The Elusive Interleaving Effect: Why Doesn't Interleaving Improve Learning From Examples in Statistics?

Robert S. Ryan
Steven R. Howell
Dale W. Kappus
Mara E. Wilde

Kutztown University

Presented at SARMAC IX, New York, NY, June 27 – June 29, 2011.

Correspondence:

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

The Elusive Interleaving Effect: Why Doesn't Interleaving Improve Learning From Examples in Statistics?

People sometimes learn better from examples if the examples are presented in an interleaved, rather than a blocked, format. However, a previous study using statistics examples failed to replicate the effect, and also resulted in very poor learning overall (Ryan et al., 2010). In the present study, we attempted to produce the interleaving advantage with statistics examples by providing enhanced training to improve performance. We found that the enhanced training dramatically improved performance, but only on an immediate test. Furthermore, there was still no interleaving advantage. We discuss whether requiring participants to try to generate the relevant features, followed by feedback, may enable us to replicate the interleaving advantage.

When people try to learn different concepts by studying several examples of each one, they sometimes benefit from seeing examples of different concepts interleaved, rather than having all the examples of each concept presented together in a block. For example, interleaving examples of mathematics problems facilitates college students' ability to recall the correct equation for the problem (Rohrer & Taylor, 2007). Similarly, interleaving examples of paintings by different artists facilitates learning their painting styles (Kornell & Bjork, 2008).

Both of the studies cited above required subjects to learn to discriminate perceptual categories. In Rohrer and Taylor's (2007) first experiment, subjects had to learn which steps of a mathematical procedure to apply to a letter permutation problem depending on how many total characters there were, how many different letters were among the characters, and how often each letter was repeated. For example, in the problem *abbccc*, the correct permutation formula is to form a fraction with 6! in the numerator (because there are 6 characters). The denominator of the formula should consist of, 1! (because there is a single *a*), times 2! (for the 2 *b's*), times 3! (for the 3 *c's*). In Rohrer and Taylor's (2007) second experiment, subjects had to learn formulas to find volumes of various three dimensional geometric shapes. They also had to learn which formula went with which shape based on a picture of the shape. In Kornell and Bjork (2008) subjects had to learn to associate the name of a painter with a particular style of painting. Thus, interleaving helped with learning perceptually distinguishable categories, but it was not clear whether this finding also applied to conceptually distinguishable categories.

We were interested in whether we could apply the findings from Rohrer and Taylor (2007) and Kornell and Bjork (2008) to help students in an actual classroom setting learn concepts in the domain of statistics. For example, one of the most important concepts that statistics students have to learn is not only how to do various statistical procedures, but also which statistical procedure is the appropriate one to apply in a given research situation. Knowing which procedure to apply requires distinguishing between research situations that have different critical features, such as having only two conditions versus having more than two conditions, or having a between subjects design versus having a within subjects design. Therefore, the features that distinguish the categories to be learned in a real classroom setting in statistics are more conceptual than the features of the categories in the previously cited lab studies, which are more perceptual. It is not clear whether training by interleaving rather than blocking examples of those conceptually defined categories will have the same beneficial effect as interleaving examples of perceptually defined categories. Indeed Rohrer and Taylor (2007) claimed that giving students many examples of a problem requiring a repeated measures *t*-test might teach them *how* to do the test, but not how to recognize *when* to use it. Determining whether or not this was true was the aim of a series of studies, the first four of which were reported in Ryan et al. (2010). We begin by summarizing those studies, and then we present the study that followed them, which is the focus of this paper.

Ryan et al. (2010) 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. The subjects were students in several statistics classes. They participated in a training session followed by an immediate acquisition test at the beginning of the semester. However, there was also to be a later test of retention. Of course, later in the semester these students were to receive formal classroom instruction in the same task for which they had been trained in the experiment. Therefore, the training, immediate test, and also the retention test given a few weeks later were all administered before the formal classroom instruction. Then at the end of the semester, after all the formal instruction had been provided, a final test was administered. The final test provided a way to also examine whether the training method affected how much the participants benefited from the formal classroom instruction.

In all of Ryan et al.'s (2010) experiments the materials consisted of a training booklet and the three tests that occurred at different intervals after the training. The training booklet contained several descriptions of research situations along with the statistical procedure that would be appropriate. (see Appendix A for an example of the one of the descriptions used in Experiments 1 and 2). The tests consisted of several items describing a research situation just as had been done in the training materials. The subjects' test task was to select the correct statistical procedure for the research situation described. The early and late retention tests were the same as the immediate test but with different examples (see Appendix B for an example of one of the test items from Experiment 1).

In the first experiment, the interleaving subjects performed slightly better than the blocked subjects on all of the tests. However, averaged across the tests the main effect of interleaving had a significance level of p = .078. Out of the three tests, the difference between the conditions was the largest on the first retention test. A two tailed *t*-test on just the first retention test showed that the the advantage of interleaving (mean percent correct = .33) over blocking (mean percent correct = .21) was reliable, with a significance of *p* = .007. However, performance on the immediate and first retention tests was in the 20% to 30% range. Even after formal training, the performance on the last test was only in the 40% range. There was a main effect of time of test, with the effect coming from the increase in performance from the first retention test to the last test, the one after the formal training. However, given the scant evidence for an interleaving advantage and the generally low performance in the first experiment, three more experiments were conducted.

In the second experiment, Ryan et al. (2010) decreased the number of different kinds of research situations from six to four, and increased the number of examples of each type from four to six. Otherwise, the second experiment was the same as the first. The interleaving subjects performed slightly worse than the blocked subjects on all of the tests, however, the difference was not even close to significant. Overall performance was slightly better than in the first experiment, but only on the immediate test. And, even so, performance was only in the 30% to 50% range. There was a main effect of time of test, with the biggest difference being the drop from the 50% range on the immediate test, to the 30% range on the early retention test. Performance rose to the 40% range on the last test.

In the third experiment, Ryan et al. (2010) returned to training the subjects with four each of six different types of research situations. They also increased the number of training sessions from one to three. Each training session was exactly like those in the previous experiments, including being followed by an immediate test. In addition one other change was made, which was to the training materials. The descriptions of the research situations used in training were made shorter and simpler, and the terms “independent measures” and “repeated measures” were used consistently, instead of sometimes using those terms and sometimes using terms such as “two sample *t*-test”, and “paired *t*-test”. (see Appendix A for an example of the one of the descriptions used in Experiments 3 and 4) Otherwise, the third experiment was the same as the first two. The interleaving subjects performed slightly worse than the blocked subjects on all of the tests except for the last one. However, neither the main effect of interleaving nor the time by interleaving interaction were significant. There was a main effect of time of test. Performance rose from roughly the 20% range in the first immediate test, to the 30% range on the second, to the 40% range on the third, then back down to the 30% range on the early retention test, and back up to the 40% range on the final test.

So far, Ryan et al. (2010) had found virtually no credible evidence of an interleaving effect, and overall performance was still at a level that would be a failing grade in an actual classroom situation. So, at this point, they decided to do an experiment designed only to raise performance, and not to test the interleaving effect.

Of the two changes made previously, lowering the number of types of research situation from six to four had at least improved immediate performance more than had giving three training sessions. Therefore, for fourth experiment used only four types of research situation and only one training session. However, to make the training a little shorter, the subjects were presented with only four each of the four types of training situation instead of six. Also, the fourth experiment used the shorter, simpler, and more consistent wording that had been used in the third experiment.

The fourth experiment, however, also had another major change. It was designed to test the possible effect of providing the subjects with the names of the relevant features that determined which statistical procedure was correct for a given research situation. We did not manipulate interleaving; all of the examples of research situations were presented in blocked format. For a control condition, the subjects received just a description of a research situation, as they had in all of the previous experiments. In the experimental condition, in addition to the usual description of a research situation, the subjects were explicitly told the features that determined which statistical procedure to use. In addition, the instructions for the training were changed to reflect the emphasis on relevant features. The instructions for all of the subjects included an explanation of the importance of trying to recognize the relevant features of the research situation and of trying to associate the right features with the right statistical procedure. They were given a description of the category induction task used in Kornell and Bjork's (2008) painting styles study to use as an example of how to induce categories. However, only the description plus features subjects were told what the relevant features were, and only during their training task, not during the instructions. The description only subjects, on the other hand, were not told the relevant features during the training, but, instead, were encouraged to figure them out and to write them down.

Across all three tests, the description plus features subjects performed slightly better than the description only subjects, but this main effect did not reach significance at the .05 level (p = . 088). There was a main effect of time of test with performance dropping from the 50% to 60% range on the immediate test to the 30% to 40% range on the early retention test, and rising back to the 50% to 60% range on the last test.

In the fourth experiment, at least on the final test, performance for the first time rose to a level that would be at least passing in an actual classroom, although it would be a D, and it did so without interleaving the training examples. This led us to the experiment that we report here, in which we incorporated many of the beneficial characteristics of the previous experiments, and also manipulated interleaving.

Method

Participants

The subjects were 164 college students in an introductory statistics course. There were 118 who reported their gender as female, 41 as male, and 5 who did not report their gender. There were 2 freshmen, 38 sophomores, 77 juniors, and 41 seniors. Six did not report their year in college. They were in eight different sections of the course, with a mean class size of 20.5. Class sizes ranged from 17 to 25.

Design and Conditions

This experiment was a 2 by 2 completely randomized design. We called the first factor information. One group, called the description-only group, received instructions for their training that emphasized that they should use the examples to learn to associate each type of research situation with the appropriate statistical procedure. However, the instructions said nothing about features of each research situation (see Appendix C for the training instructions for the description-only group). Their training examples were the same as those used in Experiment 4 (see Appendix A). They consisted only of the paragraph describing the research situation along with the appropriate statistical procedure.

The other group, called the description-plus-features group, received training that emphasized the importance of learning to recognize what features of the research situation determined the correct statistical procedure to use. The Kornell and Bjork (2008) painting styles study was used an example of how to do the task. The instructions explained that the descriptions of the research situations would provide them with the critical features (see Appendix D). Their training examples consisted of the paragraph describing the research situation along with a statement of what the critical features were and which statistical procedure was appropriate for those features (see Appendix E).

The second factor was interleaving. The blocked group received all four of the descriptions of one type of research situation consecutively before receiving all four of the next type, and so on. We counterbalanced the order of the blocks in a Latin Square design. The interleaving group received their descriptions interleaved in a within subjects randomized blocks design. The randomized blocks were created so that each of the four types of research situation occurred once, but in a semi-random order, in each block before appearing again in the next block. Thus, within a block, the same type of research situation never followed itself. Also, the semi-random ordering within blocks was constrained to the extent that the same type of research situation never followed itself by being the last member of one block and the first member of the next block. Thus, for the interleaved subjects, after receiving a description of one type of research situation, they always received a different type next. All the subjects in the interleaved group received the same order of interleaved descriptions.

Materials

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

Training. The training booklet contained 16 descriptions of research situations. There was one description on each page about a quarter of a page in length. There were four types of research situations and there were four examples of each type. Each type of research situation required a certain statistical procedure. The four statistical procedures were the independent *t*-test, the repeated measures *t*-test, the independent measures ANOVA, and the repeated measures ANOVA. Each paragraph was labeled at the top to indicate the correct statistical procedure for that example. At the end of the example, the description stated what procedure the researcher used in the study (see Appendix A for an example of the descriptions).

Tests. We used an immediate test, a retention test, and a final test. The immediate test had five 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. Below the test item were the four statistical procedures from which to choose. Finally, there were instructions for the subject to indicate whether they had just guessed, and, if they thought they knew the correct answer, to try to indicate what features of the research situation enabled them to select their answer (see Appendix F for an example of the test items). Since there were five items in the test and only four choices for the correct answer, 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 retention test and the final test 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 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 final test occurred at the end of the semester after all the formal instruction had been provided.

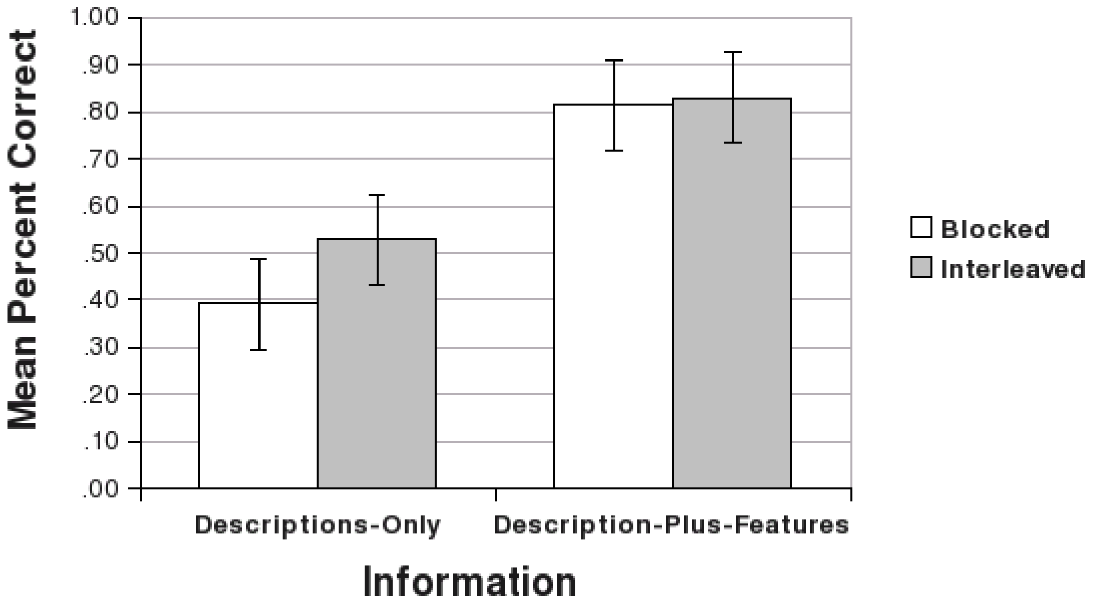
Training. The subjects were not timed. In some of the previous studies in Ryan et al. 2010, we had instructed the subjects to study each training example for one minute. However, when we saw the low performance in the first studies we became concerned that maybe one factor contributing to that low performance was that sometimes a subject finished studying an example before other subjects and their mind wandered while they were waiting for the others to finish. Therefore, we were now allowing the subjects to study at their own pace. We instructed them to try to study enough so that they learned which procedure goes with which research situation, but not so much that they got bored or frustrated. We instructed them to study the paragraphs in the order they were presented. We asked them to not move on to a new paragraph until they were finished studying the one they were on, and to not look back at any previous paragraphs. Finally, we instructed them that it was not a problem if they went through the pages at a faster or slower pace than someone else. Rather, we told them that it was important that they move at a pace that was comfortable for them. We told them that if they finished earlier than some others, they should just wait for the others to finish. We told them that if they were taking longer to finish than others, they should not feel that they had to hurry to get done.

Tests. The instructions for all of the tests were the same. We instructed the subjects to read every paragraph carefully and to select the statistical procedure they thought was correct. We told them to answer all the questions on the test even if they had to guess. We told them to work through all the items in order and not to go back to any previous items. The tests were not timed.

Results

First, we analyzed the results from each of the tests separately with a two factor, between subjects ANOVA. We used each subject's percent correct as the dependent measure and information and interleaving as the two factors.

Figure 1. Mean Percent Correct on the Immediate Test

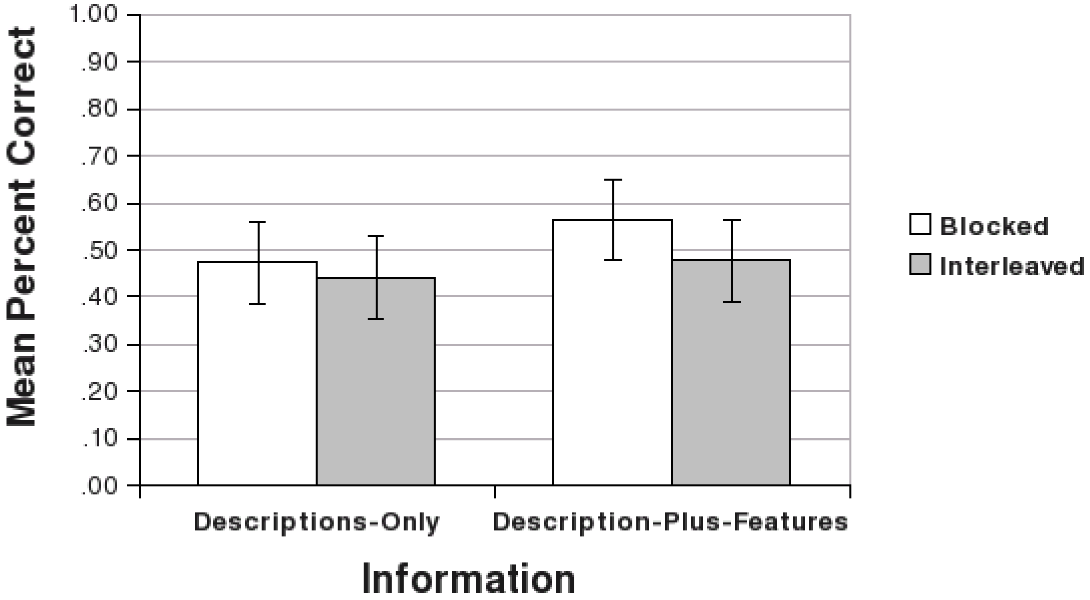


Note: Error bars are MSE

Immediate Test

As shown in Figure 1, on the immediate test, there was a large benefit of adding the features to the descriptions, $F(1, 160) = 55.19$, $p < .001$, $\eta^2 = .26$. There was a slight, but not significant, benefit of interleaving, $F(1, 160) = 2.44$, $p = .12$, $\eta^2 = .02$. The information by interleaving interaction was not significant, $F(1, 160) = 1.58$, $p = .21$, $\eta^2 = .01$.s A separate contrast for the interleaving advantage in the descriptions-only condition was significant, *Scheffe's F* (1, 160) = 4.23, $p = .0412$, $\eta^2 = .03$, although it should be remembered that doing such a test lacked the justification of an information by interleaving interaction.

Figure 2. Mean Percent Correct on the Retention Test

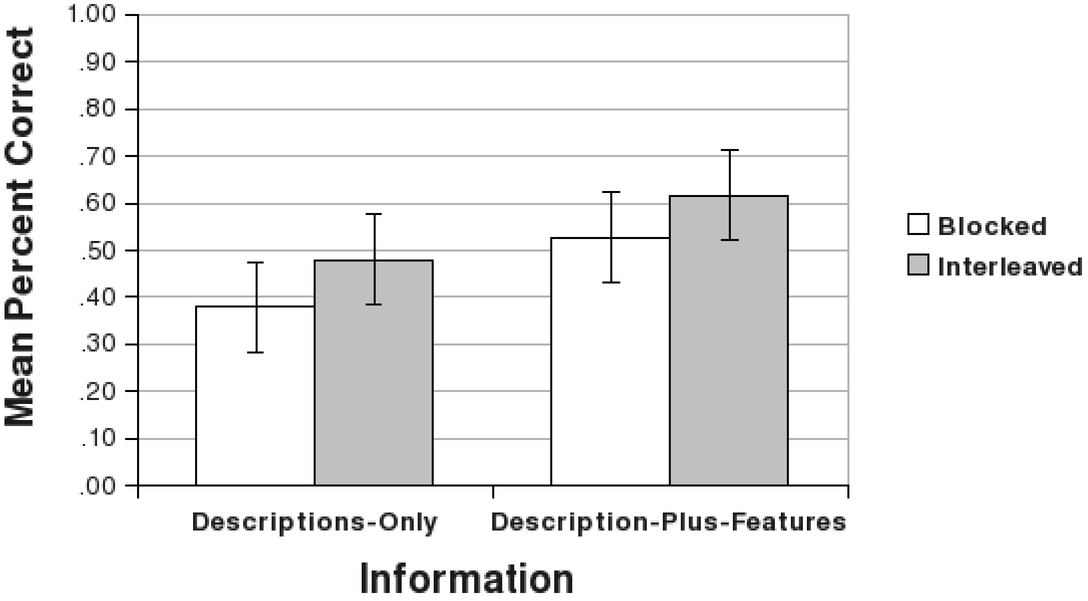


Note: Error bars are MSE

Retention Test

As shown in Figure 2, the advantage for adding features seen on the immediate test was not retained. There were no significant main effects or interactions on the retention test.

Figure 3. Mean Percent Correct on the Final Test



Note: Error bars are MSE

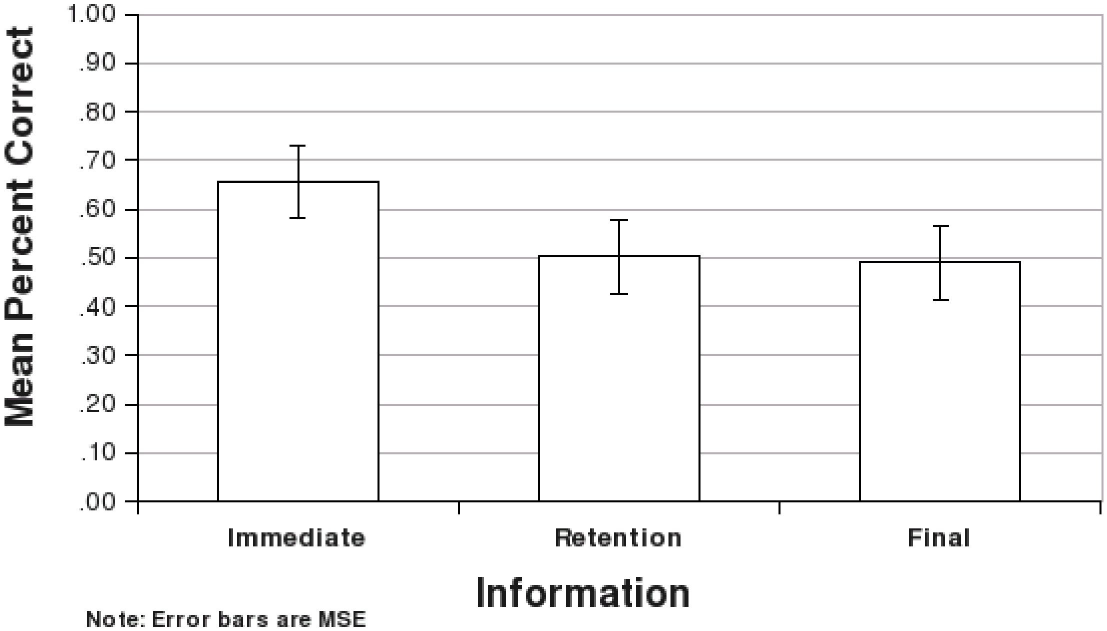
Final Test

As shown in Figure 3, there were at least numerical advantages for both adding features to the descriptions, and interleaving. The advantage of features was significant, $F(1, 121) = 6.43$, $p = .012$, $\eta^2 = .05$. The interleaving advantage, although it was about 9 or 10 percentage points, was not quite significant, $F(1, 121) = 2.90$, $p = .091$, $\eta^2 = .02$. There was no interaction, $F < 1$.

Overall ANOVA

We conducted a mixed model overall ANOVA using time of test as a within subjects factor and information and interleaving as between subjects factors.

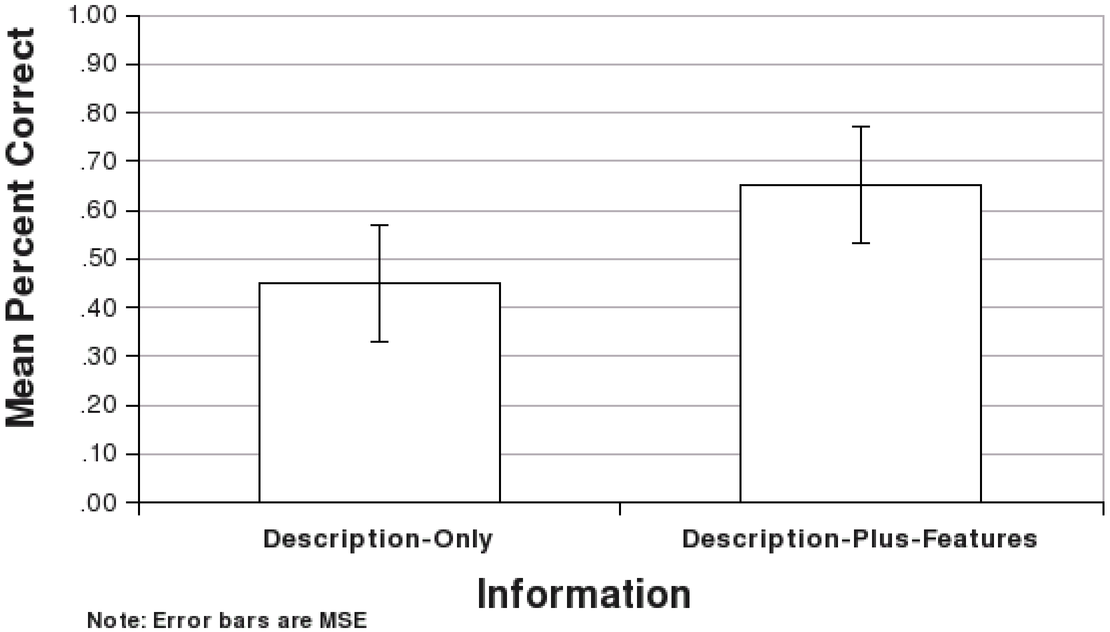
Figure 4. Mean Percent Correct Across Tests



Note: Error bars are MSE

Time of test. As shown in Figure 4, the overall ANOVA revealed a main effect of time, $F(2, 220) = 12.76$, $p < .001$, $\eta^2 = .10$. Separate ANOVA's showed that performance was better on the immediate test than on the retention test, $F(1, 138) = 18.25$, $p < .001$, $\eta^2 = .12$, and better on the immediate test than on the final test $F(1, 117) = 26.74$, $p < .001$, $\eta^2 = .19$. There was no significant difference in performance between the retention test and the final test. $F < 1$.

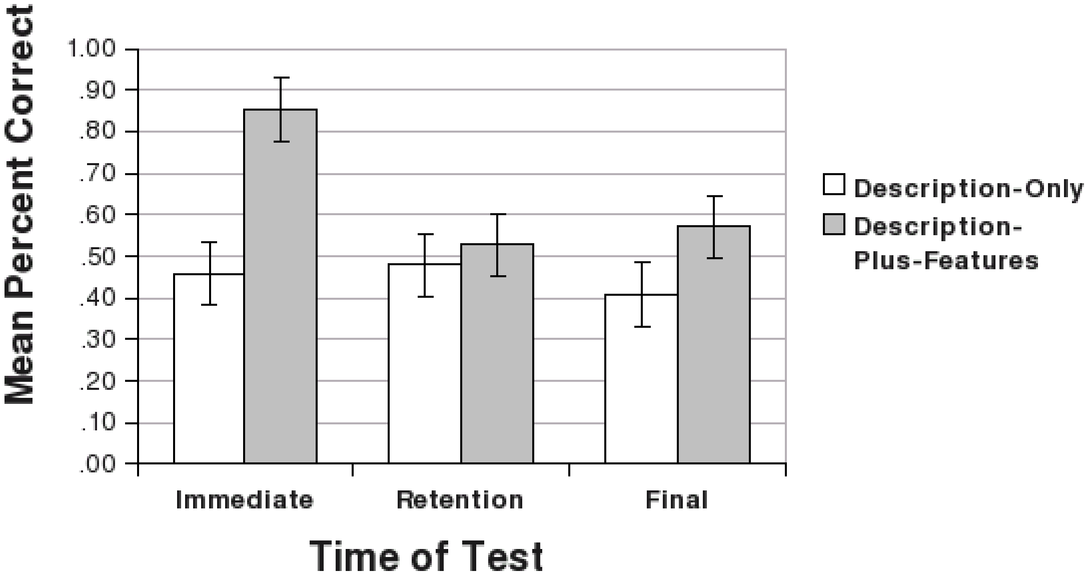
Figure 5. Mean % Correct as a Function of Information



Note: Error bars are MSE

Overall advantage of features. As shown in Figure 5, there was an overall advantage of having the features presented along with the descriptions, $F(1, 110) = 29.21$, $p < .001$, $\eta^2 = .21$.

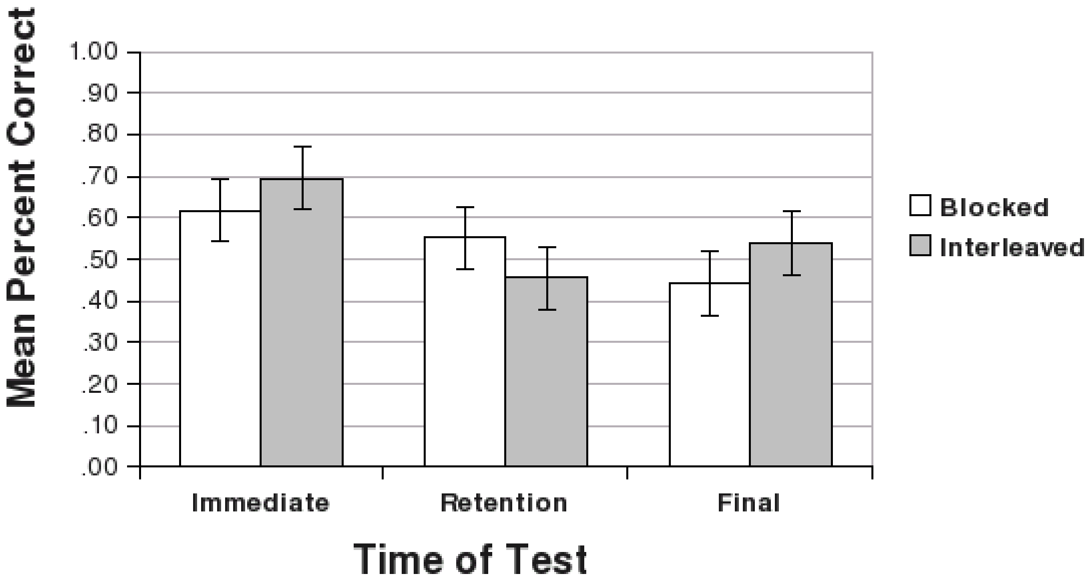
Figure 6. Mean % Correct as a Function of Time And Information



Note: Error bars are MSE

Overall time by information interaction. As shown in Figure 6, there was a time by information interaction in which adding the features to the descriptions was beneficial on the immediate test, but not on the retention test or the final test, $F(2, 220) = 11.68$, $p < .001$, $\eta^2 = .10$.

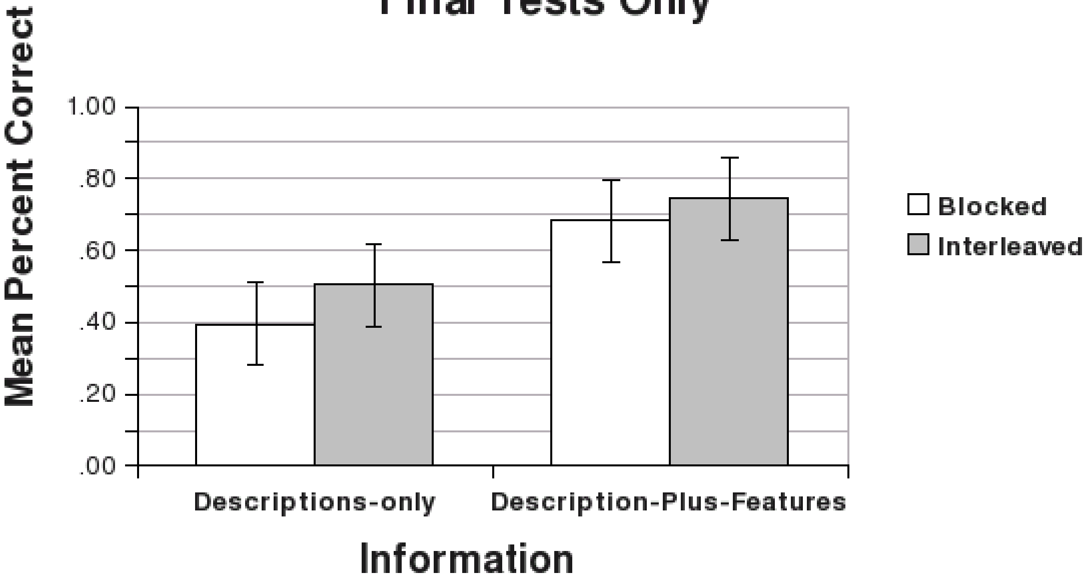
Figure 7. Mean % Correct as a Function of Time And Interleaving



Note: Error bars are MSE

Overall time by interleaving interaction. As shown in Figure 7, there was a time by interleaving interaction in which there was an interleaving advantage for the immediate and final tests, but not for the retention test, $F(2, 220) = 4.13$, $p = .017$, $\eta^2 = .04$.

Figure 8. Mean % Correct as a Function of Information And Interleaving, Collapsed Across Immediate and Final Tests Only



Note: Error bars are MSE

Main effects for just the immediate and final tests combined. The time by interleaving interaction shown in Figure 7 suggested that there might be a main effect of interleaving if we considered just the immediate and final tests. Therefore, we conducted a separate mixed ANOVA using only the immediate and final tests as a within subjects time of test factor, and information and interleaving as between subjects factors. Similar to the ANOVA using all three tests as a time of test factor, this ANOVA also revealed a main effect of time, and a time by information interaction, $F(1, 117) = 26.74$, $p < .001$, $\eta^2 = .19$ and $F(1, 117) = 9.86$, $p = .002$, $\eta^2 = .08$ respectively. Also, as shown in Figure 8, there was a significant benefit of features, $F(1, 117) = 36.67$, $p < .001$, $\eta^2 = .24$ and there was an almost significant benefit of interleaving, $F(1, 117) = 3.78$, p = .054, $\eta^2 = .03$.

Discussion

The current study, like the four that preceded it in Ryan et al. (2010), did not provide convincing evidence of an interleaving advantage for learning and retaining the knowledge of which statistical procedure to use for various research situations. The interleaving advantage seen on the retention test in the first experiment in Ryan et al. (2010) did not replicate on any of the subsequent experiments in that study, nor did it replicate in the current study. Thus, it is probably best to conclude that that result was a Type 1 error.

In the current study, there was at least some evidence of an interleaving advantage, but on the immediate test, not the retention test. Among the subjects who had to infer the critical features for distinguishing the categories rather than having them provided to them, although their performance was much worse than those who were given the features, it was better if their examples were interleaved. However, this finding has to be interpreted cautiously, because the evidence came from a separate contrast that was not justified by an interaction, and is therefore subject to the criticism of coming from unjustified multiple testing.

The final test showed a small numerical advantage of interleaving for benefiting from the formal instruction, but it fell short of being statistically significant at the .05 level. Finally, an overall analysis also showed a small numerical advantage of interleaving. However, it was revealed only by collapsing over the immediate and final test, thus ignoring the retention test, and it also fell short of the .05 significance level.

In each of the cases cited above where there was at least weak evidence of an advantage of interleaving, there was a larger, and statistically significant, advantage of providing the subjects with the critical features. Thus, the difficulty of learning which statistical procedure to use in which research situation may lie in being able to recognize what it is about a research situation that determines what statistical procedure applies. Learning to associate the features with the correct procedure, on the other hand, may be relatively easier. Future research should directly test that question.

Why doesn't interleaving have a stronger effect on students' ability to learn the relevant features of research situations, and to retain that knowledge, when it has been shown to be beneficial for other kinds of category learning tasks? Perhaps it is because the features of the statistics examples are conceptual, rather than perceptual. The students needed to recognize that some research situations have only two treatment conditions, whereas other have more than two. The students may have lacked sufficient prior knowledge about research design to have a well developed concept of even what is meant by treatment conditions. Thus, they may not have recognized examples of them. Even if they had, without a fairly well developed understanding of research design it may not have occurred to them that the number of treatment conditions had any relevance.

The idea of treatment conditions is one that is fairly restricted to the domain of research, with which the students have not had a lot of experience at this point. Even more so, the concept of between versus within subjects treatments is one that would be very unfamiliar to these students. Even if they had encountered the basics of research design in a prior course, such as introductory statistics, they may have only been taught about experimental and control conditions at a general level. They may not have been taught the more detailed information that conditions can be manipulated either between or within subjects. Even if they had, such information would be so removed from their everyday experience that it may have been poorly understood and easily forgotten. The lack of such prior knowledge would make it very difficult to recognize the abstract idea of manipulating conditions between versus within subjects from the concrete examples of the research situations.

The lack of relevant prior knowledge may also explain why, even when provided with the features so that they did not have to infer them, students benefitted only on an immediate test. Having nothing to connect these newly encountered concepts to, they were not able to retain them over the four to six weeks between the immediate and retention tests.

Expecting students who lack relevant prior knowledge to infer that knowledge from examples might be asking a lot. However, one possible way to boost the effectiveness of interleaving examples in statistics might be to use feedback. One way to do that could be to have the students try to initially guess the correct statistical procedure for each training example, and then provide them with the correct procedure. This method could encourage students to look for the characteristics of the example that determined the correct procedure in order to increase their chances of selecting the correct procedure on future examples. Looking for those characteristics might enable them to begin forming concepts such as treatment conditions, between subjects manipulations, and within subjects manipulations.

Forming concepts by using feedback may be a way to at least improve immediate acquisition of the knowledge of what statistical procedure to use in what situation. However, retaining that knowledge is another matter. The studies reported here showed that even factors that improved acquisition, did not improve retention. This may be another example of the idea proposed by Schmidt and Bjork (1992), that making acquisition of newly learned material easier does not necessarily help, and may in fact actually hinder, later retention. The converse is that making acquisition difficult often improves retention. Trying to learn from feedback may be more difficult when examples are interleaved than when they are blocked. Therefore, with feedback, there may be an advantage of interleaving for retention. Future studies should examine that possibility.

However, even providing feedback may not be sufficient to enable students to form concepts such as treatment conditions and between subjects versus within subjects designs. It may be necessary to give students at least some rudimentary instructions on research design that goes beyond merely describing experimental and control conditions. It may be necessary to provide direct instruction in research design that includes the ideas of only two versus more than two treatment conditions, and the idea of using either different subjects in each treatment condition or the same subjects in all treatments. Once students have been provided with the relevant prior knowledge, then a training task that involves both interleaving examples, and providing feedback may activate that prior knowledge and enable students to connect it to the descriptions of the research situations and their appropriate statistical procedures. Conducting studies to examine these issues will inform decisions about instructional methods when teaching introductory statistics at the college level.

References

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “Enemy of Induction”? *Psychological Science, 19*(6), 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: An International Journal of the Learning Sciences*, *35*(6), 481-498.

Ryan, R. S., Howell, S. R., Shaw, H. A., Kappus, D. W., Wilde, M. E., & Crist, S. L. (2010, October). *Using interleaved examples to teach inferential statistics*. Paper presented at the 1st Annual PASSHE Potluck Psychology Conference, Indana, PA.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207-217. doi:10.1111/j.1467-9280.1992.tb00029.x

Appendix A

An Example of the Descriptions of Research Situations From the Training Materials for Experiment 1 and 2 in Ryan et al. 2010

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.

An Example of the Descriptions of Research Situations From the Training Materials for Experiments 3 and 4 in Ryan et al. 2010

Appropriate statistical procedure: 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 relaxed 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.

Appendix B

An Example of the Test Items from Experiment 1 in Ryan et al. 2010

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

The Instructions for Training for the Description-Only Subjects

Different types of research situations call for different statistical procedures. Statistics students need to learn to recognize the different types of research situations and the correct statistical procedure to use in each of the different kinds of situations.

In order to learn how to recognize the different types of research situations and the correct statistical procedure to use, it is helpful to study examples. In this experiment, you will be given training in which you will spend some time studying such examples. Specifically, you will be given 16 examples to study. Each example is in the form of a short paragraph describing a research study. The paragraph will include the name of the correct statistical test to use in that particular type of research situation. There will be four different kinds of research situations, and there will be four examples of each one. Your job will be to try to learn which statistical test goes with which research situation.

After you study, you will be given a test. The test will consist of examples similar to the ones that you studied. The examples will again be in the form of a short paragraph describing a research situation. It will be a multiple choice test. Your job will be to select the correct statistical test to go with the type of research situation described in the paragraph.

- Study the example of a type of research situation described in each paragraph.
- Notice what the appropriate statistical procedure is.
- Try to associate that type of research situation with the appropriate statistical procedure.

Appendix D

The Instructions for Training for the Description-Plus-Features Subjects

Different types of research situations call for different statistical procedures. Statistics students need to learn to recognize the features of the different types of research situations that determine which type it is, which in turn tells them which statistical procedure to use.

In order to understand how to recognize features, consider the example of people trying to learn to recognize paintings by the artist's style. To do that, they would have to notice the features of the style. For example, they would have to notice whether the brush strokes were short or long, whether the colors were bright or dark, and so on. Then, they would have to associate those features with that painter. Later, if they encountered a new painting, they could notice the features, and, if they could remember which artist's style had those features, then they could name the artist, even though they had never seen that painting before.

The examples of different types of research situations you are about to study will tell you the features to notice, and they will tell you what statistical procedure to use. Try to associate the features with the statistical procedure so that when you are tested with new examples you will be able to recognize the features and therefore to identify the correct statistical procedure to use.

- Study the example of a type of research situation described in each paragraph.
- Notice what the appropriate statistical procedure is.
- Try to associate that type of research situation with the appropriate statistical procedure.

Appendix E

An Example of the Descriptions of Research Situations for the Description-Plus-Features Subjects

Appropriate statistical procedure: 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 relaxed 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.

Features:

This situation calls for a *t* test because there were only two groups of test scores, not three or more groups.

It calls for an Independent-measures test because each group of scores came from a different group of subjects.

Appendix F

Example of a Test Item

A group of researchers wanted to determine whether people have better comprehension for stories that they hear verbally or stories that they read. Twenty subjects performed the following procedure. Each subject listened to a recorded voice narrate a brief story. Then they were given a comprehension test to see how much detail about the story they could remember. Next, they read a very similar story printed on a sheet of paper and were tested for their comprehension of that story. The researchers calculated the average comprehension score for the story that the subjects had listened to and for the story that the subjects had read. To determine whether the average comprehension scores were significantly different, the researchers calculated a _____________.

a. Independent-measures *t* test

b. Repeated-measures *t* test

c. Independent-measures ANOVA

d. Repeated-measures ANOVA

“Please indicate below how you made your choice. Indicate if you just guessed. If there were some features of the research situation that enabled you to make your choice, then indicate what the features were.”