Does Interleaving Rather Than Blocking Facilitate Learning Examples in Statistics?

Robert S. Ryan
Steven R. Howell
Heather A. Shaw

Kutztown University

Presented at the 22nd Annual Convention of the Association For Psychological Science, Boston, MA, May 27 – May 30, 2010.

Correspondence:

Robert S. Ryan
Box 730, Psychology Department
Kutztown University
Kutztown, Pennsylvania 19530
rryan@kutztown.edu

Does Interleaving Rather Than Blocking Facilitate Learning Examples in Statistics?

Participants studied either blocked or interleaved examples of appropriate uses of statistical tests. Acquisition was not affected, but interleaving benefited retention. Attempts to raise typically low performance were unsuccessful and did not replicate the benefit. We consider why performance was low and whether raising it would reinstate the interleaving benefit.

Rohrer and Taylor (2007) found that interleaving rather than blocking examples of geometry problems facilitated college students' ability to recall the correct equation for the problem. Similarly, Kornell and Bjork (2008) found that interleaving examples of paintings by different artists facilitated learning their painting styles.

We hypothesized that interleaving examples of different research situations might facilitate the ability of statistic students to learn the correct statistical procedure to use in each situation. Our subjects participated in a training session followed by an immediate acquisition test. However, we also wanted to test retention, and because we conducted the experiment in a statistics course, rather than in a lab, at some point in the course participants were to receive formal classroom instruction in the same task for which they had been trained in the experiment. Therefore, a few weeks after the training and immediate test, but before the formal classroom instruction, we gave them an early retention test.

In addition, we administered a late retention test at the end of the semester after all the formal instruction had been provided. This enabled us to also examine whether the training method affected how much the participants benefited from the formal classroom instruction.

Experiment 1

Method

Participants. The participants were 63 university undergraduates in a behavioral statistics course who completed the entire experiment.

Materials. The materials consisted of a training booklet and three tests that occurred at different intervals after the training.

Training. The training booklet contained 24 descriptions of research situations. There was one description on each page about a half page in length. There were six types of research situations and there were four examples of each type. Each type of research situation required a certain statistical test. The six statistical tests were the two sample t test, the paired t test, the one way ANOVA, the repeated measures ANOVA, the chi square test, and correlation. Each paragraph was labeled at the top to indicate the correct statistical test for that example. At the

end of the example the description stated what test the researcher used in the study (see Appendix A for an example of the descriptions).

In the blocked condition the four examples of each different type of research situation occurred on consecutive pages. In the interleaved condition they were arranged so that the same type of research situation did not occur twice in a row.

Tests. We used an immediate test, an early retention test, and a late retention test. Each test had nine test items. Each test item was a description of a research situation similar to those in the training booklet. However, there was no label provided at the top. Also, at the end of the test item where the correct statistical test was provided in the training booklet, there was a blank line (see Appendix B for an example of the test items). The participants were given six statistical tests as possible answers from which to choose. Since there were nine items in the test and only six choices, the participants were instructed that some of the statistical tests could occur as the correct answer more than once or could have not occurred at all. This was done so that the participants could not use the process of elimination. The early and late retention tests were the same as the immediate test but with different examples.

Procedure. The training and immediate tests occurred in the first week of the semester. The early retention test occurred four to six weeks after the immediate test but before the formal instruction on the statistical procedures used for the experimental materials. The late retention test occurred at the end of the semester after all the formal instruction had been provided.

Training. Prior to the experiment the training booklets were arranged into alternating blocked and interleaving booklets so that there would be approximately the same number of blocked and interleaved participants in each of four classes of statistics students. Before the training, participants were given a consent form that informed them that they could decline to participate by simply not performing the task. For the training, the participants studied each research situation for one minute. They worked through the items in the order in which they were presented in the training booklet, and they did not return to any previous items.

Tests. The instructions for all of the tests were the same. We instructed the participants to read every paragraph carefully and to select the statistical procedure they thought was correct. The participants were told they had to answer all the questions on the test even if they had to guess. We instructed them to work though all the items in order and that they were not permitted to go back to any previous items. The tests were not timed, but they had to work quickly enough to finish before their class period ended. If the participant was done with the test early, they were asked to sit quietly and wait until everyone else was done.

Results

Table 1 shows the mean scores on the immediate test, early retention test, and late retention test as a function of training condition. There was only one significant difference. There was a significant advantage of interleaving in the early retention test, t (58, two tailed) = 2.81, p = .007. There was no effect of the training condition on the immediate test, t (61, two tailed) = .381, n.s., nor on the late retention test, t (61, two tailed) = .876, n.s.

Table 1

Mean Percent Correct in the Blocked and Interleaved Condition for the Immediate, Early Retention, and Late Retention Tests in Experiment 1.

	Training Condition	
Test	Blocked	Interleaved
Immediate	31	33
Early Retention	21	33
Late Retention	41	46

Discussion

Interleaving the examples, rather than blocking them, did not result in better acquisition, but it did result in better retention. Also, it did not affect the benefit of learning the examples from the formal classroom instruction. Perhaps more importantly, especially from a practical application standpoint, the advantage in retention for interleaving was only relative to the blocked condition, rather than being good performance in an absolute sense. All of the test performance was at such a low level that it would result in a failing grade if this had been an actual classroom assessment. Furthermore, with performance so low it is surprising that the blocked participants performed significantly lower than the interleaved participants on the early retention test. Therefore, we wished to replicate this result before drawing any conclusions.

Experiment 2

In Experiment 1 the performance was very low. Therefore, in Experiment 2 the number of different kinds of research situations was decreased and the number of examples of each kind was increased. We hypothesized that giving the participants more practice with each kind of research situation might improve learning. This might enable us to uncover an effect of interleaving on the immediate test, a larger benefit of interleaving on the early retention test, and, perhaps, an effect of interleaving on benefiting from the formal classroom instruction.

Method

Participants. The participants were 108 university undergraduates in a behavioral statistics course who completed the entire experiment.

Materials and Procedure. In the training booklet instead of having six types of research situations and four examples of each type, the types of research situations were decreased to four by eliminating the chi square test and correlation, and the number of examples of each type was increased to six. The procedure for Experiment 2 was exactly the same as for Experiment 1.

Results

As shown in Table 2 performance was still very low. There were no significant differences in performance between the interleaved and blocked conditions in either the immediate test, early retention test, or late retention test.

Table 2

Mean Percent Correct in the Blocked and Interleaved Condition for the Immediate, Early Retention, and Late Retention Tests in Experiment 2.

	Training Condition	
Test	Blocked	Interleaved
Immediate	29	29
Early Retention	26	23
Late Retention	32	29

Discussion

Contrary to what we expected, increasing the amount of practice with each kind of research situation by simply providing six examples of four types, rather than four examples of six types, did not improve learning. The beneficial effect of interleaving on the early retention test seen in Experiment 1 did not replicate. Therefore, we speculated that a greater increase in the amount of practice might be needed to produce the desired result. Also, we hypothesized that another contributing factor to the low performance was that the descriptions of the research situations may have been too difficult for our participants to read.

Experiment 3

In Experiment 3 more changes were made in order to try to raise performance. The paragraphs describing the research situations were simplified, we changed the labels for some of the statistical tests, and we used three training sessions instead of just one.

Method

Participants. The participants were 75 university undergraduates in a behavioral statistics course who completed the entire experiment.

Materials and Procedure. Reducing the number of types of research situations from six to four had failed to improve performance in Experiment 2. Therefore, in Experiment 3, we went back to six types of research situations and four examples of each type, as in Experiment 1. We also made three other changes. First, we made the descriptions shorter, easier to read, and equal in length (Appendix C shows how the example provided in Appendix A was changed). Second, we changed the labels we used for some of the statistical tests. In the two prior experiments, we had called the first four tests: the two sample t test, the paired t test, the one way ANOVA, and the repeated measures ANOVA. In Experiment 3 we called them: the independent-measures t test, the repeated –measures t test, the independent-measures ANOVA, and the repeated-measures ANOVA. We believed that highlighting that these four tests could be distinguished on two dimensions (i.e., whether they were independent or repeated and whether they were t tests or ANOVA's) would make it easier for the participants to learn them. Third, there were three training sessions on the same training examples, each followed by an immediate test, instead of just one. The training sessions were scheduled in three successive weeks early in the semester. Otherwise, the procedure in Experiment 3 was the same as that in Experiment 2.

Results

As shown in Table 3, there were again no significant differences in performance between the interleaved and blocked conditions in any of the tests. However, performance did increase across the three immediate tests.

Table 3

Mean Percent Correct in the Blocked and Interleaved Condition for the Immediate 1, 2, and 3, Early Retention, and Late Retention Tests in Experiment 3.

	Training Condition	
Test	Blocked	Interleaved
Immediate 1	23	26
Immediate 2	38	31
Immediate 3	41	40
Early Retention	36	32
Late Retention	38	40

Discussion

Again, contrary to our expectation, performance on the retention tests was still at a level that was so low that it would be considered failing for purposes of actual classroom evaluation. The beneficial effect of interleaving on the early retention test seen in Experiment 1 still did not replicate.

General Discussion

This series of studies did not provide any convincing evidence that interleaving rather than blocking examples during training affected people's ability to learn which statistical test should be used in various research situations. However, what is a greater cause for pessimism than the lack of an interleaving effect is the extremely low performance on all of the tests. In the studies that inspired the current study (Kornell & Bjork, 2008; Rohrer & Taylor, 2007) performance of interleavers on immediate tests was in the 60% to close to 80% range. In our studies, on the tests that occurred immediately after an initial 24 minutes of studying the examples, performance ranged from 23% correct to 33%. When participants were given two more opportunities to study the same items, performance gradually increased but only to a maximum of 41%. Perhaps even more discouraging, when we tested the participants after they had received formal classroom instruction, the maximum performance was only 46% - not even half of the items correct.

On the positive side, it is helpful to learn what does not work. Our task was a category induction task similar to that Kornell and Bjork's (2008) painting study. Such tasks require the participants to notice which features of the examples determined the correct category. However, Kornell and Bjorks's stimuli had perceptual features that may have been more easily noticed than ours. Therefore, it may have been obvious to Kornell and Bjork's participants that they were to perform a category induction task by noticing the perceptual features. Rohrer and Taylor (2007) , on the other hand, used mathematical problems as stimuli and they gave their participants a tutorial in how to do the problems before the participants practiced them. Our stimuli did not involve solving mathematical problems, but the features of our stimuli were abstract concepts such as the number of conditions, and whether there were the same or different subjects in those conditions rather than the perceptual characteristics of paintings. Therefore, our participants may have needed advance instruction on the nature of a category induction task and how to do it.

Furthermore, in one of Kornell and Bjork's (2008) studies, the participants test task involved recall, whereas in the other, it involved the usually easier task of recognition. Notably, for the more difficult recall task Kornell and Bjork provided immediate feedback during the test. The test task for our participants was a recall task. Therefore, immediate feedback during the test might be another strategy that would be effective for raising our participants' performance.

In our future studies we will provide advance instruction in how to do a category induction task as well as feedback during training. With these improvements we hope to raise performance on our task, which may finally enable us to more effectively examine the role of interleaving versus blocking on learning which statistical test to use for different research situations.

References

Kornell, N, & Bjork, R. A. (2008). Learning concepts and categories: Is spacing "The enemy of induction"? *Psychological Science, 19,* 585-592. doi: 10.1111/j.1467-9280.2008.02127.x

Rohrer. D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35,* 481-498. doi: 10.1007/s11251-007-9015-8

Appendix A

An Example of the Descriptions of Research Situations From the Training Materials for Experiment 1

Two sample *t* test

A group of researchers wanted to determine whether studying while relaxed results in better learning than studying while anxious. Each of a group of 100 subjects was randomly assigned to one of two conditions. In one condition, called the relaxation condition, the subjects engaged in a relaxation technique before studying a chapter in a history text. In the other condition, called the anxiety condition, in order to make them anxious, the subjects were told that they would have to give a speech about what they learned to an audience. Then they also studied the history text. The subjects were all very similar in important characteristics such as their natural tendency to be anxious, their age, IQ, motivation to learn, etc. They all studied the same chapter for the same amount of time. The conditions of study were exactly the same for both groups except for their relaxation versus anxiety having been manipulated. After they studied, they were given a test on the history chapter. The researchers calculated the average test scores for the two groups. To determine whether the average test scores were significantly different, the researchers calculated a two sample *t* test.

Appendix B

An Example of the Test Items from Experiment 1

A group of researchers wanted to determine whether aromatherapy while studying results in better learning than studying without pleasant aromas. A group of 110 subjects was recruited. Each subject was randomly assigned to one of two conditions. In one condition, called the Perfume condition, the subjects studied a chapter in an anthropology text while a mild pleasant scent was released continuously into the room. In the other condition, called the Normal condition, the subjects studied the same chapter in a normal, relatively scent-free room. The subjects were all very similar in important characteristics such as their olfactory sensitivity, their age, tolerance of scents, IQ, motivation to learn, reading ability, etc. They all studied the same chapter for the same amount of time. The conditions of study were exactly the same for both groups except for the aroma of the room having been manipulated. After they studied, they were given a test on the anthropology chapter. The researchers calculated the average test scores for the two groups. To determine whether the average test scores were significantly different, the researchers calculated a _____.

a. Two sample *t* test

b. Paired *t* test

c. One way ANOVA

d. Repeated measures ANOVA

e. Chi square test

f. Correlation

Appendix C

An Example of How the Descriptions of Research Situations From Experiment 1 Were Simplified For Experiment 3

Independent-measures *t* test

A group of researchers wanted to determine whether studying while relaxed results in better learning than studying while anxious. Each of a group of 100 subjects was randomly assigned to one of two conditions. For the subjects assigned to the relaxation condition, they first engaged in a relaxation technique. Then they studied a chapter in a history text and took a test on the chapter. For the subjects assigned to the anxiety condition, first they were told that they would have to give a speech about what they learned to an audience. Then they studied the chapter and took the test. The researchers calculated the average test scores for the two groups. To determine whether the average test scores were significantly different, the researchers calculated an independent-measures *t* test.