Using Interleaved Examples to Teach Inferential Statistics

Robert S. Ryan
Steven R. Howell
Heather A. Shaw
Dale W. Kappus
Mara E. Wilde
Samantha L. Crist

Kutztown University

Presented at the 1st Annual PASSHE Potluck Psychology Conference, Indiana, PA, October 22 - October 23, 2010

Correspondence:

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

Using Interleaved Examples to Teach Inferential Statistics

Previous research suggests that interleaving examples, such as examples of painter's styles, leads to better learning than presenting them in blocks. Our four experiments, however, examined factors affecting learning in the domain of statistics. We examined statistics students' ability to learn which statistical test to use by studying examples of research situations. The first three studies failed to provide convincing evidence that interleaving the examples during studying resulted in any better learning than presenting them in blocks, but performance was very low. Therefore, in order to try to raise performance, our fourth experiment used only blocked examples, but manipulated whether we focused students' attention on the features of the research situation that determined which statistical test to use. This manipulation improved their performance. Future studies will use this manipulation to examine whether an interleaving advantage emerges when performance is sufficiently raised.

Studying examples is an important way for people to learn. For example, LeFevre and Dixon (1986) showed that people prefer to learn from examples rather than from instructions. Furthermore, they can apply what they learned from training examples to similar test examples (Anderson, Fincham, & Douglass, 1997). However, does the acquisition and retention of knowledge gained from examples depend on how they are presented? For example, in teaching students to learn several different concepts, teachers could present examples in *blocks*, that is, several examples of one concept, followed by several examples of another. But they could also present them *interleaved*, that is, the examples of the different concepts could be mixed so that each example of a particular concept was always followed by an example of a different concept.

Rohrer and Taylor (2007) shed some light on this question by showing that people learned to solve geometrical problems better if they were trained with interleaved rather than blocked examples. However, Kornell and Bjork (2008) questioned whether the benefit of interleaving would extend to a different kind of learning, specifically, category induction. They hypothesized that for a task such as learning to categorize painting styles, people might benefit from being able to compare several examples of the same style presented one right after the other. Therefore, they hypothesized that in this case blocking might be superior to interleaving. However, they found that interleaving was superior in this case as well.

Kornell and Bjork (2008) showed that interleaving rather than blocking examples is beneficial for a category induction task. However, their materials, painting styles, were perceptual. We wished to examine whether the benefit of interleaving would extend to inducing conceptual, rather than perceptual, categories. We chose concepts in statistics as materials in order to be able to examine a different type of category learning, and also to examine it in a context that would have immediate educational applications.

We report the results of four experiments. The training materials were examples of different types of research situations and which inferential statistics procedure was appropriate to analyze the data from that particular situation. We trained students in our introductory statistics course with the examples before they received formal classroom instruction in the statistical procedures used in the examples. The training was followed by an immediate acquisition test and a retention test a few weeks later. Also, in order to examine whether interleaving examples affected how much benefit students obtained from their formal instruction, there was also a final test that occurred at the end of the course, and therefore, after the formal classroom instruction in the statistical procedures.

In the first experiment, we found some evidence for a benefit of interleaving, but only on the retention test. Furthermore, performance was at such a low level that it would not even be a passing grade for an actual classroom test. In the next two experiments, we changed the training materials and instructions in efforts to improve performance above floor level, in hopes that the improvement might allow a stronger interleaving effect to be found. When neither set of changes produced either the desired improvement or any interleaving effect, we conducted a fourth experiment, in which our goal was just to improve performance. We made an even larger change in the materials and instructions, but without manipulating blocking versus interleaving. In that experiment, we manipulated whether or not the subjects received explicit information about the features of the examples that determined their category membership, but all the examples were presented in blocks.

Experiment 1

Method

Participants. The participants were 63 Kutztown University undergraduate students, all of whom were enrolled in an introductory statistics course. About half of the participants were assigned to the blocked condition (*n* = 30) and the rest (*n* = 33) were assigned to the interleaved condition.

Materials. Data was collected to assess the effects of certain types of training on the immediate tests that followed. A training packet was provided to the subjects, it consisted of different research situations and the statistical procedure that was appropriate to analyze it. There were two different types of training packets. One packet used an interleaving method another packet used a blocked method. However, both conditions included research situation where the appropriate inferential statistics procedure was either; two sample t test, chi square, correlation, paired t test, repeated measures ANOVA or one way ANOVA. Each procedure was represented four times in the packet. Also, each packet consisted of 24 different research situations all of which were approximately a paragraph long. The retention test was used to study the effects of the training from the training packet. The same test was given for the immediate retention, early retention and late retention. The test consisted of 9 multiple choice questions. A paragraph that explained a research situation acted as the question and the possible correct answers were the appropriate statistical procedure in multiple choice form. In order to keep the training and testing standardized a script was utilized by the investigators. The script provided what to do and what to say while administering the training packets and the following retention tests.

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 Kutztown University undergraduates in an introductory 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 Kutztown 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 Three Immediate Tests, the Early Retention, and the 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.

Experiment 4

Methods

Participants. The participants were 40 Kutztown University undergraduate students, all of whom were enrolled in an introductory statistics course.

Materials and Procedure. In this experiment we did not include interleaving as a factor. Instead, we made two changes. The first change was made in order to try to raise the performance of all the subjects. We believed that our subjects might benefit from an explanation of the nature of a category induction task. Therefore, in all the subjects' training instructions, we described Kornell and Bjork's (2008) painting study and compared it to the task they were about to perform.

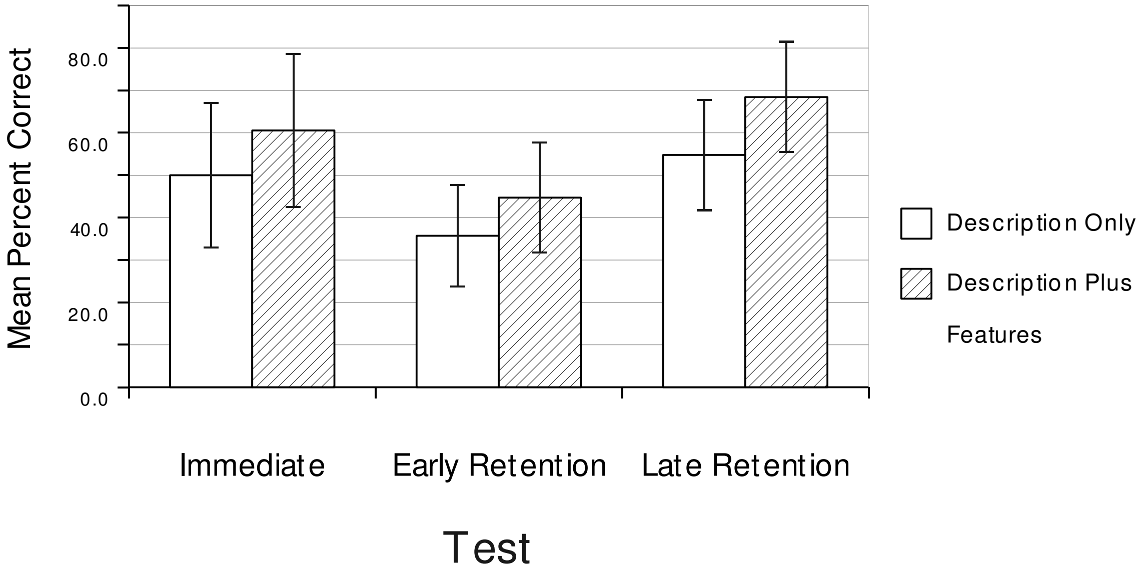
The second change involved manipulating how the subjects thought about two things. First, we wanted them to think about what features of the research situation determined which statistical test to use. For example, having two conditions, with different subjects in each condition, would call for an independent samples *t* test. Second, we wanted them to think about how they would determine what those features were. For example, if the research situation said that one group of subjects studied in dim light, whereas a different group of subjects studied in bright, that would enable them to determine that there were two conditions, with different subjects in each condition.

For our manipulation, subjects in an example only condition were explicitly given only the example of the research situation. In addition however, they were encouraged to generate both the features and an explanation of how they would determine them. Subjects in an examples plus features condition were explicitly given the example of the research situation, and they were also explicitly given the features. They were then encouraged to generate an explanation of how the features had been determined.

Results

A MANOVA was used to analyze the performance on the immediate, early, and late retention tests. As shown in Figure 1, the subjects' performance decreased from the immediate test to the early retention test, and then improved from the early retention (before formal training) to the late retention (following formal training). The changes in performance across tests were significant, $F(2,76) = 5.07$, $p = .009$. The main effect of the training condition was marginally significant, $F(1,38) = 3.07$, $p = .088$.

Figure 1. Mean Percent Correct For the Description Only Condition And the Description Plus Features Condition Across the Immediate, Early Retention, And Late Retention Tests.



Discussion

Adding the features to the examples resulted in marginally better performance on the all subsequent tests. Additionally, there was a significant change in performance across both conditions from the immediate to the late retention test.

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 was a greater cause for pessimism than the lack of an interleaving effect was 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 first two studies, on the tests that occurred immediately after the initial studying of 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 of 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, apparently the participants in our first three studies needed advance instruction on the nature of a category induction task and how to do it.

Fortunately, by adding instruction on how to do category induction we did raise performance on all retention tests. Hopefully, this will enable us to more effectively examine the role of interleaving versus blocking on statistical learning in future studies.

Another factor that may also improve performance is providing feedback. For example, 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.

Furthermore, given that the materials our subjects are trying to learn is conceptual, rather than perceptual, it may be more important to provide the feedback during the concept learning task than during the test. That, in fact, may enable us to finally demonstrate an interleaving effect for our statistics materials (Doug Rohrer, personal communication, 10/04/10).

References

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 932-945.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "Enemy of Induction"? *Psychological Science, 19,* 585 - 592.

LeFevre, J. & Dixon, P. (1986). Do written instructions need examples? *Cognition and Instruction, 3,* 1-30.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35,* 481-498.