Feedback And Interleaved Examples Improve Category Induction in Statistics

Robert S. Ryan
Steven R. Howell*

Kutztown University

Correspondence:

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

* Steven R. Howell's current affiliation is Keystone College

Steven R. Howell
307 Ward Hall
Keystone College
College Rd.
La Plume, PA 18440
steven.howell@keystone.edu

Feedback And Interleaved Examples Improve Category Induction in Statistics

(Note: Kornell and Bjork's (2008) painting study is referred to in the method of Experiment 4, but is not described there. It had been described in the introduction that I had originally written for previous versions of this paper)

Experiment 1

The goal of Experiment 1 was to provide a simple, direct test of whether interleaving versus blocking examples during training would facilitate learning which statistical procedure to use in a given research situation. We tested immediately after training, and at two later times.

Method

Participants. The participants were university undergraduates in several sections of an introductory statistics course for the behavioral sciences in the Fall semester of 2008. There were 60 participants who completed the entire experiment. [1]

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

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

In the blocked condition ($N = 29$) the four examples of each different type of research situation occurred on consecutive pages to form each block. The blocks were presented in the same order for all participants. In the interleaved condition ($N = 31$) the examples were presented in a completely randomized order. The randomization resulted in one type of research situation, the one requiring the paired *t* test, occurring twice in a row one time. There were no other instances of any type of research situation occurring twice in a row.

Tests. We used an immediate test, a delayed retention test, and a final test. Each test had nine test items, although, as explained below, only six were scored. 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

1 In all experiments reported in this paper, in order to report all results in a consistent fashion, we report the results for only those participants who completed the entire experiment. This also results in the participants' being more similar to one another because they are the stronger students who did not drop out of the course or have a tendency to miss classes.

top. Also, at the end of the test item where the correct statistical procedure had been provided in the training booklet, there was a blank line (see Appendix B for an example of the test items).

The participants were given the six statistical procedures as possible answers from which to choose. Because there were nine items in the test and only six choices, the participants were instructed that some of the statistical procedures 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. In fact, in each test, every statistical procedure did occur as the correct answer at least once, and there was one procedure that occurred twice, and one that occurred three times. Which statistical procedure was doubled or tripled was different for the different tests, so that across the three tests, every statistical procedure was either doubled or tripled once. We scored only the first occurrence of any item that occurred more than once. The delayed test and the final test were the same as the immediate test but with different examples.

Procedure. At the beginning of the semester the students participated in the training session followed by the immediate test to assess their initial acquisition of the material. However, there was also to be a delayed test to assess their 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, not only the experimental training and immediate test, but also the delayed test, which was 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 examine whether the training method affected how much the participants benefited from the formal classroom instruction.

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 sections of statistics students. Before the training, students were given a consent form that informed them that they could decline to participate by simply not performing the task. For the training, in order to randomly assign the students to conditions, the booklets were simply passed out to the students in the order in which they were seated. The students 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 through 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 proportion correct on the immediate, delayed, and final tests as a function of training condition. A mixed model ANOVA with time of test as a within subjects factor and training condition as a between subjects factor revealed a main effect of time of test, F (2, 116) = 10.24,

$p < .001$, $\eta^2 = .15$ [2]. Pairwise comparisons [3]showed that performance on the final test was superior to that on the immediate test ($p = .006$) and the delayed test ($p < .001$), but there was no significant difference between the immediate and delayed tests ($p = .145$).

Performance in the interleaved condition was marginally significantly greater than that in the blocked condition, F (1, 58) = 3.22, $p = .078$, $\eta^2 = .05$. Although there was no interaction between condition and time of test ($F < 1$), pairwise comparisons showed that performance in the interleaved condition was significantly greater than in the blocked condition for the delayed test ($p = .007$), but not for either of the other tests ($p's > .05$).

Table 1

Mean Proportion Correct (SE) in the Blocked and Interleaved Condition for the Immediate, Delayed, and Final Tests in Experiment 1

	Training Condition		
Test	Blocked	Interleaved	Total
Immediate	.30 (.043)	.34 (.041)	.32 (.030)
Delayed	.21 (.032)	.33 (.031)	.27 (.022)
Final	.41 (.045)	.45 (.044)	.43 (.031)
Total	.31 (.027)	.38 (.026)	.34 (.019)

Discussion

These data suggest that interleaving the examples, rather than blocking them, may have produced a slight advantage. However, if that is the case, then interleaving did not result in better acquisition or learning the examples from the formal classroom instruction, but rather better retention. Also, the superior retention on the delayed test appears to be driven by especially poor performance in the blocked condition, rather than by especially good performance in the interleaved condition.

One possible explanation for the especially poor performance in the blocked condition on the delayed test may involve attention to the task. As the participants in the blocked condition worked through their training examples, they may have noticed that each example of a particular type was followed by several more of the same type. The examples may have been long and detailed enough to cause their attention to the repeated examples of the same type to wane. Although they may have learned enough from the examples to which they did attend to do reasonably well on the immediate test, their lack of attention to the others may have resulted in the lower retention on the delayed test.

2 All eta squares reported in this paper were calculated as the SS for the effect divided by the sum of the SS for the effect plus the SS for the error for that effect.

3 All pairwise comparisons are the pairwise comparisons that SPSS provides with the compare subcommand of the emmeans command. The SPSS output for the pairwise comparisons contains the following note:" Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments)".

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. Therefore, before drawing any conclusions about the practical value of interleaving examples for this type of material, we wished to explore ways to improve performance, and to attempt to replicate the effect more convincingly.

Experiment 2

The results of Experiment 1 were not as expected. First, there was only a marginal benefit of interleaving across the three tests. This could have been due to the overall low performance. Second, on the one test, the delayed test, where there was a significant difference between the training conditions, the difference might have been best characterized as more a detriment of blocking than a benefit of interleaving.

We hypothesized that the low performance could have resulted simply from insufficient training on the examples. We hypothesized that the possible detriment of blocking on the delayed test could have resulted from the blocked participants' attention waning due to the length and complexity of the examples. The first issue we wished to address in Experiment 2 was the overall low performance in Experiment 1. We did not address the possibility of waning attention until Experiments 3 and 4. 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, and that this might enable us to uncover an effect of interleaving on all of the tests.

Method

Participants. The participants were university undergraduates in several sections of an introductory statistics course for the behavioral sciences in the Spring semester of 2009. There were 98 participants 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. In the blocked condition ($N = 52$), all participants received the blocks in the same order, as in Experiment 1. In the interleaved condition ($N = 46$), rather than completely randomizing the order of the examples, which could have resulted in some procedure occurring twice in a row as had occurred in Experiment 1, we created six within subjects blocks, within which the order of the procedures was randomized. This resulted in no procedure occurring twice in a row. However, this time our randomization did result in one of the 24 possible orders occurring in two of the blocks, although those blocks were not adjacent. All participants received the same set of randomized blocks in the same order.

Reducing the types of research situations from six to four resulted in a change in the tests. In Experiment 2, the test consisted of only six items, with only one of them doubled. Otherwise, the training booklets and tests were the same as in Experiment 1. The procedure for Experiment 2 was the same as for Experiment 1.

Results

The data for Experiment 2 were analyzed in the same way as in Experiment 1. As shown in Table 2, performance on the immediate test was better than it had been in Experiment 1. On the other two tests it was about the same as in Experiment 1, although the especially low performance of the blocked participants on the delayed test did not reappear. Nevertheless, performance was still very low. More importantly, there was no significant effect of interleaving in either direction ($F < 1$). In fact, the blocked participants actually did numerically better than the interleaved participants on all the tests.

A mixed model ANOVA revealed a main effect of time of test, F (2, 192) = 9.65, $p < .001$, $\eta^2 = .09$, but no interaction ($F < 1$). Pairwise comparisons showed that performance on the immediate test was significantly higher than on the delayed test ($p < .001$) and marginally higher than on the final test ($p = .065$). Also, performance on the final test was higher than on the delayed test ($p = .016$).

Table 2

Mean Proportion Correct (SE) in the Blocked and Interleaved Condition for the Immediate, Delayed, and Final Tests in Experiment 2

	Training Condition		
Test	Blocked	Interleaved	Total
Immediate	.52 (.040)	.51 (.042)	.52 (.029)
Delayed	.35 (.034)	.34 (.036)	.34 (.025)
Final	.46 (.043)	.42 (.046)	.44 (.031)
Total	.44 (.024)	42 (.026)	.43 (.018)

Discussion

Although 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 improve immediate acquisition slightly from what it was in Experiment 1, it did not improve delayed retention, nor did it improve learning the examples from the formal classroom instruction. Contrary to our expectation, interleaving was not beneficial for immediate acquisition, nor was it beneficial for the final test. Furthermore, the poor performance in the blocked condition on the delayed test that had been seen in Experiment 1 did not replicate, suggesting that it may have been a Type I error, rather than being due to waning attention.

Given that the absolute level of performance was still so low as to represent a failing grade in an actual classroom situation, we speculated that a greater increase in the amount of practice might be needed to both raise performance generally, and to reveal a benefit of interleaving. Also, we hypothesized that, although the length and complexity of the examples may not have contributed to waning attention specifically in the blocked condition on the delayed test, it still could have been

another contributing factor to the low performance. Accordingly, we made several changes for Experiment 3.

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 university undergraduates in several sections of an introductory statistics course for the behavioral sciences in the Fall semester of 2009. There were 64 participants 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 as much as we had wished. Also, where there was a slight improvement, it was mostly only in immediate acquisition, not as much in retention, and it did not result in revealing an interleaving benefit.. Therefore, in Experiment 3, we went back to training with six types of research situations and four examples of each type, as in Experiment 1, and the tests were as in Experiment 1. In the blocked condition (N = 33), we created one order of the blocks by randomizing the order of the different procedures, and then we counterbalanced the order of the blocks by using a Latin square in which each procedure occurred in each serial position once and only once. However, each procedure did not necessarily follow a different procedure in each order. Each participant was randomly assigned to one of the six possible orders produced by the Latin square. In the interleaved condition (N = 31), as in Experiment 2, we created within subjects blocks, within which the order of the procedures was randomized. However, this time, there were 720 possible orders of the six procedures, and we used four different randomly chosen orders. This resulted in no instances where any procedure occurred twice in a row, nor did any of the 720 possible orders occur more than once. All participants received the same set of four within subjects randomized blocks in the same order.

We also made three other changes. First, we made the descriptions shorter, easier to read, and equal in length. Second, we changed the labels for some of the statistical procedures.[4] In the two prior experiments, we had called the first four procedures 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 (Appendix C shows how the example provided in Appendix A was changed). We believed that highlighting that these four procedures 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. Otherwise, the materials in Experiment 3 were the same as in Experiment 2.

Third, instead of just one training session, there were three training sessions on the same training examples, each followed by an immediate test. The training sessions were scheduled in three

4 We made the changes in the labels at the top of the example, and where the same terms appeared in the tests as alternative choices. However, we inadvertently did not change the labels at the end of the text of the examples. This oversight was corrected in Experiment 4.

successive weeks early in the semester. Otherwise, the procedure in Experiment 3 was the same as that in Experiment 2.

Results

The data for Experiment 3 were analyzed in the same way as in Experiments 1 and 2. As shown in Table 3, there was no effect of training condition ($F < 1$). There was a significant effect of time of test, F (4, 248) = 7.10, $p < .001$, $\eta^2 = .10$. Pairwise comparisons showed that performance on all of the tests other than the first immediate test was better than it was on the first immediate test *(p's* < .005*).* Both the third immediate test and the final test approached being significantly higher than both the second immediate test and the delayed test (*p's* = from .071 to .115). There was no interaction ($F < 1$).

Table 3

Mean Proportion Correct (SE) in the Blocked and Interleaved Condition for the Immediate 1, 2, and 3, Delayed, and Final Tests in Experiment 3

	Training Condition		
Test	Blocked	Interleaved	Total
Immediate 1	.24 (.037)	.22 (.038)	.23 (.027)
Immediate 2	.38 (.041)	.29 (.043)	.33 (.030)
Immediate 3	.40 (.050)	.38 (.052)	.39 (.036)
Delayed	.35 (.035)	.32 (.036)	.33 (.025)
Final	.37 (.043)	.41 (.044)	.39 (.031)
Total	.35 (.028)	.32 (.029)	.34 (.020)

Discussion

Increasing the number of training sessions provided increases in performance across the immediate tests. By the third immediate test, performance was higher than what was seen on the immediate test in Experiment 1, but it was lower than that seen on the immediate test in Experiment 2. Also, as in Experiments 1 and 2, performance dropped on the delayed test. Furthermore, performance on all of the tests was still at a level that was so low that it would be considered failing for purposes of actual classroom evaluation. Finally, there was also still no benefit of interleaving. Thus, we concluded that in order to raise performance in an effort to uncover an interleaving benefit, we needed to do something more than just the changes we had made in Experiments 2 and 3.

The highest performance we had seen thus far had occurred on the immediate test in Experiment 2, in which the participants had received six instances of each of four types of procedures. Thus, even though that change had not resulted in better retention or learning from the formal training, we decided to reverse ourselves once again and use just the four types of procedures in Experiment 4. In addition, however, we made further changes to the nature of the training. Also, for Experiment 4 we temporarily focused only on increasing performance overall, and did not include interleaving as a factor.

Experiment 4

In Experiment 4 we attempted only to improve overall performance to a level that would be considered at least a passing grade in an actual classroom situation, rather than to attempt to produce an interleaving benefit. We did so by providing participants with what we believed would be a stronger instructional manipulation.

Method

Participants. The participants were university undergraduates in several sections of an introductory statistics course for the behavioral sciences in the Spring semester of 2010. There were 62 participants who completed the entire experiment.

Materials and Procedure. In this experiment we did not include interleaving as a factor. Instead, all participants received the training examples in blocks. We continued to be concerned about keeping the participants on task during the training. Just as we had suspected that the length and complexity of the examples could have resulted in poor attention to the task, we were also concerned that being required to study 24 examples, each for one whole minute, may have been having a similar effect. Therefore, for Experiment 4 the participants received only 16 training examples, not 24, as in Experiments 1, 2, and 3. We used the four types of procedures that we had used in Experiment 2. We presented four instances of the independent-measures *t* test, the repeated-measures *t* test, the independent-measures ANOVA, and the repeated-measures ANOVA), and we used those terms to label the procedures throughout the materials, including at the end of the text of the examples, thus correcting the oversight mentioned in the footnote in Experiment 3. By changing back to training with only four types of research situations, rather than six, we also changed the tests back to only five items with one doubled, as in Experiment 2.

In addition, again in an effort to insure attention to the task, rather than instructing the participants to study each training example for one minute, we allowed them to work 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 randomly assigned participants to the four different orders of the blocks determined by a Latin square. This Latin square design, like that in Experiment 3 resulted in each procedure occurring once, and only once, in each serial position, but each procedure did not follow a different procedure in each of the orders.

We also made two other 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 manipulated whether we explicitly gave subjects certain critical information or whether they were only instructed to try to induce that information from the examples on their own. First, we manipulated whether we explicitly told them what features of the research situation determined which statistical procedure to use. For example, for some participants, if the example were for an independent measures *t* test, we explicitly told them that the number of groups of data was only two, whereas if the example were for an independent measures ANOVA, we explicitly told them that the number of groups of data was more than two. Second, we manipulated whether we explicitly told

them how to induce the critical features from the examples. For example, in one case, where the example was for an independent measures *t* test, some subjects were explicitly told that because the experiment had only an anxious condition and a relaxed condition, there were only two groups of data. In another case, where the example was for an independent measures ANOVA, some subjects were explicitly told that because the experiment tested five different types of fertilizer, there were more than two groups of data.

The critical information manipulation resulted in three conditions. First, in an examples-only condition (N = 21) the participants saw just the example. Below the example was the label "Features:", and they were instructed to write down below that label, if they could, what features of the example determined which statistical procedure to use. Below that was a label "How you figure out the features:", and they were instructed to write down an explanation, if they could, for how one could figure out what the determining features were.

Second, in an examples-plus-features condition (N = 19) they saw the example, and below the example, after the label "Features:" we explicitly told them what the critical features were (as explained above). Below that was the label "How to figure out the features:", and they, like the participants in the examples-only condition, were instructed to write down an explanation, if they could, for how one could figure out what the determining features were.

Third, in an examples-plus-features-and-explanation condition (N = 22), they saw the example, and below the example, after the label "Features:" we explicitly told them what the critical features were, and also, after the label "How to figure out the features:" we explicitly told them how to figure out the features (as explained above).

Other than as explained above, the materials and procedures for Experiment 4 were the same as in Experiment 3.

Results

The data for Experiment 4 were analyzed in the same way as in Experiments 1, 2, and 3. We present the analysis of the data first as a function of time of test and then as a function of condition. In both cases we present the data for the factor first averaged across the other factor and then within each level of the other factor.

Performance as a function of time of test, across all conditions. As shown in Table 4, there was a large main effect of time of test, F (2, 118) = 13.60, $p < .001$, $\eta^2 = .19$. Pairwise comparisons showed that there was a significant .21 drop in performance from the immediate test to the delayed test ($p < .001$), and then there was a significant .27 improvement from the delayed test to the final test ($p < .001$). However, there was no significant difference between the immediate and final tests (p =.225). The performance on the final test resulted, for the first time, in performance that would be at least a passing grade (above .60) in an actual classroom situation.

Performance as a function of time of test, within each condition. In the examples-only condition, the only significant pairwise comparison was the .19 improvement from the delayed test to the final test ($p = .034$). Similarly, in the examples-plus-features condition, the only significant difference was a .23 improvement from the delayed test to the final test ($p = .013$). In this case, however, the absolute level of performance was above .60 on the immediate test, and, although it dropped .16 on the delayed test, it went back up the the highest level seen in all of the experiments so far, to .68, on the final test. In the examples-plus-features-and-explanation condition, performance on the immediate test was .60, then dropped a significant .32 on the delayed test ($p = .001$), but then

improved a significant .39 on the final test ($p < .001$), thus, as in the examples-plus-features condition, reaching an above .60 level.

Performance as a function of training condition, across all tests. Performance in the examples-plus-features condition and in the examples-plus-features-and-explanation condition was numerically higher than in the examples-only condition, although this effect did not reach statistical significance across all three condition, F (2, 59) = 1.54, p = .223, η^2 = .05. However, a set of orthogonal contrasts showed that performance in the examples-plus-features condition was marginally higher than in the examples-only condition (p = .085). On the other hand, performance in the examples-plus-features-and-explanation condition was not even marginally higher than in the examples-only condition (p = .408). There was no difference between the examples-only condition and the other two conditions combined (p = .137).

Performance as a function of training condition, within each test. On none of the tests was there a significant difference between conditions across all three conditions ($p's > .05$). For the immediate and final tests, there were no significant pairwise comparisons. However, for the delayed test only, adding the explanations to the examples-plus-features condition resulted in an almost significant .17 drop in performance, (p = .055).

Table 4

Mean Proportion Correct (SE) in the Three Training Conditions for the Immediate, Delayed, and Final Tests in Experiment 4

	Training Condition			
Test	Examples Only	Examples Plus Features	Examples Plus Features and Explanation	Total
Immediate	.50 (.085)	.61 (.090)	.60 (.083)	.57 (.050)
Delayed	.36 (.058)	.45 (.061)	.28 (.057)	.36 (.034)
Final	.55 (.061)	.68 (.064)	.67 (.060)	.63 (.036)
Total	.47 (.044)	.58 (.046)	.52 (.043)	.52 (.025)

Discussion

The changes we made for Experiment 4 resulted in higher performance, when averaging across all three training conditions, than in any of the previous experiments on all three tests. Even with the drop in performance from the examples-plus-features condition to the examples-plus-features-and-explanation condition on the delayed test, the overall performance on that test was still higher than on any of the delayed tests so far. Thus, giving all the participants an explanation of the nature of a category induction task was beneficial.

These changes also resulted in some cases where performance, for the first time in all of the experiments so far, reached at least a barely passing grade for an actual classroom situation (i.e., .60 or

above). That occurred averaged across all three conditions for the final test, which measured how well the participants learned from the formal instruction that occurred later in the semester, after the experiment had been conducted. But it also occurred on the immediate test in both conditions where we explicitly told the participants which features of the example determined which procedure to use. Thus, the explicit information about which features to focus on, like the explanation of the nature of a category induction task, was also at least marginally beneficial. However, explicitly telling the participants how to induce the critical features from the examples did not add any additional benefit. In fact, at least in terms of the delayed test, it was detrimental. This could have been because that additional information began to confuse the participants, although clarifying that issue will remain for future research.

Having achieved the goal of raising performance, we were now ready to re-introduce the interleaving manipulation. In Experiment 5, we used what we had learned from the previous experiments to strengthen our manipulation as much as possible in hopes that doing so would enable us to reveal an interleaving advantage.

Experiment 5

In Experiment 5, in order to provide a stronger manipulation, rather than giving all the participants an explanation of the nature of a category induction task, we manipulated that factor, along with manipulating whether we provided the critical information about the features of the research situation that determined which statistical procedure to use. We also manipulated interleaving.

Method

Participants. The participants were university undergraduates in several sections of an introductory statistics course for the behavioral sciences in the Fall semester of 2010. There were 114 participants who completed the entire experiment.

Design, materials, and procedure. This experiment was a 2 by 2 by 3 design. We called the first factor features. 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 which features of each research situation could be used to do so (see Appendix D for the training instructions for the description-only group). Their training examples were the same 16 examples used in Experiment 4. 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 as 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 E). 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 F).

The second factor was interleaving. For the blocked participants, the examples were presented in a Latin square design. However, this Latin square design was arranged so that not only did every procedure occur once, and only once, in each serial position, but also, each procedure followed a different procedure in each order.

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.

Crossing the features and interleaving factors resulted in four conditions, with the following numbers of subjects who completed all the tests: (a), description-only/blocked *(N* = 28) (b), description-only/interleaved *(N* = 33) (c), description-plus-features/blocked *(N* = 27), and (d) description-plus-features/interleaved *(N* = 26).

The third factor, time of test, was the same as in all the previous experiments. All the materials and procedures, except as described above, were the same as in Experiment 4.

Results

Overall performance. As shown in Figure 1, a mixed model ANOVA using test performance as the dependent measure, time of test as a within subjects factor, and features and interleaving as between subjects factors, showed that there was a large main effect of time of test, F (2, 220) = 12.76, $p < .001$, $\eta^2 = .10$. Pairwise comparisons showed that performance on the immediate test was significantly higher than on both the delayed and the final tests *(p's* < .001*).* There was no difference between the delayed and the final tests *(p* = .729*).*

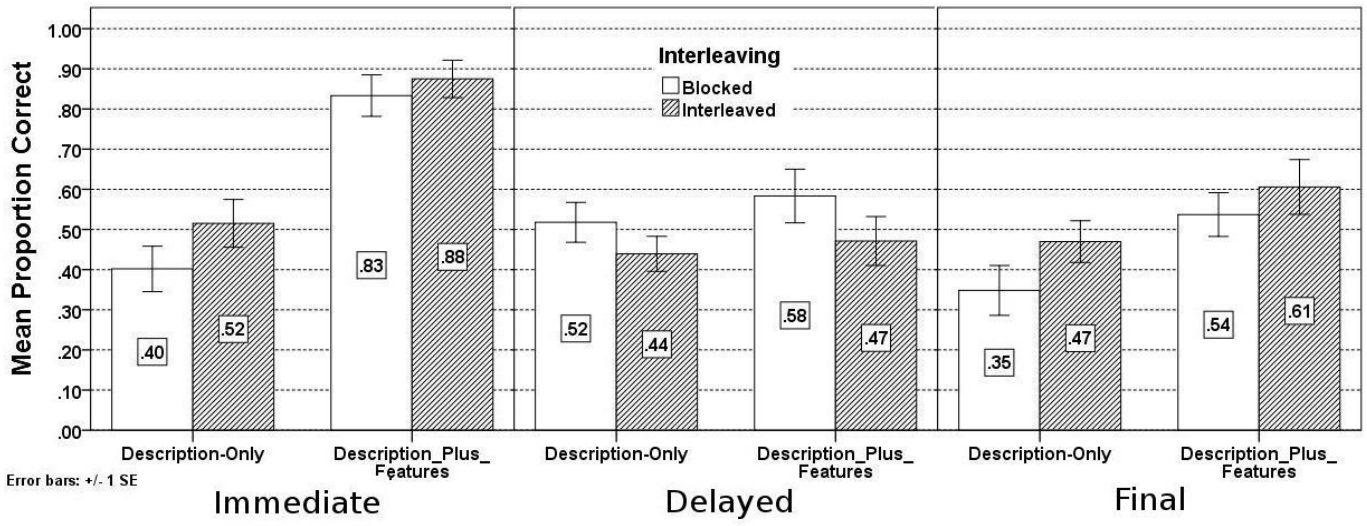

Figure 1. Performance on the three tests as a function of information and interleaving in Experiment 5.

Overall, there was no main effect of interleaving, $F < 1$, but there was a large and significant main effect of features, F (1, 110) = 29.21, $p < .001$, $\eta^2 = .21$. There was no three-way interaction between time, features, and interleaving. However, there were two, two-way interactions. First, adding features was not as effective for the delayed test as for the two others, F (2, 220) = 11.68, $p < .001$, $\eta^2 = .10$. Pairwise comparisons showed that adding features produced a significant advantage for immediate acquisition *($p < .001$)* and for the final test ($p = .007$), but not for the delayed test *($p = .379$).* Second, in a similar pattern, interleaving was beneficial for both the immediate and final tests, but not for the delayed test, F (2, 220) = 4.13, $p = .017$, $\eta^2 = .04$. This interaction was due to interleaving being as least somewhat beneficial for the immediate and final tests, but somewhat detrimental for the delayed test, although pairwise comparisons did not show any of those effects for the individual tests to be significant *($p's > .05$)* . Across the three tests, there was no interaction between features and interleaving.

We also performed separate two way ANOVA's on each test alone, with features and interleaving as between subjects factors.

Performance on the immediate test alone. On the immediate test alone, there was a large benefit of features, $F(1, 110) = 51.57$, $p < .001$, $\eta^2 = .32$. There was a slight, but not significant, benefit of interleaving, $F(1, 110) = 1.98$, $p = .16$, $\eta^2 = .02$. The features by interleaving interaction was not significant ($F < 1$). Although there was a slight numerical advantage of interleaving for the description-only condition, a pairwise comparison showed that it was not quite significant *($p = .135$).*

Performance on the delayed test alone. On the delayed test alone, the advantage of features seen on the immediate test was not retained ($F < 1$). For interleaving, the direction of the data reversed from the slightly beneficial effect seen on the immediate test to a slightly, but not quite significantly, detrimental effect, $F(1, 110) = 3.00$, $p = .086$, $\eta^2 = .03$. There was no interaction ($F < 1$).

Performance on the final test alone. On the final test alone, there were at least numerical advantages for both features and interleaving. The advantage of features was significant, $F(1, 110) = 7.55$, $p = .007$, $\eta^2 = .06$. The interleaving advantage was not quite significant, $F(1, 110) = 2.59$, $p = .111$, $\eta^2 = .02$. In the description-only condition, there was a .12 interleaving advantage which approached significance ($p = .135$). In the description-plus-features condition, interleaving provided only a .07 numerical advantage, which was not significant ($p > .05$). There was no interaction ($F < 1$).

Performance on just the immediate and final tests combined. The time by interleaving interaction shown in Figure 1 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 features 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, $F(1, 110) = 22.74$, $p < .001$, $\eta^2 = .17$, and a time by features interaction, $F(1, 110) = 11.21$, $p = .001$, $\eta^2 = .09$. Also, there was a significant benefit of features, $F(1, 110) = 37.94$, $p < .001$, $\eta^2 = .26$, and there was an almost significant benefit of interleaving, $F(1, 110) = 3.63$, $p = .059$, $\eta^2 = .03$.

Discussion

Training participants in how to do a category induction task, and explicitly telling them which features of the research situation determined which statistical procedure to use enabled them to perform better on the immediate test than in any of the previous experiments. Although this advantage was not retained over a delay of a few weeks, performance on the delayed test was, like performance on the immediate test, at least better than in any of the previous experiments. Also, strengthening the features manipulation did not result in convincingly revealing an interleaving advantage, as we has hoped.

The one place in Experiment 5 where we saw a marginally significant interleaving advantage was on the immediate and delayed tests combined. However, it is notable that in this case, as in the instance in which we saw an interleaving advantage that was significant at the .05 level (on the delayed test in Experiment 1), the effect again looked like it may have been driven mainly by poor performance in the blocked condition, rather than by exceptionally good performance in the interleaving condition. 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 was the very potential problem that we sought to alleviate for all participants by the changes we made in Experiments 3 and 4. Thus, 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 it might be possible to capitalize on this positive effect of interleaving 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 Experiment 6, 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.

Experiment 6

Method

Participants. The participants were university undergraduates in several sections of an introductory statistics course for the behavioral sciences in the Spring and Fall semesters of 2011. There were 188 participants who completed the entire experiment.

Design, materials, and procedure. The experiment was a 2 by 2 by 3 design. The first factor was feedback. The feedback factor actually manipulated several of the factors that had been effective for raising performance in the previous studies. The materials and procedure for the no-feedback group were the same as for the description-only group in Experiment 5. The materials and procedure for the feedback group were the same as for the description-plus-features group in Experiment 5 with the following addition. 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 participants were 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 on a following page. 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.

The second factor was interleaving. The materials for the blocked participants were the same as in Experiment 5. The materials for the interleaved participants were presented in a within subjects randomized blocks design as in Experiment 5. However, this time we used the same randomized order within each block and the same order of the blocks for the interleaved condition as for the blocked condition. This resulted in one instance where one of the examples of a particular statistical procedure occurred twice in a row, once at the end of one randomized block, and again at the beginning of the next. All the subjects in the interleaved group received the same order of interleaved descriptions.

Crossing the feedback and interleaving factors resulted in four conditions, with the following numbers of subjects who completed all the tests: (a), no-feedback/blocked *(N* = 46) (b), no-feedback/interleaved *(N* = 52) (c), feedback/blocked *(N* = 43), and (d) feedback/interleaved *(N* = 47).

The third factor, time of test, was the same as in all the previous experiments. All the materials and procedures, except as described above, were the same as in Experiment 5.

Results

Overall performance. As shown in Figure 2, a mixed model ANOVA using test performance as the dependent measure, time of test as a within subjects factor, and feedback and interleaving as between subjects factors, showed that there was a large main effect of time of test, $F(2, 368) = 22.73$, $p < .001$, $\eta^2 = .11$. Pairwise comparisons showed that performance on the immediate test was

significantly higher than on both the delayed and the final tests *(p's* < .001*)*. There was no difference between the delayed and the final tests *(p* = .307*)*.

There was a significant main effect of interleaving, $F(1, 184) = 12.36$, $p = .001$, $\eta^2 = .06$. However, there was no main effect of feedback, $F(1, 184) = 2.16$, $p = .143$, $\eta^2 = .01$. There was no two way interaction between feedback and interleaving ($F < 1$). There was no two way interaction between feedback and time of test ($F < 1$). However, there was a two way interaction between interleaving and time of test, in which there was a positive effect of interleaving on both the immediate and delayed tests, but not on the final test, $F(2, 368) = 13.34$, $p < .001$, $\eta^2 = .07$. Pairwise comparisons showed that the the .26 advantage of interleavers over blockers on the immediate test was significant ($p < .001$), and their .09 advantage on the delayed test was almost significant ($p = .052$). All of the above, however, is qualified by a three way interaction between feedback, interleaving, and time of test, $F(2, 368) = 3.62$, $p = .028$, $\eta^2 = .02$.

In order to further examine the three way interaction, we conducted separate two way ANOVA's on each test alone, with feedback and interleaving as between subjects factors.

Performance on the immediate test alone. On the immediate test alone, there was a large, positive effect of interleaving, $F(1, 184) = 32.91$, $p < .001$, $\eta^2 = .15$. There was also an almost significant positive effect of feedback, $F(1, 184) = 3.71$, $p = .056$, $\eta^2 = .02$. There was no interaction ($F < 1$). Pairwise comparisons showed that the positive effect of interleaving was significant in both the no-feedback condition ($p = .001$) and feedback conditions ($p < .001$). However, the positive effect of feedback was significant in the interleaved condition ($p = .041$), whereas it was not significant in the blocked condition ($p = .485$).

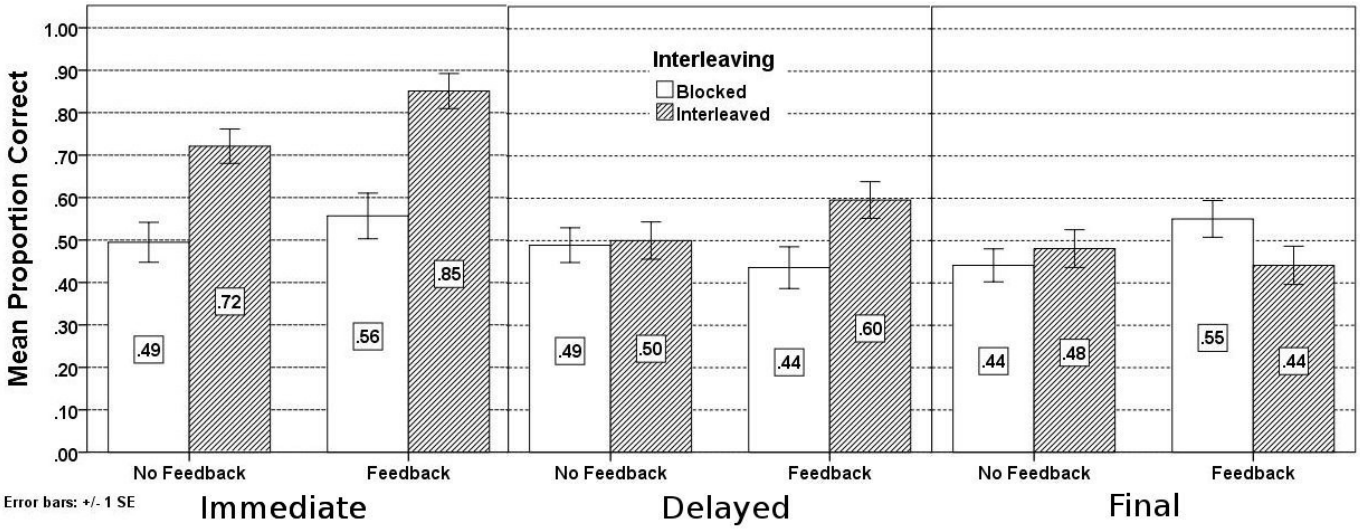

Figure 2. Performance on the three tests as a function of information and interleaving in Experiment 6.

Performance on the delayed test alone. On the delayed test alone, there was an almost significant main effect of interleaving, $F(1, 184) = 3.83$, $p = .052$, $\eta^2 = .02$, but no main effect of

feedback (F <1). There was a marginally significant interaction in which the positive effect of interleaving was numerically larger with feedback than without it, $F(1, 184) = 2.59$, $p = .110$, $\eta^2 = .01$. In spite of the lack of a significant interaction, pairwise comparisons showed that the positive effect of interleaving was not significant in the no-feedback condition ($p = .801$), but it was significant in the feedback condition($p = .014$).

Performance on the final test alone. On the final test alone, although there were no significant effects at the .05 level, there were marginally significant patterns in the data that were the reverse of those seen in the immediate and delayed tests. Neither the main effect of interleaving nor feedback were significant ($F's < 1$). However, there was a marginally significant interaction, $F(1, 184) = 2.73$, $p = .100$, $\eta^2 = .01$. Pairwise comparisons showed that there was slight, but not significant, numerical advantage of interleaving in the no-feedback condition *($p = .455$)*, and a slightly, and marginally significant disadvantage of interleaving in the feedback condition *($p = .118$)*.

Performance on just the immediate and delayed tests combined. Finally, given that on both the immediate and delayed tests the effect of interleaving was numerically greater with feedback than without it, we did one more mixed ANOVA on just the immediate and delayed tests combined. Performance on the immediate and delayed tests was the dependent variable, time of test was a within subjects factor, and interleaving and feedback were between subjects factors. Performance on the immediate test was significantly higher than on the delayed test, $F(1, 184) = 26.81$, $p < .001$, $\eta^2 = .13$. There was a significant .18 advantage of interleaving, $F(1, 184) = 24.92$, $p < .001$, $\eta^2 = .12$ and a marginally significant .06 advantage of feedback, $F(1, 184) = 2.33$, $p = .129$, $\eta^2 = .01$.

There was a marginally significant interleaving by feedback interaction in which interleaving was more helpful in the feedback than in the no-feedback condition, $F(1, 184) = 2.62$, $p = .107$, $\eta^2 = .01$. Pairwise comparisons showed that there was a significant .12 advantage of interleaving in the no-feedback condition *($p = .016$)* and a significant .23 advantage of interleaving in the feedback condition *($p < .001$)*. Also, there was virtually no effect of feedback in the blocked condition ($p = .949$) compared to a significant .11 advantage of feedback in the interleaved condition ($p = .023$).

There was a significant time by interleaving interaction in which interleaving was more beneficial on the immediate test than on the delayed test, $F(1, 184) = 9.36$, $p = .003$, $\eta^2 = .13$. Pairwise comparisons showed that there was a significant .26 advantage of interleaving on the immediate test ($p < .001$), compared to a marginally significant .09 advantage on the delayed test ($p < .052$). There was no time by feedback interaction, $F(1, 184) = 1.46$, $p = .228$, $\eta^2 = .01$. There was no three way interaction ($F < 1$).

Discussion

In Experiment 6, the advantage of interleaving finally was clearly evident. Also, as predicted, combining feedback and interleaving was more beneficial than either alone. There were three places in the data where the benefit of combining feedback and interleaving was evident at least by simple effects tests. On the delayed test only, the feedback by interleaving interaction was marginally significant, and the benefit of interleaving was significant with feedback, but not without it. Also, on the immediate test alone, although the feedback by interleaving interaction was not significant, there was a significant benefit of feedback with interleaving, but not without it. On the immediate and

delayed tests combined, the feedback by interleaving interaction was marginally significant, and, as on the immediate test alone, there was a significant benefit of feedback with interleaving, but not without it.

Combining interleaving with feedback resulted in high performance on the immediate test, similar to the high performance in the description-plus-features condition on the immediate test in Experiment 5. On the immediate test in Experiment 6, the interleavers with feedback would have received a B (85%). Even without feedback the interleavers would have received a C (72%) on the immediate test. However, given that the training for the no-feedback/interleaving participants was the same as for the description-only/interleaving participants in Experiment 5, whose performance was at 52% correct, it remains unclear why these participants did so much better. Their having done so much better, however, along with the benefit of the larger sample size, produced a significant interleaving benefit, whereas there was none in the description-only condition on the immediate test in Experiment 5. On the delayed test, similar to the immediate test, the highest performance was in the interleaving with feedback condition.

As with the previous experiments, however, the story in terms of absolute level of performance is different for retention. All of the participants would have received a poor grade for their delayed test. In fact, only the interleaving with feedback participants would have received a passing grade, and a just barely passing grade at that (.60).

Performance on the final test was about at the same low level as it had been in the previous experiments. In fact, it was lower than in Experiment 4, and there was not even a hint of an interleaving advantage, as their had been in Experiment 5.

General Discussion

By implementing interleaving examples under the right conditions, we were able to demonstrate its benefit for learning to associate the correct statistical procedure with a given research situation. However, this series of experiments showed that the benefit of interleaving in this situation has the limitation that it only occurs if the right conditions are met regarding the nature of the training examples and the training procedure. In Experiment 6, we found that learning, at least for an immediate test, was facilitated by interleaving if, among other conditions, (a) there were only four types of examples, (b) if the labels for the statistical procedures identified the types along two dichotomous dimensions, (c) if the examples were not too complex, and (d) if the participants were allowed to study them at their own pace.

In addition to those conditions, we also examined the impact, alone, and combined with interleaving, of three other aspects of the training. We examined the effects of (a) instructions about the nature of an inductive reasoning task, (b) provision of the relevant features for the task, and (c) immediate feedback on performance during training. Providing the relevant features was especially helpful. In Experiment 5, both with and without interleaving, providing the relevant features, when combined with instructions about the nature of an inductive reasoning task, but without feedback, had a large impact on immediate acquisition. However, this benefit did not extend to retention. On the other hand, in Experiment 6, we found that when feedback was added, there was not only an almost

significant benefit on immediate acquisition, but, furthermore, we were able to demonstrate an interleaving benefit for retention, although to a very small degree.

Why was there only convincing evidence of interleaving when so many conditions were met in regard to the nature of the training materials and procedures? The first four conditions all were related to enhancing encoding of the examples by reducing their complexity and by enhancing the students' attention to the examples. But most of those enhancements had been made by Experiment 3, and yet even with three training session in Experiment 3, we did not see performance on any test (even the third immediate test) that would be acceptable in a classroom situation, and we did not see any benefit of interleaving.

Statistics instructors are aware of the exceptional difficulty for students of acquiring and retaining the kind of conceptual categories required for learning in statistics. There are many reasons why learning in statistics is so hard for students. First, there is the abstract nature of the material. 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. Second, there is the hypothetical reasoning required. For example, consider dealing with probability distributions to do a hypothesis test about an effect. The real situation could be that the effect of interest actually exists. But it could be that it does not. There is no way of knowing which situation is the actual state of affairs. However, doing the test requires reasoning about the probability distribution that would apply if one of those situations was, in fact, the true state of affairs. Third, the reasoning is conterfactual. One hopes to convince others that the true state of affairs is that the effect does exist, yet, to do the test, one must think of the implications of the situation in which it does not. Fourth, the reasoning involves multiple negatives. The evidence that one hopes to provide is data that one can say would not be likely to have been obtained if the effect of interest did not exist. Fifth, the reasoning has to be indirect. The evidence is not about the likelihood that the effect exists, it is about the unlikelihood of getting certain data if the effect did not exist.

However, the participants' task in our experiments did not require any of those kinds of reasoning. Rather, it required recognizing that certain features of a research situation determine that a certain test should be used. The features were the number of conditions and whether the conditions were between or within subjects. But these are concepts with which these students are very unfamiliar. 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. This lack of familiarity with the relevant features may have made it very difficult for the participants to induce them from the examples.

According to Koedinger, Corbett, and Perfetti's (2012) *Knowledge, Learning, and Instruction* (KLI) *Framework*, feature focusing is an instructional principle that should aid the learning processes of induction and refinement. It was only in Experiment 4, in which we provided the relevant features along with instructions in how to do an induction task, that we saw performance rise to a passing grade.

However, even then, the passing performance was only on an immediate test or after formal instruction, not on a delayed test before formal instruction. The lack of relevant prior knowledge may explain why, even when provided with the features so that they did not have to infer them, students benefited only on an immediate test. Having nothing to which to connect these newly encountered concepts, they were not able to retain them over the four to six weeks between the immediate and retention tests.

What would be required to produce retention at an academically acceptable level? According to the KLI framework, the instructional principles of spacing and multiple testing should aid in the memory and fluency processes needed to produce better retention. Interleaving provides some spacing of learning. However, it was only when we added feedback in Experiment 6 that we saw a convincing benefit of interleaving. According to Koedinger, Corbett, and Perfetti's (2012), the instructional principle of timely feedback should be another factor that aids induction and refinement processes. However, if it enhanced the effectiveness of the interleaving, which may aid in memory processes, then that might explain why, in Experiment 6, we saw some retention to a passing level, and a benefit of interleaving on the retention test when combined with feedback.

Thus, although the KLI framework is a recent development, it may provided a way to design instructional interventions to improve learning even in such difficult domains as statistics. For example, learning about inferential statistics, the difficulties with which were described above, might benefit from instructional principles that aid in understanding and sense making processes. According to the KLI framework, those processes are aided by principles such as prompted self-explanation and accountable talk. Examining those possibilities appears to be a promising area for future research.

Appendix A

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

Two sample *t* test

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

Appendix B

An Example of the Test Items from Experiment 1

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

a. Two sample *t* test

b. Paired *t* test

c. One way ANOVA

d. Repeated measures ANOVA

e. Chi square test

f. Correlation

Appendix C

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

Independent-measures *t* test

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

Appendix D

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 E

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 F

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.