The Effect of Distributed Practice in Undergraduate Statistics Homework Sets: A Randomized Trial

Bryan R. Crissinger
University of Delaware

Journal of Statistics Education Volume 23, Number 3 (2015),
www.amstat.org/publications/jse/v23n3/crissinger.pdf

Key Words: Massed practice; Cognitive theory; Interaction; Hawthorne effect

Abstract

Most homework sets in statistics courses are constructed so that students concentrate or "mass" their practice on a certain topic in one problem set. Distributed practice homework sets include review problems in each set so that practice on a topic is distributed across problem sets. There is a body of research that points to the efficacy of distributed practice for developing a variety of skills from word recall to surgical techniques. A trial was conducted in several sections of a business statistics course where students were randomly assigned to either have massed practice homework sets or distributed practice homework sets. The two groups were then compared on the course assessments. The results show some evidence for the efficacy of distributed practice homework sets, although this effect may be modified significantly by the instructor or by a Hawthorne effect.

1. Introduction

A recent report by the National Research Council (2013) cites "exponential increases in the amount of data" and the need for undergraduates and graduates to be able to "make inferences about the world from data." The report also raises the question of whether calculus is an appropriate prerequisite course for many STEM undergraduates. Might we read between the lines to infer the need for higher levels of statistical ability in more students going forward?

One of the most effective methods in the author's experience as a student in math classes was the use of distributed practice in the homework sets. Each homework set had some problems from

that day's lesson and some review problems from previous lessons. In this way practice on a topic was distributed across homework sets. That experience helped to motivate this study.

Careful thinking about all types of assessments in statistics courses is the subject of recommendation 6 of the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) college report (American Statistical Association 2005). The report's guidelines around assessment include coordinating assessments with class activities, using a variety of assessments, and designing instruments that assess students' understanding of statistical ideas and concepts, not just those that focus on procedures and computation. Homework is shown first in a list of assessment types, but little guidance is given as to whether homework assignments should regularly include review problems.

Problem sets in statistics textbooks tend to consist of problems taken from the same section, covering the same content. For example, in the textbook used in the author's course (McClave, Benson, and Sincich 2014), the chapter on probability includes a section on finding the expected value and standard deviation for a discrete random variable and a section on the normal distribution. None of the problems from the section on the normal distribution require students to revisit the standard deviation for a discrete random variable. Only in the chapter review homework set will students again see problems about that topic. By the time of the final exam, students may not have worked a problem requiring expected value and standard deviation for discrete random variables in over a month. The author has found no statistics textbooks which consistently deliver review problems in each homework set.

In general, we can define distributed practice (DP) of a skill as activities which develop the skill spread across practice sessions. Massed practice (MP) of a skill is defined as a set of activities which develop the skill performed in one practice session (Rohrer and Taylor 2006). In the context of homework sets, DP homework sets include problems on both new and old topics so that practice on a topic is spread across several homework sets. MP homework sets include only problems on new topics so that practice on the topic occurs in only one homework set.

Two theories as to why DP on a skill may be superior to MP include the "forgetting mechanism" and memory consolidation during sleep. The forgetting mechanism refers to the effect of the spacing of practice sessions with sufficient time between them so that "full processing" of the material must be repeated each time practice on the skill is done (Krug, Davis, and Glover 1990). Since there may be more cognitive engagement each time, it is thought that perhaps learning is more efficient and retention is increased. The same theory is also proposed in the context of learning motor skills in dance (Batson and Schwartz 2007). MP, on the other hand, may deceptively reinforce the notion that since successive practices become a bit easier, that the skill is being mastered, which may lead to "attention attenuation" (Dempster 1991; Kornell, Castel, Eich, and Bjork 2010). In a homework set context, students do a few similar problems, which they are able to do at the time, and believe they understand the material which discounts to them the value in concentrating fully on the rest of the homework set. The student develops a false sense of confidence in her abilities as she finds she is able to correctly work several problems on the same topic in one sitting, and thus concludes she has mastered the content. The student reasons that, "Since I remember it so well, why pay more attention to it?"

The forgetting theory then would seem to imply that MP might allow short-term retention but not good long-term retention (Fishman, Keller, and Atkinson 1968; Willingham 2002). In one sense, MP is a form of cramming and perhaps MP homework sets reinforce cramming for tests since the "do it all at once" approach seems to work just fine on homework sets (Dempster 1991). William James advocated against cramming and for DP over a century ago:

> "You now see why 'cramming' must be so poor a mode of study. Cramming seeks to stamp things in by intense application immediately before the ordeal. But a thing thus learned can form but few associations. On the other hand, the same thing recurring on different days, in different contexts, read, recited on, referred to again and again, related to other things and reviewed, gets well wrought into the mental structure. This is the reason why you should enforce on your pupils habits of continuous application." (James 1901, p. 129)

DP in homework is designed to enforce these habits of continuous application. "Spaced repetitions, [compared to massed repetitions], are likely to encourage exactly the kinds of constructive mental processes, founded on effort and concentration, that teachers hope to foster" (Dempster 1991).

There is a body of research supporting the theory that the spacing of practice in a DP approach allows for rest and for memory consolidation during sleep. Fischer, Hallschmid, Elsner, and Born (2002) demonstrated a benefit of sleep on motor skills. In a study of piano performance, significant accuracy gains occurred when practice sessions were separated by 24 hours (spanning one sleep cycle) but not for shorter intervals between practices (Simmons 2012). Both Gais, Plihal, Wagner, and Born (2000) and Karni, Tanne, Rubenstein, Askenasy, and Sagi (1994) showed that sleep increased performance on a task requiring recall of a visual image shown to subjects for only a fraction of a second. Walker (2005) gives an extensive summary of the cognitive theory and research on sleep and procedural memory, the kind of long-term memory responsible for knowing how to do things like play a piano, ride a bike, balance a checkbook, or work a statistics problem. He states that, "…evidence for the reliance of procedural memory on sleep in humans has been incredibly robust…" (Walker 2005, p. 54).

DP problem sets then, which allow time for students to forget how to do problems encountered earlier in the course, require fuller cognitive engagement for many of those review problems, and allow time for the effect of sleep on procedural memory, may be more effective than MP problem sets in developing the skills we want students to develop. The purpose of this study was to determine if DP homework sets would be more effective than MP homework sets with respect to learning outcomes in an introductory statistics course. Rohrer and Taylor (2006) suggest a study design much like this one.

2. Literature Review

There is a rich body of literature on the comparison between DP and MP methods in a wide variety of disciplines, but very few studies in the context of a statistics course. There are no relevant studies reported in the *Journal of Statistics Education* (1993 to March 2014), the *Statistics Education Research Journal* (May 2002 to May 2014), and *Technology Innovations in*

Statistics Education (2007 to May 2014). There were also no relevant hits on the search terms "homework," "distributed practice," "cumulative practice," or "interleaving" in *Teaching Statistics* (1979 to May 2014).

Much of the literature is comprised of studies done by behavioral scientists and education researchers. The following list gives examples of studies finding evidence for DP over MP: elementary math (Agodini, Harris, Thomas, Murphy, and Gallagher 2010; Good and Grouws 1979), junior high general science (Reynolds and Glaser 1964), college algebra (MacDonald 1984), introductory statistics (Bude, Imbos, Wiel, and Berger 2011; Smith and Rothkopf 1984), word recall (Cull 2000; Rohrer and Taylor 2006), piano performance (Duke and Davis 2006; Simmons 2012), high school physics (Grote 1995), reading comprehension (Krug et al. 1990), spelling (Fishman et al. 1968), laparoscopic technique (Mackay, Morgan, Datta, Chang, and Darzi 2002), eighth-grade algebra (Holdan 1985), and bird classification (Wahlheim, Dunlosky, and Jacoby 2011). To be fair, there are some studies that did not find a significant difference between DP and MP (Horine 1983; Mayfield and Chase 2002) and one study of French pronunciation skills that found that MP is superior to DP (Carpenter and Mueller 2013). This list is not based on an exhaustive literature review, however, and so there may be some studies on DP in statistics that are not listed here.

These studies are examples of ones with good design features. They employ either random assignment to treatments, matching, or covariate adjustment to control for confounding variables. If randomization is used, the data are analyzed at the same level as the randomization. For example, if randomization is done at the class level, the data analysis is also performed at the class level, not the student level. Using the student as the unit of analysis when classes are randomly assigned may underestimate, perhaps significantly, the variance estimates of effects (Bloom 2005, p. 125) and therefore lead to misleading conclusions regarding efficacy.

3. Study Design and Methods

Many studies of the efficacy of educational interventions are observational and not well-designed to address the question of whether the intervention caused a change in outcomes. This study employs randomization as a control for other potential explanations for differences in outcomes between study groups. Those of us who teach statistical methods impress upon our students the need for random assignment to groups, where possible, in order to allow for cause-effect conclusions. In clinical trials of a new drug, subjects are expected to be randomly assigned to the study groups: "Ordinarily, in a concurrently controlled study, assignment is by randomization, with or without stratification" (21 CFR 314.126 b4 2014). The rationale is that

> [b]ecause the groups do not differ systematically from one another at the outset of the experiment, any differences between them that subsequently arise can be attributed to the intervention or treatment rather than to preexisting differences between the groups. Random assignment also provides a means of rigorously determining the likelihood that subsequent differences could have occurred by chance rather than because of differences in treatment assignment (Bloom 2005, p. 1).

Students were recruited for this study from three sections of an introductory business statistics course at a large east-coast research university in the fall of 2014. The course is required for many majors in the College of Business and Economics but also satisfies requirements for many other non-STEM majors. The textbook used is the business statistics text by McClave et al. (2014). This study has been granted exempt status by the university's Institutional Review Board.

Out of 309 students who had initially enrolled in the three sections, 241 (78%) consented to participate. Informed consent was obtained in writing with a form (see Appendix) which was distributed on the first day of class. The study was explained to the class and the informed consent form was read aloud which included language that let students know that the distributed practice treatment was hypothesized to be more effective. Incentive to participate was provided in the form of a 2% extra credit bonus. Students who did not wish to participate could also earn the 2% bonus by scoring at least 8 questions correct on the 20-item Goals and Outcomes Associated with Learning Statistics (GOALS) assessment of learning outcomes in a first statistics course administered at the end of the course (Sabbag, Garfield, and Zieffler 2015). The somewhat arbitrary threshold of at least 8 questions correct was intended to require students to take the assessment somewhat seriously but also to make allowance for some material assessed by the instrument that was not covered in the course.

Within one week from the first day of class, participants were randomly assigned to either the DP group or the MP group. Randomization was implemented using SAS/STAT® software. Participants' names were sorted according to a random number having a uniform distribution on [0, 1]. The top half of the sorted list was assigned to one group and the bottom half of the list was assigned to the other group. The names of each group (MP or DP) were also determined at random.

Each homework set in the DP group consisted of 10 questions; the first 5 questions were taken from the current section and the last 5 questions were taken from previous sections. Some of the review problems were problems students had not seen before and some were recycled from previous homework sets. Each homework set in the MP group consisted of 10 questions taken from the current section. All of the DP problems also appeared in the MP problem sets but there were some MP problems which did not appear in the DP problem sets.

A total of 27 homework sets were assigned throughout the course covering topics from summarizing and collecting data, probability, and one-sample inference. All problems in the MP group were new problems but some problems in the DP group homework sets were repeated in later homework sets. Table 1 shows the frequencies with which problems in the MP group sets were used in the DP group sets. About 79% of the MP problems appeared exactly once in the DP problem sets. Twenty-six problems appeared more than once and 31 problems appeared only in the MP problem sets. The 31 MP-only problems were those in later problem sets where there was insufficient time to assign them in later DP problem sets. For example, there were five problems in the last MP problem set alone which could not appear in any DP problem sets. However, overall, the two sets of 270 problems completed by each group by the end of the course were quite similar: 88.5% of the MP problems were also completed by students in the DP group.

Table 1. Frequencies of MP problems used in the DP problem sets.

Number of times used	Frequency	Percent
0	31	11.5%
1	213	78.9%
2	21	7.8%
3	5	1.9%

Review problems for the DP problem sets were chosen from earlier MP problem sets. Table 2 shows how topics were distributed across subsequent DP problem sets. For example, 5 problems from the chapter 1 MP problem set were repeated in the next DP problem set (2.1) and 3 (some assigned for the second time) were repeated in problem sets which were more than 10 problems sets after chapter 1 (e.g., 4.3 or later).

Table 2. Number and distribution of MP problems used in later DP problem sets.

	Distance to next DP problem set						
Problem Set	Next	2	3	4 – 5	6 – 10	> 10	Total
Ch. 1: Data and statistical thinking	5	2	2	1	2	3	15
2.1: Describing qualitative data	3	2	1	1	2	1	10
2.2: Graphs for qualitative data	1	2	2	1	2	1	9
2.3: Measures of center	1	1	1	1	1	1	6
2.4: Measures of variability	1	1	1	1	2	1	7
2.5: Empirical rule, Chebyshev's rule	1	1	1	1	2	1	7
2.6: Percentiles and z-scores	1	1	1	1	1	1	6
2.7: Boxplots and detecting outliers	1	1	1	1	1	1	6
2.10: Graphical distortions	1	0	0	1	1	1	4
4.1: Two types of random variables	1	1	0	1	1	1	5
4.2: Discrete distributions	1	1	1	0	1	1	5
4.3: Binomial distribution	1	1	1	1	1	0	5
4.6: Normal distribution	1	1	1	1	1	0	5
4.7: Assessing normality	1	0	1	0	2	0	4
5.1: Sampling distributions	1	1	1	0	2	0	5
5.3: Distribution of sample mean	1	0	1	1	1	0	4
5.4: Distribution of sample proportion	0	1	1	0	2		4
6.2: z-interval for μ	0	1	0	1	2		4
6.3: t-interval for μ	1	1	0	1	1		4
6.4: z-interval for p	1	1	0	1	0		3
6.5: Determining sample size	0	1	0	1	0		2
7.2: Setting up hypotheses	1	1	0	0			2
7.3: P-values	1	1	1	0			3
7.4: z-test for μ	1	1	0				2
7.5: t-test for μ	1	1					2
7.6: z-test for p	1						1
7.7: Chi-square test for σ^2							0

An investigation into the balance of the two groups at randomization on many demographic characteristics is shown in Table 3 and Table 4. P-values in Table 3 were computed using either the chi-square test or Fisher's Exact Test (denoted with asterisks). Fisher's Exact Test was used when at least 25% of the expected cell counts were less than 5. Comparisons in Table 4 were conducted using Welch's t-test and Satterthwaite's degrees of freedom. For those students who had taken at least one math course at the university, the highest level math course was observed as well as the grade in that course.

In order to eliminate any instructor effects on the outcomes, the protocol was written to exclude from participation any students not registered in one of the sections of the course taught by the author. Two weeks after randomization, however, the author's department reassigned one of the three sections to another instructor. Therefore, results are shown two ways: an intent-to-treat analysis which uses data on all 241 participants who were randomized (Gupta 2011) and a secondary analysis which uses data on only the 156 students enrolled in the two sections for which the author continued as instructor.

Table 3. Categorical characteristics of participants at randomization.

	Intent-to-Treat Analysis			Secondary Analysis		
	MP (n=120)	DP (n=121)	p-value	MP (n=78)	DP (n=78)	p-value
Gender						
Female	38%	47%	0.1686	41%	47%	0.4202
Male	62%	53%		59%	53%	
Class						
Freshman	4%	13%	0.0865	6%	19%	0.1177
Sophomore	73%	68%		74%	65%	
Junior	18%	13%		15%	12%	
Senior	6%	6%		4%	4%	
College						
Arts/Sciences	12%	15%	0.5972	13%	15%	0.4640
Bus/Econ	66%	67%		64%	69%	
Other	23%	18%		23%	15%	
STEM						
Yes	11%	11%	0.9821	12%	13%	0.8066
No	89%	89%		88%	87%	
Ethnicity						
Asian	7%	5%	0.2194	8%	6%	0.5501*
Black	8%	3%		7%	3%	
Hispanic	11%	9%		12%	11%	
White	70%	82%		67%	78%	
Other	3%	1%		5%	2%	
Highest Math						
College Algebra	5%	2%	0.3078*	3%	0%	0.1749*
Pre-Calculus	11%	6%		12%	4%	
Calculus I	77%	78%		75%	78%	

Calculus II	5%	10%		6%	10%	
Calculus III	3%	5%		3%	7%	
Grade Highest Math						
A	18%	24%	0.7484	22%	28%	0.1869*
B	32%	33%		31%	35%	
C	28%	22%		29%	12%	
D	7%	10%		5%	10%	
F	3%	2%		3%	3%	
Other	13%	10%		11%	13%	
Repeat Course						
Yes	5%	3%	0.5392*	1%	1%	1.0000*
No	95%	97%		99%	99%	
Teaching Assistant						
1	35%	35%	0.6863	26%	24%	0.9455
2	20%	17%		17%	19%	
3	29%	35%		34%	36%	
4	16%	13%		23%	20%	
Instructor						
A	65%	64%	0.9305			
B	35%	36%				

Table 4. Numeric characteristics of participants at randomization.

	Intent-to-Treat Analysis			Secondary Analysis		
	MP (n=120) mean (SD)	DP (n=121) mean (SD)	p-value	MP (n=78) mean (SD)	DP (n=78) mean (SD)	p-value
Age	19.8 (0.9)	19.8 (0.9)	0.9113	19.7 (0.8)	19.8 (1.0)	0.6258
SAT (Math)	611.6 (74.0)	625.0 (64.2)	0.1447	616.1 (82.3)	630.3 (67.2)	0.2616
SAT (Verbal)	588.3 (84.7)	578.6 (81.7)	0.3852	587.2 (88.5)	579.3 (78.9)	0.5759
SAT (Writing)	579.0 (80.6)	583.1 (80.1)	0.6988	579.2 (79.3)	586.6 (81.7)	0.5881
GPA	2.98 (0.51)	3.09 (0.53)	0.1047	2.99 (0.55)	3.12 (0.57)	0.1572

Homework sets were delivered and graded using MathXL®, the homework engine developed by Pearson Education, Inc., which is the companion online homework system for the course textbook. Students' overall homework average was counted as 5% of their course grade.

4. Dropouts

From the time of randomization in the first week of class to the final exam, there were 20 students who dropped out of the study. Five dropped out of the course shortly after randomization, 8 changed their status to audit during the semester, 4 officially withdrew from the course, and 3 withdrew unofficially. Unofficial withdrawals are defined as students who were enrolled in the course but did not complete many assignments, including the final exam. A comparison of the numbers of dropouts is shown in Table 5; the p-values were computed using Fisher's Exact Test.

Table 5. Dropouts, auditors, and withdrawals by group.

	Intent-to-Treat Analysis			Secondary Analysis		
	MP (n=120)	DP (n=121)	p-value	MP (n=78)	DP (n=78)	p-value
Early Dropouts	1	4	0.3698	1	4	0.3669
Auditors	4	4	1.0000	2	2	1.0000
Official Withdrawals	3	1	0.3698	3	0	0.2452
Unofficial Withdrawals	2	1	0.6219	2	1	1.0000
Total	10	10	1.0000	8	7	1.0000

5. Outcome Measures

The primary outcome measure was the score on the instructor-developed final exam for the course consisting of 21 multiple choice questions each worth 4 points and a three-part free response question worth 16 points, graded according to a rubric. The final exam was cumulative but weighted more heavily on the material covered after the second exam (sample size determination for parameter estimation through hypothesis testing). Sixty percent of the exam covered inference topics, 16% covered probability topics, and 24% covered descriptive statistics and data collection topics. Table 6 shows the percent correct on each of the 21 multiple choice questions in the group of study participants.

Table 6. Final exam multiple choice item difficulty.

Item #	Item Description	% Correct
1	Identify the hypotheses for a test for one proportion	67.9
2	Interpret the p-value for a test for one proportion	45.7
3	Compute a binomial probability	47.1
4	Compute a p-value for a test for one mean	80.1
5	Use z-scores to compare values in two distributions	90.5
6	Identify the parameter symbol in a test for one mean	68.3
7	Compute the z-statistic for a test for one mean	78.3
8	Identify lack of random sampling as a flaw in a study	94.6
9	Find a percentile in a normal distribution	85.1
10	Compute the sample size for estimating a proportion	73.8
11	Identify the symbol that does not have a sampling distribution	20.8
12	Interpret the endpoints of a confidence interval for a mean	85.1
13	Compute the probability for an event involving a sample mean	48.0
14	Identify the type of variable given a summary of categorical data	72.4
15	Find the point estimate of a population proportion	71.0
16	Make the correct decision in a test for one proportion using the p-value	67.9
17	Choose which one of two histograms shows more variation	91.9
18	Choose correct p-value given three different computer outputs	91.9
19	Identify the characteristic of a confidence interval related to a test result	40.3
20	Identify the false statement concerning the inner fences of a boxplot	48.9
21	Recognize a rejection of a true null hypothesis as a Type I error	86.0

Other outcomes observed were scores on the two instructor-developed midterm exams. Exam 1 was multiple choice in format and covered topics in descriptive statistics and some probability. Exam 2 covered topics from the binomial, normal, and sampling distributions through parameter estimation and had a 66-point multiple choice section and a 34-point free response section, graded according to a rubric.

Students met once a week for a 50-minute computer lab taught by a teaching assistant who led students through data analysis or probability activities, depending on the current topic in lecture. The data analysis activities were designed to introduce students to data analysis in Minitab® and Excel®. These activities were delivered through the online course management system and were very low-stakes; students had immediate feedback after submission and had unlimited submissions before the due date. Five summative, higher-stakes assessments of students' software skills, called "data assignments," were closely based on the lab activities but assigned students data sets with algorithmically-generated values so that students' answer keys were different.

Homework scores on the MathXL® homework sets were also observed.

All students enrolled in the course (regardless of instructor) took the same exams and completed the same lab activities and data assignments. In addition to the algorithmically-generated data assignments, several of the free response questions on the exams used different numbers as well.

6. Results

Table 7 shows comparisons of the treatment groups on each of the outcome measures. Comparisons were conducted using Welch's t-test and Satterthwaite's degrees of freedom. While mean scores on all of the outcome variables were higher in the DP group than in the MP group, the difference was not statistically significant on the final exam scores ($p = 0.2638$). The only clearly significant difference was on exam 1 ($p = 0.0005$). Significant differences at the 0.10 level were observed on the data assignments and homework while the difference on exam 2 closely approached the 0.10 level.

These patterns were largely similar for the ITT analysis and secondary analysis with the exception of the difference on the final exam scores; a marginally significant result was observed on the final exam in the secondary analysis, consisting of only students enrolled in the two sections of the course for which the author ultimately had responsibility.

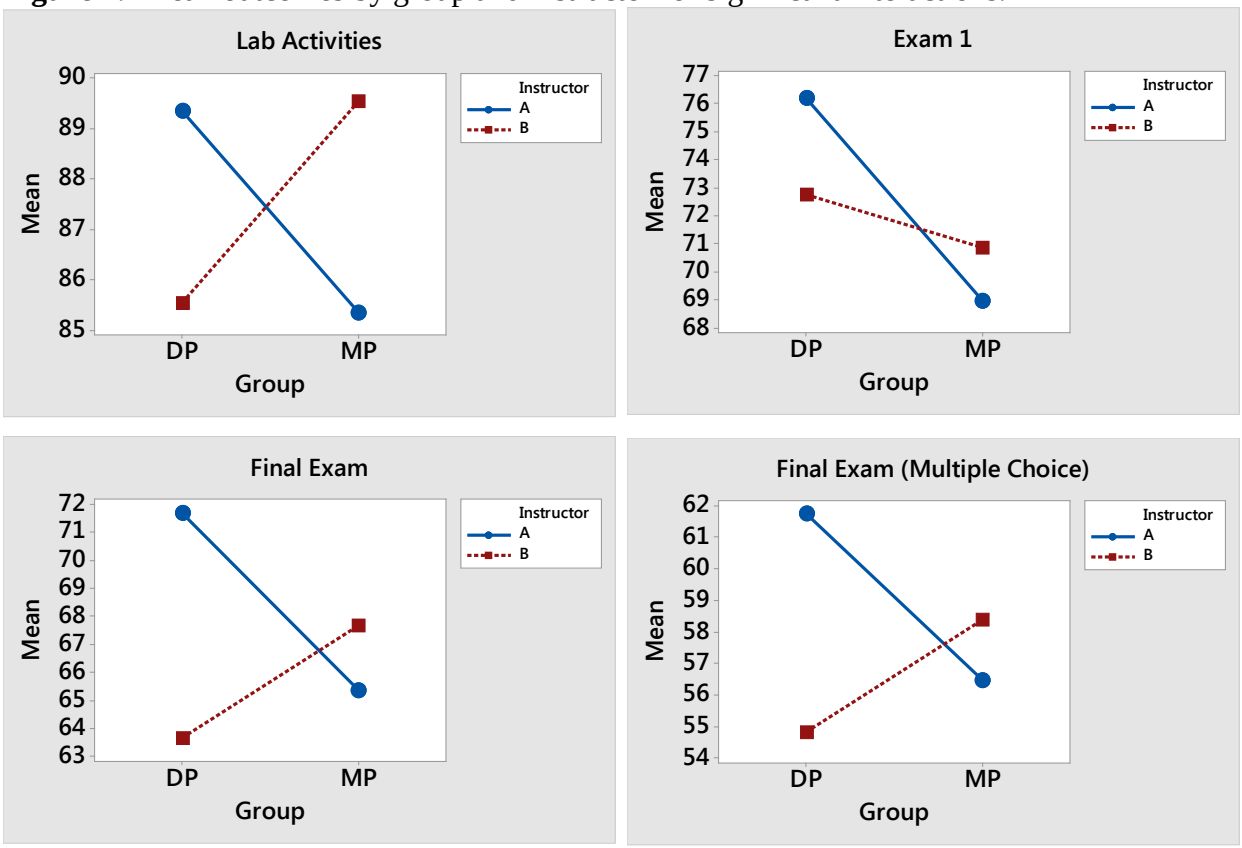
Table 7. Comparison of outcome measures.

	Intent-to-Treat Analysis			Secondary Analysis		
	MP (n=120) mean (SD)	DP (n=121) mean (SD)	p-value	MP (n=78) mean (SD)	DP (n=78) mean (SD)	p-value
Final Exam	66.2 (17.5)	68.8 (16.8)	0.2638*	65.4 (18.8)	71.7 (16.1)	0.0335
Multiple Choice	57.2 (14.7)	59.2 (14.0)	0.2818*	56.5 (15.6)	61.7 (13.3)	0.0323
Free Response	9.0 (3.7)	9.5 (3.7)	0.3122	8.9 (4.0)	9.9 (3.8)	0.1123
Exam 2	61.9 (17.2)	65.5 (15.3)	0.1004	61.9 (17.9)	66.8 (15.6)	0.0830
Multiple Choice	43.6 (11.4)	46.0 (10.0)	0.0876	43.5 (12.1)	46.7 (10.1)	0.0895
Free Response	18.4 (7.3)	19.5 (7.3)	0.2788	18.4 (7.4)	20.1 (7.7)	0.1756
Exam 1	69.6 (12.0)	75.0 (11.2)	0.0005+	68.9 (12.4)	76.2 (11.1)	0.0002
Data Assignments	70.8 (21.0)	77.1 (19.3)	0.0172	70.2 (21.8)	77.9 (19.4)	0.0232
Lab Activities	86.8 (17.2)	87.9 (16.2)	0.6065*	85.3 (19.5)	89.4 (14.5)	0.1539
Homework	82.0 (19.1)	86.4 (16.0)	0.0553	80.3 (21.1)	86.5 (15.4)	0.0388

* Significant ($p<0.10$) group x instructor interaction (disorderly)
+ Significant ($p=0.0910$) group x instructor interaction (orderly)

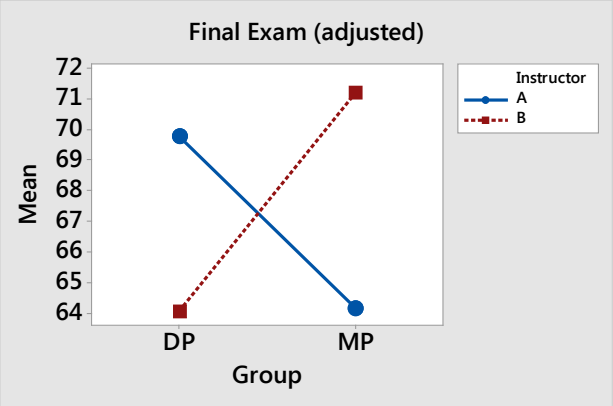
The intent-to-treat analysis yielded several group-by-instructor interactions. Interaction effects were analyzed using 2-treatment-group by 2-instructor ANOVA F-tests. The four interactions that were significant at the 0.10 level are highlighted in Table 7 and illustrated in Figure 1. In three of these interactions, the direction of the effect was different for each instructor (DP better than MP for one instructor, MP better than DP for the other instructor). They are labeled as disorderly and indicated with asterisks (*) next to the main effect p-values in Table 7. For one of the instructors, the DP group performed 6.3 points better on the final exam, on average, than the MP group, but for the other instructor, the MP group performed 4.1 points better than the DP group ($F_{1,217}$=4.77, p=0.0300). For one of the instructors, the DP group performed 5.2 points better on the multiple choice portion of the final exam, on average, than the MP group, but for the other instructor, the MP group performed 3.6 points better than the DP group ($F_{1,217}$=5.02, p=0.0261). For one of the instructors, the DP group performed 4.1 points better on the lab activities, on average, than the MP group, but for the other instructor, the MP group performed 4.1 points better than the DP group ($F_{1,232}$=3.16, p=0.0766). For the other interaction, the direction of the effect was the same for both instructors (DP better than MP) but the magnitude of the effect was different. This interaction is labeled as orderly and indicated with a plus sign (+) next to the main effect p-value in Table 7. On exam 1, the DP group performed better than the MP group for both instructors but by significantly different margins: 7.3 points better, on average, for one instructor and only 1.9 points better, on average, for the other instructor ($F_{1,230}$=2.88, p=0.0910).

Figure 1. Mean outcomes by group and instructor for significant interactions.



A 2-treatment-group by 2-instructor ANCOVA on the final exam score which adjusted for GPA and Math SAT score yielded an even more highly significant group x instructor interaction compared to the unadjusted analysis. The interaction is illustrated in Figure 2. The DP group scored 5.6 points better, on average, than the MP group for one instructor and the MP scored 7.2 points better than the DP group for the other instructor ($F_{1,202}$=12.75, p = 0.0004). The overall treatment effect averaged across the two instructors was negligible (F1,202=0.19, p = 0.6665) which is not surprising given the significant disorderly interaction between group and instructor.

Figure 2. Mean final exam by group and instructor adjusted for GPA and SAT (Math).



7. Discussion

Randomization left the two study groups fairly well-balanced on many characteristics prior to implementing the homework treatments. The only exceptions may have been class level, where the DP group had more freshman than sophomores and juniors, and GPA where the students in the DP group had a slightly higher GPA. These differences approached significance at just above the 0.05 and 0.10 levels, respectively. Accounting for GPA and Math SAT score differences in the analysis changed neither the direction nor the statistical significance/insignificance of the treatment effect on the final exam.

It is difficult to explain the group-by-instructor interactions on exam scores. Perhaps one instructor's DP students and the other instructor's MP students were somehow higher-achieving or more motivated in ways that are not captured by GPAs or SAT scores. The relationship between treatment group and homework scores seems fairly consistent for the intent-to-treat analysis and the secondary analysis, so it is difficult to justify an instructor-by-homework effect on the final exam outcome. Regrettably, it is impossible to know how things would have turned out had teaching duties not been reassigned after the start of the semester. The data on only the author's students (the secondary analysis) suggest that this DP homework treatment was effective for them on the final exam. Perhaps these data could be given more weight as the other instructor was not included in the planning and design stages of the study, though through no fault of their own. Perhaps the other instructor had lower enthusiasm for the potential benefit of DP that was somehow transmitted to that instructor's students and in turn affected the DP group differently than the MP group.

It is notable that the DP group performed significantly better than the MP group on the first exam but that the effect seemed to diminish for later exams. The same pattern is observed in both the intent-to-treat and in the secondary analyses. Since students were not blinded to the kind of homework sets they were doing, perhaps the DP students were initially more motivated than the MP students because they saw that they were in the experimental group which they were told was hypothesized to be more effective. If so, this Hawthorne effect, where students modify their behavior in response to their awareness of being observed, would have implications for any novel treatment and rigorous follow-up study would be needed to determine whether the effect would eventually disappear once the novelty wore off. In future research where students cannot be blinded to treatment, not tipping one's hand to students as to the investigator's hypothesis is recommended in order to reduce any Hawthorne effects.

Interestingly, the DP group did marginally significantly better on the data assignments. The data assignments assess software skills that are not explicitly addressed by most of the problems in the homework sets, so this difference is a bit puzzling. Again, it could be an issue of increased overall motivation spurred by students knowing they are being treated in a way that is hypothesized to be more effective.

Homework scores in the DP group were higher than in the MP group. One could surmise that this was because the DP homework sets were easier for students since they included fewer new problems and some review problems that they had previously worked. Almost 10% of the problems were used more than once across the DP problem sets.

While this study certainly provides some evidence for the efficacy of using a distributed practice approach in homework sets, it is only a first step and there is much room for further study. A fairly glaring flaw in this study is the violation of protocol in introducing a second instructor who was not trained in the study protocol. Compared to other non-experimental study designs, this error is perhaps not as egregious, but it detracts, nonetheless, from the study's integrity. As a first step in future investigations, the study should be repeated under the intended conditions: one instructor and simple random assignment to the two study groups. On the other hand, the significant group by instructor interactions highlight the potential for an instructor factor to modify any effects of DP vs. MP in a big way. Should further single-instructor studies find favorable effects of DP, next steps would be to design multi-instructor studies. In such a study, random assignment stratified by instructor would be appropriate to allow for formal comparisons of DP and MP within instructor.

Simple randomization was used in this study but perhaps a randomization stratified by GPA or SAT score would both ensure balance on these variables and allow for a more precise comparison of DP to MP, perhaps within categories of these variables. Studying the student populations at different institutions, both undergraduate and high school, would be a natural next step. Also recommend is the use of a validated, standardized assessment instrument like the Comprehensive Assessment of Outcomes in a First Statistics Course (delMas, Garfield, Ooms, and Chance 2007) to measure learning outcomes. While instructor-designed instruments are appropriate for assessing outcomes for a particular course at a particular institution, using a standardized instrument would allow comparison of research across courses and institutions. There is little research on the optimal number of questions and spacing of DP homework, so

lines of inquiry as to how many problems to include in each homework set, the ratio of new to review problems, and the spacing of review problems across DP homework sets are wide open.

Finally, if we wish to assess the efficacy of DP on longer-term retention, we will need to design studies that address long-term follow-up. Since follow-up with students is difficult after graduation, this is a tall order indeed. Even if we are able to make contact with some students years later, we likely will not capture the entire randomized participant pool, thereby introducing the potential for biases that the original randomization was designed to eliminate. Long-term retention, nevertheless, is the goal; if an intervention is effective for only a short time, we may question the ultimate utility of the intervention.

With the advent of online homework systems, it is much easier to perform randomization of homework interventions at the student level instead of at the class level and researchers are encouraged to proceed along those lines as much as possible. Where it is still necessary to use a class-level randomization, the data analyses should take the class-level clustering of observations into account. Regardless of the kind of randomization used, let's practice what we preach to our students about the importance of using randomization to determine efficacy.

Appendix

University of Delaware
Informed Consent Form

Title of Project: The Effect of Distributed Practice in Undergraduate Statistics Homework Sets: A Randomized Trial

Principal Investigator: Bryan Crissinger, Department of Mathematical Sciences

Other Investigators: Kevin R. Guidry, Center for Teaching and Assessment of Learning

You are being asked to participate in a research study. This form tells you about the study including its purpose, what you will be asked to do if you decide to participate, and any risks and benefits of being in the study. Please read the information below and ask the research team questions about anything we have not made clear before you decide whether to participate. Your participation is voluntary and you can refuse to participate or withdraw at any time without penalty or loss of benefits to which you are otherwise entitled. If you decide to participate, you will be asked to sign this form and a copy will be given to you to keep for your reference.

WHAT IS THE PURPOSE OF THIS STUDY?

The purpose of this study is to find out whether homework sets in MATH201 (Introduction to Statistical Methods) that include review problems are better than homework sets that include only new problems.

You are being asked to take part in this study because you are currently enrolled in one of my sections of MATH201 (sections 10, 14, or 15). You may be excluded from volunteering for the study if you are enrolled in another section of MATH201. Between 200 and 300 students are expected to participate.

WHAT WILL YOU BE ASKED TO DO?

All participants will do homework using MathXL, the online homework system we use for MATH201. I will assign some participants at random to do problem sets that include review problems and the other participants will do problem sets that include only new problems. This will be the only intervention. Other than the differences between the two kinds of problem sets, all other course activities will be the same for the two groups.

By participating, you also agree to allow me to use information about you that is already on file with the University of Delaware. Such information may include gender and SAT score. This information will be used to describe the group of students who participate and to compare the two homework groups. Such information will be kept confidential to the extent permitted by law.

WHAT ARE THE POSSIBLE RISKS AND DISCOMFORTS?

The primary risk to participants is a breach of academic data confidentiality. However, this risk is no more than what students encounter in a normal educational setting.

WHAT ARE THE POTENTIAL BENEFITS?

Students assigned to the homework group with review problems may potentially perform better in MATH201. Note that this is a potential benefit and is not guaranteed; it is possible that there may be no direct benefit to you. If convincing evidence based on this study shows that using review problems in homework is better, future students may also benefit. This may be true not only at the University of Delaware and in introductory statistics but at other schools and in other subjects.

HOW WILL CONFIDENTIALITY BE MAINTAINED?

We will make every effort to keep all research records that identify you confidential to the extent permitted by law. Signed informed consent forms and any other paper records will be kept in a locked filing cabinet in my campus office which is locked when not occupied. The data set used for research purposes will be de-identified; your names and ID numbers will be removed and replaced with a code number. All academic data, including that used for research purposes, will be stored in my University-maintained password-protected file system. In addition, the code file containing the links between subjects' identities and the code numbers will be encrypted. All data used for research purposes will be retained through Spring 2018.

In the event of any publication or presentation resulting from the research, no personally identifiable information about you will be shared.

Your research records may be viewed by the University of Delaware Institutional Review Board, but the confidentiality of your records will be protected to the extent permitted by law.

WILL THERE BE ANY COSTS RELATED TO THE RESEARCH?

There are no costs associated with participating in this study.

WILL THERE BE ANY COMPENSATION FOR PARTICIPATION?

Students who participate will earn a 2% extra credit bonus added to their final course average in MATH201.
Students who choose not to participate will also have the opportunity to earn the 2% bonus by taking a standardized statistical reasoning assessment at the end of the course. To qualify for the bonus, students must answer at least 8 out of around 20 questions correctly. Two possible options for the assessment are the GOALS (Goals and Outcomes of Learning Statistics) and CAOS (Comprehensive Assessment of Outcomes in a first Statistics course), both developed by researchers at the University of Minnesota in collaboration with leaders in statistics education.

DO YOU HAVE TO TAKE PART IN THIS STUDY?

Taking part in this research study is entirely voluntary. You do not have to participate in this research. If you choose to take part, you have the right to stop at any time. If you decide not to participate or if you decide to stop taking part in the research at a later date, there will be no penalty or loss of benefits to which you are otherwise entitled. Your refusal will not influence current or future relationships with the University of Delaware. As a student, if you decide not to take part in this research, your choice will have no effect on your academic status or your grade in the class.

Should you decide to stop participating before the end of the semester, you are still eligible for the 2% extra credit bonus by completing the statistical reasoning assessment (see above for details).

WHO SHOULD YOU CALL IF YOU HAVE QUESTIONS OR CONCERNS?

If you have any questions about this study, please feel free to contact me:

Bryan Crissinger
302-831-8142
crissing@math.udel.edu

If you have any questions or concerns about your rights as a research participant, you may contact the University of Delaware Institutional Review Board at 302-831-2137.

Your signature below indicates that you are voluntarily agreeing to take part in this research study. You have been informed about the study's purpose, procedures, possible risks and benefits. You have been given the opportunity to ask questions about the research and those questions have been answered. You will be given a copy of this consent form to keep.

_____ _____
Signature of Participant Date

_____
Printed Name of Participant

Acknowledgements

The author thanks the Center for Teaching and Assessment of Learning (CTAL) at the University of Delaware for providing funding for this research. Many thanks to Kevin Guidry at CTAL for providing access to students' demographic data and for his patience in answering many questions.

The randomization and data analysis for this paper were generated using SAS/STAT software, Version 9 of the SAS System for Unix. Copyright © 2002-2010 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. Figures were generated using Minitab 17 Statistical Software. Copyright © 2010 Minitab, Inc., State College, PA, USA.

Results of this research were presented at a seminar in the Department of Mathematical Sciences, University of Delaware on May 14, 2015.

References

Agodini, R., Harris, B., Thomas, M., Murphy, R., and Gallagher, L. (2010), "Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders," Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

American Statistical Association (2005), *Guidelines for Assessment and Instruction in Statistics Education*, available at http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf.

Bloom, H. (ed.) (2005), *Learning More from Social Experiments: Evolving Analytic Approaches*, New York: The Russell Sage Foundation.

Batson, G., and Schwartz, R. E. (2007), "Revisiting the Value of Somatic Education in Dance Training Through an Inquiry into Practice Schedules," *Journal of Dance Education*, 7, 47 – 56.

Bude, L., Imbos, T., Wiel, M. W., and Berger, M. P. (2011), "The Effect of Distributed Practice on Students' Conceptual Understanding of Statistics," *Higher Education*, 62, 69 – 79.

Carpenter, S. K., and Mueller, F. E. (2013), "The Effects of Interleaving Versus Blocking on Foreign Language Pronunciation Learning," *Memory & Cognition*, 41, 671 – 682.

Cull, W. L. (2000), "Untangling the Benefits of Multiple Study Opportunities and Repeated Testing for Cued Recall," *Applied Cognitive Psychology*, 14, 215 – 235.

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2007), "Assessing Students' Conceptual Understanding After a First Course in Statistics," *Statistics Education Research Journal*, 6, 28 – 58.

Dempster, F. (1991), "Synthesis of Research on Reviews and Tests," *Educational Leadership*, 48, 71 – 76.

Duke, R. A., and Davis, C. M. (2006), "Procedural Memory Consolidation in the Performance of Brief Keyboard Sequences," *Journal of Research in Music Education*, 54, 111 – 124.

Fischer, S., Hallschmid, M., Elsner, A. L. and Born, J. (2002), "Sleep forms memory for finger skills," *Proceedings of the National Academy of Sciences USA*, 99(18), 11987–91.

Fishman, E. J., Keller, L., and Atkinson, R. (1968), "Massed Versus Distributed Practice in Computerized Spelling Drills," *Journal of Educational Psychology*, 59, 290-296.

Gais, S., Plihal, W., Wagner, U., and Born, J. (2000), "Early sleep triggers memory for early visual discrimination skills," *Nature Neuroscience*, 3(12), 1335–39.

Good, T. L., and Grouws, D. A. (1979), "The Missouri Mathematics Effectiveness Project," *Journal of Educational Psychology*, 71, 355 – 362.

Grote, M. G. (1995), "Distributed Versus Massed Practice in High School Physics," *School Science and Mathematics*, 95, 97 – 101.

Gupta, Sandeep K. (2011), "Intention-to-treat concept: A review," *Perspectives in Clinical Research*, 2, 109 – 112.

Holdan, E. G. (1985), "A Comparison of the Effects of Traditional, Exploratory, Distributed, and a Combination of Distributed and Exploratory Practice on Initial Learning, Transfer, and Retention of Verbal Problem Types in First-Year Algebra (Doctoral dissertation, Pennsylvania State University)," *Dissertation Abstracts International*, 46, 2542A.

Horine, J. (1983), "The Cumulative Effects of Incremental Practice on the Retention of Mathematical Rules (Doctoral dissertation, Florida State University)," *Dissertation Abstracts International*, 44, 372A.

James, W. (1901), *Talks to Teachers on Psychology: And to Students on Some of Life's Ideals*, New York: Holt.

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J., and Sagi, D. (1994), "Dependence on REM sleep of overnight improvement of a perceptual skill," *Science*, 265(5172), 679–82.

Kornell, N., Castel, A. D., Eich, T. S., and Bjork, R. A. (2010), "Spacing as the Friend of Both Memory and Induction in Young and Older Adults," *Psychology and Aging*, 25, 498 – 503.

Krug, D., Davis, T. B., and Glover, J. A. (1990), "Massed Versus Distributed Repeated Reading: A Case of Forgetting Helping Recall?" *Journal of Educational Psychology*, 82, 366 – 371.

MacDonald, C. J. (1984), "A Comparison of Three Methods of Utilizing Homework in a Precalculus College Algebra Course (Doctoral dissertation, The Ohio State University)," *Dissertation Abstracts International*, 45, 164A.

Mackay, S., Morgan, P., Datta, V., Chang, A., and Darzi, A. (2002), "Practice Distribution in Procedural Skills Training: A Randomized Controlled Trial," *Surgical Endoscopy*, 16, 957 – 961.

Mayfield, K. H., and Chase, P. N. (2002), "The Effects of Cumulative Practice on Mathematics Problem Solving," *Journal of Applied Behavior Analysis*, 35, 105 – 123.

McClave, J. T., Benson, P. G., and Sincich, T. (2014), *Statistics for Business and Economics*, 12th Edition, New York, NY: Pearson Education, Inc.

National Archives and Records Administration (2014), "Food and Drugs: Adequate and Well-Controlled Studies," *Code of Federal Regulations*, Title 21.

National Research Council (2013), "The Mathematical Sciences in 2025," Washington, D.C.: The National Academies Press.

Reynolds, J. H., and Glaser, R. (1964), "Effects of Repetition and Spaced Review upon Retention of a Complex Learning Task," *Journal of Educational Psychology*, 55, 297 – 308.

Rohrer, D., and Taylor, K. (2006), "The Effects of Overlearning and Distributed Practice on the Retention of Mathematics Knowledge," *Applied Cognitive Psychology*, 20, 1209 – 1224.

Sabbag, A., Garfield, J., and Zieffler, A. (July 2015), "Quality assessments in statistics education: A focus on the GOALS instrument," in *International Association of Statistics Education (IASE) Satellite Conference Proceedings, Advances in Statistics Education: Developments, Experiences and Assessment*. Available at http://iase-web.org/documents/papers/sat2015/IASE2015%20Satellite%2041_SABBAG.pdf.

Simmons, A. L. (2012), "Distributed Practice and Procedural Memory Consolidation in Musicians' Skill Learning," *Journal of Research in Music Education*, 59, 357 – 368.

Smith, S. M., and Rothkopf, E. Z. (1984), "Contextual Enrichment and Distribution of Practice in the Classroom," *Cognition and Instruction*, 1, 341 – 358.

Wahlheim, C. N., Dunlosky, J., and Jacoby, L. L. (2011), "Spacing Enhances the Learning of Natural Concepts: An Investigation of Mechanisms, Metacognition, and Aging," *Memory & Cognition*, 39, 750–763

Walker, M. P. (2005), "A Refined Model of Sleep and the Time Course of Memory Formation," *Behavioral and Brain Sciences*, 28, 51 – 104.

Willingham, D. T. (2002), "How We Learn. Ask the Cognitive Scientist: Allocating Student Study Time. 'Massed' versus 'Distributed' Practice," *American Educator*, 26, 37 – 39.

Bryan Crissinger
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716
crissing@udel.edu
302-831-8142
