Summative versus formative assessments in teaching physiology to biomedical engineering students: a comparison of outcomes

Dr. William H Guilford, University of Virginia

Will Guilford is an Associate Professor of Biomedical Engineering at the University of Virginia. He is also the Undergraduate Program Director for Biomedical Engineering, and the Director of Educational Innovation in the School of Engineering. He received his B.S. in Biology and Chemistry from St. Francis College in Ft. Wayne, Indiana and his Ph.D. in Physiology from the University of Arizona. Will did his postdoctoral training in Molecular Biophysics at the University of Vermont under David Warshaw. His research interests include novel assessments of educational efficacy, the molecular basis of cell movement, and the mitigation of infectious diseases.

Dr. Brian P. Helmke, University of Virginia

Brian Helmke is currently Associate Professor of Biomedical Engineering at the University of Virginia. He received the B.S.E. in bioengineering from the University of Pennsylvania, the B.S.Econ. from The Wharton School of the University of Pennsylvania, and the Ph.D. in bioengineering from the University of California, San Diego. Brian's research interests include cardiovascular physiology, cellular mechanobiology, and nanotechnology-based biomaterials. He is also interested in technology-enhanced teaching and in experiential learning for undergraduates in science and engineering.

Summative versus formative assessments in promoting learning of physiology by biomedical engineering students: a comparison of outcomes

Testing plays three roles in education. First, it serves a motivational role by holding students accountable for their work.

Second, testing serves an assessment function, not only for the purpose of assigning grades ("summative assessment") but also for providing feedback to students to guide their learning ("formative assessment"). Formative assessment has been broadly defined:

> "Practice in a classroom is *formative* to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited." [1]

There is, in fact, no formally recognized definition of the term. Perhaps because of its broad and uncertain definition, it remains uncertain how efficacious formative assessment is in improving student learning [2], [3]. Despite this, formative assessment is common in modern educational practice, particularly in hybrid learning paradigms [4].

Third and finally, summative testing *intrinsically* improves learning [5]. The latter is called the "testing effect." While a comprehensive review of the literature on the testing effect (see [6]) is beyond the scope of this manuscript, it is known to affect not only the retention and recall of knowledge, but also that of manual skills [7].

With the uncertainty surrounding the effectiveness of formative assessment, it certainly remains an open question as to whether it (frequent low- or no-stakes testing) or the testing effect (frequent high-stakes testing) is more effective in promoting learning.

We compared formative assessment to summative assessment via their effects on learning in two sections of a course in human physiology for biomedical engineering students. One section of the course (control) used weekly quizzes between each of four exams, with students receiving the higher of two scores – either the exam score, or the average score of the quizzes. The other section of the course (experimental) used frequent, low-stakes, primarily formative assessments to help students gauge their own learning between each of four exams. Learning outcomes were assessed through a physiology concept inventory administered on the first and last days of the course, and through a subset of questions on each of the four exams that were common between the two course sections. The data showed a trend toward higher overall exam scores and post-course retention and recall in the section taught using only summative assessments compared to the section that used formative assessment. The differences, however, were not significant. These data suggest that well-structured formative assessments can perform nearly as well in inducing the testing effect as frequent, higher-stakes formative assessments.

Intervention and Methods

Our interventions were made in two sections of a first-semester course in physiology for biomedical engineering students, both taught within a biomedical engineering department. We structured our course roughly around the textbook "Textbook of Medical Physiology" by Guyton and Hall [8]. The major topical areas covered were:

1. Fundamentals of cell biology, ion gradients, and excitability
2. Muscle contraction
3. Cardiac physiology
4. Vascular physiology
5. Respiratory physiology

The course was divided into two sections. Students were free to choose either section. Students had no *a priori* knowledge of whether or how the sections would differ, other than who was to be the instructor.

Both sections were divided into four units of 3-4 weeks' duration, each with an associated exam. The final exam of the course was not "comprehensive," but rather covered only the final unit of the course material.

Course section 1 (control, with the testing effect presumably induced)

Students in section 1 of the course took quizzes weekly that covered the previous two lectures (one week) of material. A typical quiz consisted of 10 multiple choice or short answer questions, and students were given 10 minutes at the beginning of class to complete it. The timing of these weekly quizzes was not arbitrary. Spitzer showed that the magnitude of the testing effect depends on how proximal the testing is to the studying – sooner is better, with impressive results out to 1 week delay between study and first testing [9].

Every 3-4 weeks, the weekly quiz was followed by a full class period set aside to deliver a unit exam. These were in every way similar in format to the quizzes, but consisted of 30-33 questions. The online service QuestionPress (*www.questionpress.com*) was used to deliver and automatically score these summative assessments. Students ordinarily received their quiz scores, but not the answer key, immediately after completing the quiz. On days when a quiz was followed immediately by a unit exam, students received their scores as well as the answer key immediately after taking the quiz.

For each quarterly unit of the course, students were credited either (a) their exam score, or (b) the mean of their quiz scores, whichever was higher toward their final course grade. They were therefore not obligated to take both the quizzes and the exams, but could instead opt not to take the exam if they were satisfied with their weekly quiz scores. 96% of their grade in this section was based on these summative assessments of knowledge and comprehension, while the remaining 4% was based on class participation.

The comparatively high-stakes assessments in this section were assumed to induce the testing effect. Indeed, this summative assessment structure was very similar to that of McDaniel and coworkers [10]. They found significantly improved correct answers on unit exams (every 3 weeks) as a result of weekly quizzes, in contrast to weekly reading assignments.

Course section 2 (intervention, with formative assessment)

Students in section 2 of the course participated in a blended lecture/active learning format that included frequent formative assessments in addition to quarterly summative unit exams. Quarterly unit exams, which were substantially similar to those in the control section, comprised 60% of the students' grades, and formative assessments comprised 40%, as outlined below. In fact, on every exam some questions were identical between the two sections (see "Exam questions," below).

Daily class discussion activities (7.5% of the grade) included group discussions, thought questions, and concept checks designed to help students self-evaluate their understanding of concepts and learning objectives of the day. Questions were administered using the online tool QuestionPress (*www.questionpress.com*), and students received full credit for answering questions thoughtfully, whether or not their answers were correct. Answers were used in aggregate by the instructor to correct misunderstandings, stimulate problem-solving, and reinforce key course concepts.

Daily flashcard quizzes (7.5% of the grade) were designed to provide practice memorizing vocabulary and numerical values commonly used in physiology. Quizzes were administered online using the Tests & Quizzes tool in the university's learning management system, Collab (based on Sakai). Quizzes were released before each class day to help students prepare for class discussions, and students could take the quizzes as often as they liked throughout the semester. Correct answers were provided automatically at the end of each quiz attempt, and students received full credit for completing the quizzes at least once.

Weekly practice quizzes (15% of the grade) served as "exam simulators" to help students practice for the quarterly unit exams. Practice quiz questions integrated vocabulary and concepts into mechanisms and control systems in a manner very similar to exam questions. Practice quizzes were administered and automatically graded by the Tests & Quizzes tool in Collab. Students could complete practice quizzes as often as they liked before each unit exam, and the highest score from each quiz was used to compute the final grade.

Exploration activities (10% of the grade) were team activities that included discussion questions relating online resources to course concepts. The learning objectives were to help students recognize mechanistic relationships, find and evaluate physiology facts online and in published literature, and appreciate historical development of the physiology field. Short reports answering the discussion questions were graded by the instructors.

Overall, these categories of formative assessments served to promote frequent interaction with course material and to stimulate questions and discussion during class periods. The goal was to determine whether these frequent activities improved students' performance on quarterly exams to a degree similar to the well-established testing effect.

Exam questions

Fifty-eight (58) of the exam questions given across the semester were identical between the two sections of the course, and allowed direct section-to-section comparisons.

Concept inventory

Students were given a brief, *ad hoc* Pre-Post physiology concept inventory (see Appendix A) to judge their improvement over the course of the semester. For completing the concept inventory students were awarded class participation credit.

Statistics

Concept inventory results within each of the two course sections were compared, end of course to the beginning, by paired t-test. Effect sizes were calculated as Cohen's *d*. All comparisons between the sections were made by one-way ANOVA.

Results

Concept inventory gains did not differ between the two course sections

The two sections of the course did not differ in their original concept inventory scores ($p = 0.716$). The mean concept inventory scores at the beginning of the semester were 43±12 (N=68) and 42±12 (N=36, mean ± σ), respectively, for the control and formative assessment sections. Only one student transferred between the two sections of the course after the start of classes.

Both sections showed increases in concept inventory score by the end of the semester, with means of 73±9 and 67±16, respectively. These increases were significant at the $p<10^{-9}$ level.

We found a nearly significant difference between the two sections at the end of the semester, with the testing effect section tending toward higher scores than the formative assessment section ($p=0.052$, $d=0.4$). This difference, however, may be greater or lesser when the starting scores are considered for each individual student.

We therefore calculated a difference score for the concept inventory for each student across the span of the semester (end of semester minus the beginning of semester score, or

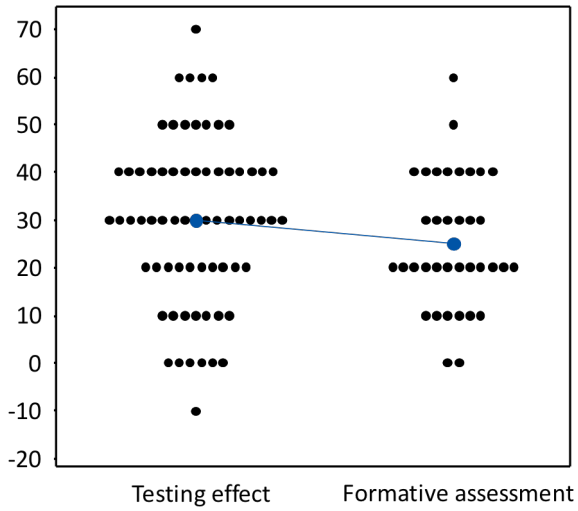

Figure 1: Difference scores for the post-pre-physiology concept inventory. Individual data points are shown in black, while the blue data points indicate the mean of that section. Note that the majority of the data are positive, indicating substantial gains in the concept inventory over the span of the semester for both sections

"post-pre"). While there was a trend in the post-pre-scores in favor of the control group performing better (Figure 1), the difference was not significant by ANOVA (p=0.185). These difference scores were 30±16 for the control section, and 25±14 for the formative assessment section. Further, the effect size was small (Cohen's d=0.28).

Post-pre-gains were not evenly distributed, however, as a function of beginning of semester concept inventory scores. Not surprisingly, students who scored lower at the beginning of the semester had more room for gain than did students who scored well at the beginning of the semester. Thus the slope of post-pre gain as a function of pre-test score was negative in both sections (Figure 2). However, the intercept – the expected post-pre-gain if the pre-test score is zero – was significantly higher ($p=5\times10^{-5}$) in the control section of the course (73±6 s.e.) than in the formative assessment section (38±7 s.e.). This suggests that students who had less pre-course knowledge of physiology tended to retain more knowledge in the section with presumed strong testing effect than in the formative assessment section of the course.

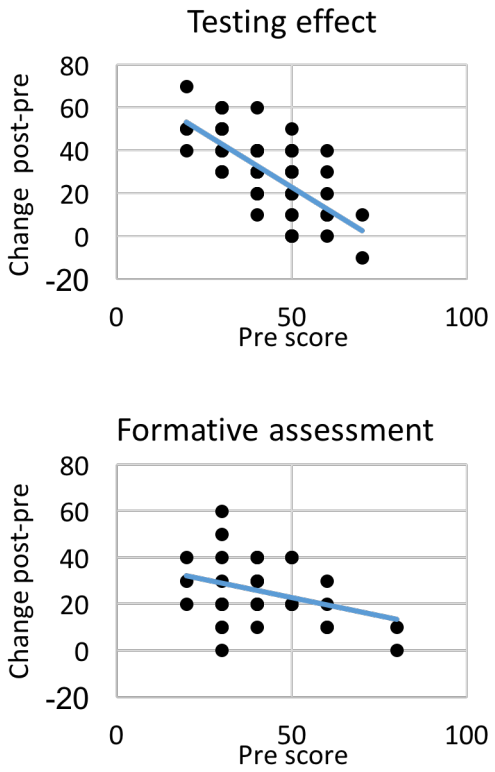

Figure 2: Post-pre course change in concept inventory scores as a function of pre-course score.

Accurate responses to mid-term exam questions did not differ between the two sections

We compared the performance of students on 58 unit exam questions that were identical between the two sections of the course. On a 0-3 scale, students averaged 2.2±0.6 for these questions in the testing effect section, compared to 2.0±0.7 for the formative assessment section. This difference was not significantly different by paired t-test (p=0.074).

There was a significant (p=0.005) but weak (correlation = 0.38) positive relationship between the individual exam question scores in the two sections (Figure 3). This relationship included a non-zero intercept (1.1±0.3, p=0.002), indicating that on low-scoring, and presumably more difficult

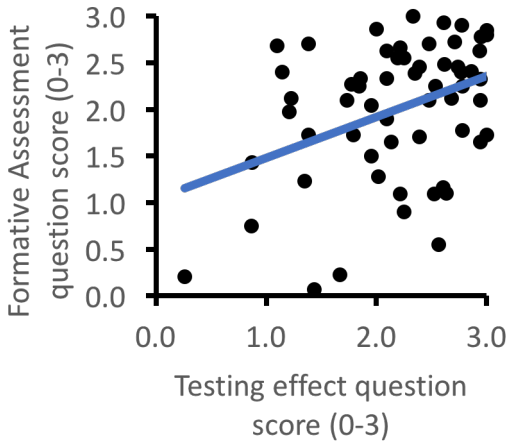

Figure 3:The relationship between matched exam question scores between the two sections. The line shows the linear regression on these data.

questions, the formative assessment section tended to perform better than the testing effect section.

One can also appreciate from the graph and from the weak correlation that there were some large differences in how students in the two sections performed on any given question. To further explore this phenomenon, we classified each exam question as either testing retention/recall, or comprehension of material. We analyzed these two sets of questions separately. We found that while students in the two sections performed similarly on comprehensive questions ($p=0.98$, $d=0.005$), students in the testing effect group performed significantly better in retention and recall than did those in the formative assessment group ($p=0.011$) and with moderate effect size ($d=0.53$).

Discussion

We conclude that well-structured formative assessments perform as well in inducing the testing effect as frequent, higher-stakes formative assessments in boosting student comprehension. While our data showed a trend toward improved exam scores in the control (testing effect) group compared to the formative assessment group, the differences were small and not statistically significant unless questions emphasizing pure retention and recall were viewed in isolation. Retention and recall specifically were boosted by the testing effect relative to formative assessment. One might achieve even higher gains in retention and recall by increasing the frequency of summative assessment from once per week to twice per week, since shortening the delay from learning to first assessment from 7 days to 1 has been reported to improve recall by approximately 10% [9].

There are two obvious explanations for the relative similarities in student learning between the two sections of the course. First, it is impossible to completely separate formative from summative assessments. Any assessment that is required of students must include a summative aspect, and every summative assessment can be used by students for formative purposes. In fact, students in the formative assessment section were asked in class discussions to identify points of confusion when reviewing the exam solutions. Thus, formative assessments may themselves induce the testing effect.

Alternatively, formative assessments may induce learning by causing students to recognize, evaluate, and react to the assessment or the course material [11]. That is, it is a reflective exercise. Detailed, but not superficial, reflection on learning has been associated with significant learning gains [12].

The formative assessments were part of a broader educational strategy to enhance student the learning experience of the student. For example, student feedback regarding “exploration activities” in section 2 revealed students’ appreciation for connecting in-class discussions and textbook readings to current events and to research in physiology and biomedical engineering. Although we did not measure student motivation, we hypothesize that student motivation would be increased when provided with opportunities for engagement with authentic biomedical problems [13]. Thus, retention/recall and comprehension may not be the only relevant metrics of learning [14].

Literature Cited

[1] P. Black and D. Wiliam, "Developing the theory of formative assessment," *Educ. Assess. Eval. Account.*, vol. 21, no. 1, 2009.

[2] K. E. Dunn and S. W. Mulvenon, "A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education," *Pract. Assess. Res. Eval.*, vol. 14, no. 7, Mar. 2009.

[3] R. E. Bennett, "Formative assessment: a critical review," *Assess. Educ. Princ. Policy Pract.*, vol. 18, no. 1, pp. 5–25, Feb. 2011.

[4] M. Horn and H. Staker, "Formative Assessment Is Foundational to Blended Learning," *THE Journal*, 14-Nov-2012. [Online]. Available: https://thejournal.com/articles/2012/11/14/formative-assessment-is-foundational-to-blended-learning.aspx. [Accessed: 25-Jan-2017].

[5] A. I. Gates, *Recitation as a factor in memorizing*, vol. 40. New York: The Science press, 1917.

[6] H. L. Roediger and J. D. Karpicke, "The Power of Testing Memory: Basic Research and Implications for Educational Practice," *Perspect. Psychol. Sci.*, vol. 1, no. 3, pp. 181–210, Sep. 2006.

[7] C. B. Kromann, M. L. Jensen, and C. Ringsted, "The effect of testing on skills learning," *Med. Educ.*, vol. 43, no. 1, pp. 21–27, Jan. 2009.

[8] J. E. Hall, *Guyton and Hall Textbook of Medical Physiology*, 13th ed. Elsevier Health Sciences, 2015.

[9] H. F. Spitzer, "Studies in retention," *J. Educ. Psychol.*, vol. 30, no. 9, pp. 641–656, 1939.

[10] M. A. McDaniel, J. L. Anderson, M. H. Derbish, and N. Morrisette, "Testing the testing effect in the classroom," *Eur. J. Cogn. Psychol.*, vol. 19, no. 4–5, pp. 494–513, Jul. 2007.

[11] B. Bell and B. Cowie, "The characteristics of formative assessment in science education," *Sci. Educ.*, vol. 85, no. 5, pp. 536–553, Sep. 2001.

[12] M. Menekse, G. Stump, S. J. Krause, and M. T. H. Chi, "The Effectiveness of Students' Daily Reflections on Learning in an Engineering Context," presented at the 2011 ASEE Annual Conference & Exposition, 2011, p. 22.1451.1-22.1451.10.

[13] W. Newstetter and P. Benkeser, "Learning Assessment In Problem Based Learning For Bme Students," presented at the 2002 Annual Conference, 2002, p. 7.801.1-7.801.8.

[14] W. Guilford, A. Blazier, and A. Becker, "Integration of Academic Advising into a First-year Engineering Design Course and Its Impact on Psychological Constructs," 2015, p. 26.995.1-26.995.13.

Appendix A – Concept Inventory

This concept inventory is designed to cover the first semester of a two-semester sequence in human physiology, including excitability, muscle contraction, the cardiovascular system, and the respiratory system. Correct answers are shown in *italics*.

The process of keeping internal conditions constant is called:

- hemostasis
- *homeostasis*
- steady state
- equilibrium

"Active transport" will always result in:

- *a higher concentration gradient*
- a lower concentration gradient
- a higher concentration of the transported substance inside the cell
- improved oxygen delivery to tissue

At the peak of the action potential, the membrane potential is:

- exactly at the Na^+ equilibrium potential
- close to but more positive than the Na^+ equilibrium potential
- *close to but less positive than the Na^+ equilibrium potential*
- exactly at 0 mV
- the same as the resting membrane potential

The atrioventricular valves open during:

- ventricular systole
- *ventricular diastole*
- atrial diastole
- both atrial and ventricular systole

If the heart's natural pacemaker fails to fire, then:

- no blood would enter the atria
- no blood would enter the ventricles
- *the node on the floor of the right atrium would act as a secondary pacemaker*
- the node on the floor of the left ventricle would act as a secondary pacemaker
- the person would die within minutes

The exchange of gases and nutrients between blood and tissues is a major function of:

- arterioles

- arteries
- *capillaries*
- veins

Which of the following parameters, if doubled, would cause the largest increase in flow of blood through a vessel?

- Blood pressure
- Vessel length
- *Vessel diameter*
- Blood viscosity

Which of the following statements about skeletal and cardiac muscle is true?

- force increases as velocity of shortening increases
- *there is an optimal muscle length for force generation*
- nerve impulses activate contraction
- both muscle types can generate sustained contractions

Carbon dioxide:

- has no effect on hemoglobin
- *is carried by hemoglobin*
- influences hemoglobin only through pH changes
- impedes release of oxygen from hemoglobin

At which of the following times in the respiratory cycle is the intrapleural pressure most negative?

- *just after the beginning of inhalation (inspiration)*
- just before the end of inhalation
- just after the beginning of exhalation (expiration)
- just before the end of exhalation