RESEARCH

Effects of a Composite Examination Method, Which Assesses Multiple Courses on the Same Day, on the Performance of PharmD Students in a Basic Pharmacokinetics Course

Reza Mehvar, PharmD, PhD, Chapman University School of Pharmacy, Irvine, California

Richard Beuttler, PsyD, Chapman University College of Health and Behavioral Sciences, Irvine, California

Corresponding Author:
Reza Mehvar, Department of Biomedical and Pharmaceutical Sciences, Rinker Health Science Campus
Chapman University, 9401 Jeronimo Road, Irvine, CA 92168
Tel: 714-516-5490. Fax: 714-516-5481
E-mail: mehvar@chapman.edu

ABSTRACT

Objective. To investigate the effects of multicourse, composite examinations on the performance of students in a Pharmacokinetics course.

Methods. A linear, mixed-effects model was used to analyze individual students' performances in identical daily quiz (N=55) and examination (N=47) questions in a Pharmacokinetics course at two schools of pharmacy (Texas Tech University, N=162 and Chapman University, N=79). The entire course was taught by the same instructor at both institutions. The only difference between the two courses was that the method of administration of examinations was different between the Texas Tech (traditional, individual course examinations) and Chapman (multicourse, composite examinations) cohorts.

Results. Students' scores (mean ± SE) in the identical daily quiz questions, which were administered in an identical manner to students in both schools, were the same for Texas Tech (74.1 ± 0.82) and Chapman (74.7 ± 1.26) students. However, the grades of students in the multicourse examinations in the Chapman group (82.0 ± 1.09) were significantly (p<0.0001) lower than those in the individual course examinations in the Texas Tech group (90.6 ± 0.49). This difference amounted to an effect size of 1.15, indicating a large difference between the two cohorts in terms of their examination scores. The mixed-effects model revealed a negligible difference (0.622%) between the two student cohorts in terms of their academic

abilities but showed a substantial effect (9.40%) for the examination format in favor of single course assessment.

Conclusions. When compared to the traditional, individual course examination, composite examination significantly reduces the grades of students in a Pharmacokinetics course.

Keywords: composite examination, multicourse examination, integrated examination, student performance, basic pharmacokinetics

INTRODUCTION

Most schools of pharmacy in the United States assess students' performances in different courses using separate, course-specific examinations during each semester. Additionally, many schools of pharmacy use some form of cumulative, progress assessments annually or before the start of experiential education to demonstrate that students retain the required foundational knowledge and skills and are ready for advanced pharmacy practice experiences (APPE).[1-3] Indeed, the newly-in-place standards (Standards 2016) from the Accreditation Council for Pharmacy Education (ACPE) now require that all the US schools of pharmacy use Pharmacy Curriculum Outcomes Assessment (PCOA) at the conclusion of the didactic curriculum. PCOA is a nationally standardized examination that is developed by the National Association of Boards of Pharmacy and provides an assessment of student performance in foundational knowledge in various domains of biomedical, pharmaceutical, social/administrative/behavioral, and clinical sciences.[4]

Traditionally, course-specific assessments include administration of one or more mid-term (or within term) examinations plus a final examination for each course throughout and at the end of each semester, respectively. However, the effectiveness of this method of assessment with regard to content retention and integration of materials across different courses has been questioned. Therefore, efforts have been made to improve the assessment methods used in pharmacy schools in the United States. For example, Medina et al[5] incorporated a biannual integrated examination during the first 3 years of the pharmacy curriculum. The integrated examinations were administered twice a year during the final

examination periods and consisted of questions from all the required courses during that semester. The integrated examinations were embedded in the final examination of the pharmacy practice course series offered in each semester and accounted for 10% of each course's final course grade. The authors concluded that the integrated examinations improved the culture of assessment and faculty's understanding of the curriculum. However, further studies are needed to determine whether such integrated examinations would affect student learning outcomes, performance, or retention of content.

To mitigate the scheduling challenges associated with multiple examinations for several courses during the semester and the apparent limitations of traditional, course-specific assessments, the leadership of the new Chapman University School of Pharmacy (Chapman) planned its assessments based on simultaneous, multicourse examinations at regular intervals. Under this plan, a composite examination (CE), consisting of questions from all the courses in that trimester, would be administered every 2-3 weeks, resulting in a total of five examinations for the entire trimester. There are theoretical and conceptual arguments in favor of CE versus individual course examinations, including ease of scheduling, accommodation for more frequent testing, improved study habits, increased content retention, and similarity to board examinations. However, to date, there are no studies that assess potential differences between the two methods in learning and performance of students in individual courses. Indeed, other than a very recent publication by McDonough et al,[6] reporting attitudes and perceptions of students towards a CE implemented at the University of Tennessee Health Science Center College of Pharmacy, the literature on this subject is practically non-existent. As an instructor of Basic Pharmacokinetics, the primary author is interested in investigating the effects of composite examinations on the performance of students in this discipline, which is the objective of the current study. Comparison of performance of students in CE versus individual course examinations would be logistically very difficult. However, the primary author is in a unique position to investigate this topic because he has gathered substantial performance data for individual students in daily quizzes and periodic examinations at Texas Tech University Health Sciences Center, where traditional, course-specific examinations were used, before teaching the same (identical) course at Chapman University, where CE is used. The hypothesis of the

study was that the examination format (individual versus multicourse) would not affect the performance of students in learning basic pharmacokinetics principles.

METHODS

Basic Pharmacokinetics was offered as a three credit hour course to students in two schools of pharmacy, one with individual course examinations (Texas Tech; 2013-2014 academic year) and another with CE (Chapman; 2015-2016 academic year). At both schools, the entire course was taught by the same instructor (RM). The format of the course, which is based on the principles of active learning and substantial engagement of students before, during, and after the class sessions, is described in detail before.[7-10] The course contents, sequence of topics, total number of lectures per course, number of lectures per each topic, length of class sessions (75 min), and course resources (eg, online tools for assignmnts,[11] quizzes and examinations,[12] and simulations[10]) were identical for both classes. Additionally, every-day quizzes were administered in an identical manner to both cohorts at the end of each class session. However, the method of administration of a total of five examinations was different between the two cohorts. At Texas Tech, students received traditional, single course examinations throughout the semester on days that there were no other major examinations scheduled. However, at Chapman, there were five examination days throughout the trimester (every 2-3 weeks) when a composite examination, consisting of questions form all the courses offered during the trimester, would be administered during one block of time.

There were 162 students at Texas Tech and 79 students at Chapman enrolled in the course in 2013-2014 and 2015-2016 academic years, respectively. The quiz and examination questions were based on an online program described before in the *Journal*,[12] which creates individualized questions for each student by incorporating some random parameters in a question with identical structure for all students, but with different pharmacokinetic and/or dosing parameters. For example, a question asking all students to estimate the plasma half life of a drug after intravenous administration of the drug would have different plasma concentration-time courses and, hence, a different half life value for each student. Additionally,

multiple-choice, conceptual questions are drawn from multiple dynamic scenarios where all the choices are potentially possible depending on the dynamic scenario randomly selected for each student.[12] The questions are drawn from an examination question bank consisting of ~500 dynamic (individualized) questions. The interested readers may consult the published article[12] for more details about the question bank and examples of dynamic scenarios. Whereas some of the quiz and examination questions were the same for both Texas Tech and Chapman students, other questions were different. Therefore, for both daily quizzes and examinations, only questions that were the same for both cohorts were identified and used for comparison of performance between the two cohorts. There were a total of 55 quiz questions and 47 examination questions that were the same in both cohorts and were, therefore, included in the analysis.

To statistically compare the performance of students in quiz and examination questions between Texas Tech and Chapman students, we used a stepwise, linear, mixed-effects model. Mixed-effects model analysis is a preferred method of analysis when there are correlated data due to grouping of subjects (eg, Texas Tech and Chapman students) or repeated measurements on each subject over time (daily quizzes and regular examinations throughout the semester/trimester). An advantage of the model is that it allows the use of individual students as their own control to account for differences in the academic abilities of students. The model uses both fixed and random effects in the same analysis. The fixed effects in our model were school (X_{School}, assigned 0 for Chapman and 1 for Texas Tech), type of test (X_{Type}, assigned 0 for quiz and 1 for exam), and the interaction of school and type of test (X_{School} ´ X_{Type}, assigned 1 for Texas Tech exam and 0 for others). The random effect was individual students' intercept shift as a result of their academic differences from the group (b_i). The full model, incorporating all the fixed and random effects, is shown in the following equation:

$$\boldsymbol{Y_{ij} = \beta_0 + X_{ijSchool}\beta_{School} + X_{ijType}\beta_{Type} + \left(X_{ijSchool} \times X_{ijType}\right)\beta_{School\times Type} + b_i + \in_{ij}} \quad (1)$$

where Y_{ij} is the score of the i[th] student at the j[th] measurement; $\boldsymbol{\beta_0}$ is the population intercept; $\boldsymbol{\beta_{School}}$, $\boldsymbol{\beta_{Type}}$, and $\boldsymbol{\beta_{School\times Type}}$ are the fixed effect coefficients for school, type of test, or school-type of test

interaction, respectively; and $\in_{ij}$ refers to the unaccounted error. This method uses individual student's performance on each of the 55 quiz and 47 examination questions to create a separate intercept for each student to control for each student's academic starting point.

For determination of the best model to describe the data, a stepwise forward model selection method was used. Briefly, the simplest model (containing intercept only without any of the fixed effects) was used as the starting point and was statistically compared with progressively more complex models with the addition of one fixed effect at a time to finally arrive at the full model. Variations of the tested models are listed in Table 1. Progressively complex models were compared using both likelihood ratios and pairwise comparisons based on Pearson's Chi-Squared analysis.

The likelihood ratio test compares the likelihood of two models, which are different from each other only by the presence or absence of one factor, to predict the observed data. For example, the ability of a model with intercept only (Model 1, Table 1) to predict the observed results is compared with a model that incorporates both intercept and School (Model 2, Table 1), with School being the only differentiating factor. In the next step, the likelihood of accurate prediction of the observed data using a model that incorporates intercept and School (Model 2, Table 1) is compared with a model that incorporates intercept, School, and Test Type (Model 3, Table 1), with the Test Type being the differentiating factor. This process is continued until the likelihood of all the possible factors (including School-Test Type interaction) on the predictability of the model is determined. In addition to the likelihood ratio test, the Pearson's Chi-Squared analysis compares the probability (*p* value) of the pairwise comparison between any of the above two models when the effect of addition or removal of only one factor is tested. Basically, the Chi-Squared analysis of two models determines whether the addition or removal of a factor significantly affects the ability of the model to predict the observed values. For likelihood ratios, Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) were used as penalized likelihood criteria, and the model with the lowest AIC or BIC (highest likelihood) was selected. For Chi-Squared analysis a *p* value of <0.05 was considered significant.

After the best model fit was obtained based on the above criteria, post-hoc, two-tailed t-tests were also used to compare differences between the following groups using a Bonferroni correction for *p* values to control for type I error: Chapman Quizzes versus Texas Tech Quizzes, Chapman Exams versus Texas Tech Exams, Chapman Quizzes versus Chapman Exams, and Texas Tech Quizzes versus Texas Tech Exams. In addition to *p* values, we also estimated the effect sizes when comparing the groups. Recently,[13] it has been argued in the literature related to education that merely citing a *p* value for determination of significance is not enough, and *p* values should be associated with an appropriate measure of magnitude of the difference between the groups, which is called effect size. However, absolute effect sizes (absolute differences between two groups), do not account for the variability in the data. Therefore, one of the commonly used effect size calculation methods, called Cohen's *d* effect size, is estimated by dividing the mean differences between the two groups by the standard deviation of the data. This method transforms the absolute differences into standard deviation units. Here, we estimated Cohen's *d* effect sizes for comparison of the magnitude of the effect for different groups. Analysis of data was performed using the "lmer" function in the "lme4" package by Bates[14] in the R-Project for statistical computing.

The study was screened and deemed exempt from formal review by Texas Tech University Health Sciences Center Institutional Review Board (IRB) for the Protection of Human Subject and was also approved by the Chapman University IRB.

RESULTS

The likelihood of the five models used for description of data and the pairwise comparisons of the progressively more complex models are presented in Tables 2 and 3, respectively. As demonstrated in Table 2, addition of School, Test Type, and their interactions progressively decreased both AIC and BIC values (ie, an increase in likelihood), indicating that the full model (Model 5, Table 1) best describes the data. This conclusion was also confirmed by the pairwise comparison of progressively more complex models, presented in Table 3, as addition of each of the fixed effects to the model significantly improved

the model predictability of the observed data. Overall, based on both statistical methods, the full model (Equation 1) was chosen for description of our data.

The estimates of intercept and coefficients of the fixed effects and their variabilities (SE and coefficient of variation or CV) for School, Test Type, and School × Test type interaction, based on the full model (Model 5), are presented in Table 4. To provide numeric examples of Equation 1 and demonstrate how these estimates fit the observed data, the predicted grades for quizzes and exams for Texas Tech and Chapman students were calculated, which are listed below:

$$Predicted\ Chapman\ Quiz\ Grade = 74.7 + (-0.622 \times 0) + (7.14 \times 0) + (9.40 \times 0) + b_i = 74.7 + b_i$$

$$Predicted\ Texas\ Tech\ Quiz\ Grade = 74.7 + (-0.622 \times 1) + (7.14 \times 0) + (9.40 \times 0) + b_i = 74.1 + b_i$$

$$Predicted\ Chapman\ Exam\ Grade = 74.7 + (-0.622 \times 0) + (7.14 \times 1) + (9.40 \times 0) + b_i = 81.8 + b_i$$

$$Predicted\ Texas\ Tech\ Exam\ Grade = 74.7 + (-0.622 \times 1) + (7.14 \times 1) + (9.40 \times 1) + b_i = 90.6 + b_i$$

where b_i is the individual student's shift from the intercept, and the numbers 74.7, 74.1, 81.8, and 90.6 are the predicted average grades for Chapman quiz, Texas Tech Quiz, Chapman Exam, and Texas Tech exam, respectively. These predicted values are identical to the observed averages. The analysis also generated b_i values for individual students, which are not shown here.

The model-generated interaction plots, demonstrating mean ± SE values of grades, are demonstrated in Figure 1, and the pairwise comparisons of different groups and their associated effect sizes are presented in Table 5. There were no significant differences (p=1.00) between the two cohorts of students in term of their performance in 55 quiz questions (Figure 1 and Table 5); the mean ± SE of quiz scores were 74.1 ± 0.8 and 74.7 ± 1.3 for the Texas Tech and Chapman students, respectively. However, the grades of students in the 47 examination questions in the Chapman group (82.0 ± 1.1) were significantly (p<0.0001) lower than those in the Texas Tech group (90.6 ± 0.5) (Figure 1 and Table 5). This difference amounted to a Cohen's d effect size of 1.15 (Table 5). Additionally, there were significant differences between Quiz and Exam grades for both Texas Tech (effect size of 1.92) and Chapman (effect size of 0.704) groups (Table 5).

Figure 2 demonstrates plots of differences between the Texas Tech and Chapman students' grades in each of the 55 quiz (left panel) and 47 exam (right panel) questions. Whereas the differences in the quiz scores were randomly distributed around the line of zero (no statistical difference), a large majority of examination score differences were negative, indicating a significant ($p< 0.0001$) bias towards lower examination grades for the Chapman students.

DISCUSSION

Composite examination (CE) is a new assessment method to test student performances in individual courses using a single test administered at regular intervals, which contains questions from all the courses offered during that semester. Composite examination is different from, and does not replace, progress or milestone assessments,[1,2] which are normally administered at the end of each semester, academic year, and/or before students start their APPE. Additionally, CE is different from integrated assessments, where the assessment questions integrate different disciplines or courses together.[5] Although CE mixes questions from several courses, each question is specific to a particular course. One potential advantage of CE is that it allows more frequent testing during the semester by reducing the total number of individual examinations necessary for traditional, single course examinations. Educational research has shown that testing can decrease the normal memory decline and improve retention of newly-learned material.[15,16] Additionally, it has long been established that studying at spaced intervals, as opposed to cramming in one session, improves long-term retention or memory, when the total study time is the same.[17-19] Interestingly, one recent study[19] showed that although sleep restriction significantly reduced the recall of learned materials after cramming, it did not negatively affect retention of materials learned over spaced intervals. Therefore, compared with traditional, single course examinations, CE may improve learning and/or increase the retention of content in individual courses by potentially allowing more frequent testing and spaced learning.

Despite theoretical advantages of CE over single course examinations, there is no report in the literature investigating how CE might affect the performance of students in various courses that it covers. A very recent study[6] on the perception of students about the application of CE in a school of pharmacy reported that 41-44% of students thought CE increases knowledge retention, while only 12-19% of students believed that separate course examinations lead to greater content retention. Pre- and post-CE surveys also revealed a significant decrease in the number of students who described their study habit as cramming (29% before CE and 11% after CE). However, taking a CE examination on several subjects together, as opposed to a single course examination, may negatively affect the performance of students in

questions related to each course, a subject that has not yet been addressed. Indeed, our results (Figures 1 and 2 and Table 5) clearly show that CE has a substantial negative effect on the grades of students in the pharmacokinetics questions. The effect size for the difference between the Chapman (CE) and Texas Tech (individual course examination) students in terms of performance in the examination questions was 1.15, which means the performance of Chapman students in the examination questions was on average 1.15 standard deviations lower than that of Texas Tech students. Cohen[20] suggested that effect sizes of 0.2, 0.5 and 0.8 represent small, medium, and large effects, respectively. Therefore, an effect size of 1.15 is substantial and can have significant impact on the overall grades of students in the course, depending on the weight of CE in the overall course grades.

Comparison of groups also showed substantial effect sizes for the differences between the quiz and examination grades (Table 5), with higher grades achieved in the examinations for both Texas Tech (effect size of 1.92) and Chapman (0.704) students. This observation is consistent with the fact that quizzes were administered at the end of each class session and were related to the topic presented in the same day. Therefore, quizzes reflect the first exposure of the students to the topic. However, for examinations, students had additional opportunities to study the topics after the class session and before taking examinations. The much lower quiz-examination effect size for Chapman students (0.704), compared with that for the Texas Tech students (1.92), observed in our studies (Table 5), is due to the School x Type of Test interaction detected by our mixed-effects model (Tables 3 and 4), indicating poorer performance of Chapman students in the composite examinations.

An ideal study design to compare CE and course-specific examinations would divide the same student cohort into two subgroups, with identical treatments for both subgroups in terms of instruction but with one subgroup subjected to CE and the other to individual examinations. That type of study, however, is very difficult, if not impossible, to implement both logistically and academically. Instead, we used two student cohorts at two schools of pharmacy who were subjected to an identical instructional method in a Pharmacokinetics course taught by the same instructor and identical instructional resources, but with different examination methods. Because we used two different student cohorts, one may argue that the

lower performance of students in the CE at Chapman, compared with the performance of Texas Tech students in the individual course examinations, might be due to a difference in the academic abilities of the two student cohorts. However, our data clearly refute this argument because the performances of students in daily quizzes, which were administered in an identical fashion to both cohorts, were nearly identical for the two cohorts (74.1 ± 0.82 and 74.7 ± 1.26 for the Texas Tech and Chapman students, respectively). Furthermore, our mixed-effects model is capable of accounting for differences in the academic abilities of students between the two cohorts. This is determined by the School coefficient ($\boldsymbol{\beta_{School}}$), which was –0.622 (Table 4), indicating that the academic abilities of the Texas Tech students were marginally (0.622%) lower than those of Chapman students. If the differences in the quiz and examination grades were only due to differences in the academic abilities of the students, one would expect parallel slopes for the two cohorts with different starting points, as opposed to a lower slope for the Chapman students observed in Figure 1. Indeed, the significant (p<0.0001) interaction coefficient ($\boldsymbol{\beta_{School \times Type}}$) of 9.40% detected by our model (Table 4) is an indication of a lower performance of Chapman students in the composite examinations, regardless of possible differences between the two student cohorts in their academic abilities, which in this case was marginal.

The findings of our study need to be interpreted in the context of some limitations. First, our study had a relatively small sample size of the daily quiz (N=55) and examination (N=47) questions that were identical for both cohorts. Additionally, caution should be exercised when extrapolating our observation of the negative effects of CE on the performance of students in pharmacokinetics examination questions to other courses. This is because a major perceived advantage of CE is that by allowing more frequent testing, CE improves study habits,[6] which is expected to improve learning. However, this was not true for our study because the number of examinations at both institutions was the same for the Pharmacokinetics course. It is likely, though, that for most courses, CE would allow for more frequent testing, thus potentially mitigating the negative effects of CE on the performance of students observed in our studies. Because we started our new program at Chapman with implementing CE, we could not compare the frequency of testing before and after implementation of CE for other courses, which were

newly developed at Chapman. However, McDonough et al[6] reported that the frequency of testing for all of their courses was increased after implementation of CE. Therefore, although our results with the pharmacokinetics course cannot be directly extrapolated to other courses, one may expect that in the absence of more frequent testing, CE may have a negative effect on the performance of students.

Because of its theoretical and perceived advantages over the traditional, single-course examinations,[6] composite examination may be attractive to many schools or colleges of pharmacy that are in the process of revising their curricula, including their delivery and scheduling. However, these perceived advantages have yet to be documented using actual performance data in future studies. Major questions remaining to be answered are performance of students in other courses, which are included in the CE, and the actual effects of CE on the long-term retention of contents. The true advantage of CE over traditional examinations in terms of learning outcomes, aside from scheduling preferences, would be if it indeed increases the retention of learned material over time. In that case, even a reduction in the individual course grades, as observed here with Pharmacokinetics, may be justified because in the long-term, learning would be improved. However, a reduction in the course grade because of CE, if observed with other courses as well, may require adjustments to the current standards for passing individual courses or overall progression of students, which are established within the context of individual course examinations.

CONCLUSIONS

Performances of two cohorts of students in daily pharmacokinetics quizzes, which were administered in an identical manner to both cohorts, were identical. However, students' performances in the examinations were significantly lower in the cohort that received composite examinations as opposed to traditional, individual course examinations. These data indicate that when compared to the traditional, individual course examination, composite examination significantly reduces the grades of students in a Basic Pharmacokinetics course. Further studies are needed to determine whether these results may be extrapolated to other disciplines in pharmacy curricula.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Siu Fun Wong from the Chapman University School of Pharmacy for her assistance in scheduling and managing the composite examinations.

REFERENCES

1. Brady RL, Coker AO, Copeland JT, et al. The validation of a case-based, cumulative assessment and progressions examination. *Am J Pharm Educ.* 2016;80(1):Article 9.
2. Plaza CM. Progress examinations in pharmacy education. *Am J Pharm Educ.* 2007;71(4):Article 66.
3. Mehvar R, Supernaw RB. Outcome assessment in a PharmD program: The Texas Tech experience. *Am J Pharm Educ.* 2002;66(3):243-253.
4. Gortney JS, Bray BS, Salinitri FD. Implementation and use of the pharmacy curriculum outcomes assessment at US schools of pharmacy. *Am J Pharm Educ.* 2015;79(9):Article 137.
5. Medina MS, Britton ML, Letassy NA, Dennis V, Draugalis JR. Incremental development of an integrated assessment method for the professional curriculum. *Am J Pharm Educ.* 2013;77(6):Article 122.
6. McDonough SL, Alford EL, Finks SW, Parker RB, Chisholm-Burns MA, Phelps SJ. Student pharmacists' perceptions of a composite examination in their first professional year. *Am J Pharm Educ.* 2016;80(1):Article 4.
7. Mehvar R. Development and evaluation of a quasi problem-based, objective-driven learning strategy in introductory and clinical pharmacokinetics courses. *J Pharm Teaching.* 1999;7(1):17-29.
8. Mehvar R. A participation requirement to engage students in a pharmacokinetics course synchronously taught at a local and distant campus. *Am J Pharm Educ.* 2010;74(7):Article 118.
9. Mehvar R. The importance of active learning and practice on the students' mastery of pharmacokinetic calculations for the intermittent intravenous infusion dosing of antibiotics. *BMC Med Educ.* 2012;12(1):116.
10. Mehvar R. Effects of simulations on the learning of pharmacokinetic concepts. *Curr Pharm Teach Learn.* 2012;4(4):278-284.
11. Mehvar R. On-line, individualized, and interactive pharmacokinetic scenarios with immediate grading and feedback and potential for use by multiple instructors. *Am J Pharm Educ.* 1999;63:348-353.
12. Mehvar R. Creation of a dynamic question database for pharmacokinetics. *Am J Pharm Educ.* 2000;64(4):441-445.
13. Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. *J Grad Med Educ.* 2012;4(3):279-282.
14. Bates DM. Ime4: Mixed-Effects Modeling with R. New York: Springer; 2010.
15. Bangert-Drowns RL, Kulik JA, Kulik C-LC. Effects of frequent classroom testing. *J Educ Res.* 1991;85:89-99.
16. Roediger HL, 3rd, Karpicke JD. The Power of testing memory: Basic research and implications for educational practice. *Perspect Psychol Sci.* 2006;1(3):181-210.
17. Putnam AL, Sungkhasettee VW, Roediger HL, 3rd. Optimizing learning in college: Tips from cognitive psychology. *Perspect Psychol Sci.* 2016;11(5):652-660.
18. Carpenter SK, Cepeda NJ, Rohrer D, Kang SHK, Pashler H. Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educ Psychol Rev.* 2012;24(3):369-378.
19. Huang S, Deshpande A, Yeo SC, Lo JC, Chee MW, Gooley JJ. Sleep restriction impairs vocabulary learning when adolescents cram for exams: The need for sleep study. *Sleep.* 2016;39(9):1681-1690.
20. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed: Routledge; 1988.

Table 1. Characteristics of Statistical Models Used for Analysis of Data

Model #	Intercept	School	Test Type	School × Test Type Interaction
1	+	–	–	–
2	+	+	–	–
3	+	–	+	–
4	+	+	+	–
5	+	+	+	+

Table 2. Best-Fit Model Based on Likelihood Ratios

Model	Degree of Freedom	AIC	BIC
1 (Intercept Only)	3	3735	3747
2 (Plus School)	4	3724	3741
3 (Plus Test Type)	4	3490	3506
4 (Plus School & Test Type)	5	3478	3499
5 (Full Model)	6	3430	3455

Table 3. Best-Fit Model Based on Pairwise Comparison of Progressively More Complex Models Using Pearson's Chi-Squared Analysis

Model Comparisons	Difference	Chi Square	Degree of Freedom	*p* Value
2 vs 1	School	12.2	1	<0.001
3 vs 1	Test Type	247	1	<0.0001
4 vs 2	Test Type	248	1	<0.0001
4 vs 3	School	13.3	1	<0.001
5 vs 4	Interaction	50.9	1	<0.0001

Table 4. Estimates of Intercept and Coefficients of the Fixed Effects for the Full Model (Model 5)

Parameter	Estimate	SE	CV (%)
Intercept ($\boldsymbol{\beta_0}$)	74.7	1.04	1.39
School ($\boldsymbol{\beta_{School}}$)	–0.622	1.27	1.71
Test Type ($\boldsymbol{\beta_{Type}}$)	7.14	1.02	1.25
Interaction ($\boldsymbol{\beta_{School\times Type}}$)	9.40	1.24	1.37

Table 5. Cohen's *d* Effect Sizes and Post-Hoc, Pairwise Comparisons (Two-Tailed t Test) of Different Groups

Pairwise Comparison	Cohen's *d* Effect Size (95% Confidence Intervals)	*p* Value[a]
Chapman Quizzes vs Chapman Exams	0.704 (0.380 – 1.03)	<0.0001
Texas Tech Quizzes vs Texas Tech Exams	1.92 (1.66 – 2.19)	<0.0001
Chapman Quizzes vs Texas Tech Quizzes	0.0600 (-0.210 – 0.330)	1.00
Chapman Exams vs Texas Tech Exams	1.15 (0.86 – 1.44)	<0.0001

[a] The *p* values are Bonferroni adjusted.

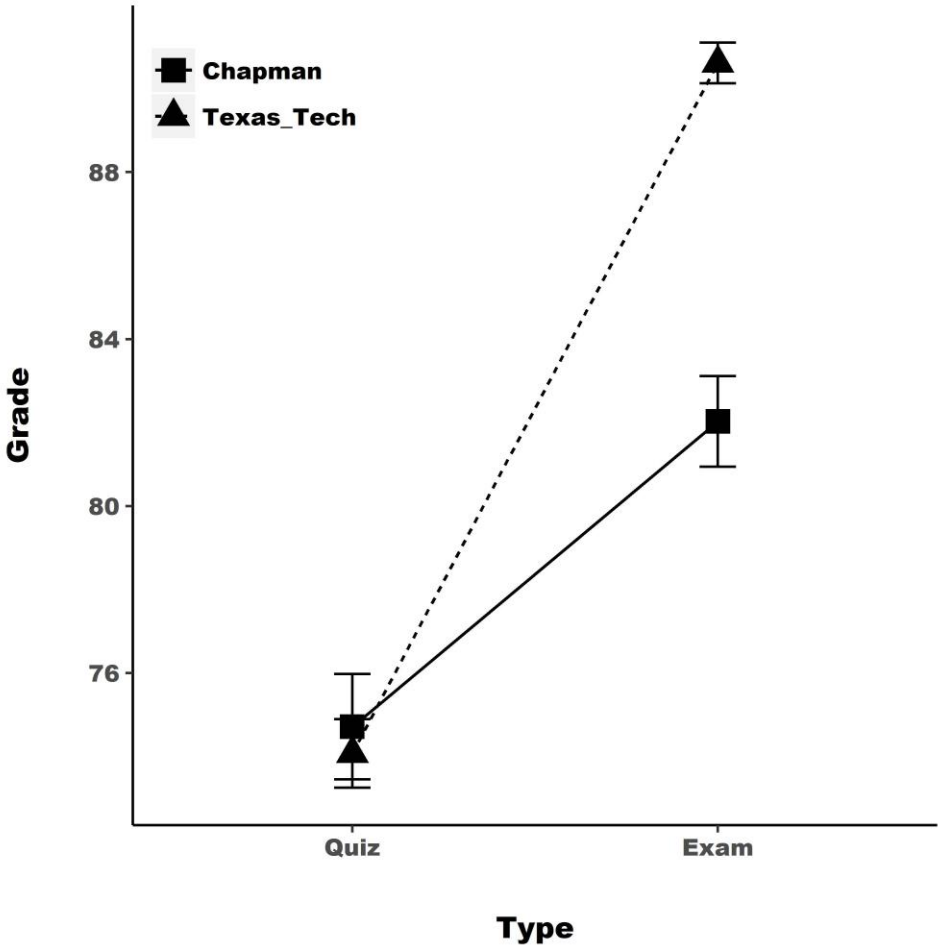

Figure 1. Grades of Chapman and Texas Tech Students in Quiz and Examination (Exam) Questions. Mean (± SE) values are presented for grades of 162 Texas Tech students and 79 Chapman students in 55 quiz and 47 examination questions administered throughout the semester. Whereas the quiz questions were administered in an identical manner to both student cohorts, the examinations were administered differently (single course examination for Texas Tech students versus multiple course, composite examination for Chapman students).

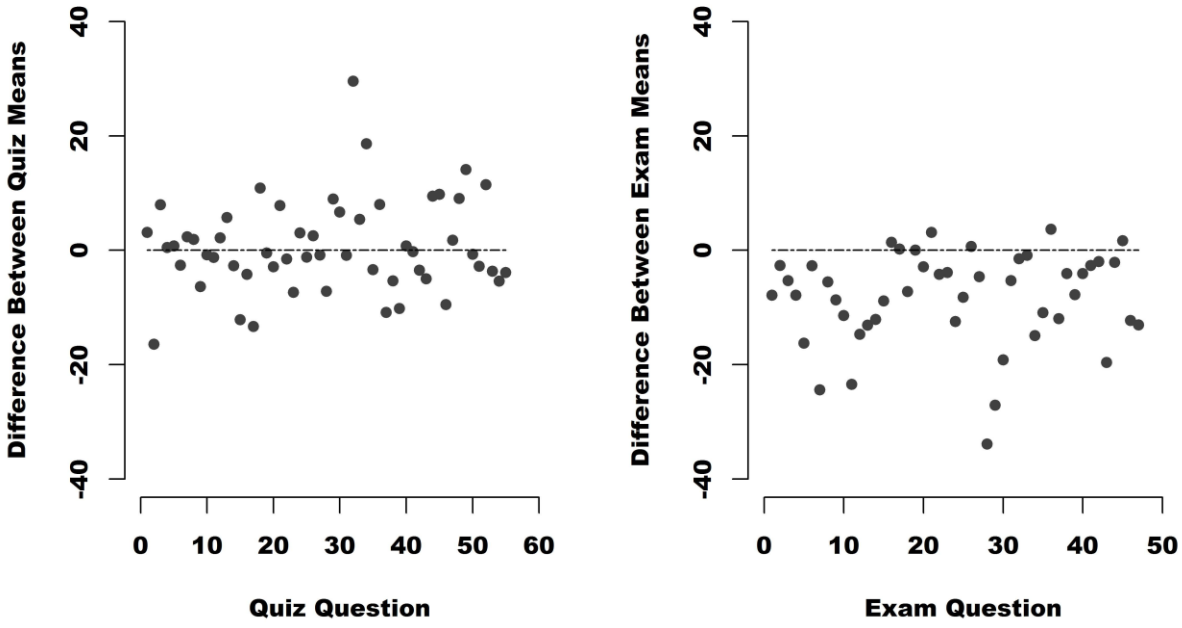

Figure 2. The differences between the Chapman and Texas Tech Students in Their Grades for Quiz (Left Panel) and Examination (Right Panel) Questions. Grades of Texas Tech students were subtracted from those of Chapman students for each of 55 quiz (left panel) and 47 examination (right panel) questions. Each symbol represents one question, and the horizontal line indicates the line of no difference in grades between the two student cohorts.