A Study of the Testing Effect in an Engineering Classroom

Dr. Monica H. Lamm, Iowa State University

Monica Lamm is an Associate Professor of Chemical and Biological Engineering at Iowa State University.

Miss Shuting Yan, Iowa State University
Clark R. Coffman
Dr. Carly L. Manz, Iowa State University

Department of Genetics, Development and Cell Biology, Iowa State University

Dr. Robert D Reason, Iowa State University

Dr. Reason is professor of education in the School of Education at Iowa State University.

A Study of the Testing Effect in an Engineering Classroom

Introduction

This research paper describes a study that examines a testing effect intervention deployed in an engineering classroom setting. The testing effect is based on the premise that learning is improved when students engage with newly acquired information by challenging themselves to answer questions about the content instead of using other means of interacting with the content, such as rereading a text. The testing effect has been established in laboratory research studies [1]. To translate this finding into educational practice, classroom research studies [2]-[6] aim to define the conditions for which the testing effect remains robust in authentic classroom settings. In the classroom domain, a testing effect intervention often consists of low- or no-stakes quizzing with feedback during the learning period, followed by a summative assessment at the end of the unit. Previous investigations have studied the impact of conditions, such as the question type (identical or related; definitional or application), the quiz participation incentives, and the quiz delivery patterns on the testing effect outcome for all participants in the study. Nguyen & McDaniel [7] review several classroom studies aimed at improving student learning through the use of quizzing.

Student motivation is an important factor to consider in educational settings. The achievement goal theory uses a 2 x 2 framework to characterize an individual's reasons for wanting to establish their competence [8]. One dimension represents an individual's tendency to pursue absolute competence (mastery) or normative competence (performance). The other dimension represents the meaning that an individual learner attaches to the attainment of competence. An individual can view competence either as a positive, successful outcome to attain (approach) or as a negative, failed outcome to avoid (avoidance). This framework led Elliot & McGregor [8] to establish four achievement goal constructs: Mastery approach, Mastery avoidance, Performance approach, and Performance avoidance. A brief summary of each construct is as follows. The Mastery approach construct describes a learner who derives satisfaction from the attainment of learning goals. The Mastery avoidance construct describes a learner whose aim is to reach learning goals because of a desire to avoid failing. The Performance approach construct is one in which a learner derives satisfaction from performing better than their peers. In contrast, the Performance avoidance construct is one in which a learner aims to attain learning goals so as not to perform the worst among their peers. Geller *et al.* [9] provide further elaboration about the implications of achievement goals as related to individual differences in learners.

In this study we examined the impact of a testing effect intervention used in an authentic engineering classroom setting. Building on previous literature that establishes positive testing effect outcomes in classrooms, we sought to investigate how the outcomes derived from the use of a testing effect intervention coincides with individual differences in learner motivation. The testing effect intervention used in this study was a system of weekly online quizzes in an engineering course for second-year students. The research questions (RQ) addressed in this study are:

RQ 1: Does participation in weekly online quizzes coincide with achievement as measured through exam scores?

Hypothesis: Participation in the online weekly quizzes will correlate positively with exam scores.

RQ 2: Does performance on weekly online quizzes correlate with pre-existing student achievement?
Hypothesis: Scores for the online weekly quizzes will correlate positively with a student's grade point average prior to taking the course.

RQ 3: Does performance on weekly online quizzes correlate with student motivation?
Hypothesis: Scores for the online weekly quizzes will correlate with student motivation goals, as defined by the achievement goals framework [8].

Methods

Participants. Participants were recruited from 77 students enrolled in a material and energy balances course at a large, public university in the Midwestern United States. The lead author of this article taught the course. Another investigator from the research team visited the class at the start of the semester and invited students to participate in the research study. Informed consent was obtained and only students who consented to allow their coursework performance and grade point average (GPA) to be used in this study were included in the analysis reported here. Twelve students opted not to participate and 3 students dropped the course, leaving 62 participants in the study. All students enrolled in the course were invited to complete a study goals survey near the end of the term. Nineteen out of the 62 students in the study opted not to submit a survey, leaving 43 students in the sample. All participants were enrolled in the chemical engineering major in the College of Engineering.

Course setting. The material and energy balances course is a required course for second-year chemical engineering majors. The course included three content modules: (1) material balances, (2) volumetric properties and multiphase systems, and (3) energy balances. Following the content modules, there were two modules devoted to case studies. The course met three times per week for 50 minutes over a 15-week semester. There were 10 class meetings dedicated to instructional activities for each of the three content modules. Other class meetings were dedicated to scheduled exams, review of information following exams, and case studies.

The course was structured using the team-based learning instruction method [10], [11]. Approximately three hours of out-of-class work per class meeting was expected. This included assigned reading, homework problems, optional online quizzes, and preparing for exams. During each class period, the learning activities included mini-lectures by the instructor to clarify concepts from the reading, and application exercises that were solved by student learning teams. The application exercises were designed to provide the students with practice applying the concepts they were learning to the solution of engineering problems. These exercises increased in complexity and difficulty as the content module progressed. At the end of each content module an individual examination was administered.

Materials and design. The intervention was introduced to all course participants and was a component of the routine course learning activities throughout the semester. We did not use a

control group (i.e., a group for which there was no intervention) because this study took place in a classroom and we presumed that all students would benefit to some degree from the intervention (i.e., we did not want to withhold the learning benefit from a subgroup of students). For this reason, this study is considered to be what Freeman *et al*. [12] define as a "second-generation research" study. That is, we are using findings from educational psychology to inform changes in course design and testing hypotheses about the impact of these changes on student learning in an engineering course. In this context, instead of using a control group, we controlled for individual differences in student motivation and student metacognition in our analysis. Our primary goal was to identify the characteristics of participants for whom low-stakes quizzing positively impacted learning outcomes. The participant characteristics of interest included quiz-taking patterns (score on first quiz attempt and number of quiz attempts before a unit exam), calibration error (measured as the participant's predicted weekly quiz score minus their actual weekly quiz score), pre-existing academic achievement (measured as GPA prior to enrolling in the course), and achievement goals (measured with the Achievement Goals Questionnaire [8]). The data collected was analyzed in aggregate and individually.

Within each of the three content modules in the course, there were three online study quizzes that were made available on the learning management system (Blackboard Learn). All questions were multiple-choice format. Each study quiz contained 3 or 6 conceptual questions, and 1 or 2 challenge questions. See Geller *et al*. [13] (Figure 1) for a representative question sequence that maps to the learning objectives for a lesson. The challenge questions were usually story problems that required students to apply more than one concept to arrive at an answer. At the conclusion of each content module, a summative unit exam was administered. The format for all questions asked on the unit exams was free response. The unit exams contained story problems that required multiple analysis steps. The duration of each unit exam was 50 minutes. The unit exams were graded and scored by the course instructor.

The Achievement Goals Questionnaire [8] was administered to students after the third summative unit exam. A diagram of the data collected during the semester is shown in Figure 1. Table 1 summarizes the variables used in this study.

Procedure. The study quiz was released to the students immediately after class on Friday and was due by Sunday at 10 pm. The quizzes were not scored for correctness. Students received one point per answer for the first attempt for every quiz question that they answered prior to the deadline. Once a student submitted the quiz, they received a performance report that showed each question, the response selected for each question, the correct response to the question, and feedback from the instructor to explain the reasoning behind the correct answer. Once a weekly online quiz was released, it remained available to students for the duration of the semester and students could return to the quizzes as often as they wished to review the content. There was no additional course incentive for completion of the study quizzes after the initial deadline.

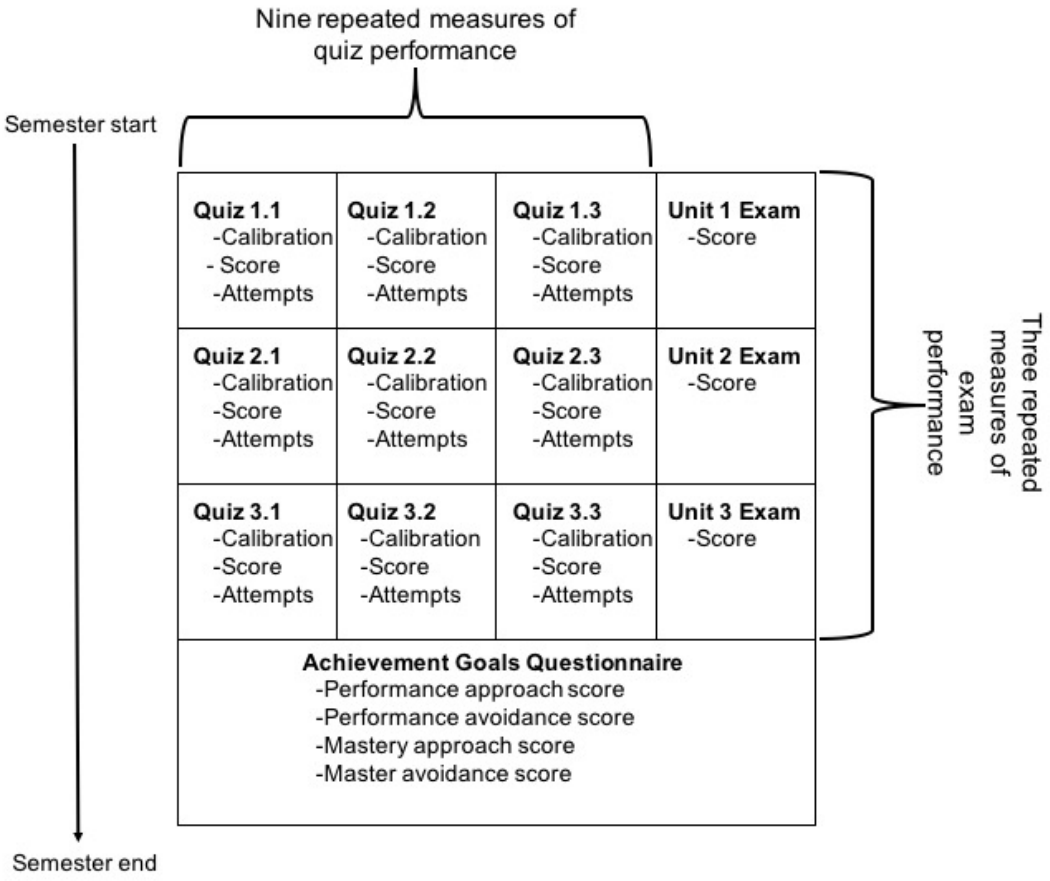

Figure 1. Diagram showing the data collected during the study**.** Each assessment or instrument is shown in boldface. The data collected appear as a bulleted list.

Analysis using a linear mixed effects model. We used the *lmer* program from the *lme4* package [14] for estimating fixed and random coefficients. This package is supplied in the *R* system for statistical computing [15] under the GNU General Public License (Version 2, June 1991). Here we describe the process for model specification and evaluation to develop a model for the exam score. A similar process was used to develop a model for the quiz scores.

Exam scores were recorded for three units in the course. A unit exam score earned by a particular subject was only included in the model if the subject completed the three accompanying quizzes in that unit. This resulted in a data set with 125 exam scores, collected from 43 subjects.

We specified four models to determine the fixed or random effect of the unit in two groups, where each group has two models of increasing complexity with respect to the parameterization of the subject-related variance/covariance matrix. The first model (M1) is a random-intercept model, allowing only for between-subject variance in the exam score. The *lmer* specification for this model is given by

Table 1. Summary of variables used in the study.

Variable	Scale	Definitions	Reason for inclusion
Unit exam score	[0,1]	Fraction of points earned on assessment.	Summative measure of student performance upon completion of a unit
Quiz score average	[0,1]	Average quiz score for the unit; fraction of questions answered correctly	Formative measure of student performance within a unit
Quiz attempts average	[0,6]	Average number of attempts per quiz in a unit	Measure of participation frequency
Quiz calibration error	[0,1]	|Predicted quiz score – actual quiz score|; magnitude	Measure of student metacognition
Performance approach	[0,21]	Derives satisfaction from performing better than peers	Measure of student motivation
Performance avoidance		Seeks to avoid performing the worst among peers	
Mastery approach		Derives personal satisfaction from the attainment of learning goals	
Mastery avoidance		Maintains personal expectation that mistakes or failure are to be avoided	
GPA	[0,4]	Student grade point average prior to enrolling in the course	Measure of pre-existing academic achievement

M1<- lmer(Unit exam score ~ GPA + Quiz score average + Quiz attempts average + Performance approach + Performance avoidance + Mastery approach + Master avoidance + Unit + (1|Subject))

where the (1|Subject) term denotes the variance for the intercept over subjects. In this model, Unit is set to be a fixed effect.

The second model (M2) is a random-intercept-and-slopes model given by

M2<- lmer(Unit exam score ~ GPA + Quiz score average + Quiz attempts average + Performance approach + Performance avoidance + Mastery approach + Master avoidance + Unit + (1|Subject) +(0+Unit|Subject))

Here, the terms for the Subject's random effect for intercept and slope are listed separately to specify them as independent of each other (i.e., with zero covariance). The presence of the "0" in the random slope term suppresses the default estimation of the intercept and its covariance with the Unit. We estimate the significance of each variance component by checking the decrease in goodness of fit due to its exclusion from model M1.

The other two models (M3 and M4) treat the Unit as a random effect:

M3<- lmer (Unit exam score ~ GPA + Quiz score average + Quiz attempts average + Performance approach + Performance avoidance + Mastery approach + Master avoidance + (1|Unit) + (1|Subject))

M4<- lmer (Unit exam score ~ GPA + Quiz score average + Quiz attempts average + Performance approach + Performance avoidance + Mastery approach + Master avoidance + (1|Unit) + (1|Subject) +(0+Unit|Subject))

All models were fit by maximum likelihood (ML). For assessment of relative differences in goodness of fit, the *lmer* program provides the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the log-likelihood (logLik), and, in the case of model comparisons, the χ^2-distributed likelihood ratio and its associated *p*-value. The AIC (= −2 logLik + 2 n_{param}) and BIC (= −2 logLik + n_{param}log N_{obs}) values correct the log-likelihood statistic for the number of estimated parameters and, in the case of BIC, also for the number of observations. That is, we use them as a guide against overfitting during the process of model selection.

The best model selected from the above four models (M3) was refit using the less biased restricted maximum likelihood (REML). The Satterthwaite method [16] was used to estimate denominator degrees of freedom for *t* statistics and obtain more conservative *p*-values for fixed effect variables. The packages *pbkrtest* [17] and *lmerTest* [18] were used to calculate *p*-values with the Satterthwaite method. The *p*-values of random effect variables were generated from likelihood ratio test using the *anova* function in *lme4* and using the default approach of refitting the model with maximum likelihood. The package *sjstats* [19] was used to obtain intraclass correlation coefficient (ICC) values for random effects.

Results and discussion

Relationship between exam performance and weekly quiz participation (RQ1). Based on previous literature, we hypothesized that students' performance on unit exams would be predicted by prior academic achievement (GPA), the number of quiz attempts in the unit, the quiz score on the first attempt for each quiz offered in the unit, and the four achievement goal constructs measured by the Achievement Goals Questionnaire [8]. To test this hypothesis, we performed a linear mixed effect analysis and developed the following model

Unit exam score = Intercept + GPA + Quiz score average + Quiz attempts average + Performance approach + Performance avoidance + Mastery approach + Mastery avoidance + (1|Subject) + (1|Unit)

Table 2. Summary of linear mixed effect model for predicting unit exam performance.

Fixed Effects

Parameter	Estimate	*SE*	*p*-value
(Intercept)	0.354	0.135	0.0142
GPA	0.0879	0.0204	1.15×10^{-4} ***
Quiz score (unit average)	0.120	0.0807	0.141
Quiz attempts (unit average)	0.00675	0.0185	0.716
Performance approach	-6.33×10^{-4}	0.00245	0.797
Performance avoidance	-0.00391	0.00317	0.225
Mastery approach	-0.00139	0.00452	0.760
Mastery avoidance	0.00411	0.00227	0.0791

Random Effects

Parameter	Variance	*SD*	ICC	*p*-value
Subject (intercept)	0.00318	0.0564	0.125	0.0771
Unit (intercept)	0.01111	0.1054	0.437	9.07×10^{-12}***
Residual	0.01112	0.1055		

Model fit statistics

REML criterion at convergence	-120.3

Note: The data set contains 125 exam scores and 43 subjects. There were four instances where a subject did not complete the quizzes for a given unit; in those cases the subject's exam score was not included in the data set. SE (standard error), SD (standard deviation), ICC (intraclass correlation coefficient), REML (restricted maximum likelihood)
*** $p<0.001$

where the fixed effects are GPA, Quiz score average, Quiz attempts average, and the four achievement goal constructs (Performance approach, Performance avoidance, Mastery approach, Mastery avoidance). The random effects in the model are the Subject and the Unit. The results of the analysis with this model are shown in Table 2. We checked for multicollinearity among fixed effect random variables by calculating the variance inflation factor (VIF) and found that a very low level of multicollinearity was present (VIF<1.5). The analysis shows that GPA is the only significant factor that predicts unit exam scores. The Unit itself is a significant random effect and its variance is nearly equivalent to the variance of the residual error in the model.

The fact that GPA is a significant predictor of unit exam scores is not surprising, given that GPA is a global indicator of a student's ability to perform on course assessments where the stakes are high. It is interesting that neither quizzing performance nor quizzing participation had a significant effect on unit exam scores. One explanation may be that the unit exams questions were in the free response format, which is a different format than the weekly quiz questions that

were delivered as multiple-choice questions. In future studies, it would be of interest to map specific quiz questions to the specific problem-solving decisions or steps that are used to solve a free response question on a unit exam. Using a higher resolution mapping between unit exam questions and weekly quiz questions may provide more insight about student learning than the lower resolution mapping at the level of a course unit that was used in this study.

Relationship between weekly quiz performance, pre-existing student achievement, and student motivation (RQ2 and RQ3). Based on previous literature, we hypothesized that students' performance on the weekly quizzes would be predicted by prior academic achievement (GPA), the number of quiz attempts in the unit, the magnitude of student quiz score calibration error (|predicted quiz score – actual quiz score|) on the first attempt for each quiz, and the four achievement goal constructs measured by the Achievement Goals Questionnaire [8]. To test this hypothesis, we performed a linear mixed effects analysis and developed the following model

Quiz score = Intercept + GPA + Quiz attempts + Quiz calibration error + Performance approach + Performance avoidance + Mastery approach + Mastery avoidance + (1|Subject) + (1|Quiz)

where the fixed effects are GPA, Quiz attempts, Quiz calibration error and the four achievement goal constructs (Performance approach, Performance avoidance, Mastery approach, Mastery avoidance). The random effects in the model are the Subject and the Quiz. The results of the analysis are shown in Table 3. We checked for multicollinearity among fixed effect random variables by calculating the variance inflation factor (VIF) and found that a very low level of multicollinearity was present (VIF<1.5). The analysis shows that the magnitude of the Quiz calibration error and the Performance approach goal have a significant effect on quiz scores. The Subject and Quiz are significant random effects, yet they do not explain more variance than the residual error in the model.

Student GPA is not a significant factor in how well a student performs on the weekly quizzes. This lack of correlation may be due to the fact that the quizzes were low stakes and were seen by students as a means to test what they knew without the need for studying or cramming. In future studies, it would be interesting to collect data on student perceptions of the weekly quizzes and whether they viewed the quizzes as an opportunity to test what they knew or an opportunity to practice and reinforce course concepts.

The sign on the coefficient for the magnitude of the Quiz calibration error term is negative and indicates that when a student's prediction of their quiz score is more accurate (a lower magnitude error) the student score on the actual quiz is higher. This finding is consistent with research studies in other subject areas that have examined the relationship between a student's awareness of his or her own learning and the student's academic performance [20]. We also examined the direction of calibration error (predicted quiz score – actual quiz score) and found a similar effect in the model when the magnitude of error was replaced with directional error (model not shown). There were very few instances of subjects underestimating their quiz score for a given weekly quiz.

The coefficient on the Performance approach term in the model is positive and indicates that students who are motivated by doing better than their peers have higher weekly quiz scores. As educators, our wish is for student academic success to align with the Mastery approach goal. Falling short of that, however, we note that the Performance approach goal is positioned within the positive valence of achievement goal theory. This means that the more students view the attainment of course knowledge as a desirable outcome the better they perform on the quizzes.

Conclusions

In this study the factor most associated with students' unit exam performance was found to be GPA prior to enrolling in the course. Participation in weekly quizzing had no significant effect on unit exam performance. We propose that examining the relationship between exams and weekly quizzes at a higher resolution may provide more insights about student learning. For example, a future study could align particular components of the unit exam to specific weekly questions, as an improvement over alignment at the course unit level.

Concerning performance on the weekly quizzes, we found that the magnitude of student calibration error is inversely proportional to student scores on weekly quizzes. This is consistent with the literature that has shown that students who are aware of their own learning perform better academically. We also found that the weekly quiz scores have a positive correlation with students' Performance approach goals.

As a final remark we note that the pedagogical approach of team-based based learning used in this study is similar in many respects to the interteaching method used in the study reported by Saville *et al.* [21]. Interteaching is a strategy whereby students complete preparatory guides before class, discuss their responses with a partner during class, and then listen to a short clarifying lecture from the instructor [22]. Saville *et al.* [21] found that adding postdiscussion quizzes to a course taught using the interteaching method did not improve students' exam scores. The absence of testing-enhanced learning in this context was attributed to the instructors' use of an already effective teaching strategy that promotes learning [7]. In the context of the study reported here, the primary factor associated with exam performance was student GPA prior to enrolling in the course. This suggests that instructors may wish to use other measures, such as student engagement with content or student perception of the testing intervention, as a means to evaluate the impact of the testing effect in classrooms where evidence-based teaching practices, such as small group learning, are the norm.

Table 3. Summary of linear mixed effect model for predicting quiz performance.

Fixed Effects				
Parameter	Estimate	*SE*		*p*-value
(Intercept)	0.835	0.107		1.30×10^{-9}***
GPA	-0.0132	0.0197		0.509
Quiz attempts	-0.00976	0.0104		0.347
Quiz calibration error	-0.778	0.0318		$<2\times10^{-16}$ ***
Performance approach	0.00620	0.00233		0.0116*
Performance avoidance	-0.00605	0.00303		0.0537
Mastery approach	0.00375	0.00434		0.394
Mastery avoidance	4.68×10^{-4}	0.00222		0.834
Random Effects				
Parameter	Variance	*SD*	ICC	*p*-value
Subject (intercept)	0.00530	0.0728	0.313	1.28×10^{-13}***
Quiz (intercept)	0.00122	0.0350	0.0721	2.80×10^{-5}***
Residual	0.0104	0.102		
Model fit statistics				
REML criterion at convergence	-460			

Note: Data set contains 348 quiz scores and 43 subjects. Some subjects did not complete all nine quizzes. SE (standard error), SD (standard deviation), ICC (intraclass correlation coefficient), REML (restricted maximum likelihood)
* $p<0.05$, *** $p<0.001$

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant DUE-1504480. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors acknowledge Shana Carpenter for development of the study design, Shuhebur Rahman for data collection, and Patrick Armstrong for helpful discussions. Monica Lamm and Shuting Yan are grateful to Yalin Rao for guidance concerning the data analysis.

References

[1] Roediger, H. L., III & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210.

[2] Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: Effects of incentives on student participation and performance. *Advances in*

Physiology Education, 31, 253–260.

[3] McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a Web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, 1, 18–26.

[4] McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exam. *Applied Cognitive Psychology*, 27, 360-372.

[5] McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied,* 21, 370-382.

[6] Carpenter, S. K., Rahman, S., Lund, T. J. S., Armstrong, P. I., Lamm, M. H., Reason, R. D., & Coffman, C. R. (2017). Students' use of optional online reviews and its relationship to summative assessment outcomes in introductory biology. *CBE-Life Sciences Education*, 16, ar23:1-9.

[7] Nguyen, K., & McDaniel, M. A. (2015). Using Quizzing to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology*, 42, 87-92.

[8] Elliot, A. J., & McGregor, H. A. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology*, 80, 501-519.

[9] Geller, J., Toftness, A. R., Armstrong, P. I., Carpenter, S. K., Manz, C. L., Coffman, C. R., & Lamm, M. H. (2017). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory*, DOI:10.1080/09658211.2017.1397175

[10] Michaelsen, L. K., Knight, A. B., & Fink, L. D. (Eds.) (2004). *Team-based learning: A transformative use of small groups in college teaching.* Sterling, VA: Stylus Publishing.

[11] Sibley, J., Ostafichuk, P., Roberson, B., Franchini, B., & Kubitz, K. A. (2014). *Getting started with team-based learning.* Sterling, VA: Stylus Publishing.

[12] Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111, 8410-8415.

[13] Geller, J., Carpenter, S. K., Lamm, M. H., Rahman, S., Armstrong, P. I., & Coffman, C. R. (2017). Prequestions do not enhance the benefits of retrieval in a STEM classroom. *Cognitive Research: Principles and Implications*, 2:42, 1-13.

[14] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015b). lme4: Linear mixed-effects models using Eigen and S4. R packageversion1.1-8. Retrieved from http://CRAN.R-project.org/package=lme4

[15] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org/

[16] Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6, 309–316.

[17] Halekoh, U., & Højsgaard, S. (2014). pbkrtest: Parametric bootstrap and Kenward Roger based methods for mixed model comparison. R package version 0.4-2.

[18] Kuznetsova, A., Brockhoff, P., & Christensen, R. (2014). LmerTest: Tests for random and fixed effects for linear mixed effect models. R package, version 2.0-3.

[19] Lüdecke, D. (2018). *sjstats: Statistical Functions for Regression Models*. R package version 0.14.1, https://CRAN.R-project.org/package=sjstats.

[20] Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to realize their own incompetence. *Current Directions in Psychological Science*, 12, 83-87.

[21] Saville, B. K., Pope, D., Lovaas, P., & Williams, J. (2012). Interteaching and the testing effect: A systematic replication. *Teaching of Psychology*, 39, 280-283.

[22] Boyce, T. E. & Hineline, P. N. (2002). Interteaching: A strategy for enhancing the user-friendliness of behavioral arrangements in the college classroom. *The Behavior Analyst*, 25, 215-226.