Final version published as: Bazelais, P., Lemay, D. J., & Doleck, T. (2019). The Effects of Testing on Pre-University Science Students' Academic Outcomes in an Electricity and Magnetism Course. International Journal of Engineering Education, 35(6), 1667–1672.

The Effects of Testing on Academic Outcomes of College Students in an Electricity and Magnetism Course*

PAUL BAZELAIS[1], DAVID JOHN LEMAY[1] and TENZIN DOLECK[2]
[1] McGill University, 3700 McTavish Street, Montreal, Quebec, H3A 1Y2, Canada. E-mail: paul.bazelais@mail.mcgill.ca; david.lemay@mail.mcgill.ca
[2] University of Southern California, 3470 Trousdale Pkwy, Los Angeles, CA 90089, USA. E-mail: doleck@usc.edu

Testing can influence student learning outcomes by influencing their approach to study and to learning. It is important to understand the influence of testing on students' learning outcomes to optimize instruction. We examine the role that testing played in a science course, to examine the effect of testing on retention and performance on a standardized final exam. This study compared two sections—experimental condition with testing (N = 35) and comparison condition with homework (N = 39)—of an Electricity and Magnetism course in a pre-university program to explore the role of the testing effect, that is, whether taking a test aids subsequent learning and retention. Results indicated that the students in the experimental group had a higher final exam average and greatest achievement gains. Our findings corroborate previous research and suggest that the traditional homework-based instructional strategy is a less effective approach for science learning or later retention compared to an instructional approach incorporating regular testing. Implications of these findings and the importance of testing in science instruction are also discussed.

Keywords: testing effect; academic performance; learning outcomes; STEM education; pre-university students; electricity and magnetism course

1. Introduction

Being assessed for our performance has become the norm in many facets of our lives. Indeed, assessment is an integral part of teaching and learning today. Testing is generally used as an assessment tool to measure student knowledge and skill and to award certification. Thus, the implications of testing assume great importance. Whereas tests provide a convenient means to measure what students have learned related to a specific topic or over a course, research suggests that testing can be beneficial in bolstering student retention of information [1–3]. Frequent testing has also been associated with better performance, enhanced retrieval of information, and in promoting retention [4–6]. Testing can even serve as a source of further learning as well [7, 8]. Taking a test, students are also engaged in learning-related cognitive processes [9]. While research suggests that testing might be beneficial in retention of material compared to other learning strategies (such as restudying), students may be unaware of the benefits of testing; for example, in a study of students' study behaviors, researchers [10] found that relatively few students engage in self-testing, suggesting that "students lack metacognitive awareness of the testing effect when they monitor their own learning" (p. 477). Further, researchers [5] noted that testing is generally infrequent and commonly perceived as a bother in higher education. Likewise, it is also argued that there is insufficient opportunity to test students given some teachers' approaches to teaching which allocate much of the classroom time to the transmission of information. Thus, understanding the influence of testing on learning and performance is relevant to both instructors and learners.

The term *testing effect*, refers to a cognitive phenomenon where "students who take a test on material between the time they first study and the time they take a final test remember more of the material than students who do not take an intervening test" [11, p. 392]. Whereas the testing effect can be traced to early works by scholars such as Abbott [12] and Spitzer [13], Glover's [11] influential article spurred renewed interest in the phenomenon [5, 6, 14–17]. In recent years, scholars have increasingly underscored the importance of testing as key to the improvement of teaching and learning.

Much of this recent work has relied on lab-based and experimental studies, using various stimulus materials, learning tasks, test formats, and questionnaires to substantiate the testing effect [9, 10, 15, 18–24]. Relatively little research has focused on investigating the testing effect in natural educational contexts where testing is an integral part of the instructional design, although some studies have evaluated the effect of repeated testing on course performance [14, 25]. More research is needed to demonstrate the efficacy of the testing effect in more relevant natural educational contexts. To better understand the testing effect outside of the lab setting, we compared two semester-long courses in Electricity and Magnetism to assess whether testing

is indeed a beneficial instructional strategy. The aim of the present study was to extend the work on the testing effect by investigating how testing affects student learning outcomes.

2. Methodology

2.1 Research context

This comparative case study contrasts two sections of the Electricity and Magnetism course. A traditional course with online homework and formative feedback served as the control group. In contrast, a course that implemented robust testing with no assigned homework served as the treatment group. To reduce teacher bias, each of the two sections were taught by a different instructor with identical content, including three required unit tests (e.g., test 1 on week 5, test 2 on week 10, and test 3 on week 15) and a standardized final exam. Both the three required unit tests and the standardized final exam were identical to both sections.

Rather than assigning weekly homework (as was the case with the control group), the students in the treatment group were quizzed 12 times during the 15-week semester with no assigned homework outside of the classroom. The weekly quizzes were designed to replace the out of the classroom assignments or homework with an added focus on time-on-task and peer formative feedback. In addition, the quizzes were used to reinforce key concepts; for example, the quizzes combined both conceptual and word problems (as was the case for the online homework) that exhibited a similar level of difficulty as the three required unit tests and the final exam. These weekly quizzes integrated the concept of ''interleaving'' a process that mixes both new and old concepts together [26]. Research shows that the effect of interleaving and testing with formative feedback produces better results and enhances long-term retention [9, 26–28]. This is similar to the notion of cognitive elaboration discussed by Bradshaw and Anderson [29] wherein repeated exposure helps to deepen understanding. The inclusion of opportunities for peer formative feedback is supported by the literature on instructive academic conversations supporting the kind of cognitive elaboration that encourages deeper learning [30, 31] as well as the literature on instruction based on peer interactions [32], and peer tutoring [33]. Students in the treatment group worked on the quizzes for fifteen minutes individually, and then an additional ten minutes were allocated to group discussion where they could work together on the quizzes with any of their peers. This was done to allow students to co-construct and share their understanding and to give each other formative feedback. The quizzes were used to assess students' understanding as well as to promote peer learning through formative feedback. The quizzes were used not only to measure what students learned and understood from the lecture content but also to assess whether they were able to retain that information throughout the semester. Table 1 illustrates the two conditions in the present study. For each condition, the outcome measures (such as quizzes, unit tests, homework, and standardized final exam (FX)) are also listed. Both instructors followed the same structure and used the same content. The only difference between the two groups was the use of homework and quizzes.

2.2 Study participants and procedure

The target population for this study is second-semester college physics students at an English *Collège d'enseignement général et professionnel* (CEGEP) in Montreal, Quebec (for a primer on CEGEPs, see [34]). The sample (n = 74, 62% males, 38% females) was drawn primarily from two sections of the Electricity and Magnetism physics course. The treatment group consisted of $n = 35$ students (66% males, 34% females), whereas, the control group consisted of n = 39 students (59% males, 41% females).

Participants were informed of the confidential nature of the study and the data. They were assured that study results would not be linked to any student's name or student ID number. The data was not analyzed until the final grades were submitted. The participants of this research gave their consent to the researcher to assess and measure teaching and learning effectiveness using their aggregate quizzes, homework, unit tests and final exam marks in the course.

3. Analysis and results

The data were analyzed using within-sample paired t-tests to determine whether significant shift

Table 1. Summary of Methodology

Sections	Condition	Outcome Measures
Treatment group	Lecture format with testing effect	Quizzes, unit tests, FX
Control group	Lecture format with online homework	Homework, unit tests FX

Note. Each section was taught by a different instructor.

Table 2. Overall shift between unit tests and final exam scores for each group

Sections	Tests avg % (*SD*)	FX avg. % (*SD*)	Shift % (*SD*)	*t*-test		
				t	*p*	*ES*
Treatment group	74.52 (15.58)	75.10 (14.18)	0.576 (10.04)	0.339	0.736	–
Control group	71.34 (10.68)	65.35 (16.57)	–5.99 (9.91)	–3.77	0.001	0.43

occurred within and between the two groups (treatment and control) on the unit tests average (M = 74.52%, SD = 15.58) and the final exam average (M = 75.10%, SD = 14.18). A similar procedure was employed in the study [35]. The trivial shift between the unit tests average and the final exam average was not statistically significant for the treatment group (see Table 2), demonstrating that students in the treatment group performed slightly better in the standardized final exam as compared to their unit tests average. In contrast, the negative shift for the control group was statistically significant, [$t(69)$ = –3.77, p = 0.001, ES = 0.43], demonstrating that students in the control group were less successful on the standardized final exam by 6 points (see Table 2). The effect size (ES), based on the standard Cohen's d [36], is significant if $d > 0.2$, indicating that the observed changes have, both statistical and practical significance. The within-sample paired t-test difference between the treatment and the control group shows that robust testing can lead to better retention and performance in a standardized final exam in a college science course.

As indicated in Table 2, the overall unit tests average was slightly higher for the treatment group (M = 74.52%, SD = 15.58), as compared to the control group (M = 71.34%, SD = 10.68). An independent t-test analysis showed that this difference was not statistically significant, [$t(72)$ = –1.01, p = 0.315]. However, this trivial achievement gain demonstrates that both in-class quizzes or out of class assignment can be as successful as a measure to assess student learning and progress in the course. Furthermore, an independent t-test showed that both the overall final exam average [$t(72)$ = –2.70, p = 0.009] and the shift [$t(72)$ = –2.83, p = 0.006] between the two groups were statistically significant (see Table 3), thereby suggesting that robust testing and peer formative feedback enhance later retention and performance in science.

3.1 Gender differences

The data were first analyzed with an independent t-test to evaluate the overall effects (both groups combined) of achievement gains and overall academic performance by gender. Both males and females do not exhibit any change (see Table 4) nor did the independent t-tests reveal any overall difference across all examined variables, suggesting that both males and females students were essentially the same before and after the standardized final exam.

The data were further analyzed with an independent t-test to determine whether gender differences exist within each group on achievement gain, pre and post final exam. The test did not reveal any significant difference between genders across all variables for the control group (see Table 5). Although the final exam average for the control group were higher for males compared to females, these differences were not statistically significant (see Table 5).

The independent t-test also revealed no significant gender differences in the treatment group

Table 3. Overall difference between the treatment and the control group

Sections	Unit tests avg. % (*SD*)	FX avg. % (*SD*)	Shift % (*SD*)
Treatment group	74.52 (15.58)	75.10 (14.18)	0.576 (10.04)
Control group	71.34 (10.68)	65.35 (16.57)	–5.99 (9.91)
t-test	–1.01	–2.70*	–2.83*

Note. * $p < 0.05$.

Table 4. Overall unit tests and final exam average, and shift by gender for both groups combined

Gender	Unit test avg. % (*SD*)	FX avg. % (*SD*)	Shift % (*SD*)
Males (n = 46)	72.05 (13.68)	69.79 (14.34)	–2.25 (8.89)
Females (n = 28)	74.16 (12.60)	70.23 (19.01)	–3.92 (12.70)
t-test results	p = 0.510	p = 0.911	p = 0.508

Note. An independent t-test was not significant across all variables.

Table 5. Average unit tests, final exam, and shift for the control group

Gender	Unit test avg. % (*SD*)	FX avg. % (*SD*)	Shift % (*SD*)
Males (n = 23)	70.52 (10.02)	66.91 (12.14)	–3.60 (7.20)
Females (n = 16)	72.52 (11.81)	63.09 (21.67)	–9.43 (12.31)
t-test results	p = 0.571	p = 0.530	p = 0.071

Note. No significant differences exist between genders across all variables $p > 0.05$.

Table 6. Average unit tests, final exam, and shift by gender for the treatment group

Gender	Unit tests avg. % (*SD*)	FX avg. % (*SD*)	Shift (%) (*SD*)
Males (n = 23)	73.58 (16.65)	72.67 (15.99)	–0.903 (10.29)
Females (n = 12)	76.34 (13.80)	79.75 (8.59)	3.41 (9.31)
t-test results	p = 0.626	p = 0.098	p = 0.233

Note. An independent t-test analysis shows no significant differences between the genders, $p > 0.05$.

across all the examined variables (see Table 6). Nontrivial differences existed between the genders overall. Female students had higher final exam average (M = 79.75%, SD = 8.59) than male students (M = 72.67%, SD = 15.99) in the treatment group. In addition, they also had a higher achievement gain (M = 3.41 %, SD = 9.31) than males (M = –0.903%, SD = 10.29). Females scored 3.41 points higher on the final than they did on unit tests while males scored 0.903 points lower. This significant difference suggests that on average female students might have benefited more from the testing effect.

4. Discussion

The findings reported above suggest that traditional homework-based instructional strategy is a less effective strategy for science learning and later retention. Higher final exam average and achievement gains for the treatment group affirm that the implementation of a testing design coupled with peer learning can be successful and have a positive impact on student learning outcomes in STEM, especially for females. Poor retention rates in STEM programs remains an imperative challenge for pre-university college programs [34, 37] as poor student retention limits the conduit of graduates into STEM careers. Struggling to complete introductory STEM courses can engender disappointment and lack of motivation in students, and eventually lead to plummeting retention and graduation rates in STEM programs [34, 38]. Enhancing performance in introductory science courses could potentially increase confidence and self-efficacy and ultimately lead to higher retention rates in STEM education.

These findings are interpretable within the approaches to learning literature [39–43]. In this perspective, it is theorized that students' approaches to studying and learning are influenced by instructional decisions and the learning environment. Two broad orientations have been documented, the surface and deep approaches. The surface approach is associated to performance, whereas the deep approach is oriented toward understanding. Further, deep learning is associated with better attitudes towards the learning environment, greater motivation, and generally better study approaches, but, most importantly, with deeper knowledge and better outcomes in the long-run. Whereas, surface learning is associated with a shallower knowledge base, less favorable attitudes towards the learning environment, less curiosity and motivation for self-directed learning and lower self-efficacy [44].

Teacher and content-centered approaches [41], such as exhibited in the traditional homework-based instructional strategy employed in the control group have been shown to promote surface approaches. However, student-centered approaches, such as those implementing peer learning [33], have been shown to increase long-term retention of knowledge and skills over traditional instructional approaches [44]. Most importantly, student-centered instructional approaches are associated with increased student success in terms of pass-rates. Thus, it is important to examine the facilitating and the discouraging factors that increase deep learning approaches. A recent study [45] examined the effect of perceived workload and task complexity on education undergraduates, finding no effect for perceived workload on the students' approaches to learning, but a significant effect for perceived task complexity, finding lack of information as a consistently discouraging factor for deep learning and an encouraging factor for surface approaches. These findings provide context for our results and suggest that if the testing effect is real, it may be attributable to other factors related to

the implementation of the testing strategy, namely to peer learning, as peer tutoring has shown one of the most robust effects for achievement gains [33]. Overall, the phenomenographic approaches to learning research suggests that the instructional situation is a multi-factorial environment where many enabling and hindering factors interact to create the conditions that influence students' approaches to learning and to studying. To foster deep learning, and increase student retention, teachers and faculty must be sensitive to student perceptions of the learning environment and its effect on their approaches to learning and study.

Our study suggests that the traditional homework-based instructional strategy is a less effective strategy for science learning and later retention and that a testing design with peer learning can deliver improved learning outcomes. Such results are important because they can help inform pedagogical development initiatives, optimize learning environments to support students, and help to improve student learning outcomes in STEM.

4.1 Limitations

This quasi-experimental study is limited by its design and its non-randomized convenience sample. Uncontrolled factors could have influenced the results, for instance, teacher individual differences, group composition and peer learning could be confounding factors, among others. Although our findings might not be generalizable, the convergence of findings from other studies suggests that the testing effect is indeed real, and can explain difference achievement differences in pre-university STEM education.

4.2 Conclusions and future directions

High attrition rates in pre-university STEM college programs motivate the search for instructional strategies that can increase student retention and achievement, and eventual success in STEM careers. Our study suggests that a testing design coupled with peer learning can improve student learning outcomes. However, more research is needed, particularly large-scale randomized controlled trials that can help eliminate confounding factors and establish to what extent the testing effect is real and not attributable to other factors in the instructional situation.

References

1. J. Glover, The "testing" phenomenon: Not gone but nearly forgotten, *Journal of Educational Psychology*, **81**(3), pp. 392–399, 1989.
2. T. Kuo and E. Hirshman, Investigations of the Testing Effect, *The American Journal of Psychology*, **109**(3), pp. 451–464, 1996.
3. M. Wheeler and H. Roediger, Disparate Effects of Repeated Testing: Reconciling Ballard's (1913) and Bartlett's (1932) Results, *Psychological Science*, **3**(4), pp. 240–246, 1992.
4. R. Bangert-Drowns, J. Kulik and C. Kulik, Effects of Frequent Classroom Testing, *The Journal of Educational Research*, **85**(2), pp. 89–99, 1991.
5. H. Roediger and J. Karpicke, Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention, *Psychological Science*, **17**(3), pp. 249–255. 2006a.
6. H. Roediger and J. Karpicke, The Power of Testing Memory: Basic Research and Implications for Educational Practice, *Perspectives on Psychological Science*, **1**(3), pp. 181–210, 2006b.
7. M. McDaniel and R. Fisher, Tests and test feedback as learning sources, *Contemporary Educational Psychology*, **16**(2), pp. 192–201, 1991.
8. C. Rowland, The effect of testing versus restudy on retention: A meta-analytic review of the testing effect, *Psychological Bulletin*, **140**(6), pp. 1432–1463, 2014.
9. P. Agarwal, J. Karpicke, S. Kang, H. Roediger and K. McDermott, Examining the testing effect with open- and closed-book tests, *Applied Cognitive Psychology*, **22**(7), pp. 861–876, 2008.
10. J. Karpicke, A. Butler and H. Roediger III, Metacognitive strategies in student learning: Do students practice retrieval when they study on their own?, *Memory*, **17**(4), pp. 471–479, 2009.
11. J. Glover, The "testing" phenomenon: Not gone but nearly forgotten, *Journal of Educational Psychology*, **81**(3), pp. 392–399, 1989.
12. E. E. Abbott, On the analysis of the factors of recall in the learning process, *Psychological Monographs*, **11**, pp. 159–177, 1909.
13. H. F. Spitzer, Studies in retention, *Journal of Educational Psychology*, **30**(9), pp. 641–656, 1939.
14. M. McDaniel, J. Anderson, M. Derbish and N. Morrisette, Testing the testing effect in the classroom, *European Journal of Cognitive Psychology*, **19**(4–5), pp. 494–513, 2007.
15. S. Carpenter, H. Pashler, J. Wixted and E. Vul, The effects of tests on learning and forgetting. *Memory & Cognition*, **36**(2), pp. 438–448, 2008.
16. P. Delaney, P. Verkoeijen and A. Spirgel, Spacing and Testing Effects, Psychology of *Learning and Motivation*, **53**, pp. 63–147, 2010
17. J. Karpicke and W. Aue, The Testing Effect Is Alive and Well with Complex Materials, *Educational Psychology Review*, **27**(2), pp. 317–326, 2018.
18. A. Butler and H. Roediger, Testing improves long-term retention in a simulated classroom setting, *European Journal of Cognitive Psychology*, **19**(4–5), pp. 514–527, 2007.
19. C. Johnson and R. Mayer, A testing effect with multimedia learning. *Journal of Educational Psychology*, **101**(3), pp. 621–629, 2009.
20. N. Kornell and L. Son, Learners' choices and beliefs about self-testing, *Memory*, **17**(5), pp. 493–501, 2009.
21. S. Lipowski, M. Pyc, J. Dunlosky and K. Rawson, Establishing and explaining the testing effect in free recall for young children, *Developmental Psychology*, **50**(4), pp. 994–1000, 2014.
22. M. McDaniel, H. Roediger and K. Mcdermott, Generalizing test-enhanced learning from the laboratory to the classroom, *Psychonomic Bulletin & Review*, **14**(2), pp. 200–206, 2007.
23. T. Toppino and M. Cohen, The Testing Effect and the Retention Interval, *Experimental Psychology*, **56**(4), pp. 252–257, 2009.
24. P. Verkoeijen, S. Bouwmeester and G. Camp, A Short-Term Testing Effect in Cross-Language Recognition, *Psychological Science*, **23**(6), 567–571, 2012.
25. M. Vojdanoska, J. Cranney and B. Newell, The testing effect: The role of feedback and collaboration in a tertiary classroom setting, *Applied Cognitive Psychology*, **24**(8), pp. 1183–1195, 2009.
26. D. Rohrer, R. F. Dedrick and S. Stershic, Interleaved practice improves mathematics learning, *Journal of Educational Psychology*, **107**(3), pp. 900–908, 2015.
27. O. Adesope, D. Trevisan and N. Sundararajan, Rethinking

the Use of Tests: A Meta-Analysis of Practice Testing, *Review of Educational Research*, **87**(3), pp. 659–701, 2017.
28. S. Pan and T. Rickard, Transfer of test-enhanced learning: Meta-analytic review and synthesis, *Psychological Bulletin*, **144**(7), pp. 710–756, 2018.
29. G. L. Bradshaw and J. R. Anderson, Elaborative encoding as an explanation of levels of processing, *Journal of Verbal Learning and Verbal Behavior*, **21**, pp. 165–174, 1982.
30. T. Elizabeth, T. L. Ross Anderson, E. H. Snow and R. L. Selman, Academic Discussions: An Analysis of Instructional Discourse and an Argument for an Integrative Assessment Framework, A*merican Educational Research Journal*, **49**, pp. 1214–1250, 2012.
31. K. Murphy, I. A. Wilkinson and A. O. Soter, Instruction based on discussion, in R. Mayer & P. Alexander (eds), *Handbook of Research on Teaching and Learning*, New York: Routledge, pp. 382–407, 2011.
32. K. R. Wentzel and D. E. Watkins, Instruction based on peer interactions, in R. Mayer & P. Alexander (eds) *Handbook of Research on Teaching and Learning*, New York: Routledge, pp. 322–343, 2011.
33. A. Graesser, S. D'Mello and W. Cade, Instruction based on tutoring, in R. Mayer & P. A. Alexander (eds) *Handbook of research on learning and instruction*, New York: Routledge, pp. 408–306, 2011.
34. P. Bazelais, D. J. Lemay and T. Doleck, How Does Grit Impact Students' Academic Achievement in Science? *European Journal of Science and Mathematics Education*, **4**(1), pp. 33–43, 2016.
35. P. Zhang, P. L. Ding and E. Mazur, Peer Instruction in introductory physics: A method to bring about positive changes in students' attitudes and beliefs, *Physical Review Physics Education Research*, **13**(1), pp 1–10, 2017.
36. J. Cohen, A power primer, *Psychological Bulletin*, **112**, pp. 155–159, 1992.
37. X. Chen, *STEM Attrition: College Students' Paths Into and Out of STEM Fields (NCES 2014-001)*, National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC, 2013.
38. J. Gasiewski, M. Eagan, G. Garcia, S. Hurtado and M. Chang, From Gatekeeping to Engagement: A Multicontextual, Mixed Method Study of Student Academic Engagement in Introductory STEM Courses, *Research in Higher Education*, **53**(2), pp. 229–261, 2011.
39. M. Baeten, K. Struyven and F. Dochy, Student-centred teaching methods: Can they optimise students' approaches to learning in professional higher education? *Studies in Educational Evaluation*, **39**(1), pp. 14–22, 2013.
40. J. Biggs, Enhancing teaching through constructive alignment, *Higher Education*, **32**(3), pp. 347–364, 1996.
41. L. Gow and D. Kember, Conceptions of teaching and their relationship to student learning, *British Journal of Educational Psychology*, **63**, pp. 20–23, 1993.
42. F. Marton and R. Säljö, On qualitative differences in learning: I-Outcome and process, *British Journal of Educational Psychology*, **46**(1), pp. 4–11, 1976.
43. K. Trigwell and M. Prosser, Improving the quality of student learning: The influence of learning context and student approaches to learning on learning outcomes, *Higher Education*, **22**(3), pp. 251–266, 1991.
44. M. Prosser and D. Sze, Problem-based learning: student learning experiences and outcomes, *Clinical Linguistics & Phonetics*, **28**(1–2), pp. 131–142, 2014.
45. E. Kyndt, F. Dochy, K. Struyven and E. Cascallar, The perception of workload and task complexity and its influence on students' approaches to learning: A study in higher education, *European Journal of Psychology of Education*, **26**(3), pp. 393–415, 2011.

Paul Bazelais is a PhD candidate at McGill University and an instructor at John Abbott College in Montreal, QC, Canada.

David John Lemay, PhD is a research associate with the Douglas Mental Health Institute, McGill University, QC, Canada.

Tenzin Doleck is a postdoctoral research scholar at the University of Southern California, USA.