Feedback And Interleaved Examples Improve Category Induction in Statistics

Robert S. Ryan
Steven R. Howell

Kutztown University

Presented at the 24th Annual Convention of the Association For Psychological Science, Chicago, IL, May 24 – May 27, 2012.

Correspondence:

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

Feedback And Interleaved Examples Improve Category

Induction in Statistics

People sometimes learn better from examples if the examples are presented in an interleaved, rather than a blocked, format. However, several previous studies using statistics examples failed to replicate the relative interleaving advantage, and also resulted in a very poor absolute level of performance overall. On the other hand, in two of those studies, explicitly presenting the defining features of the examples did raise performance (Ryan et al., 2011). In the present study, we again made the defining features explicit, but in the form of immediate feedback. Crossing interleaving with feedback resulted in an advantage of both interleaving and feedback on an immediate test and an advantage of interleaving on a delayed retention test. Also, interleaving was significantly more effective with feedback than without it. Finally, there was also a better absolute level of performance for the interleavers on the immediate test, but not on the delayed test. Future research should examine whether factors such as delayed feedback, retrieval practice, and the spacing effect, which have been shown to improve retention in other domains, might do so also in statistics.

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.

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 applies 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 were more perceptual. It is not clear whether training by interleaving rather than blocking examples of those conceptually defined categories would 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, Howell, and Shaw (2010) with a fifth reported in Ryan, Howell, Kappus, and Wilde (2011). We begin by summarizing those studies, and then we present the study that followed them, which is the focus of this paper.

Ryan and his colleagues 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 (Ryan et al., 2010; Ryan et al., 2011). 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; 2011) 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. The tests consisted of several new items similar to the training items, but without the appropriate statistical procedure indicated. The subjects' test task was to select the correct statistical procedure for the research situation described from among several alternatives. The early and late retention tests were the same as the immediate test but with different examples.

In the first experiment in Ryan et al. (2010), 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, five more experiments, for a total of six 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 training 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”. 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, the 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, during the training, the subjects received just a description of a research situation, as they had in all of the previous experiments. In the experimental condition, during the training, in addition to the usual description of a research situation, the subjects were explicitly told the features that determined which

statistical procedure to use. Importantly, however, the instructions given to *all* the subjects *before* 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 pre-training 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 immediate test and on the final test (after formal training) 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 to the fifth experiment (Ryan et al., 2011) which incorporated many of the beneficial characteristics of the fourth experiment, strengthened the feature-providing manipulation, and also manipulated interleaving.

In the fifth experiment (Ryan et al., 2011) a 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, unlike the description-only control group in the fourth experiment, the instructions said nothing about features of each research situation and the instructions did not give them 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. For this group, the descriptions of the research situations did not provide them with the critical features.

A description-plus-features group, received training that not only emphasized that they should use the examples to learn to associate each type of research situation with the appropriate statistical procedure, but also emphasized the importance of learning to recognize what *features* of the research situation determined the correct statistical procedure to use. Furthermore, for this group, the Kornell and Bjork (2008) painting styles study *was* used an example of how to do the task. Finally, for this group the descriptions of the research situations provided them with the critical features. This features factor was crossed with the interleaving factor that had been used in the first three experiments.

On the immediate test, there was a large and significant ($p < .001$) advantage for the description-plus-features group. Those subjects achieved the highest performance we had seen so far, a little above 80% correct. However, there was only a small and not significant ($p = .12$) advantage for interleaving. There were no significant effects on the retention test. On the final test there was again an advantage for the features group that was significant ($p = .012$) but it was smaller than it had been on the immediate test. And, again, there was a small, but not significant ($p = .091$) advantage for interleaving.

Across all five experiments we saw the best performance from the subjects who were either told to try to learn the defining features, were given the defining features, or both. In the fourth experiment, actually giving subjects the features led to the highest immediate test performance we had seen up to that point *(M* = 61%), but which was not significantly better at the .05 alpha level when compared to the performance of a control condition (*M* = 50%) in which, although the subjects were not actually given the defining features, they were at least told about their importance. In the fifth experiment, the description-plus-features subjects, who were both told about the importance of the features and actually given the features, performed significantly better on the immediate test (*M* = 82%) than the description-only control subjects (*M* = 46%) , who were neither given the features, nor even told about their importance.

In the fifth experiment, unlike the fourth, in which there was no interleaving manipulation, the aforementioned advantage for the features condition was collapsed across interleaved and blocked presentation. However, breaking down the means by both the features and the interleaving factor reveals that, although there was no advantage of interleaving for the features subjects (83% for interleaved versus 81% for blocked), there was an interleaving advantage for the description-only subjects (53% for interleaved versus 39% for blocked). Although examining this interleaving advantage for the description only subjects separately was not justified by a significant interaction, nevertheless a *Scheffe's F* showed it to be significant at *p* = .041.

The interleaving advantage for the description only subjects in the fifth experiment was the second instance in which an interleaving advantage was found. It is notable that in this case, as in the first instance (on the immediate test in the first experiment) the effect looked more like poor performance for the blocked subjects than exceptionally good performance for the interleaving subjects. It is not hard to see how blocked presentation might have the disadvantage that once subjects catch on to the fact that the examples they are going to see are similar to the ones they have just seen, they may begin to allow their attention and task engagement to wane. This suggests that it might be best to think of interleaving as a factor that prevents the detrimental effect of blocked presentation by encouraging the subject to be more engaged in the learning task because each example that they see is different from the one they just saw.

If this is the case, then this positive effect of interleaving might be especially useful if, during training, the subjects had to generate the correct statistical test for the given research situation, followed by receiving feedback (Doug Rohrer, personal communication, 10/04/10). In a blocked presentation, after receiving feedback on the first of several items of a given type about both the correct statistical test and the defining features, the subjects would know which statistical test to generate for several items, until the type of item changed. Thus, they would not have to pay much attention to what the defining features were. But in an interleaved presentation, they would be forced to try to learn which statistical test to generate by learning the defining features because they would not know what type of item was coming next. Therefore, in the experiment reported here, we crossed the interleaving factor with whether or not we required the subject to generate the correct statistical test followed by feedback about both the correct statistical test and the defining features.

Method

Participants

The subjects were 326 college students in an introductory statistics course. There were 238 who reported their gender as female, 61 as male, and 27 who did not report their gender. There were 29 who

reported their year in college as freshmen, 63 as sophomores, and 124 as juniors. There were 70 who reported their year in college as seniors, with 60 reporting that they were in year 4, eight in year 5, and two in year 6. There were 39 who did not report a year in college. Across two semesters (Spring 2011 and Fall 2011), there were 14 sections of Statistics ranging in size from 18 to 26 students per section, with an average of 23. There were 35 subjects who did not report their age. The age of the majority (266) of the reporting subjects was between 18 and 23, with a few (25) varying degrees older. Their mean age was 21.42 with a standard deviation of 4.95, the median and modal age was 20, the minimum age was 18, and the maximum 58.

Design and Conditions

The experiment was a 2 by 2 by 3 design. The first factor was interleaving. The second factor was feedback. The feedback factor actually manipulated several of the factors that had been effective for raising performance in the previous studies. Both of those factors were manipulated between subjects. The third factor, which was within subjects, was time of test.

Materials

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

Training. For all subjects, 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.

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.

Both the blocked and interleaved groups were randomly subdivided into no-feedback and feedback groups. The no-feedback group received instructions for their training that only emphasized that they should use the examples to learn to associate each type of research situation with the appropriate statistical procedure. However, their instructions said nothing about features of each research situation and the instructions did not give them 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. Each training example was labeled with the appropriate statistical procedure at the top, and at the end of the description the name of the correct statistical procedure was re-iterated as the one that the researchers used. Therefore, the subject did not have to guess the correct statistical procedure. After

each example, the subject simply moved on to the next example without receiving any feedback (Appendix A shows the instructions and a training example for the no-feedback condition).

The feedback group received instructions for their training that not only emphasized that they should use the examples to learn to associate each type of research situation with the appropriate statistical procedure, but also emphasized the importance of learning to recognize what *features* of the research situation determined the correct statistical procedure to use. Furthermore, for this group, the Kornell and Bjork (2008) painting styles study *was* used an example of how to do the task. Their training examples were not labeled with the appropriate statistical procedure at the top. At the end of the descriptions, the name of the correct statistical procedure was left blank and the name of the four possible statistical procedures were listed. The subject was instructed to select whatever they believed was the correct statistical procedure. Of course, they would be guessing on the first one. But feedback was provided after each example. The feedback provided the example again, but this time with both the name of the correct statistical procedure and an explanation of what features of the research situation determined which procedure was correct. For the feedback group, the examples were printed on only one side of the pages with a masking page in between to prevent the subject from seeing the feedback through the page before they made their response (Appendix B shows the instructions, a training example, and the feedback for the feedback condition).

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 C 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, in the later studies, we allowed the subjects to study at their own pace.

The manipulation of the interleaving factor was accomplished by the order in which the training examples were presented. However, the manipulation of the feedback factor was accomplished by the instructions for the training as well as by the training examples themselves. However, because the

subjects were randomly assigned to the four conditions within each class and it was important for the subjects to be unaware of the instructions for the condition to which they were not assigned, we could not read the instructions aloud. Therefore, we emphasized to the subjects that different subjects had different instructions and that it was very important that they read them very carefully. If they had questions, they were instructed to call the experimenter over to them to ask there question and receive their answer privately.

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 pointed out the instruction that said that if there were some features of the research situation that enabled them to make their choice, then they should try to indicate what the features were. 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.

As in the five preceding studies, the immediate test was right after the training, which occurred at the beginning of the semester. The retention test was in about the fourth week of the semester, before formal instruction on the statistical procedures used in the experiment. The final test occurred at the end of the semester, after all the formal instruction.

Results

Across all three tests, there was a main effect of interleaving, $F(1, 285) = 17.40$, $p < .001$, $\eta^2 = .058$. The main effect of feedback was almost significant, $F(1, 285) = 3.75$, $p = .054$, $\eta^2 = .013$. There was also a large effect of test time, with performance declining linearly over the three tests, $F(2, 570) = 83.65$, $p < .001$, $\eta^2 = .227$. However, as shown in Figure 1, there was a three way interaction between feedback, interleaving, and time of test, $F(2, 570) = 3.60$, $p = .028$, $\eta^2 = .012$. All eta squares were calculated as the SS for the effect divided by the SS for the effect plus the SS for the error for that effect.

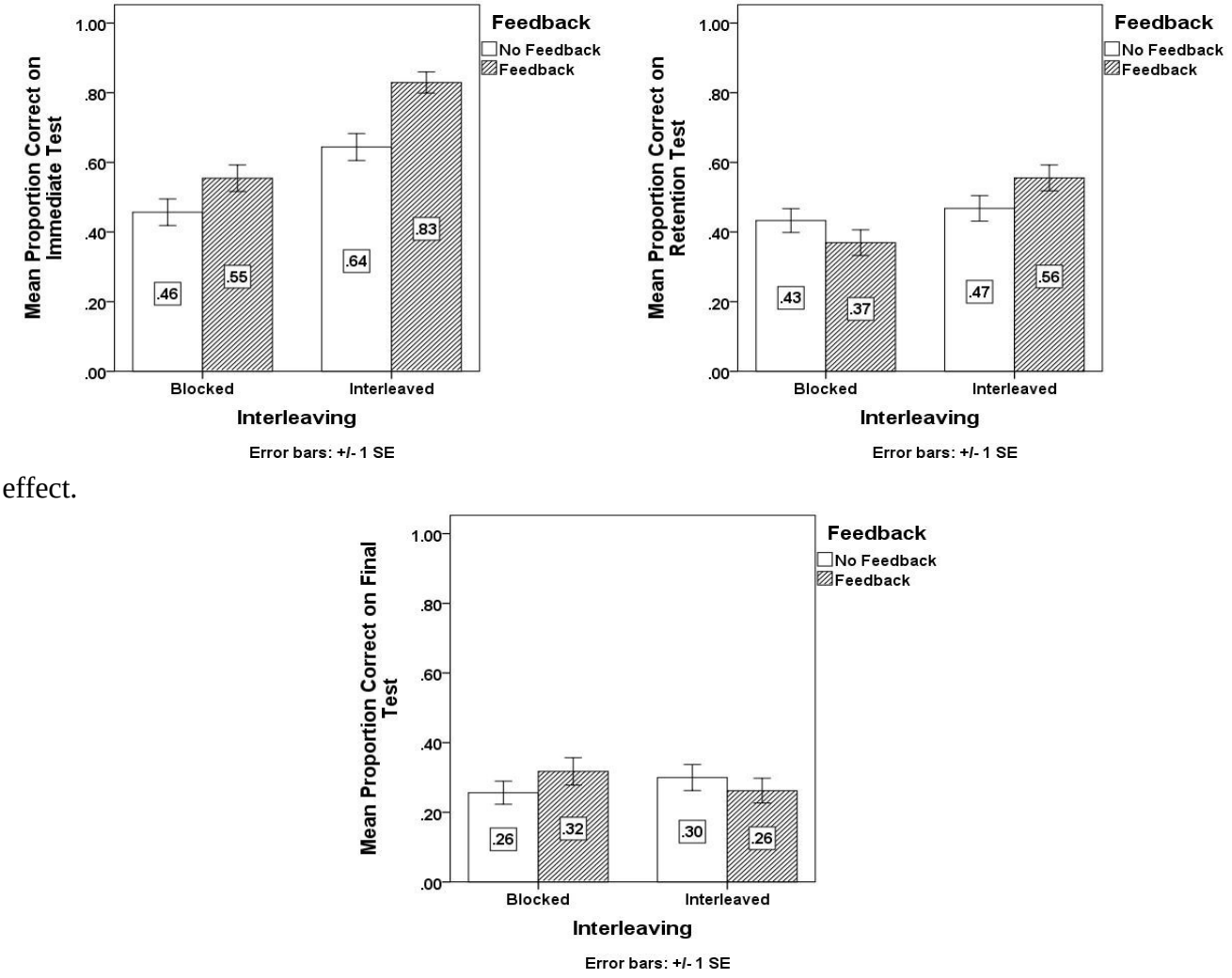

Figure 1. *Percent Correct on the Immediate, Retention, And Final Tests as a Function of Feedback and Interleaving.*

There was also a two way interaction between interleaving and time of test. There was a positive effect of interleaving on both the immediate and retention tests, but not on the final test, $F(2, 570) = 11.73$, $p < .001$, $\eta^2 = .040$. The interaction between feedback and time of test was almost significant. There was a positive effect of feedback on the immediate test, but not on the retention test or the final test, $F(2, 570) = 2.92$, $p < .055$, $\eta^2 = .010$. Over all three tests, there was no interaction between feedback and interleaving, $F(1, 285) = 1.12$, $p > .05$.

In order to further examine the three way interaction, we conducted separate anovas on each test. On the immediate test, there was a positive effect of both interleaving and of feedback, $F(1, 322) = 40.04$, $p < .001$, $\eta^2 = .111$, and $F(1, 322) = 15.00$, $p < .001$, $\eta^2 = .045$, respectively. The effect of interleaving was numerically greater with feedback, although this interaction did not reach significance, $F(1, 322) = 1.43$, $p > .05$.

On the retention test, there was an interaction between interleaving and feedback in which there was a positive effect of interleaving with feedback, but not without it, $F(1, 286) = 4.33$, $p < .038$, $\eta^2 = .015$. There was also a main effect of interleaving, but not of feedback, $F(1, 286) = 9.53$, $p < .003$, $\eta^2 = .031$, and $F(1, 286) < 1$, respectively. On the final test, there were no main effects or interactions.

Finally, given that on the immediate test the effect of interleaving was numerically greater with feedback than without it, and on the retention test the same pattern emerged, but as a significant interaction, we did one final anova on just the immediate and retention tests combined. In that anova, for the immediate and retention test combined, the greater effect of interleaving with feedback than without it did produce a significant interaction, $F(1, 286) = 4.52$, $p < .034$, $\eta^2 = .016$.

Discussion

As predicted, adding immediate feedback during training did result in strengthening the positive effect of interleaving. This pattern was seen at a significant level on the immediate and retention tests combined and on the retention test alone, and it was seen numerically on the immediate test alone.

Combining feedback with interleaving resulted in the highest performance seen so far in this series of experiments, at least on the immediate test. All of the interleavers would have received at least a passing grade (above 60%), and the interleavers with feedback would have received a B (83%).

As with the previous experiments, however, the story is different for retention. As before, all of the subjects would have received a failing grade for their retention after four weeks. What is even more distressing is that the performance on the final test, which occurred after formal instruction on the learning materials, was the worst that it has been out of all of the experiments so far. Schmidt and Bjork (1992) argued that instructional manipulations that improve retention are the very ones that impair later retention and transfer. This experiment suggests that the factors that we used to improve initial acquisition not only did not carry over to a four week retention test, but may also have actually impaired the retention of formal instruction outside of the experiment.

This series of studies has illustrated the exceptional difficulty for students of acquiring and retaining the kind of conceptual categories required for learning in statistics. There are many possible reasons why learning in statistics is so hard for students. Not only is the material very abstract, but also the abstraction is higher order. For example, consider trying to learn the properties of a probability distribution. Properties such as variability are not only abstract ideas, but they are properties of a mathematical object, a distribution, that is itself an abstraction. As another example of the difficulty,

dealing with probability distributions to do a hypothesis test requires reasoning about a hypothesis that one takes as true for purposes of the test, while trying to show that the evidence suggests that the hypothesis is not true. So the reasoning is not only hypothetical, but also hypothetically counterfactual.

Furthermore, even the concrete examples that instructors try to use to help students learn statistics are very unfamiliar. In our studies, we used examples in which we explicitly told the students, for example, that *t* tests are used for experiments with only two conditions, whereas an anova is used for an experiment with more than two conditions. We also told them that independent measures tests are used when different subjects participate in the conditions, whereas repeated-measures tests are used when the same subjects participate in all conditions. However, because our subjects may have had only a little exposure, if any, to the idea of experimental designs, they may have only poorly understood what we meant by conditions in an experiment.

Thus, it seems that in statistics, every possible cognitive factor that creates difficulty for students is at work. Therefore, it should not be surprising that in this domain, acquisition of concepts, and especially retention, would require bringing every possible *helpful* cognitive factor into play. Interleaving examples in itself was not sufficient. Interleaving with immediate feedback has, so far, been shown to improve acquisition. With feedback, interleaving has been shown to be helpful for retention compared to with no interleaving, but not very helpful compared to the performance at acquisition.

What would be required to produce retention at an academically acceptable level? There are other factors that have been shown to improve learning and retention in laboratory studies, and to some extent also in the classroom. Chief among them is practicing retrieval. Practicing retrieval can be afforded by extra testing, by delaying feedback, and by spacing of practice. Further research needs to examine how all of these factors can be built into instructional methods in order to improve learning and retention in the domain of statistics.

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*(6), 585–592. doi:10.1111/j.1467-9280.2008.02127.x

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

Ryan, R. S., Howell, S. R., Kappus, D. W., & Wilde, M. E. (2011, June). *The elusive interleaving effect: Why doesn't interleaving improve learning from examples in statistics?* Paper presented at the SARMAC IX, New York, NY.

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, Indiana, PA.

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.

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

The instructions and one training example for the no-feedback condition.

INSTRUCTIONS FOR THE TRAINING

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 A (cont.)

Example 1.

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

The instructions, one training example, and the feedback for the feedback condition (the actual materials were printed on the front side only and included a masking page between the training example and the feedback).

INSTRUCTIONS FOR THE TRAINING

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.

In this training, you will study examples of different types of research situations. Different types will have different features. Each type of research situation requires you to use a particular statistical procedure. For example the research situation may require you to use a statistical procedure called an “independent measures *t* test”, or a “repeated measures *t* test, or an “independent measures ANOVA”, or a “repeated measures ANOVA”.

After you study each example, you will be asked to indicate which of those four statistical procedures should be used. For the very first example you study, you will be totally guessing what features you should be looking for, and which statistical procedure to select. However, after you make your guess, you will be given feedback on the next page to tell you what the relevant features were, and which statistical procedure is to be used when a research situation has those features. Then you will study another example in which you will do the same steps. This will be repeated for a total of 16 examples.

Each time you study an example, you should look for the features, select a statistical procedure, and, when you get the feedback, you should try to learn from it. You should try to learn enough so that by the end of the training you can recognize what the relevant features of the research situation are and remember which statistical procedure to use in that research situation. Later you will be tested.

Appendix B (cont.)

Example 1.

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 a/an _____ .

a. Independent-measures *t* test

b. Repeated-measures *t* test

c. Independent-measures ANOVA

d. Repeated-measures ANOVA

Appendix B (cont.)

Feedback for Example 1.

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 C

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."