual exam scores were combined to form a calibration score. We used a formula similar to the one used and developed by Hacker et al. (2008). Whereas, Hacker et al. asked students to predict the number of items they would correctly recall, we asked students to predict the percentage of items they would correctly recall. This difference explains why the formula we used contains 100 in the denominator (because 100% is the maximum percentage correct) and the original formula used total items in the denominator.

$$\left(1 - \frac{|Prediction - Grade|}{100}\right) \times 100 \tag{1}$$

Using this formula, students could have calibration scores from 0 to 100, where 0 indicates total inaccuracy and 100 indicates perfect accuracy. For example, a student predicting they would earn an "86" with an actual score of "86" would have a calibration score of 100. Note that the calibration formula we have used yields the same pattern of results as comparable analyses using difference scores; we have chosen to report only the calibration scores to be consistent with previous studies.

In the syllabus and in class students were instructed that they could earn two percentage points of extra credit if they correctly predicted their exam score for each of four mid-term exams. Students earned extra credit if they predicted any version of the correct grade (e.g., they received extra credit if they predicted a "B" and earned a "B-" on the exam). Individual student grades were posted using an online course system (i.e., Blackboard Learning System) where students could see their total grades (either containing extra credit if it was earned, or not). Additionally, following each exam period, the instructor displayed the average exam grade and average exam prediction for the entire class and encouraged students to reflect on the accuracy of their own predictions.

Results

To determine whether or not metacognitive accuracy improved over the course, we first analyzed students' calibrations scores on all four exams using repeated measures ANOVA. Note that reliability analyses indicated an acceptable level of reliability between exams (α=.749). Results revealed that overall, students did not improve $F(2.77, 310.07)=2.55$, $MSE=81.35$, $p=.06$ (using Greenhouse-Geisser correction). Descriptive statistics indicate there was a slight numerical improvement from exam 1 to exam 2.

More consistent with the literature though, we also examined how exam performance effects metacognitive calibration. We created high and low performance groups and entered calibration scores in a repeated measures ANOVA with performance group as a between subjects variable. Performance groups were decided by students' average score on all four exams (their overall course performance). Low-performers' exam scores ranged from 57 to 78 and high-performers' scores ranged from 78 to 93. Results showed a main effect of performance group as expected, $F(1, 111)=8.06$, $MSE=285.72$, $p=.01$, $\eta_p^2=.07$, but no interaction between performance group and exam on metacognitive calibration, $F(3, 333)=.24$, $MSE=29.67$, $p=.87$, $\eta_p^2<.01$). Thus, calibration scores of high performers were higher than the calibration scores of low performers but neither group improved calibration across the four exams (see Table 1 or Fig. 1).

We were also interested in how asking students to predict their exam scores might affect their exam performance. If making predictions encourages students to be more proficient at self-monitoring and they then regulate their study as in Thiede et al. (2003), then exam scores might improve across the four exams. We again used performance group (low or high) as a between subjects variable to see if there were exam score improvements. The main effect of performance group was highly significant as would be expected, $F(1,111)=201.44$, $MSE=74.90$, $p<.001$, $\eta_p^2=.65$, showing that, by definition, high-performing students had better exam grades than low-performing students. However, the interaction between performance group and exam was non-significant, $F(3, 333)=1.05$, $MSE=52.34$, $p=.37$, $\eta_p^2=.01$, indicating that both groups were similarly affected by repeated testing. Therefore, neither the students' calibration scores nor their grades improved over time.

It is possible that the feedback provided for students was not salient or specific enough for them to make adjustments to their predictions. The feedback simply consisted of presenting students with their performance and encouraging students to try to improve their calibration in subsequent exams. Under these conditions there was no improvement in calibration or exam performance, thus corroborating Hacker et al. (2008). In Experiment 2 we attempted to change the type of feedback that was given to students by providing explicit and concrete strategies for improving calibration. We also examined student's self-reports of how (if at all) they used the feedback.

Experiment 2

Method

Participants Eighty-one students in a cognitive psychology course taught by the one of the authors at Texas A&M University participated in the study.

Table 1 Experiment 1: mean grade, prediction and calibration score for each exam by performance group

	Low performing			High performing		
Exam	Prediction	Grade	Calibration	Prediction	Grade	Calibration
1	78.56 (.86)	72.31 (1.26)	91.50 (.95)	84.42 (.69)	84.85 (.73)	94.38 (.54)
2	78.76 (.78)	75.25 (.91)	93.43 (.56)	85.40 (.77)	85.00 (.85)	94.35 (.58)
3	77.55 (.84)	70.82 (.94)	91.22 (.83)	84.68 (.73)	81.87 (.78)	93.46 (.52)
4	78.40 (.72)	73.30 (1.24)	91.17 (.95)	85.27 (.75)	86.13 (.83)	93.17 (.58)

Higher calibration scores (0–100) indicate more accuracy. Standard errors are shown in parentheses

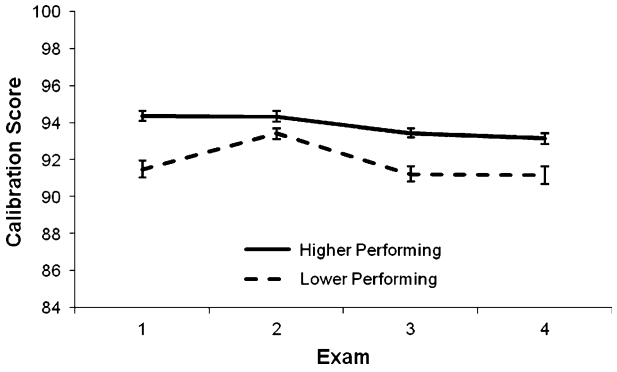

Fig. 1 Calibration scores for high and low performing students in Experiment 1. *Error bars* represent standard error

Design and procedure The design and procedure were identical to those used in Experiment 1 with the following exceptions. In this experiment, students were given more explicit feedback regarding the accuracy of their exam predictions. In addition to posting student's individual exam scores, each student's predicted exam score was posted online to remind students of the prediction and finally, a total exam score was posted that included the extra credit if the student was accurate. In all, a student could view three pieces of information regarding each exam, their raw score, their prediction, and their total score with the extra credit added if the prediction was accurate.

Students were also given in-class feedback regarding their predictions. As before, students saw the grade and prediction data for the entire class. However, in this study, the instructor described the data and encouraged students to improve their metacognitive calibration so they could earn the extra credit. The instructor further specified that there are two ways to increase calibration given that students are generally overconfident. That is the students could either lower their predictions or they could attempt to raise their exam scores. Following the final exam, students completed the post-course survey. One major goal of the survey was to identify how students used the prediction feedback and to examine if there were differences for the two performance groups in reported use of the feedback.

Results

As expected, students were generally overconfident, meaning that their exam score predictions were higher than their exam scores. As in the previous experiment, we first examined whether students improved their calibration across exams using a repeated-measures ANOVA. Given that we used the same exams from Experiment 1 to Experiment 2 we expected the results from reliability analyses to be comparable for both experiments, indeed exams in Experiment 2 had an acceptable level of reliability (α=.746). Results showed that, overall, students' calibration scores did not change across exams, $F(3,213)$= 1.98, MSE=27.98, p=.19, η_p^2=.03. We then examined the influence of performance group on calibration using a median split based on final course grade. Grades for low-performing students ranged from 60 to 78 and grades for high-performing students ranged from 79 to 94 (see Table 2). In contrast to the results from Experiment 1, calibration scores improved for low-performing students. Results from the ANOVA showed that there was a significant interaction between performance group and calibration score, $F(3, 210)=5.13$, MSE=26.44, $p<.01$, η_p^2=.07, showing that lower performing students successfully improved their

Table 2 Experiment 2: mean grade, prediction and calibration score for each exam by performance group

	Low performing			High performing		
Exam	Prediction	Grade	Calibration	Prediction	Grade	Calibration
1	81.78 (.99)	72.05 (1.14)	89.48 (.98)	85.58 (.93)	84.85 (.91)	95.60 (.60)
2	78.78 (.88)	73.90 (1.33)	92.00 (.93)	84.47 (.81)	84.21 (1.30)	93.60 (.92)
3	78.49 (.82)	74.58 (1.48)	91.82 (.89)	83.75 (.88)	86.75 (1.35)	92.35 (.83)
4	79.00 (1.10)	72.95 (1.23)	92.00 (.82)	85.14 (96)	84.50 (1.09)	95.31 (.59)

Higher calibration scores (0–100) indicate more accuracy. Standard errors are shown in parentheses

calibration scores over time but higher performing students did not (see Table 2 or Fig. 2). Planned comparisons for the low-performing students indicated that exams 2–4 calibration scores were higher than exam 1 calibration scores, $F(1, 35)=4.27$, $p<.05$, $\eta_p^2=.11$, $F(1, 35)=4.45$, $p<.05$, $\eta_p^2=.11$, $F(1, 35)=9.10$, $p<.01$, $\eta_p^2=.21$, with an *MSE* of 78.01 for all comparisons. For the high-performing students there was no improvement in calibration, only exam 3 calibration scores were significantly worse than exam 1 calibration scores, $F(1, 35)=10.03$, $MSE=49.9$, $p<.01$, $\eta_p^2=.23$. The lack of a significant improvement for high-performing students may be due to the fact that their calibration scores were very high (i.e., calibration scores of 92 and above). Given that a calibration score of 100 is perfect accuracy, high performers have little opportunity to be overconfident. Similarly, Krueger and Mueller (2002) argued in their analyses regression effects may actually account for asymmetric errors in calibration for low and high performing participants. In the context of the present studies, high performing students' performance on the exam is high (i.e., around 85% for all exams), and low performing students' exam performance was low (around 75% for all exams) so high performers have less room than low performers to make predictions that are above their performance.

To examine whether students' exam scores improved across exams, we conducted a mixed ANOVA with performance group as the between subjects variable and exam (1–4) as the repeated measures variable and exam performance as the dependent variable. The main effect of performance group was highly significant as would be expected, $F(1, 70)=133.63$, $MSE=76.87$, $p<.01$, $\eta_p^2=.66$ but there was no interaction between performance group and repeated exams, $F(3, 210)=.44$, $MSE=57.97$, $p=.72$, $\eta_p^2<.01$. Thus, low-performing students were able to improve their metacognitive accuracy, at least from exam 1 to exam 2, but this did not lead to an improvement in grades.

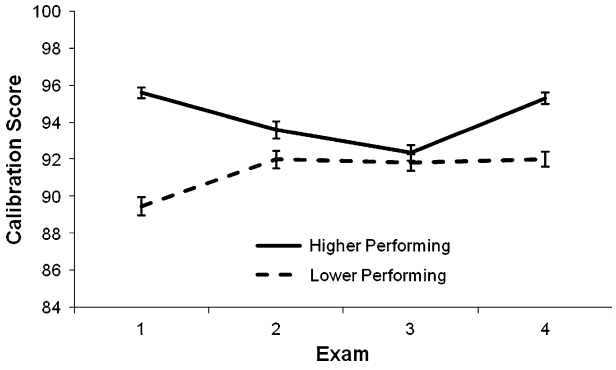

Fig. 2 Calibration scores for high and low performing students in Experiment 2. *Error bars* represent standard error

In this experiment, we also asked students at the end of the course how they had used the exam prediction feedback. Students had several response options—they could respond that they were not influenced by the feedback, that they studied more or less, that they raised/lowered their exam prediction, or other. Students' responses are summarized in Table 3 by performance group. Results from this questionnaire suggest that students made use of the exam prediction feedback. For example, the most common response for low performers was that they studied more or lowered their prediction, which are both effective uses of the feedback. High performers most often responded that the feedback had no influence on them. Given that high-performers' calibration scores were at near ceiling to begin with, no influence must be deemed as an effective use of the feedback. High-performers' second most common response was that they studied more, an ineffective use of use of the feedback. One possible explanation for this finding is that high performing students are also highly conscientious and will always report studying more. Alternatively, there was a range in exam performance for high performers, and thus students in the lower range of this group could be actually study more rather than lowering their prediction. Together, the pattern of results from Experiment 2 suggests that students used the feedback appropriately to lower their predictions and to increase study. Further, the observed data suggest that the failure to find improved exam performance was not a result of students failing to attend to the feedback. Rather, it appears that lowering predictions improved calibration but the reported increased study time did not affect performance.

General discussion

The current experiments examined whether students could improve their metacognitive calibration when they were given incentives and feedback. The common result from both experiments was that students were overconfident, low-performing students even more so than high-performing students. As such, the current findings are consistent with the ever-expanding literature showing that people are mostly overconfident in their self assessments (e.g., Dunning et al. 2004; Kelemen et al. 2007; Kruger & Dunning 1999). Experiment 1 showed that when students had the opportunity to earn extra credit for accurate predictions and were given feedback regarding their performance, they were not able to improve their metacognitive calibration. In Experiment 2, when feedback was made more explicit and concrete, low-performing students did improve their calibration. Post-exam questions indicated that students used the feedback appropriately, suggesting that the failure to find improved exam performance was not a result of students failing to attend to the feedback.

More important, we showed that improving metacognition is possible. Using a combination of explicit and concrete feedback, multiple exams and multiple opportunities to make exam predictions in addition to tangible incentives, low-performing students improved their calibration accuracy from the first to the second exam. The fact that these

Table 3 Frequencies counts for Student's responses to the question "What influence did the exam prediction feedback have on you?"

	No influence	Studied more	Studied less	Lowered prediction	Raised prediction	Other
Low performers	5	14	0	16	2	1
High performers	15	8	1	5	1	7

conditions allowed low-performing students to improve their metacognitive accuracy is notable as previous studies found that, if metacognition can be improved it is typically only the high-performing students who benefit, and early exam improvements are not always evaluated (Nietfeld et al. 2005, Hacker et al. 2000, 2008).

The finding that lower performers were only able to improve their metacognitive accuracy from exam one to exam two suggests that there may be limits on how much low-performing students can improve their calibration. Given that low-performing students are predominantly overconfident, they would need to lower their exam predictions substantially to improve their calibration. Low performing students may not lower their predictions much beyond a certain point for self perseveration reasons, because they "just want to think good things about themselves while denying bad things" (Ehrlinger et al. 2008, p. 107). In fact some have suggested that a certain amount of overconfidence may serve as an adaptive behavior because overestimation could offer motivation for students to try harder in the future (Gramzow et al. 2008). We assumed our inclusion of extra credit points for accurate self-assessment would circumvent this issue, however, two points extra credit may not be enough for some students to override this type of behavior.

Another interpretation for the lack of continued improvement in calibration for low performers and no improvement for high performers is that providing extrinsic incentives for students to improve their calibration may not be a very effective method to change student behavior. Deci et al. (2001) suggested that previous meta-analyses show that extrinsic rewards can undermine intrinsic motivation. In the present study, we provided extra credit points for students if their performance predictions were accurate. This incentive, according to some researchers, might have reduced students' intrinsic motivation to improve their performance predictions. However, in the present study low performing students did make early improvements in their calibration and their subsequent calibration remained stable—that is, there was little evidence for declines in accuracy due to reduced intrinsic motivation. It is possible though that reduced intrinsic motivation could contribute to the lack of further improvement.

In addition to only showing improvements from exam 1 to exam 2, low performing students only improved calibration by lowering their predictions. They did not, or could not, improve calibration by improving their performance. According to the double-curse explanation of low performing students abilities (see Ehrlinger et al. 2008), low performers' inadequacy at the task prevents them from doing well and also precludes their awareness of what they do not know. Thus, although they can be trained to lower their predictions, they may not be easily trained to improve their performance. Improving performance may require more specific individually tailored feedback. The feedback in the current studies simply indicated that, in general, low performing students were overconfident. The instructor stressed the idea that calibration could be improved by improving exam performance and/or by lowering exam predictions. However, the feedback was not tailored to each individual in the sense that it did not inform students of which concepts that student understood and which ones they did not understand, a critical failure of understanding according to the double-curse explanation. Further, because there were new concepts tested on each exam, low performing students may have found themselves in essentially the same situation on subsequent exams, having little idea which concepts they understand and which concepts they do not understand. Thus, even though they may have been able to use the general feedback to some extent, they may not have been able to improve calibration indefinitely because they could not accurately assess how well they understood new information. In essence, these students were still operating on incomplete metacognitive information.

There may be ways to tailor feedback to lower performing students to help them improve their metacognitive accuracy. Testing is likely a critical variable in training metacognition. If students had been required to take the practice exams as in the Hacker et al. (2000) study and were diligent about taking them (rather than perhaps only looking at the answers) they could have received valuable feedback about their understanding of the material and their exam preparation, including item-by-item feedback. This sort of information would be helpful, at least in instances where those concepts are likely to be retested like they would be on a final exam. As recent work suggests, testing may be a critical factor for learning because of the metacognitive information that it can provide. Karpicke and Roediger (2008) showed that multiple retrieval opportunities enhanced participants' long-term retention of Swahili-English word pairs. In their study, asking participants to take a test, or practice retrieving word meanings from memory, enhanced long-term retention even more so than additional study—a finding that is commonly referred to as the "testing effect." The testing effect is relevant here because if students tested themselves while they studied for an exam it would provide valuable information about how well they know the material. Students could use this information to inform their decisions to study some material over other material (see Dunlosky et al. 2002).

There are multiple avenues for future research in the area of improving metacognitive monitoring and control performance in the classroom. For example, to address one limitation of the current study, future research could track students' predictions across exams. Tracking individual student's predictions would provide detailed student-specific information about the type of adjustments students are making to their predictions. It would also reveal the pattern of over- or underconfidence and might indicate critical periods for metacognitive monitoring improvement on an individual student basis (i.e., from the first to second exam). Tracking students from exam to exam would also provide useful information about the characteristics of underconfident students, a group that has received less research focus. Although the focus of the current experiment was to improve all students' metacogntive monitoring and control performance, the results indicated that only the low performers recalibrated. Future research might also examine manipulations that would encourage underconfident students (usually high performers) to increase their predictions to become more calibrated. For instance, high performers might also benefit from individually-tailored feedback about the biases in their predictions.

Beyond the goals of improving metacognition and exam performance, there may be other pedagogical reasons to ask students to predict their exam performance, particularly before taking an exam. Anecdotally, it seemed as though asking students to predict their grade before taking the exam forced them to assess their own readiness for the exam, independent of the their assessment of the exam itself. In other words, to make this assessment, students had to consider their own preparedness for the exam, which may have increased their sense of responsibility for their performance in the class. Compared to previous semesters, the instructor noted that fewer students approached her claiming that they were surprised by the grade they earned and that they "don't understand why they got a C on the exam." Thus, there may be good pedagogical reasons to ask students to try to accurately predict their exam performance beforehand. In addition, a somewhat unexpected positive outcome from the present research was that students actually reported enjoying predicting their grades. They indicated on evaluation forms and in person that they enjoyed learning about the concept of metacognition through the hands-on experience of predicting their grades in addition to lecture and the readings.

References

Alicke, M. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*, 1621–1630.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation: reconsidered once again. *Review of Educational Research, 71*, 1–27.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69–106.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: SAGE Publications.

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 68–92). Cambridge: Cambridge University Press.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*, 98–121.

Everson, H. T., & Tobias, S. (1998). The ability to estimate knowledge and performance in college: a metacognitive analysis. *Instructional Science, 26*, 65–79.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*, 906–911.

Gramzow, R. H., Willard, G., & Mendes, W. B. (2008). Big tales and cool heads: academic exaggeration is related to cardiac vagal reactivity. *Emotion, 8*, 138–144.

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: the effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*, 101–121.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170.

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

Kelemen, W. L., Winningham, R. G., & Weaver, C. A., III. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology, 19*, 689–717.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*, 180–188.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121–1134.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*, 7–28.

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