11-1-2014

A Subgroup Analysis of the Impact of Self-testing Frequency on Examination Scores in a Pathophysiology Course

Peter C. Panus
East Tennessee State University, panus@etsu.edu

David W. Stewart
East Tennessee State University, stewardw@etsu.edu

Nicholas E. Hagemeier
East Tennessee State University, hagemeier@etsu.edu

Jim C. Thigpen
East Tennessee State University, thigpen@etsu.edu

Lauren Brooks
East Tennessee State University

Citation Information

Panus, Peter C.; Stewart, David W.; Hagemeier, Nicholas E.; Thigpen, Jim C.; and Brooks, Lauren. 2014. A Subgroup Analysis of the Impact of Self-testing Frequency on Examination Scores in a Pathophysiology Course. *American Journal of Pharmaceutical Education*. Vol.78(9). https://doi.org/10.5688/ajpe789165 ISSN: 0002-9459

A Subgroup Analysis of the Impact of Self-testing Frequency on Examination Scores in a Pathophysiology Course

RESEARCH

A Subgroup Analysis of the Impact of Self-testing Frequency on Examination Scores in a Pathophysiology Course

Peter C. Panus, PhD, PT,[a,b] David W. Stewart, PharmD, BCPS,[a] Nicholas E. Hagemeier, PharmD, PhD,[a] Jim C. Thigpen, PharmD, BCPS,[a] Lauren Brooks, SDPT[b]

[a]Bill Gatton College of Pharmacy, East Tennessee State University, Johnson City, Tennessee
[b]College of Clinical and Rehabilitative Health Sciences, East Tennessee State University, Johnson City, Tennessee

Submitted December 04, 2013; accepted March 25, 2014; published November 15, 2014.

Objective: To determine if the frequency of self-testing of course material prior to actual examination improves examination scores, regardless of the actual scores on the self-testing.
Methods: Practice quizzes were randomly generated from a total of 1342 multiple-choice questions in pathophysiology and made available online for student self-testing. Intercorrelations, 2-way repeated measures ANOVA with post hoc tests, and 2-group comparisons following rank ordering, were conducted.
Results: During each of 4 testing blocks, more than 85% of students took advantage of the self-testing process for a total of 7042 attempts. A consistent significant correlation ($p \leq 0.05$) existed between the number of practice quiz attempts and the subsequent examination scores. No difference in the number of quiz attempts was demonstrated compared to the first testing block. Exam scores for the first and second testing blocks were both higher than those for third and fourth blocks.
Conclusion: Although self-testing strategies increase retrieval and retention, they are uncommon in pharmacy education. The results suggested that the number of self-testing attempts alone improved subsequent examination scores, regardless of the score for self-tests.

Keywords: pharmacy education, self-testing, formative assessment, active learning

INTRODUCTION

Assessment, in its many forms, permeates pharmacy curricula. The Accreditation Council for Pharmacy Education (ACPE) places particular emphasis on assessment of programmatic and student outcomes and encourages use of multiple types of assessments in its standards and guidelines.[1] From a student perspective, examinations comprised of multiple-choice items are perhaps one of the most commonly employed means by which individual performance is assessed, both in a formative and summative manner (eg, taking a practice examination and licensure examination, respectively). Examinations can be short, long, low-stakes (or no-stakes), high-stakes, and can be called by several names, including quizzes, tests, and assessments. Pharmacy students have opined that testing serves primarily to assess the amount of material learned in pharmacy courses and to assign grades.[2] While valid examinations can be used to assess student outcomes, research indicates examinations serve another purpose less commonly perceived by test givers and test takers alike: examinations can improve learning.[3,4]

The testing effect has been one of the topics of interest for improving learning in the cognitive psychology arena for centuries.[5,6] Specifically, researchers have examined multiple aspects of testing as it relates to retrieval (accessing stored information) and retention (keeping information in memory).[3,4,7-15] Karpicke, Roediger, Bjork, among others, noted increased long-term retention after repeated retrieval (eg, self-testing) compared to repeated studying (eg, rereading notes).[4,8,10,14] Intermittently testing events in learning environments also facilitated learning more than repeated studying did.[7,9,16,17]

Despite the ubiquitous nature of examinations in schools and colleges, to our knowledge, no research has been conducted to evaluate the relationship between self-testing attempts and course outcomes in pharmacy curricula. Hagemeier and Mason examined student pharmacists' perceptions of testing and found that only 7.3% of respondents perceived tests as events that improve learning.[2] The same students, when given scenarios to gauge the extent to which they would employ self-testing strategies as opposed to restudying strategies, indicated a lack of awareness in their responses that self-testing can

Corresponding Author: Peter C. Panus, PhD, Bill Gatton College of Pharmacy, Box 70594, East Tennessee State University, Johnson City, TN; Tel: 423-439-8789. Fax: 423-439-6350. E-mail: panus@etsu.edu.

improve learning more than restudying can. Overall, their research indicated many students did not employ strategies that optimally promoted learning, a finding similar to what was reported by Karpicke.[18,19]

ACPE Standard 11 indicates colleges and schools should employ teaching and learning strategies that "enable students to transition from dependent to active, self-directed, lifelong learners." [1] Structuring student opportunities to self-test through development of a practice self-test item bank is one method of promoting strategies that optimize learning. Frequent testing can mediate student learning by providing students with formative feedback from the examinations and can result in material retention for longer periods.[20,21] Therefore, self-test banks allow students to test frequently and simultaneously encourage the use of retrieval techniques that promote retention.

Previously, we developed and implemented a structured self-test bank of questions for a required first professional year (P1) pathophysiology course. Published results documented benefits of self-testing, specifically showing a positive correlation between performance on self-testing activities and performance on subsequent course examinations.[22] To our knowledge, no data exists evaluating the frequency of self-testing attempts, irrespective of student performance, and the relationship to course outcomes as measured by examination scores. The purpose of this manuscript is to report an analysis of the frequency of self-testing compared to examination scores.

METHODS

Pathophysiology was a 17-week, 4-credit hour, required course taken by doctor of pharmacy students during their first professional year. The course had 4 examinations, equally spaced and weighted throughout the semester. The material covered on each examination only pertained to that presented after the previous examination, and no examination was comprehensive. There were also 8 required quizzes, equally spaced and weighted throughout the course, accounting for 20% of the final grade. In 2011, a self-testing/quizzing component was added to the course to allow students to practice retrieval of course information. After content was discussed in class, a graded quiz was given. Following the quiz, a subsequent practice pool of questions was released to the students using the university's course management software, Desire-2-Learn (D2L) (Version 8.3, Kitchener, ON, Canada). During this time, students had unlimited access to the question pool, and each practice quiz took a maximum of 30 minutes and consisted of 15 questions drawn randomly from the pool. When the practice quiz was submitted to D2L, the quiz was scored immediately, and the students were allowed to see the percentage of questions answered correctly. To provide feedback on areas of student deficiency prior to the examination, students were able to view incorrectly answered questions with all potential answer choices, but the correct answer choice was not indicated. While taking the practice quizzes and reviewing incorrect quiz responses, students used the Respondus Lockdown Browser (Respondus, Redmond, WA). This browser prevents copying and pasting or similar activities. The D2L software proctoring the quizzes was set for auto-submit when the time period for the quiz ended. Thus, the D2L system would submit the quiz at the end of the 30 minutes even if a student had not completed it. All questions for the examinations, required quizzes, and practice quiz pools were written by one of the investigators, who was also the course director and lecturer. As this was not a prospective investigation, the questions were written in real time as the content was covered during lecture or in the appropriate sections of the textbook. All questions were written prior to the practice quizzes and only grammatical or formatting changes were made to questions while the practice quiz was available to students. A discussion board on D2L provided clarification to students on content issues related to the questions, and highlighted "bad" questions, and gave correct information to students. The investigation received approval from the East Tennessee State University Institutional Review Board prior to data collection and analysis.

All data were organized using Excel 2007. Data analyses were completed using IBM SPSS Statistics 19 (IBM Inc., Armonk, NY).[23,24] Two-way repeated measures Analysis of Variance (ANOVA) with interaction was conducted. The 2 main effects for the 2-way ANOVA were the variables "time" with 2 blocks and "performance" with 4 blocks. Two time blocks were compared: the first time block when the practice quizzes were available to the students and the second time block when the examinations were taken. Each performance block represented the number of attempts at the practice quizzes and the examination score, during each of the 4 examination periods. The 2-way ANOVA determined if there were any differences in number of practice quiz attempts using the question pool prior to each examination and if there was a difference in the examination scores that followed the practice quiz interval. The interaction between main effects showed the relationship between the number of attempts prior to the examination and the subsequent examination score for each of the 4 examination periods. Mauchly's test for sphericity was conducted on the performance main effect and interaction between performance and time effects. No test for sphericity was conducted on the main effect of time as only 2 blocks represented this analysis. A Greenhouse-Geisser correction was used to correct the F statistic for lack of sphericity when appropriate. Subsequently, only differences between number of practice quiz attempts or examination scores were evaluated. Group differences were examined using a 1-way ANOVA with a Welch's F test for lack of homogeneity, followed by Games-Howell post hoc multiple comparisons. A Pearson's correlation was conducted to examine individual associations between the number of practice quiz attempts and score on the subsequent examination. Finally, a 2-group comparison was conducted to determine whether there were differences in examination grades based on number of practice quiz attempts. For each examination, the students' scores and number of practice quiz attempts were rank-ordered based on the number of practice quiz attempts prior to the examination. The rank orders were then divided into cohort upper and lower 50th percentiles based on number of attempts. Comparison of examination scores between the upper and lower 50th percentiles were conducted with a 2-tailed *t* test for independent samples with equal variance not assumed. All graphs were created using Slide Write Plus for Windows Version 5.0 (Advanced Graphics Software Inc., Encinitas, CA), and data was reported as arithmetic mean plus or minus standard error of the mean.

RESULTS

Seventy-nine students were enrolled in the course in the spring semester of 2011. Table 1 provides the descriptive statistics associated with the number of practice quiz attempts using the question pool on D2L. The number of questions in each self-testing quiz bank varied by examination from 247 to 431 questions. The number of days the practice quiz question pool was open also varied from 14 to 19. Total number of self-attempts by the students for the practice quizzes prior to the examination increased from the first to the second testing block, and subsequently declined. The range of attempts by students also varied, and some students—not necessarily the same students—never accessed the practice quizzes during a given testing block, which was indicated by a zero at the bottom of the range for each examination period. However, as Table 1 documents, during testing blocks 1 through 3 more than 95% of students accessed the practice quizzes at least once, and in testing block 4 more than 85% accessed the practice quizzes.

The 2-way, repeated measures ANOVA with a Greenhouse-Geisser adjustment for lack of sphericity documented significant differences in the number of practice attempts and subsequent examination scores during the 4 testing blocks with all reported main effects for the ANOVA being significant ($p<0.05$). Additionally, during the 4 testing blocks, the number of practice quiz attempts did not result in a consistent equivalent score on the subsequent examination as documented by the significant interaction between performance and time variables. This was observed in the increase of number of attempts during the second testing block without a proportional increase in examination score, as compared to the other 3 examination blocks (Table 1).

Each of the subsequent 1-way ANOVAs was also significant ($p<0.002$) for the main effects of "Average Examination Score" and "Quiz Attempts/Student," after Welch's F test adjustment. The variable "Average Examination Score" determined the differences between the examination scores during the 4 testing blocks, and "Quiz Attempts/Student" determined the number of quiz attempts during the same 4 blocks. As there was no control in this study, the student cohort's first exposure to the self-testing concept during the first testing time block was viewed as the control (Table 1). Compared to the first time block, there were no significant differences in the number of practice quiz attempts in blocks 2 through 4; however, the number of attempts in block 3 was significantly lower than in block 2. In contrast, the scores for the first and

Table 1. Descriptive Statistics and Multiple Group Comparisons for Number of Attempts and the Corresponding Examination Score

Testing Block (N=79)	Number of Questions in Pool	Days Pool Open	Total Attempts for Class Range/Student	No. Students without Attempt	Average No. Quiz Attempts/Student	Average Examination Score
Examination 1	278	16	1695 (0-100)	3	21.5 ± 2.0 ^{a,b}	89.3 ± 0.8 ^{a}
Examination 2	247	14	2348 (0-150)	2	29.7 ± 2.7 ^{a}	90.4 ± 0.8 ^{a}
Examination 3	386	17	1384 (0-70)	2	17.5 ± 1.7 ^{b}	82.2 ± 1.0 ^{b}
Examination 4	431	19	1615 (0-90)	9	20.4 ± 2.1 ^{a,b}	83.0 ± 0.9 ^{b}

All reported main effects for the 2-way repeated measures ANOVA were significant ($p<0.05$) for the main effects variables of "performance" F(2.71, 211)=33.75 and "time" F(1, 78)=1526. A documented significant interaction existed between the variables of "performance" and "time" F(2.53, 197)=5.12. Subsequent individual 1-way ANOVAs were also significant for the main effects variables of "Examination Score" F(3, 172)=22.97 and "Quiz Attempts/Student" F(3, 171)=5.10. Groups with different letters are statistically different ($p\leq0.002$) based on 1-way ANOVA analysis.

second examinations were not significantly different, but both were significantly higher than scores for examinations 3 and 4.

A Pearson's correlation was conducted to compare practice quiz attempts to subsequent scores on examinations 1 through 4 (Table 2). A significant correlation existed between the number of practice quiz attempts and the subsequent examination score in all four testing blocks. However, this association decreased continuously as the testing blocks (ie, the semester) progressed, and was most precipitous for the last block.

Figure 1 shows the 2-group comparison for number of attempts at the practice question pool and subsequent examination score after separating the students into the upper and lower 50th percentile based on number of attempts. Prior to all examinations, the upper 50th percentile demonstrated significantly more practice quiz attempts compared to the lower 50th percentile. On examinations 1 and 3, the lower half, based on number of practice quiz attempts, scored lower on the subsequent examination. No difference was observed between the 2 groups on examination 2, and examination 4 differences only trended toward significance (*p*<0.06).

DISCUSSION

Examinations can be viewed by students and even faculty members as a necessary evil in education, representing major components of grades in coursework at all levels of the educational process. The common opinion is that examinations are required to assess student knowledge of course material content. Yet testing has the possibility of providing a learning process by improving retrieval and retention and increasing the efficacy of study time by showing students content requiring further study.[3,4,7-15,25-27] To evaluate the concept of self-testing in this analysis, we chose the number of quiz attempts by the students for comparison to examination performance. Choosing quiz attempts as the variable minimized confounding factors, such as differences in previous academic success among students taking the course.

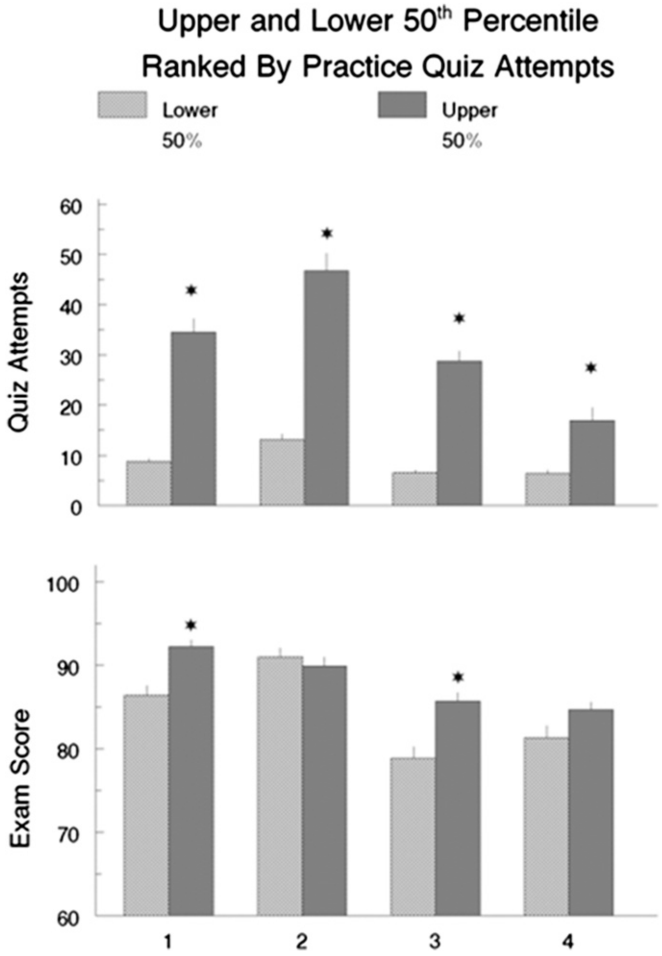

Figure 1. Comparison of upper and lower 50th percentile of class on the number of practice quiz attempts (upper graph), and subsequent examination score (lower graph) during the 4 testing blocks. The star represents when the upper and lower percentiles were significantly different (p<0.05) in either the upper or lower panels.

The value of such a learning process can go unnoticed by students. When surveyed, the majority of students both within and outside pharmacy education admitted to not using self-imposed or institutionalized self-testing compared to rereading, practice recall, or concept mapping.[2,19] Yet when compared to these alternative methods, testing provides improved retrieval and retention.[4,7,13,14,28,29] Even more surprising, students using these methods were more confident than students who self-tested, even though scores on subsequent examinations was lower for students opting not to self-test.[8,10,19,30] In contrast to previous surveys, students in this investigation did take advantage of the opportunity to self-test prior to an examination on the related content area. Seventy-nine students accessed the practice quizzes a total of 7042 times during the 17-week semester for an overall average of 22 attempts per examination per student. The difference between what students stated when surveyed about self-testing and what occurred in the present investigation may be related to ease of accessibility of such self-testing formats.

Table 2. Pearson's Correlations Between Number of Practice Quiz Attempts and Examination Score

(N=79)	Exam 1	Exam 2	Exam 3	Exam 4
Attempt = R	0.478	0.426	0.338	0.218
p value	0.001	0.001	0.002	0.05

Admittedly, there were students who never accessed the practice quizzes prior to each examination. This may have been related to the way the software program recorded access. If 2 or more students accessed the practice quizzes as a group, only 1 student would have logged into the system and would have been recorded as having taken the quiz. This obviously affected not only the potential total number of attempts by students but all subsequent analyses as well. However, when students access a testing program on their own time outside of class, there is no way to control such group participation.[31] No survey was conducted at the end of the course to determine student satisfaction with the self-testing quizzes. However, the use of these quizzes for the 4 examinations suggested that students found them beneficial. Moreover, data from a survey of medical students reported 90% of the students felt the self-testing process was beneficial for learning clinically related material.[32]

Assessment of testing on retention and retrieval has been conducted using different formats, including word pairs, free recall, and multiple choice questions.[3,4,7-11,13-15,28,29,31,33-36] Multiple choice questions have historically been thought to evaluate only recognition, and in turn result in decreased retention as compared to retrieval tests.[37-39] However, Little et al conducted research specific to multiple-choice examinations and the extent to which this commonly employed format can induce retrieval compared to recognition.[15] The researchers noted that multiple-choice examinations can induce retrieval if incorrect alternatives, or distractors, are plausible. Additionally, randomized multiple-choice questions have the advantage of facilitating recall of incorrect alternatives, which is missing from free recall and word pair assessments. In the present investigation, multiple-choice questions were used with the D2L testing program as they gave students immediate feedback when they used the practice quizzes for self-testing outside of class.

Although many investigations have examined testing under controlled laboratory settings, fewer have attempted such an investigation as a component of an actual class.[3,31,40,41] Roediger and Karpicke noted a lack of controls in the educational setting,[3] a limitation that existed in this investigation as well. Despite the lack of controls, Roediger and Karpicke concluded that the testing effect does extend to the classroom setting. In our investigation, a significant correlation existed for all 4 testing blocks between the number of attempts at self-testing via the practice quizzes and subsequent examination scores. Additionally, the upper 50th percentile, based on number of quiz attempts, scored higher on the subsequent examinations in 2 of 4 testing blocks, (examinations 1 and 3) and trended higher on examination 4 ($p<0.06$). In aggregate, these results support the value of testing, or in this case self-testing, as a learning tool in the academic setting.

Previously, investigators noted that technology limitations prevented tracking the frequency of use when self-testing outside the class room.[31] Using the Internet-based D2L program eliminated this problem. Lee, Nagel, and Gould also used D2L to assess self-testing by students in a human anatomy course in professional dentistry education.[40] The investigators concluded that self-testing, as assessed by multiple-choice quizzes, provided no consistent benefit to students on subsequent examinations; however, in that analysis only scores on the first attempt at the practice quizzes were compared to subsequent examination scores.

Previously, our team documented an association between the average practice quiz score and the related examination score.[22] The current analysis showed a similar, albeit weaker, association between the number of practice quiz attempts and subsequent examination performance irrespective of the students practice quiz scores. Both these analyses seem to be more logical predictors of students' eventual performance than simply how they performed on the first iteration of an exercise that should have been repeated multiple times.

Additionally, unsuccessful attempts, recognizing incorrect answer choices, and temporal spacing between testing attempts demonstrated a significant impact on retrieval, with equally spaced retesting superior to either retesting at expanded intervals or mass testing.[3,4,9,11] Although equally spaced testing optimizes recall, the current investigation was unable to control the interval between self-tests by students using the D2L testing program. Our results did document that more attempts correlated with a higher subsequent examination score, and that the self-testing opportunity was always open for a minimum of 2 weeks during any of the 4 separate pre-exam periods. Unfortunately, the software does not allow for minimal intervals between practice quiz attempts. Nor is systematic collection of individual attempts for statistical analysis available with the current version, so determination of intervals between attempts for each student would be difficult.

Incorrect answers during testing is also thought to enhance learning and retention, and this phenomenon may be further enhanced by delaying the feedback correcting the error.[3,13,34-36] Repeated testing of material correctly identified also increased learning and retention compared to dropping the item once correctly identified.[18] The positive benefit of continuously testing already correctly identified material is especially important, as students often stop self-testing on material once correctly identified.[4] In the current investigation, the software program parameters were set to allow students to review failed question attempts. Specifically, students were allowed to review incorrectly answered questions along with all the potential choices; however, the correct answer choice was not identified. Previous correctly and incorrectly answered questions were also kept within the random question pool, and the students had no control themselves over which questions were selected for their quizzes. This prevented students from removing correctly answered questions from the pool and repeatedly exposed students to these questions with randomization of the correct answer's position. Additionally, greater retrieval on examination occurs with testing compared to restudying when interfering tasks exist between the preparation, testing or restudying, and the examination. For students in academia such interfering tasks would be other courses.[14]

There were limitations to this investigation, some of which are unique. The number of days the self-testing practice quizzes were open prior to the examination was dictated by several factors. Each practice quiz window was open only after the course content was covered and prior to the examination date. The former was dictated by the class schedule and the latter was coordinated with the Office of Academic Affairs for the college. The variation of the self-testing practice quiz window was thus partially out of the investigators' control. The variation for number of days the self-testing practice quizzes were open was 5 days. As this investigation was not a priori, one of the investigators wrote all the questions for the practice quizzes in real time during the semester, which accounted for part of the difference in the number of questions during each testing block. Separately, these practice quizzes were an independent self-testing activity outside the class but within a larger academic curriculum for that semester. Thus, there were competing requirements from other courses within the curriculum, which may have interfered with student use of the practice quizzes during each of the 4 testing blocks. Moreover, study time for competing courses during the semester was completely out of the control of the investigators. Alternatively, students may have felt more self-confident in their ability to master the content of the course as the semester progressed. This failure to recognize their mastery of the course content may have resulted in slightly lower self-testing attempts during the third and fourth testing blocks and the subsequent scores on examinations 3 and 4.

CONCLUSION

In conclusion, in both the laboratory setting and the classroom, evidence supports the value of self-testing for retrieval and retention. This investigation presented additional data of the application of these concepts in the professional pharmacy education environment. The investigation also provided further documentation of the value and applicability of self-testing as a learning process by specifically showing a correlation between the number of self-testing attempts and subsequent examination scores.

REFERENCES

1. Accreditation Council for Pharmacy Education. Accreditation standards and guidelines for the professional program in pharmacy leading to a doctor of pharmacy degree. 2011; https://www.acpe-accredit.org/pdf/finalS2007Guidelines2.0.pdf. Accessed April 24, 2013.
2. Hagemeier NE, Mason HL. Student pharmacists' perceptions of testing and study strategies. *Am J Pharm Educ.* 2011;75(2):Article 35.
3. Roediger HLI, Karpicke JD. The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science.* 2006;1(3):181-210.
4. Karpicke J, Roediger HI. Repeated retrieval during learning is the key to long-term retention. *Journal of Memory & Language.* 2007;57:151-162.
5. Bacon F. *Novum Organum (L. Jardine & M. Silverthorne, Trans).* Cambridge, England: Cambridge University Press; 2000.
6. Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT. Improving students learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest.* 2013;14(1):4-58.
7. Nungester RJ, Duchastel PC. Testing vs review: effects on retention. *J Educ Psychol.* 1982;74(1):18-22.
8. Roediger HL, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci.* 2006;17(3):249-255.
9. Karpicke JD, Roediger HL 3rd. Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *J Exp Psychol Learn Mem Cogn.* 2007;33(4):704-719.
10. Karpicke JD, Roediger HL 3rd. The critical importance of retrieval for learning. *Science.* 2008;319(5865):966-968.
11. Karpicke JD, Bauernschmidt A. Spaced retrieval: absolute spacing enhances learning regardless of relative spacing. *J Exp Psychol Learn Mem Cogn.* 2011;37(5):1250-1257.
12. Karpicke JD, Grimaldi PJ. Retrieval-based learning: a perspective for enhancing meaningful learning. *Educ Psychol Rev.* 2012;24:401-418.
13. Kornell N, Hays MJ, Bjork RA. Unsuccessful retrieval attempts enhance subsequent learning. *J Exp Psychol Learn Mem Cogn.* 2009;35(4):989-998.
14. Halamish V, Bjork RA. When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *J Exp Psychol Learn Mem Cogn.* 2011;37(4):801-812.
15. Little JL, Bjork EL, Bjork RA, Angello G. Multiple-choice tests exonerated, at least of some charges: fostering test-induced learning and avoiding test-induced forgetting. *Psychol Sci.* 2012;23(11):1337-1344.
16. Kornell N, Bjork RA. The promise and perils of self-regulated study. *Psychon Bull Rev.* 2007;14(2):219-224.
17. Donovan JJ, Radosevich DJ. A meta-analytic review of the distribution of practice effect: now you see it, now you don't. *J Appl Psychol.* 1999;84(5):795-805.

18. Karpicke JD. Metacognitive control and strategy selection: deciding to practice retrieval during learning. *J Exp Psychol Gen.* 2009;138(4):469-486.
19. Karpicke JD, Butler AC, Roediger HL 3rd. Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory.* 2009;17(4):471-479.
20. Bangert-Drowns RL, Kulik JA, Kulik CC. Effects of frequent classroom testing. *The Journal of Educational Research.* 1991;85 (2):89-99.
21. Wheeler MA, Roediger HL. Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science.* 1992;3(4):240-245.
22. Stewart D, Panus PC, Hagemeier N, Thigpen J, Brooks L. Pharmacy student self-testing as a predictor of examination performance. *Am J Pharm Educ.* 2014;78(2):Article 32.
23. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Norwalk, CN: Appleton & Lange; 2009.
24. Field A. *Discovering Statistics Using SPSS*. 3rd ed. London, England: Sage; 2009.
25. Metcalfe J, Finn B. Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review.* 2008;15(1):174-179.
26. Izawa C. Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *J Exp Psychol.* 1970;83:340-344.
27. Nelson TO, Dunlosky J. When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: the 'delayed-JOL effect.'. *Psychological Science.* 1991;2(4):267-270.
28. Karpicke JD, Blunt JR. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science.* 2011;331(6018):772-775.
29. Bahrick HP, Hall LK. The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language.* 2005;52(4):566-577.
30. Agarwal PK, Karpicke JD, Kang SHK, Roediger HLI, McDermott KB. Examining the testing effect with open- and closed-book tests. *Appl Cognit Psychol.* 2008;22:861-876.
31. Roediger HL, Agarwal PK, McDaniel MA, McDermott KB. Test-enhanced learning in the classroom: long-term improvements from quizzing. *J Exp Psychol Appl.* 2011;17 (4):382-395.
32. Larsen DP, Butler AC, L. Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Med Educ.* 2013;47(7):674-682.
33. Pashler H, Zarow G, Triplett B. Is temporal spacing of tests helpful even when it inflates error rates? *J Exp Psychol Learn Mem Cogn.* 2003;29(6):1051-1057.
34. Butler AC, Karpicke JD, Roediger HL. Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *J Exp Psychol Learn Mem Cogn.* 2008;34 (4):918-928.
35. Butler AC, Karpicke JD, Roediger HL 3rd. The effect of type and timing of feedback on learning from multiple-choice tests. *J Exp Psychol Appl.* 2007;13(4):273-281.
36. Pyc MA, Rawson KA. Why testing improves memory: mediator effectiveness hypothesis. *Science.* 2010;330(6002):335.
37. Glover JA. The 'testing' phenomenon: Not gone but nearly forgotten. *J Educ Psychol.* 1989;81(3):392-399.
38. Foos PW, Fisher RP. Using tests as learning opportunities. *J Educ Psychol.* 1988;80:179-183.
39. McDaniel MA, Anderson JL, Derbish MH, Morrisette N. Testing the testing effect in the classroom. *European Journal of Cognitive Psychology.* 2007;19(4-5):494-513.
40. Lee LM, Nagel RW, Gould DJ. The educational value of online mastery quizzes in a human anatomy course for first-year dental students. *J Dent Educ.* 2012;76(9):1195-1199.
41. Nevid JS, Mahon K. Mastery quizzing as a signaling device to cue attention to lecture material. *Teaching of Psychology.* 2009;36 (1):29-32.

APPENDIX 1

Within Subjects Effect	Mauchly's W	Approximate Chi-square	df	Sig.	Greenhouse-Geisser
Time	1.00	.000	0	.	1.00
Performance	0.84	13.8	5	0.017	0.90
Time x Performance	0.76	21.4	5	0.001	0.84

Table 1. The main effect "time" represents the period during the attempts at the practice quizzes (1) or the period for which the examination is taken (2). The main effect "performance" represents 4 blocks during which each practice quiz period was followed by an exam. For the 2-way repeated measures ANOVA, Maulchy's test indicated that the assumption of sphericity had been violated for the main effect of performance $\chi^2(5)=13.76$, $p<0.02$ and interaction between main effects of performance and time $\chi^2(5)=21.36$, $p<0.001$. Time contains only 2 groups and thus sphericity is not an appropriate concern.

Tests of Within-Subjects Effect

Source		Type III Sum of Squares	df	Mean Square	F	Sig
Time	Sphericity Assumed	646208	1	646208	1526	.000
Error(time)	Sphericity Assumed	33020	78	423		
	Greenhouse-Geisser	33020	78	423		
Performance	Sphericity Assumed	9595	3	3198	33.8	.000
	Greenhouse-Geisser	9595	2.7	3547	33.8	.000
Error (performance)	Sphericity Assumed	22174	234	95		
	Greenhouse-Geisser	22174	211	105		
Time* performance	Sphericity Assumed	1113	3	371	5.1	.002
	Greenhouse-Geisser	1113	2.5	440	5.1	.003
Error (time*performance)	Sphericity Assumed	16952	234	72		
	Greenhouse-Geisser	16952	197	86		

Table 2. The Main ANOVA examining the variables "time," "performance," and their interaction. Sphericity was violated for the variable performance and the interaction between performance and time. As there were only 2 blocks for the variable time, sphericity was not an issue. All variable and their interactions were significant. Therefore, Greenhouse-Geisser estimates of sphericity were used to adjust the degrees of freedom: $\varepsilon = 0.90$ for performance and 0.84 for the interaction between performance and time. There were significant differences for performance $F(2.71, 211)=33.75$ and time $F(1, 78)=1526$. Finally there was a significant interaction between performance and time $F(2.53, 197)=5.12$, suggesting that the number of attempts at the practice quizzes did not result in an equivalent score on the subsequent examination during each of the testing periods.