2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023
The Harbor of Engineering Education for 130 Years
Paper ID #38667

Board 371: Relationships Between Metacognitive Monitoring During Exams and Exam Performance in Engineering Statics

Dr. Chris Venters, East Carolina University

Chris Venters is an Assistant Professor in the Department of Engineering at East Carolina University in Greenville, North Carolina, USA. He teaches introductory courses in engineering design and mechanics and upper-level courses in fluid mechanics. He earned his Ph.D. in Engineering Education from Virginia Tech in 2014, and his research primarily focuses on conceptual understanding in engineering mechanics courses. He received his M.S. in Aerospace Engineering from Virginia Tech and his B.S. in Aerospace Engineering from North Carolina State University.

Dr. Saryn Goldberg, Hofstra University

Dr. Saryn R. Goldberg is an Associate Professor of Mechanical Engineering in Hofstra University's School of Engineering and Applied Sciences. Dr. Goldberg received her Sc.B. in Engineering with a focus on materials science from Brown University, her M.S. degree in Biomedical Engineering with a focus on biomaterials from Northwestern University, and her Ph.D. in Mechanical Engineering with a focus on biomechanics from Stanford University. At Hofstra she teaches courses in mechanical engineering and materials science. Her research in engineering education focuses on the use of student question-asking to promote metacognition. She is a member of the Society of Women Engineers and the American Society of Engineering Education.

Amy Masnick, Hofstra University

Dr. Amy Masnick is an Associate Professor of Psychology at Hofstra University. Dr. Masnick received both her B.S. and Ph.D. in Human Development at Cornell University. At Hofstra she teaches courses in introductory psychology, research methods, cognitive

Kaelyn Marks, Hofstra University
Kareem Panton, Hofstra University

Relationships Between Metacognitive Monitoring During Exams and Exam Performance in Engineering Statics

INTRODUCTION

Our NSF-DUE-funded project studies whether providing students with training and practice writing questions about their confusions in an undergraduate engineering statics course supports improved course performance and metacognitive awareness. Data collection for the project includes assessing multiple measures of students' metacognition, including metacognitive monitoring during statics exams. In this current study, we focus exclusively on the monitoring data collected thus far.

Metacognitive monitoring is the process of observing one's understanding and approach while completing a learning task [1]. One way to assess students' metacognitive monitoring is to measure students' ability to accurately either predict or postdict their score on an assessment of their understanding [2], where postdiction refers to students assessing their expected score after completing a learning task. The mismatch between students' confidence estimates and their actual performance is referred to as their level of calibration [3].

Prediction of exam performance prior to answering a question is an example of calibration of comprehension, as it requires a student to provide a confidence estimate of their ability to answer a forthcoming question, while postdiction is an example of calibration of performance, as it requires a student to provide a confidence estimate of the answer that they already provided to a question [4]. A benefit of using postdiction as a measure of calibration is that students' estimates are not muddied by assumptions about or lack of familiarity with the expected learning task [4].

Studies of students' postdiction of exam performance have been carried out at the undergraduate level in a range of fields, including psychology [4-8], education [9], biology [10], physics [11], chemistry [12-13], and technology [14]. Studies that relate student performance to postdiction calibration generally find that higher-performing students are better calibrated (i.e., can more accurately estimate their score) than lower-performing students [4,6,7,9,10,12,14]. Further, students' calibration accuracy does not appear to change from the beginning of a course to the end [4,6,9,12] unless specific interventions are employed to improve students' metacognitive monitoring skills [8,11]. One exception to this trend may be students' improvement from the first exam to a subsequent exam, which may be due to students' increased familiarity with the exam format [12].

We are aware of limited examples of the study of postdiction calibration capabilities of undergraduate engineering students. Christensen et al. [15] used postdiction of exam performance as one of many metrics to evaluate student responses to statics exam questions that were either close to the course content that students studied or were more of a "stretch". Goodmann and Isaacson [16] incentivized students to accurately identify the questions on circuits exams on which they performed the best. Baisley et al. [17] asked students to postdict their performance on mechanics exam questions by having students grade them using the same rubric as the instructors. They observed that students matched the instructor-determined grades less than 50% of the time. However, the rubric required students to discern between a "minor error," a "minor logic error" and a "significant conceptual error," such that poor performance on the calibration task may have been reflective of students' inability to discern between these types of mistakes.

In this study we will examine preliminary data collected in an engineering statics course to observe whether our students follow trends observed with postdiction calibration in other fields. Specifically, we are interested in determining if:

1) High-performing students are better calibrated than low-performing students, and
2) If student calibration improves from Exam 1 to Exam 2 but does not continue to improve from Exam 2 to the Final Exam.

METHOD

Data were collected from undergraduate engineering students enrolled in engineering statics during three semesters: Fall 2021, Spring 2022, and Fall 2022. The students for the present study were from a private university located in the northeastern region of the United States. The undergraduate engineering program at the university is small, allowing for small course sizes. One instructor taught all five of the courses included in the present study.

Table 1. Summary of course characteristics.

Semester	Sections	Students Enrolled	Students Included in Study
Fall 2021	2	35	27
Spring 2022	1	26	13
Fall 2022	2	37	30

A total of 70 participants were included in the analyses, as shown in Table 1. Students who were enrolled in the courses but were excluded from the analyses are those who did not consent to have their data analyzed, did not complete all calibration responses, or were repeating the course.

Content knowledge of the course material was assessed through scores on two exams (Exam 1 and Exam 2) and a final cumulative exam (Final Exam). The exams administered each semester varied slightly to decrease the likelihood of students contaminating future students' responses. All exams were graded on a 100-point scale and included on average three multi-step questions.

Metacognitive monitoring was assessed through students' calibration on each individual exam question. During the exam, students were shown how many points each question was worth. After answering each question on the exam, students were prompted to make a postdiction estimate for how many points they thought they would receive on that question.

We calculated a calibration metric using an adaptation of the bias index described by Schraw [2]. Schraw's bias index assumes student performance is scored as either correct or incorrect while confidence is reported on a scale from 1-100%. In our case, performance and confidence are both measured using points out of a possible maximum value for each question. Therefore, we calculated the calibration for each exam question as the absolute value of the difference between their predicted and actual score, normalized by the number of points possible. For example, a student who predicted a score of 20 points and earned 15 points on a 25-point question would indicate a calibration of 0.20. We then averaged the question calibrations for each exam to calculate an exam calibration. We also calculated a semester calibration by averaging the question calibrations across all exams. By using the absolute value in these calculations, we are ignoring the directionality of students' predictions (over- or under-estimation), so that over- and under-predictions do not cancel across questions – any error between students' actual score and predicted score is maintained in the calculation. A calibration that is closer to zero indicates better alignment between predicted scores and actual scores.

RESULTS

All statistical analyses were performed using JASP version 0.16.4 software.

Descriptive statistics for the exam scores and calculated calibration scores are shown in Table 2 below. The range of scores on each exam shows the presence of very low scores in each case. Median scores higher than the corresponding mean score for each exam suggest a non-normal distribution of scores.

Table 2. Exam score and calibration statistics.

	Exam 1 Score	Exam 1 Calibration	Exam 2 Score	Exam 2 Calibration	Final Exam Score	Final Exam Calibration	Exam Average	Semester Calibration
Median	73.5	0.16	66.0	0.16	71.9	0.15	71.0	0.16
Mean	66.0	0.18	64.9	0.16	68.1	0.16	66.5	0.17
Std. Dev.	25.1	0.10	18.5	0.08	21.9	0.09	20.1	0.06
Minimum	5.0	0.02	17.0	0.03	10.0	0.04	11.8	0.06
Maximum	96.0	0.48	95.0	0.38	98.0	0.40	95.4	0.34

To investigate the distributions further, we created dot plots shown in Figure 1. Visual inspection of the distribution of exam scores and calibrations for each exam suggest that they are not normally distributed. Figure 1 also shows scatter plots of the exam calibration and exam score. The closer the exam scores are to 100, on average, the closer the calibration scores are to 0.

To answer the question of whether there is a link between exam performance and calibration accuracy, we looked for correlations between actual and estimated scores for each exam. The non-parametric test, Spearman's rho, was used for these analyses, given concerns about violations of normality. As shown in Table 3, there was a significant association between exam score and calibration, with higher exam scores linked to smaller (i.e., more accurate) calibration values. This was true for each of the three exams, and for the semester-wide exam average and average calibration.

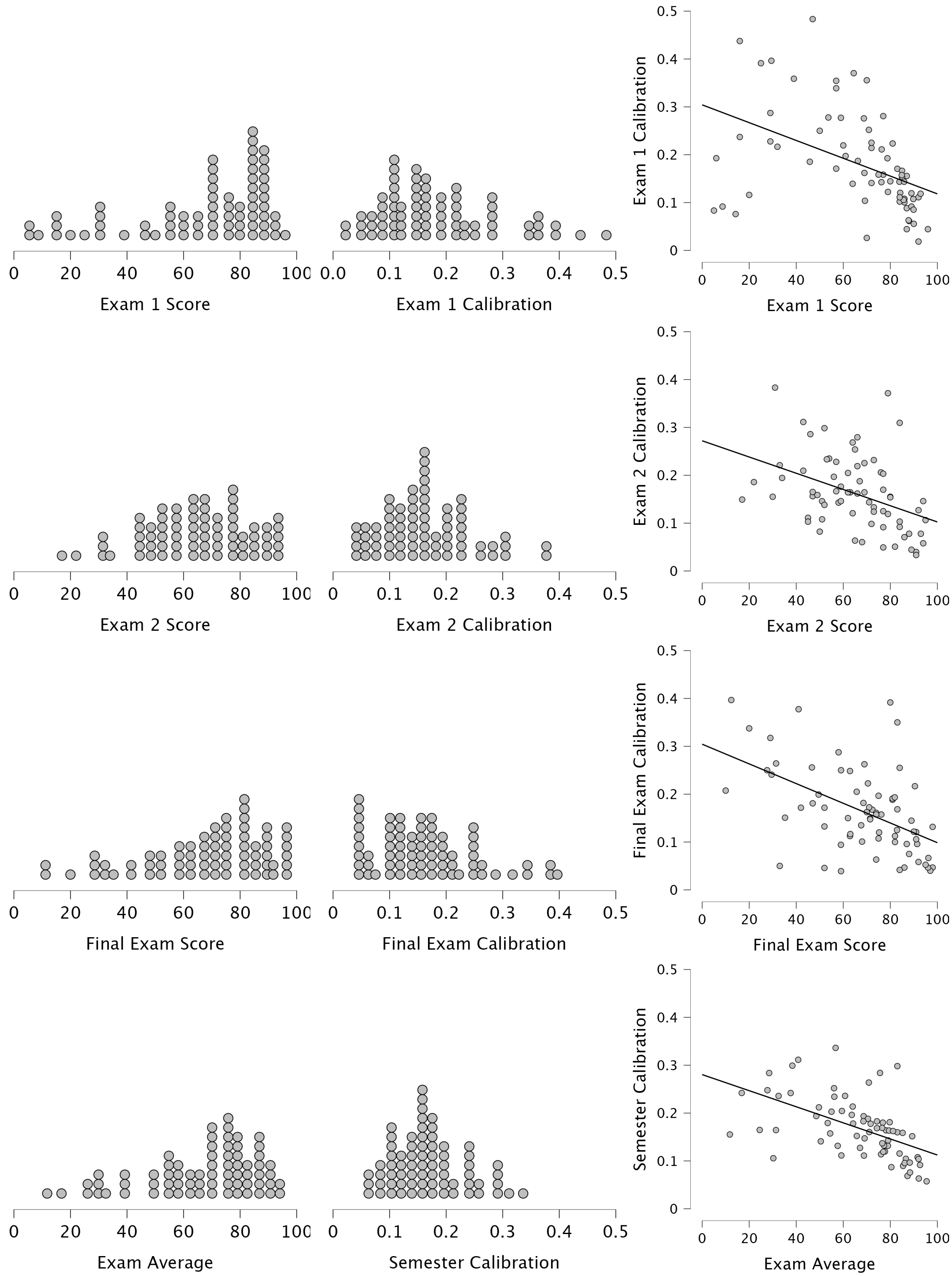

Figure 1. Dot plots showing the distributions of grades on each exam and overall semester scores are on the left. Dot plots for each student's calibration scores on each exam and for the overall semester are in the middle. Scatter plots with a regression line displaying the correlations between each student's estimated scores and actual scores are on the right.

Table 3. Correlations between exam scores and calibration scores.

	Spearman's rho	*p*	Lower 95% CI	Upper 95% CI
Exam 1 Scores with Exam 1 Calibrations	-0.594***	< 0.001	-0.728	-0.418
Exam 2 Scores with Exam 2 Calibrations	-0.454***	< 0.001	-0.622	-0.245
Final Exam Scores with Final Exam Calibrations	-0.503**	< 0.001	-0.660	-0.304
Exam Averages with Semester Calibrations	-0.625**	< 0.001	-0.750	-0.458

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The plots in Figure 1 demonstrate these patterns in graphical form. Significant correlations were found between students' exam scores and calibration scores, such that the greater a student's performance on an exam, the more accurately they can estimate their score.

Additionally, using a median split to classify students as high- or low-performing (based on exam performance) showed that there was a clear difference in calibration between the groups. With 39 students classified as high-performing, and 31 classified as low-performing, independent t-tests indicated in all exams, the high-performing students on average had lower (i.e., more accurate) calibration scores than the low-performing students. See Table 4 for details.

Table 4. Calibrations comparing low- and high-performing students.

		Mean Calibration	Standard Deviation	*t* (68)	*p*	95% CI for Cohen's *d* Lower	95% CI for Cohen's *d* Upper
Exam 1 Calibration	High-performing	0.14	0.08	-3.70	< 0.001	-1.381	-0.392
	Low-performing	0.23	0.11				
Exam 2 Calibration	High-performing	0.14	0.01	-2.50	0.015	-1.082	-0.117
	Low-performing	0.19	0.01				
Final Exam Calibration	High-performing	0.15	0.01	-2.09	0.041	-0.979	-0.021
	Low-performing	0.19	0.02				
Semester Calibration	High-performing	0.14	0.01	-4.07	< 0.001	-1.476	-0.477
	Low-performing	0.20	0.01				

To test the second research question, we ran a repeated measures ANOVA, comparing the average exam calibration across the three time points. Because of violations of sphericity, we used the Huynh-Feldt correction. This yielded a significant main effect $F(1.961, 135.318) = 1.247$, $p = 0.290$, $\eta^2 = 0.018$). Although we show a small trend of improved calibration after the first exam, the effect was so small that it was not significant. See Table 5 for details.

Table 5. Repeated measures ANOVA within-subjects effects.

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
Exam Calibrations	Huynh-Feldt	0.016	1.96	0.008	1.247	0.290	0.018
Residuals	Huynh-Feldt	0.874	135.32	0.006			

DISCUSSION

The finding that students who performed better on an exam were able to better postdict their performance is consistent with previous findings across multiple fields [4,6,7,9,10,12,14]. In particular, those students with stronger performance had much more accurate postdiction estimates. Consistent with previous findings [4,6,9,12], we did not find strong evidence of increased accuracy in postdiction performance over the course of the semester.

The preliminary data we have collected indicates that postdiction calibration capabilities of the engineering statics students studied here follow trends observed in other fields. However, a number of limitations could affect these observations. First, the sample size in this study is fairly small. Specifically, we may find that the non-significant trend of improvement from Exam 1 to Exam 2 may become significant with a larger set of participants. More importantly, in analyzing the data it became apparent that the lack of a consistent definition for calculating calibration, combined with few known studies that measure performance and confidence scores as we have, leads to some uncertainty in analysis that may impact the outcome. For example, the difference in students' postdictions and their actual exam performance can either be averaged across individual questions within an exam or summed for all questions in an exam. Further, these calculations can be performed either on the difference in the estimated and actual exam scores or on the absolute value of this difference. In future work we will explore how each of these choices affects the result of the analysis and discuss how making each choice should affect the interpretation of the result of the calculation.

Finally, the data presented here are part of an ongoing larger study that involves a metacognitive intervention aimed at improving students' question-asking abilities. The participants in this study included subjects from both the control group, who received no metacognitive intervention, and the intervention group. It is possible that the intervention impacts students' metacognitive monitoring abilities, which are reflected in their ability to accurately postdict their exam performance. As more data are collected from students in both the control and experimental conditions, the impact of the intervention on calibration ability can be more rigorously assessed.

REFERENCES

[1] J. H. Flavell, “Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry.,” *American Psychologist*, vol. 34, no. 10, pp. 906–911, Oct. 1979, doi: 10.1037/0003-066X.34.10.906.

[2] G. Schraw, “A conceptual analysis of five measures of metacognitive monitoring,” *Metacognition Learning*, vol. 4, no. 1, pp. 33–45, Apr. 2009, doi: 10.1007/s11409-008-9031-3. [Online]. Available: https://doi.org/10.1007/s11409-008-9031-3.

[3] L. M. Lin and K. M. Zabrucky, “Calibration of Comprehension: Research and Implications for Education and Instruction,” *Contemp Educ Psychol*, vol. 23, no. 4, pp. 345–391, Oct. 1998, doi: 10.1006/ceps.1998.0972.

[4] J. L. Nietfeld, L. Cao, and J. W. Osborne, “Metacognitive Monitoring Accuracy and Student Performance in the Postsecondary Classroom,” *The Journal of Experimental Education*, vol. 74, no. 1, pp. 7–28, 2005 [Online]. Available: https://www.jstor.org/stable/20157410.

[5] A. P. Gutierrez and A. F. Price, “Calibration Between Undergraduate Students’ Prediction of and Actual Performance: The Role of Gender and Performance Attributions,” *The Journal of Experimental Education*, vol. 85, no. 3, pp. 486–500, Jul. 2017, doi: 10.1080/00220973.2016.1180278.

[6] D. J. Hacker, L. Bol, D. D. Horgan, and E. A. Rakow, “Test prediction and performance in a classroom context.,” *Journal of Educational Psychology*, vol. 92, no. 1, pp. 160–170, Mar. 2000, doi: 10.1037/0022-0663.92.1.160.

[7] M. Händel and A.-K. Bukowski, “The gap between desired and expected performance as predictor for judgment confidence.,” *Journal of Applied Research in Memory and Cognition*, vol. 8, no. 3, pp. 347–354, Sep. 2019, doi: 10.1016/j.jarmac.2019.05.005.

[8] T. M. Miller and L. Geraci, “Training metacognition in the classroom: the influence of incentives and feedback on exam predictions,” *Metacognition Learning*, vol. 6, no. 3, pp. 303–314, Dec. 2011, doi: 10.1007/s11409-011-9083-7.

[9] L. Bol, D. J. Hacker, P. O’Shea, and D. Allen, “The Influence of Overt Practice, Achievement Level, and Explanatory Style on Calibration Accuracy and Performance,” *The Journal of Experimental Education*, vol. 73, no. 4, pp. 269–290, Jul. 2005, doi: 10.3200/JEXE.73.4.269-290.

[10] J. K. Knight, D. C. Weaver, M. E. Peffer, and Z. S. Hazlett, “Relationships between Prediction Accuracy, Metacognitive Reflection, and Performance in Introductory Genetics Students,” *LSE*, vol. 21, no. 3, p. ar45, Sep. 2022, doi: 10.1187/cbe.21-12-0341.

[11] N. V. Dang, J. C. Chiang, H. M. Brown, and K. K. McDonald, “Curricular Activities that Promote Metacognitive Skills Impact Lower-Performing Students in an Introductory Biology

Course," *J Microbiol Biol Educ.*, vol. 19, no. 1, p. 19.1.10, Mar. 2018, doi: 10.1128/jmbe.v19i1.1324.

[12] M. J. Hawker, L. Dysleski, and D. Rickey, "Investigating General Chemistry Students' Metacognitive Monitoring of Their Exam Performance by Measuring Postdiction Accuracies over Time," *J. Chem. Educ.*, vol. 93, no. 5, pp. 832–840, May 2016, doi: 10.1021/acs.jchemed.5b00705.

[13] M. Potgieter, M. Ackermann, and L. Fletcher, "Inaccuracy of self-evaluation as additional variable for prediction of students at risk of failing first-year chemistry," *Chem. Educ. Res. Pract.*, vol. 11, no. 1, pp. 17–24, 2010, doi: 10.1039/C001042C.

[14] H. Shih, W. Zheng, T. Pei, G. Skelton, and E. Leggette, "Integrating Self Regulated Learning Instruction In A Digital Logic Course," in *2010 Annual Conference & Exposition Proceedings*, Louisville, Kentucky, Jun. 2010, p. 15.769.1-15.769.13, doi: 10.18260/1-2—15735

[15] D. Christensen, T. Khan, I. Villanueva, and J. Husman, "Stretched too much? A case study of engineering exam-related predicted performance, electrodermal activity, and heart rate," in *The Proceedings of the 47th Annual Conference of the European Society for Engineering Education*, Budapest, Hungary, September 16-20, pp. 1481–1492.

[16] R. Isaacson and P. Goodmann, "You Bet Your Grade! Using Exams To Promote Student's Self Assessment," in *2005 Annual Conference Proceedings*, Portland, Oregon, Jun. 2005, p. 10.1482.1-10.1482.7, doi: 10.18260/1-2—15196.

[17] A. Baisley, K. Hjelmstad, and E. Chatziefstratiou, "The accuracy of self-assessment in engineering mechanics," Aug. 2022 [Online]. Available: https://peer.asee.org/the-accuracy-of-self-assessment-in-engineering-mechanics.