ASSOCIATION OF
ATLANTIC
UNIVERSITIES

ASSOCIATION
DES UNIVERSITÉS
DE L'ATLANTIQUE

Proceedings of the 2017
Atlantic Universities' Teaching Showcase

Volume 21 | Pages 23-32 | https://ojs.library.dal.ca/auts/index

ONLINE QUIZZES, METACOGNITION, AND THE TESTING EFFECT: AN OPTIMISTIC, CAUTIONARY TALE

Jennifer L. Tomes

Association of Atlantic Universities/
Association des universités de l'atlantque
Editor-in-Chief: Donovan Plumb
www.atlanticuniversities.ca

Volume 21 | *Pages 23-32*

www.atlanticuniversities.ca

ONLINE QUIZZES, METACOGNITION, AND THE TESTING EFFECT: AN OPTIMISTIC, CAUTIONARY TALE

Jennifer L. Tomes
Department of Psychology
Mount Allison University

Abstract

A program intended to improve metacognitive predictions of performance was implemented via a series of online quizzes. Students in a first-year Introductory Psychology class were given the opportunity to practice making predictive metacognitive judgments of performance on a series of online quizzes and on mid-term and final exams. There was no indication that practice making metacognitive judgments improved students' metacognitive calibration or bias scores. However, evidence of a testing effect was found, with both high and low performing students who completed more of the quizzes achieving higher marks on the final exam. Instructors who are considering implementing a manipulation or treatment in a classroom setting are encouraged to imbed the treatment within a paradigm that has known benefits to students, such as the testing effect.

Keywords

metacognition; online quizzes; testing effect; scholarship of teaching and learning

Association of Atlantic Universities/
Association des universités de l'atlantque
Volume 21 | Pages 23-32

Strong metacognitive skills are important for students as they can help students assess what they know/do not know and allocate study time and resources appropriately. Recently, researchers have focused on student metacognitive predictions of academic performance. These are prospective judgments as an individual must assess their knowledge before retrieving that information and may reflect an individual's assessment of their mastery of the material (Hacker, Bol, & Keener, 2008). These predictions of performance involve the monitoring aspect of metacognition and reflect our awareness of, and ability to, assess our cognitive processes (Dunlosky & Metcalfe, 2009). Unfortunately, our abilities to predict performance are far from perfect. Past research has demonstrated that many students make inappropriate, typically overconfident judgments about their performance on tests and other tasks (e.g., Kruger & Dunning, 1999; Miller & Geraci, 2011); further, the propensity toward overconfidence and poorly calibrated metacognitive judgments tends to be exacerbated for lower performing individuals (e.g., Hacker, Bol, Horgan, Rakow, 2000; Kelemen, Winningham, & Weaver, 2007).

Most previous studies have found that metacognitive skills are difficult to improve, often revealing modest or no improvement (e.g. Bol & Hacker, 2001; Hacker et al., 2000; Nietfeld, Cao, & Osborne, 2005). However, two recent studies have resulted in improved metacognition in classroom settings. Ryvkin, Krajc, and Ortmann (2012) report two experiments, one conducted in a classroom and the other in a controlled setting. In their first study, participants in the classroom twice predicted their performance on an upcoming mid-term (five weeks before the test and five minutes before taking the test) and predicted their performance on their final exam (immediately before the exam). Students were asked to predict their absolute performance and also their performance relative to others in the class. A "non-trivial" (p. 1016) monetary incentive was offered to the student with the most accurate prediction. Ryvkin et al. (2012) found students, and most notably low performing students, were able to reduce their miscalibration after being exposed to feedback about both the nature of the task and their absolute and relative performance. It should be noted that the feedback regarding the nature of the task was the naturally occurring experience with the course material throughout the term, and feedback regarding their absolute and relative performance was simple feedback about how accurate the predictions of their performance were. A second study, involving an experimental task, found similar results, again with the low performing students improving their metacognitive judgments the most.

Miller and Geraci (2011) conducted two studies in the classroom setting. In the first experiment, students were given additional marks (two percentage points) if they were well calibrated in their predictive judgments on four mid-terms. However, the incentive alone did not produce improvements in metacognitive ability. In their second experiment, they provided both an incentive to improve predictions and more detailed in-class feedback regarding metacognitive judgments, including strategies to improve calibration. In the second experiment, low performing students showed improved calibration in their metacognitive judgments from exam one to exam two, but calibration did not continue to improve for subsequent exams. Miller and Geraci (2011) suggested one potential reason for the lack of continued improvement is that the external reward (extra marks) may have decreased intrinsic motivation. Miller and Geraci (2011) also note that lower performing students only improved their calibration by reducing their overconfidence, not by improving their exam performance.

The intent of the current study was to build on the success of Ryvkin et al. (2012) and Miller and Geraci (2011) to attempt to improve student metacognition via a series of online quizzes and in-class assessments. The current study involved more opportunities to make metacognitive judgments and refine calibration than the previous studies that were successful in improving calibration. Miller and Geraci (2011) had four mid-term evaluations, while Ryvkin et al. (2012) had participants make three metacognitive judgments in Study 1 and two metacognitive judgments in Study 2. The current study involved three traditional course assessments (two mid-terms and a final exam), as well as nine online quizzes spaced throughout the academic term. Students were given feedback on the accuracy of their predictions versus actual performance for all quizzes and assessments similar to that found in Ryvkin et al. (2012), but unlike Ryvkin et al. (2012) and Miller and Geraci (2011) no incentives for improved or accurate predictions were given. In the current study, students were given a small percentage of their grade, a similar amount to Miller and Geraci (2011), for completing a portion of the quizzes, but the marks were not incentivized toward accurate metacognitive judgments in an attempt to maintain intrinsic motivation. It was initially predicted that the metacognitive intervention would improve calibration, especially for students who engaged in more opportunities to practice their metacognitive judgments, and also for lower performing students.

Method

Participants

One hundred and thirty-two of 188 students registered in an introductory psychology class at a Canadian university agreed to participate in the study. Six individuals were removed for not completing the mid-terms and/or the final exam and 22 individuals failed to make a prediction for one or more of the assessments and were removed from the analyses.

Procedure and Materials

In the first week, all students in the introductory psychology class were invited to participate in the study. Written informed consent was obtained to track participants’ performance and metacognitive judgments on: (a) the quizzes; (b) the class mid-terms and final exam; as well as (c) to allow access to GPA information for that academic year. Unit quizzes were available to all students in the class via an online program (Moodle). Prior to completing each quiz students were given the following metacognitive instructions: “Before beginning the quiz, take a moment to think about how well you know this topic. For this quiz there are 10 questions. Please predict how well you will do… what will your score on this test be? What percentage of the questions will you get correct? Enter a number between 0–100 to represent the percentage of items you think you will get correct based on your knowledge of this topic.” Students were instructed to check the accuracy of their prediction when they finished the quiz.

The quizzes involved multiple choice items designed to reflect key material taught in the class and tested on mid-terms and the final exam. The quizzes were designed to enhance the testing effect (Carrier & Pashler, 1992; Roediger & Karpicke, 2006). Corrective feedback was given for each item; students were told if their response was correct or incorrect and an explanation of the correct response was given. At the end of each quiz, students were given their total score and were prompted again to reflect on the accuracy of their prediction.

The mid-terms and final exam were composed of both multiple choice and short answer/identification items. The mid-terms only covered specified chapters, whereas the final exam included a cumulative section, which was made up of major concepts taught across the entire course that had all been included in the online quizzes. At the start of each class assessment, students were asked to predict their performance on the test/exam with the following statement: “Before beginning the test/exam, take a moment to think about how well prepared you are and how well you know the course material being tested.” They were asked to give a score from 0–100.[1] Once the tests were marked, students were given feedback regarding both their actual grade and their predicted grade.

---

[1] Miller and Geraci (2011) asked students to predict a letter grade which was later converted into the numerical mid-point of the letter grade range. In the current study, students were asked to predict a numerical grade for both tests and the final exam. Performance on the assessments was also evaluated as a percentage rather than a letter grade.

Results

Completion of the quizzes was quite high, with students completing an average of 6.03 of nine quizzes (*SD* = 2.275); only a single participant did not complete any of the quizzes. As the intent of the study was to assess the impact of practice completing metacognitive judgments, the following analyses include those who had completed at least one quiz per assessment period (*N* = 102). Each assessment period consisted of three quizzes and reflected the course material for the two mid-terms and final exam. Exams had a high level of reliability ($\alpha = .904$). The reliability of the quizzes was lower ($\alpha = .725$), but still acceptable.

A repeated measures ANOVA was conducted,[2] using academic performance and quiz completion as between-subject factors to assess metacognitive improvements across the term. Calibration was computed in the same manner as Miller and Geraci (2011), with scores closer to 100 representing better calibration. Students were divided into higher and lower academic performance based on a median split of overall GPA for the academic year. Similarly, students were divided into those who completed less than or more than six quizzes. There was no main effect of calibration ($F(2, 164) = 1.169$, $MSE = 30.578$, $p = .313$; $M_{\text{Test1}} = 90.646$, $SD = 9.755$, $M_{\text{Test2}} = 90.061$, $SD = 8.839$, $M_{\text{Final}} = 89.220$, $SD = 8.980$), nor did calibration interact with academic performance ($F(2, 164) = 2.258$, $MSE = 69.031$, $p = .108$) or completion of quizzes ($F(2, 164) = .400$, $MSE = 12.225$, $p = .671$). There was also no significant main effect for academic performance level ($F(1,82) = 1.577$, $MSE = 190.814$, $p = .213$), or completion of quizzes ($F(1,82) = 1.665$, $MSE = 190.814$, $p = .201$). The mean grade, predicted grade, and calibration scores can be found in Table 1. Unlike Miller and Geraci (2011), the current study found no significant differences in calibration between higher and lower performing students. As can be seen in Figure 1, although higher GPA students were consistently better calibrated than lower performing students, the difference was not statistically significant.

---

[2] Data was trimmed for two outliers to avoid violating assumptions.

Table 1: Mean Predicted Grade, Actual Grade, and Calibration Score Across Academic Performance and Quiz Completion Groups

	Low GPA						High GPA					
	Low Quiz			High Quiz			Low Quiz			High Quiz		
	Pred	Act	Cal	Pred	Act	Cal	Pred	Act	Cal	Pred	Act	Cal
Test 1	69.7	60.8	86.6	72.1	68.7	91.0	75.6	81.1	91.5	**81.7**	**83.6**	**92.4**
Test 2	65.0	53.1	85.6	69.2	64.1	90.5	71.9	73.9	91.1	**78.2**	**80.1**	**92.0**
Final	63.0	59.1	88.2	70.1	70.4	89.8	73.8	80.6	88.5	**78.0**	**87.7**	**89.7**

Note: Pred = predicted grade; Act = actual grade; Cal = calibration

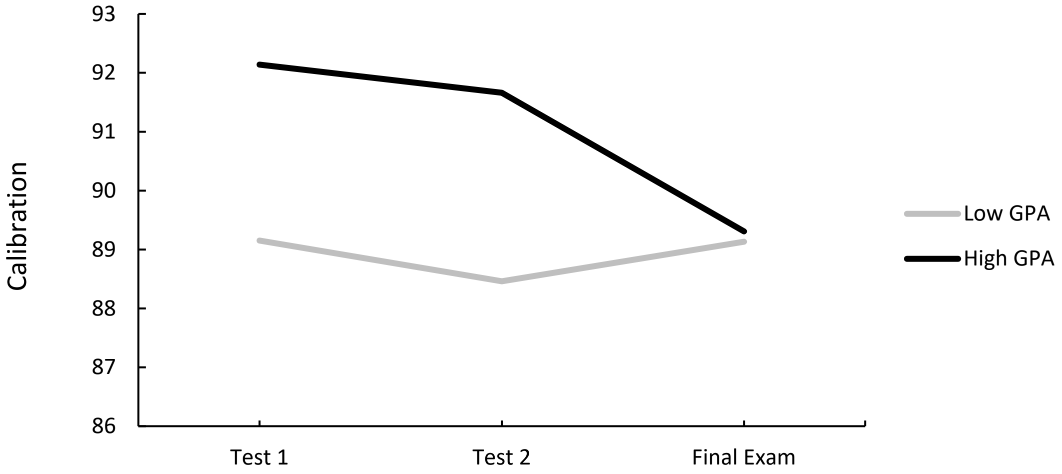

Figure 1. Calibration scores of high and low performing students across in-class assessments.

Bias was computed as predicted performance minus actual performance, with positive bias scores reflecting overconfidence and negative scores reflecting underconfidence. When examining bias, a repeated measures ANOVA was conducted using academic performance and quiz completion as between-subject factors. Overall there was a significant change in bias over time ($F(1.863, 152.758) = 12.714$, $MSE = 58.392$, $p < .001$, using a Greenhouse-Geisser correction). As can be seen in Figure 2, estimates of bias were relatively stable for the first and second tests, but dropped for the final exam; overall, students went from being slightly overconfident in their performance on the mid-term tests to being slightly underconfident in their performance on the final exam. There was no significant interaction of bias and academic performance ($F(1.863, 152.758) = .320$, $MSE = 17.400$, $p = .727$) or quiz completion ($F(1.863, 152.758) = .209$, $MSE = 11.381$, $p = .811$). There was a significant main effect for academic performance ($F(1, 82) = 11.929$, $MSE = 382.514$, $p < .001$),

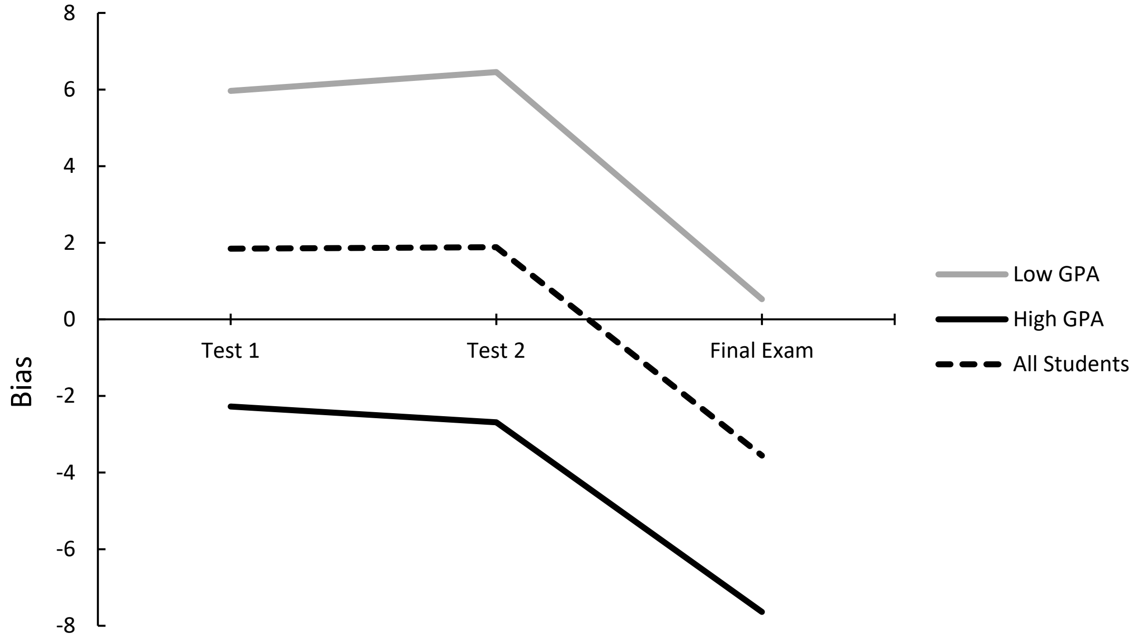

Figure 2. Bias scores of students across in-class assessments.

with higher performing students being underconfident (*M* = -4.039, *SE* = 1.837) and lower performing student being overconfident (*M* = 4.713, *SE* = 1.745). There was no significant main effect for quiz completion (*F*(1, 82) = 1.314, *MSE* = 382.514, *p* = 255).

These results suggest that the online practice with metacognitive judgments did not improve performance for either calibration or bias with respect to predictive judgments. Although there was a change in bias over time, it was not related to number of quizzes completed.

Testing Effect

A 2 (academic performance) x 2 (quiz completion) ANOVA, conducted using final exam grade as the dependent variable, revealed a significant main effect for both academic performance (*F*(1, 123) = 62.786, *MSE* = 171.978, *p* < .001) and quiz completion (*F*(1, 123) = 11.169, *MSE* = 171.978, *p* = .001) in terms of performance on the final exam. There was no significant interaction between academic performance and quiz completion (*F*(1, 123) = .231, *MSE* = 171.978, *p* = .632). Not surprisingly, the higher performing students, as assessed by overall GPA for the current year, tended to score higher on the final exam than the lower performing students, but those students who completed more of the quizzes also tended to perform better on the final exam, regardless of their academic performance (see Figure 3).

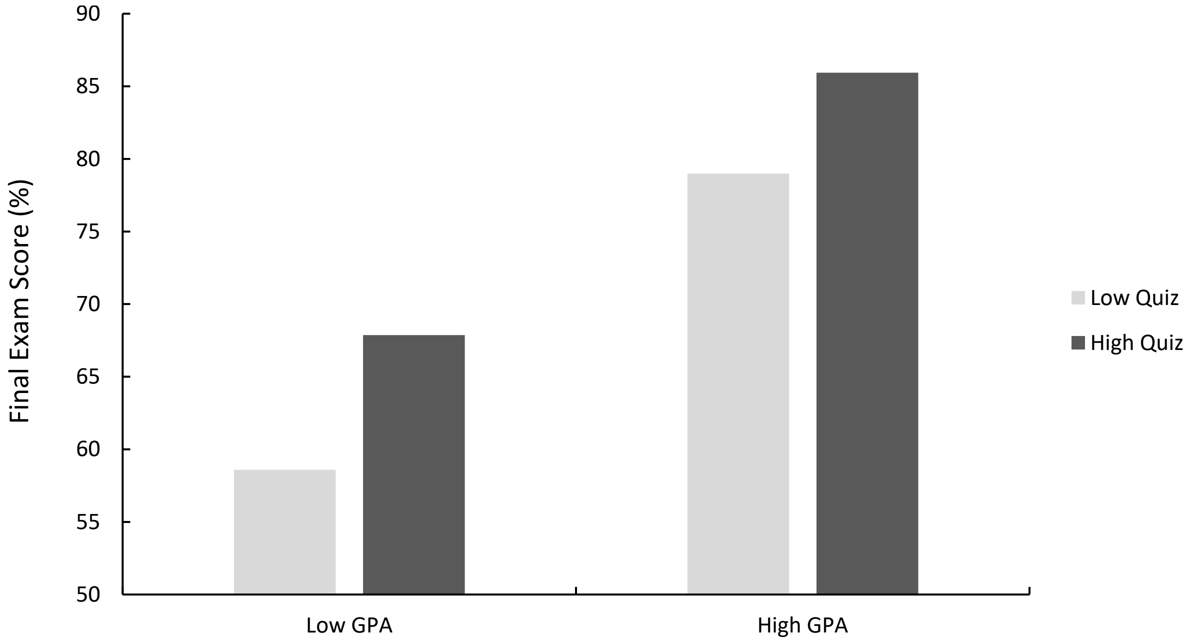

Figure 3. Mean performance on final exam by academic performance and quiz completion.

Discussion

Unlike several recent studies (Miller & Geraci, 2011; Ryvkin et al., 2012), in the current study no evidence was found of improved metacognitive predictions in spite of increased opportunities to practice making these judgments and multiple opportunities to integrate what Ryvkin et al. (2012) referred to as environmental feedback and calibration feedback. Higher performing students were slightly, but not significantly, more well calibrated than lower performing students, however there were no changes in calibration across the term that were related to either participating in the online quizzes or academic performance.

Students in the current study were given a similar amount of information and feedback as those in the Ryvkin et al. (2012) study, although over many more testing opportunities, yet they failed to show improved metacognition. One important difference between the current study and both Ryvkin et al. (2012) and Miller and Geraci (2011) involves external incentives to improve metacognition. Although students in the current study were given a similar proportion of course marks as those in the Miller and Geraci (2011) study, here the marks were tied to making metacognitive judgments and completing the quizzes. In both Miller and Geraci (2011) and Ryvkin et al. (2012) participants received external rewards (grades and monetary rewards, respectively) that were specifically tied to improvements in metacognitive performance. Future researchers interested in metacognition in the classroom are encouraged to look more closely at strong incentives for changes in metacognitive performance as it may be that these are more critical than increased opportunities for information and feedback in improving metacognitive predictions.

In spite of the non-significant findings with respect to metacognitive performance, there is an optimistic take-away message from the results of the current study. Analyses revealed evidence of a significant testing effect. Students who completed more of the online quizzes throughout the term performed better on the final exam in the course. Some caution should be used in assessing this testing effect as it does not reflect an experimental manipulation; individuals were categorized into high or low conditions based on student behaviours. However, it does reveal that those students who chose to complete more of the online quizzes benefitted regardless of their overall academic standing. Thus, although the manipulation may not have been successful in terms of significant improvements in metacognitive ability, those who engaged with the manipulation to a greater degree did show improvement in terms of their knowledge of the course material. Based on the findings in the current study, researchers who wish to implement treatment protocols to improve metacognition and/or other aspects of performance in classrooms or other real-world settings, are encouraged to embed the treatment into an effect, like the testing effect, that has been shown to have benefits for students (Roediger & Karpicke, 2006). Then, even if the targeted skill does not improve, the robust testing effect can yield academic gains for the students involved.

References

Bol, L. & Hacker, D.J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69,* 133–151.

Carrier, M. & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 632–642.

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

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.

Hacker, D.J., Bol, L., & Keener, M.C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. Bjork (Eds.), *Handbook of Memory and Metacognition* (pp. 429–455). Mahwah, NJ: Lawrence Erlbaum & Associates.

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

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

Miller, T.M. & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition & Learning, 6,* 303–314.

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.

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

Ryvkin, D., Krajc, M., & Ortmann, A. (2012). Are the unskilled doomed to remain unaware? *Journal of Economic Psychology, 33,* 1012–1031.

Author

Jennifer L. Tomes. Jennifer is an Associate Professor and Head of the Department of Psychology at Mount Allison University. Jennifer was the 2017 winner of the Tucker Teaching Award. She is a cognitive psychologist, with a specialization in human memory. Her research interests include the scholarship of teaching and learning, the development of false memories, and the long-term impact of concussion.