Practicing Retrieval of Facts in Statistics Benefits High Ability Students But Hurts Low Ability Students

Robert S. Ryan

Kutztown University

Presented at the 25th Annual Convention of the Association For Psychological Science, Washington, DC, May 23 – May 26, 2013.

Correspondence:

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

Practicing Retrieval of Facts in Statistics Benefits High Ability Students But Hurts Low Ability Students

Statistics students studied facts and then either retrieved them and re-read them once, or re-read them twice. On an easier set of facts, only the highest ability students benefited from retrieval. On a more difficult set of facts, no one benefited, and the low ability students were hurt.

Over the past several decades, cognitive psychologists have uncovered many principles of learning in laboratory studies. More recently, they have begun examining how to best apply such principles to actual classroom instructional methods. In the process, boundary conditions on their application are emerging. For example, comparing examples has been shown to help people induce a general problem solving principle ((Catrambone & Holyoak, 1989)). However, Rittle-Johnson, Star, and Durkin (2009) found that a boundary condition on problem comparison is that its usefulness varies depending on students' prior knowledge.

Practicing retrieval, a form of self-testing, has been shown to be an especially effective method of study (Karpicke, 2009). However, there is still more to be learned about the boundary conditions of this effect, especially in attempts to apply it in actual classroom practice. For example, Koedinger, Corbett, and Perfetti (2012) propose that specific types of learning processes are most applicable to specific kinds of knowledge components. According to their view, retrieval practice affects memory and fluency processes, which are most applicable to facts that can be memorized, rather than to principles that must be understood. It is not yet clear whether or not other boundary conditions, such as the effect of prior knowledge that has been shown to interact with the benefits of problem comparison, apply similarly to practicing retrieval.

In order to investigate the effect of practicing retrieval on learning statistics facts in an actual classroom setting, I designed a study using students in my Introductory Statistics classes as participants. I chose two sets of concepts, one about variability and another about two sample *t* tests, as the to-be-learned material. These concepts were chosen for two reasons. First, they are high priority learning objectives for the class, given that they are fundamental concepts upon which later concepts are based. Second, because practicing retrieval may be most effective for factual material that can be memorized (Koedinger et al., 2012), I wished to chose topics that fit that description as closely as possible. Thus, I chose topics that are typically beginning topics in Introductory Statistics, rather than a topic such as analysis of variance, whose understanding relies more on relating the new material to previously learned material. This enabled me to target facts such as “If everyone had the same score, then there would be *no* variability”, and “The researcher with the larger treatment effect is more likely to get a significant result” as the to-be-learned material.

Also, in order to investigate the possible role of ability level in learning by practicing retrieval, I divided students into, low, medium, and high ability in statistics in general. To do so, I used their average test scores on tests covering all the material to which they were exposed during the time covered by the experiment, excluding the to-be-learned material itself.

Method

Participants

The participants were 55 students in two sections of an Introductory Statistics class who participated for a small amount of extra credit in the course.

Materials and Procedure

The students were randomly divided into two groups, each of which was to participate in both an experimental condition (practicing retrieval) and a control condition (re-reading), but in opposite orders. A module on variability was taught early in the semester, and a module on *t*-tests was taught later. For each module, both groups studied a “Summary of Important Concepts” that consisted of a set of eight short paragraphs. Each paragraph described an example illustrating a target fact, and ended with stating the to-be-learned fact (see Appendix A). The study sessions consisted of a lesson followed by an activity in which the manipulation occurred.

First module. The first module (on variability) was done in the third and fourth weeks of the semester. It started with the "Summary of important concepts" during the Tuesday and Thursday classes of the third week. Tuesdays and Thursdays were 80 minute classes.

Lesson. On Tuesday of the third week, in addition to the content for that day's regularly scheduled syllabus topic, the students were presented with a 20 minute lesson on the first six of the eight concepts (on the meaning of variability, range, and interquartile range). On Thursday of the third week, again after the regular topic, they reviewed the first six concepts and were presented with the last two (on standard deviation) for 20 minutes. For both the Tuesday and Thursday lessons, all students in both conditions studied the summary for about 5 minutes. Then I gave a short lecture on the material, during which the students were allowed to ask questions and I asked them questions. Because there were two sections of statistics and students in the different sections asked different question to which I responded, the exact content of the lessons were slightly different between the two sections. However, subjects had been randomly assigned to the two conditions within the sections. Therefore, all subjects in both conditions within a section received the same lesson. All students were told that they should try to learn the material as well as possible, because in the next activity they would be asked questions about it.

Fridays were 50 minute classes. On Friday of the third week, for the entire 50 minutes, both groups received the activity in which practicing retrieval was manipulated as described below (also see Appendix B for examples of all of the activity materials). The students were told that this was to prepare them for an upcoming test.

Activity for the retrieval practice condition. For the retrieval practice condition, the students first took an open ended question quiz on the odd numbered paragraphs from the lesson, after which they were to check their answers. In order to check their answers, the next page gave them delayed feedback in which they saw the odd numbered paragraphs. They were also asked to rate how well they understood the material and to explain why the concept was easy or difficult to understand.

Next, for more retrieval practice, they took another open ended question quiz on the even paragraphs. After this, they did the same things that they did after the first quiz. In addition to using the paragraphs to check their answers they were asked to rate how likely it was that they

would get a correct answer on a test question on the paragraph, and they were asked to try to explain the concept in the paragraph in their own words.

Activity For the Control condition. For the control condition, there was no open ended quiz. Instead, the students reviewed the summary twice. On the first time, they were asked to rate, for all eight paragraphs, how well they understood each one, and to explain why the concept was easy or difficult to understand. On the second time they were asked to rate, again for all eight paragraphs, how well they thought they would do on an upcoming quiz, and to try to explain the concept in the paragraph in their own words.

Thus, the practice retrieval group worked with all eight paragraphs twice. However, one time they retrieved the paragraphs from memory by taking open ended quizzes, and the other time they actually saw the paragraphs and did ratings and explanations about them. They did the same rating and explaining activities as the control group. But they did half the ratings and explanations on one half of the paragraphs and the other half of the ratings and explanations on the other half.

The control group also worked with all eight paragraphs twice. However, they did not take any open ended quizzes. Instead, they *saw* all the paragraphs twice, once to do one half of the ratings and explanations on *all* of the paragraphs, not just half of them, and the other half of the ratings and explanation again on *all* of the paragraphs.

Test. On the following Tuesday (the fourth week) they had their test covering the first four chapters (chapter four was on variability), in which the questions for the concepts were embedded (see Appendix C for the embedded questions).

Second module. The second module, on *t* tests, occurred during the ninth, tenth, and eleventh week of the semester. Like the first module, the second module again started with a "Summary of important concepts" on a Tuesday and Thursday of the ninth week.

Lesson and activities for both conditions. On Tuesday of the ninth week, the students received a lesson on the first four concepts (on the independent measures t test and its standard error). On Thursday of the ninth week they reviewed the first four concepts and were presented with the last four (on the repeated measures t test and its standard error). Except for the difference in the number of concepts that were covered on each day, the same procedures were followed in the second module as in the first. The study activities were again done on Friday. For the second module both groups did the condition that they had not done on the first module. Except for the reversal of the conditions and the new topic, all the study activities for the second module were the same as for the first.

Test and covariate measure. In the tenth week, there was a weather related school closure on Tuesday (due to hurricane Sandy). On Thursday there was a make up lab for any students who had missed the study activities on the previous Friday. On Friday there was a lab on using SPSS to do *t* tests. On Tuesday of the eleventh week the students had their test covering chapters 10 (the independent measures t test) and chapter 11 (the repeated measures t test). As with the first module, the test questions for the concepts were embedded in that test (See appendix D). Thus, the test for the second module followed the same procedure as that for the first module, except that, due to hurricane Sandy, it was presented with an extra week's delay, rather than just a weekend delay.

The average test score for the regular items on the first four tests of the semester (not including the embedded items used for the experiment) was used as a general measure of ability in statistics. Those tests covered both the topics from which the concepts used in the experiment were drawn, plus the other topics usually covered at the beginning of an introductory statistics

for the behavioral sciences course. Among the additional topics were an introduction to research designs, frequency distributions, central tendency, *z* scores and standardized distributions, probability and the normal distribution, the distribution of sample means, the logic of hypothesis testing and the single sample *z* test, errors of inference and power, and the single sample *t* test. This ability measure was used as both a blocking variable and as a covariate in the data analysis.

Results

Of the 55 students, 36 provided data on all of the variables needed for the following analyses. The subjects were blocked on the lowest, middle, and highest thirds of the distribution of the ability measure. I conducted an analysis of variance on the ability measure using the groups into which the subjects had been randomly placed as one between subjects factor and the thirds on the ability measure as a blocking factor.

Table 1

Mean Performance on the Ability Measure as a Function of Blocks and Group

Group	Block			
	Lowest Third	Middle Third	Highest Third	Total
	M (SD) n	M (SD) n	M (SD) n	M (SD) n
Grp. 1 (Prac. Ret./Cont.)	59 (7.2) 6	74 (1.3) 7	84 (8.8) 6	72 (11.9) 19
Grp. 2(Cont./Prac. Ret.)	55 (4.5) 6	67 (4.8) 5	82 (7.5) 6	68 (12.8) 17
	57 (6.1) 12	71 (4.4) 12	83 (7.9) 12	

Table 1 shows that, as would be expected, there was a main effect of blocking by ability, F (2, 30) = 53.17, $p < .001$, MSE = 37.56. However, in spite of the random assignment, Group 1 (the group that did the practicing retrieval on the first module and the control task on the second module) had significantly more ability than Group 2, F (1, 30) = 4.17, $p = .050$, MSE = 37.56. There was no interaction between groups and the ability blocks. Thus, even within the blocks, the students' ability varied enough to justify using the ability measure scores as a covariate.

Therefore, for the next analysis I conducted an analysis of variance on the performance on the embedded test items using module as a within subjects factor, condition and the ability blocks as between subjects factors, and the ability measure as a covariate.

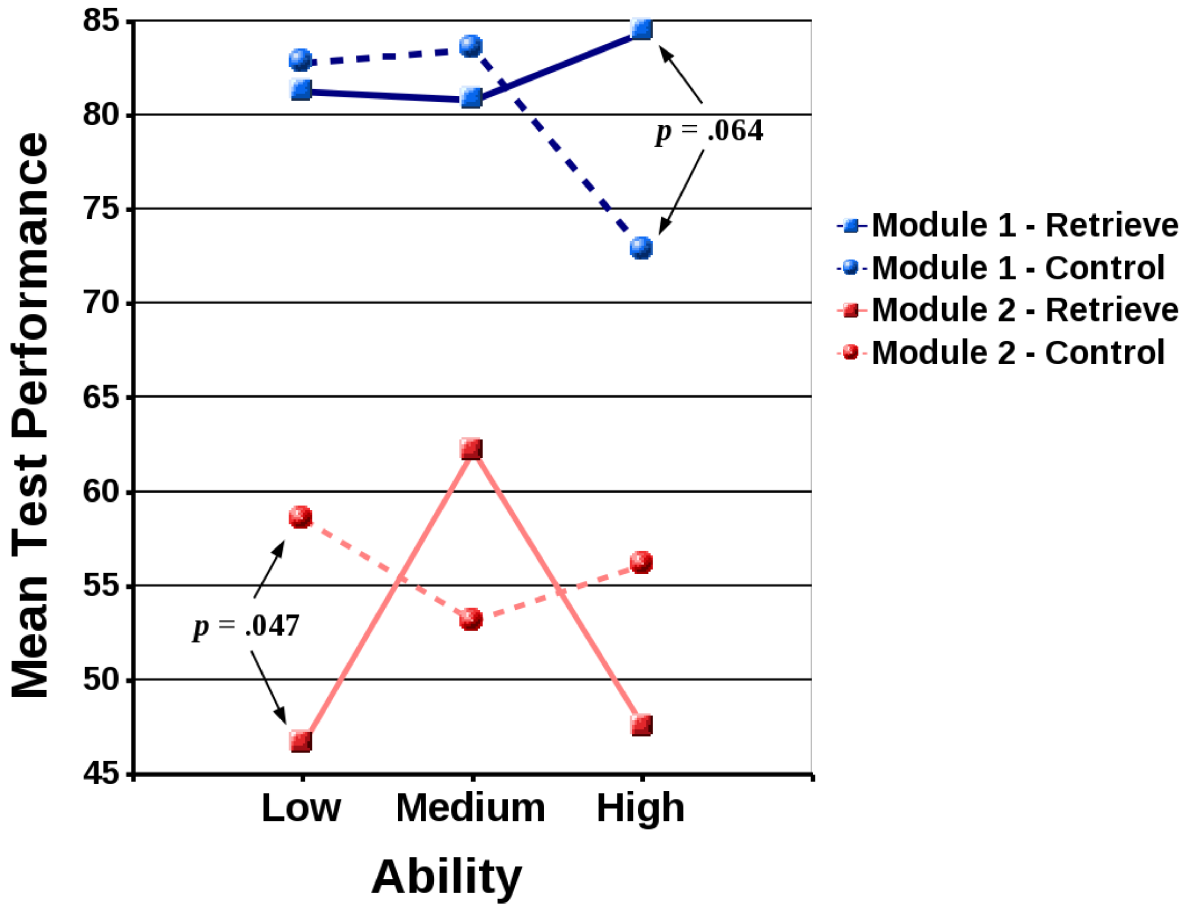

Figure 1. Test performance as a function of condition, module, and ability.

As shown in Figure 1, there were no main effects of condition or ability blocks, *F's* < 1. However, performance was significantly better on Module 1 than on Module 2, F (1, 29) = 5.89, p = .022 , *MSE* = 273.82. There were no interactions (*F's* < 1), although the three way interaction between condition, module, and ability blocks approached significance, F (2, 29) = 1.65, p = . 209 , *MSE* = 233.80.

Because the interaction between condition, module, and ability blocks approached significance, I conducted separate analyses examining the difference between the retrieval practice and control conditions for each of the six combinations of module and ability block. These were done by filtering the data on each of the ability blocks, and then conducting an analysis of covariance on performance on the embedded test items for the appropriate module using condition as a between subjects factor and the ability measure as a covariate. For the low

ability subjects, on the second module (on *t* tests), practicing retrieval resulted in significantly lower performance than the re-reading control task, F (1, 9) = 5.32, p = .047, MSE = 176.48. Although no other comparison was significant at the .05 significance level, one other comparison approached significance. For the high ability subjects, on the first module (on variability), practicing retrieval resulted in higher performance than the re-reading control task, F (1, 9) = 4.44, p = .064, MSE = 93.42.

Discussion

On the variability module, which, based on performance on the embedded questions, was the easier module, practicing retrieval had no effect on the low and medium ability students, but was helpful for the high ability students. On the t-test module, which was the more difficult module, practicing retrieval hurt the low ability students, but had no effect on the medium and high ability students.

Thus, this study showed that in this particular application of the practicing retrieval effect, practicing retrieval was not particularly effective, and where there were effects, they were different effects for different students. One thing that is not yet clear is what exactly is causing the differential effects. It could be that the students who were categorized into the different ability levels differed in prior knowledge. But they could also have differed in how they studied the to-be-learned facts. And that difference could have been because of differences in their study strategies per se, or differences in interest and motivation.

Also, why was practicing retrieval not more effective overall in this situation? One possibility is that the to-be-learned material was not really as factual in nature as I had supposed. If so, then the memory and fluency processes that Koedinger et al. (2012) theorize are affected by practicing retrieval may not have been especially important. Instead, this material may have better been considered as requiring understanding of principles. If so, then according to Koedinger et al., an instructional principle that affects understanding and sense making processes, such as prompted self-explaining may have been more beneficial. Thus, in future research, if practicing retrieval is what is manipulated, then it might be advisable to use to-be-learned learned material that is more clearly factual, rather than principled, in nature. Or, if the same materials were used, then it might be better to manipulate a different instructional event, such as self-explanation, rather than practicing retrieval.

Another consideration is that the experimental task may not have been as effective as I wished it to be at eliciting retrieval practice. Thus, in future research, regardless of which materials are used and what kind of instructional event is manipulated, attention might profitably be given to strengthening the manipulation.

References

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1147–1156. doi:http://dx.doi.org/10.1037/0278-7393.15.6.1147

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469–486. doi:10.1037/a0017341

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798. doi:10.1111/j.1551-6709.2012.01245.x

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving. *Journal of Educational Psychology*, *101*(4), 836–852. doi:10.1037/a0016026

Appendix A

The To-Be-Learned Materials

Summary of Concepts for Variability

Whenever we measure something about several individuals, we almost always get different measurements for the different individuals. For example, if we look at the heights of different people, we find that some are taller and some are shorter. If we measure the I.Q. of different people, we find that some are smarter than others. These differences are the variability in their measurements.

Variability is usually measured around the mean (or sometimes around the median). The mean represents the average individual. Some people will have scores greater than the average, some lesser. Some will be closer to the average, some further away. If everyone had the same score, then there would be *no* variability. But this seldom happens.

Suppose that in one group of people, most of their heights tend to be pretty close to the average for their group. But for another group of people, many of their heights are either a good bit taller or a good bit shorter than the average for *their* group. The second group has more variability. It doesn't matter whether the averages of the two groups are the same or different. The amount of variability is a separate issue from how large or small the average is.

So, if you have one group of tall people and a second group of shorter people, then the taller people could be more *variable* than the short people. But the taller people could also be *less variable* than the shorter people. The average height and the variability of the heights are two different things.

A simple, but not very useful, way to measure variability is to look at the difference between the highest and lowest scores in a group of scores. That's the *range.* Because it only uses two scores, it is easy to calculate. But because it ignores all the other scores, it doesn't necessarily give you a good idea of the variability of all the scores. Two groups of I.Q. scores could have the same range, yet one group could have more scores close to the mean (less variability) than the other group.

The *inter-quartile range* is a somewhat more useful measure of variability. It gives you the range from the score that cuts off the lowest 25% of the scores to the score that cuts off the highest 25%. So, it tells you how spread out the middle 50% of the scores are. It tells you how far away from the median you would (at least) have to be in order to be either in the lowest 25% of scores or the highest 25%. But it would still be possible for two groups of scores to have the same inter-quartile range, and yet for one group of scores the extremely high and low scores could be further away from the median than for the other group.

To get a measure of variability that takes into account all of the scores, you need the *standard deviation.* It gives you a good idea of the average distance that all the scores are away from the mean. Technically, it doesn't just average the distances because the distances above the mean and below it would exactly cancel each other out, giving you an average of zero every time. To avoid that problem, it calculates the average of the squares of the distances, and then takes the square root. This results in a number that is larger when the scores tend to be further away from the mean, and a number that is smaller when they tend to be closer.

The standard deviation is considered to be usually the best measure of variability because it takes into account all of the scores. If any score, or group of scores, no matter where they are located, were to mover further from the mean or closer, the standard deviation would get larger or smaller, respectively. Finally it gives you a way to specify what you mean by "close to the mean" or "far from the mean". Any score that is less than one standard deviation away from the mean can be considered fairly close, and any score more than one standard deviation away can be considered fairly far away.

Appendix A (cont.)

Summary of Concepts for Two Sample t Tests

If you give one group of subjects who have a certain average level of high blood pressure a treatment to lower their blood pressure, and you give another group who have that same level of high blood pressure a placebo treatment, then you can test whether the real treatment had any effect by looking at the difference between the average blood pressure of the groups after the treatment. If the difference is significantly greater than zero, then the treatment had an effect. So, the difference between the means measures the amount of treatment effect. To see if the treatment effect is significant, you use the independent measures t test.

The t statistic for the independent measures t test is calculated by dividing the difference between the two group means, that is, the treatment effect, by the standard error of the difference. The result will be significant if the value of the t statistic is large enough. Therefore, if two researchers each conduct an independent measures t test with the same standard error, the researcher with the larger treatment effect is more likely to get a significant result.

The standard error for the independent measures t test is calculated using the variability from the two samples. So the less variable the samples are, the smaller the standard error will be. Because the standard error is in the *denominator* of the fraction used to calculate the t statistic, the smaller the standard error, the *larger* the t statistic, and,therefore, the more likely that it will be significant. So, if two researchers each conduct an independent measures t test with the same treatment effect, the researcher whose samples have less variability in them is more likely to get a significant result.

You calculate the standard error by calculating a fraction with the variability of the two samples in the numerator, but in the *denominator* is the size of the samples. So, the larger the samples, the smaller the standard error. Because a smaller standard error is more likely to lead to a significant result, everything else being equal, larger samples are more likely to lead to a significant result.

If you measure the same subjects before and after a treatment to lower their blood pressure, you can calculate a difference score for each subject, to see how each subject changed from before to after the treatment. Then you can average the difference scores and see if the average is significantly greater than zero. If so, then the treatment had an effect. So the treatment effect is the average of the difference scores. In this case, to see if the treatment effect is significant, you use the repeated measures t test.

Once you have calculated difference scores, you use them as if they were one sample of data. Different subjects might have started with very different blood pressures, but that doesn't matter because all that goes into calculating the repeated measures t test is each subject's *change* in blood pressure, not their actual starting or ending blood pressure. Thus, the repeated measures t test eliminates from the data any variability due to individual differences between the subjects.

The t statistic for the repeated measures t test is calculated by dividing the average of the difference scores, that is, the treatment effect, by the standard error of the difference scores. Just as with the independent measures t test, you are more likely to get a significant result with a large treatment effect and a small standard error. The standard error is calculated from the difference scores. Because variability due to individual differences has been removed from the data, the standard error is likely to be smaller when you measure the same subjects twice rather than using two different groups of subjects.

The repeated measures design usually produces a smaller standard error than an equivalent study using an independent measures design. Thus, all other things being equal, the repeated measures design, because it removes variability due to individual differences, is more likely to produce a significant result than an independent measures design. However, that variability will only be removed if subjects who have relatively high scores before the treatment also have relatively high scores after it, and subjects who have relatively low scores before also have relatively low scores after. This is called having the before and after scores positively correlated. You only get the advantage of the repeated measures design if the before and after scores are positively correlated.

Appendix B

Examples of the Activity Materials

Retrieval Practice Condition

Quiz on the Odd Numbered Paragraphs

Answer the questions below. Then check your answers with the sheet provided.

1. Explain what we mean by variability.
2. Explain what we mean by the average and variability being two different things.
3. Explain why the range either is or is not a good measure of variability.
4. Explain how the standard deviation is calculated.

Instructions for Delayed Feedback on the Odd Numbered Paragraphs

Use this sheet to check how well you answered the questions. This will give you a review of the material you studied earlier.

Give yourself a percentage for how well your answer showed that you understood each concept. If you completely understood it, give yourself 100%. If you had no idea what it meant, give yourself 0%. If you had a fair, but not good, understanding of it, give yourself 50%, and so on.

Finally, write a short description of why each concept was easy or hard to understand or what you did or did not understand about it.

(The instructions were followed by the presentation of each of the four odd numbered paragraphs in the format shown below. To save space, only the first paragraph is shown here.)

Whenever we measure something about several individuals, we almost always get different measurements for the different individuals. For example, if we look at the heights of different people, we find that some are taller and some are shorter. If we measure the I.Q. of different people, we find that some are smarter than others. These differences are the variability in their measurements.

Percentage of understanding: _____

Description of why this concept was easy or hard to understand:

Appendix B (cont.)

Quiz on the Even Numbered Paragraphs

Answer the questions below. Then check your answers with the sheet provided.

1. Explain how it could happen that there is no variability.
2. Describe an example illustrating that the mean and variability are two different things.
3. Explain why inter-quartile range either is or is not a good measure of variability.
4. Explain what the standard deviation tells you.

Instructions for Delayed Feedback on the Even Numbered Paragraphs

Use this sheet to check how well you answered the questions. This will give you a review of the material you studied earlier.

For each concept, rate how likely it is, as a percent, that you would get a test question about that concept correct.

Finally, write a short explanation of the concept in your own words.

(The instructions were followed by the presentation of each of the four even numbered paragraphs in the format shown below. To save space, only the first paragraph is shown here.)

Variability is usually measured around the mean (or sometimes around the median). The mean represents the average individual. Some people will have scores greater than the average, some lesser. Some will be closer to the average, some further away. If everyone had the same score, then there would be *no* variability. But this seldom happens.

Likelihood of getting a test question correct: _____

In my own words, this concept means:

Appendix B (cont.)

Control Condition

Instructions For the First of the Two Reviews of the Summary for the Control Activity

For each concept, give a percentage of how well you understand it. If you completely understand it, give it 100%. If you have no idea what it means, give it 0%. If you have a fair, but not good, understanding of it, give it 50%, and so on.

For each concept, rate how likely it is, as a percent, that you would get a test question about that concept correct.

Finally, write a short description of why this concept is easy or hard to understand or what you do or do not understand about it..

(The instructions were followed by the presentation of each of the eight paragraphs in the format shown below. To save space, only the first paragraph is shown here.)

Whenever we measure something about several individuals, we almost always get different measurements for the different individuals. For example, if we look at the heights of different people, we find that some are taller and some are shorter. If we measure the I.Q. of different people, we find that some are smarter than others. These differences are the variability in their measurements.

Percentage of understanding: _____

Description of why this concept is easy or hard to understand:

Instructions For the Second of the Two Reviews of the Summary for the Control Activity

For each concept, rate how likely it is, as a percent, that you would get a test question about that concept correct.

Finally, write a short explanation of the concept in your own words.

(The instructions were followed by the presentation of each of the eight paragraphs in the format shown below. To save space, only the first paragraph is shown here.)

Whenever we measure something about several individuals, we almost always get different measurements for the different individuals. For example, if we look at the heights of different people, we find that some are taller and some are shorter. If we measure the I.Q. of different people, we find that some are smarter than others. These differences are the variability in their measurements.

Likelihood of getting a test question correct: _____

In my own words, this concept means:

Appendix C

The Embedded Test Questions for the First Module

1. Variability is
   a. the difference between the largest and smallest measurement.
   b. how large or how small people's measurements are.
   c. how much individuals' measurements tend to differ from the average.
   d. how large the largest measurement is.

2. In order for there to be no variability
   a. all individuals would have to have the same measurement.
   b. all individuals would have to have a very large value for their measurement.
   c. all the measurements would have to be very small.
   d. all the measurements would have to be zero.

3. In group A the measurements are 80, 81, 77, 83, and 79.
   In group B the measurements are 32, 11, 58, 24, and 45.
   a. Group A has the greater variability.
   b. Group B has the greater variability.
   c. The groups have the same variability.
   d. From this information, there is no way to tell which group has greater variability.

4. In group A the measurements are 5, 10, 15, 20, and 25.
   In group B the measurements are 45, 50, 55, 60, and 65.
   a. Group A has the greater variability.
   b. Group B has the greater variability.
   c. The groups have the same variability.
   d. From this information, there is no way to tell which group has greater variability.

5. In group A the measurements are 10, 36, 38, 40, 42, 44, 70.
   In group B the measurements are 10, 22, 25, 40, 55, 58, 70.
   a. Group A has the greater variability.
   b. Group B has the greater variability.
   c. The groups have the same variability.
   d. From this information, there is no way to tell which group has greater variability.

Appendix C (cont.)

6. Which of the following is true of the inter-quartile range?
   a. It takes into account all of the scores.
   b. It tells you how far below the median you would at least have to be in order to be in the bottom 50% of the scores.
   c. It tells you how far the highest 10% of the scores are from the median.
   d. It tells you how spread out the middle 50% of the scores are.

7. The standard deviation is calculated by.
   a. averaging all the scores together and taking the square root.
   b. averaging all of the distances of the scores from the mean and taking the square root.
   c. averaging the squares of all the distances of the scores from the mean and taking the square root.
   d. determining the difference between the highest and the lowest scores.

8. A score can be considered fairly high
   a. as long as it is at least one standard deviation above the mean.
   b. only if it is more than three standard deviations above the mean.
   c. if it is anywhere above the mean.
   d. There is no good way to tell when a score is fairly high.

Appendix D

The Embedded Test Questions for the Second Module

1. In an independent measures t test, the size of the treatment effect is indicated by
   a. the size of the mean of the treated group.
   b. the standard error.
   c. the amount of difference between the group means.
   d. the statistical significance of the test.

2. Suppose two researchers each conducted an independent measures t test using the same standard error. Researcher A might get a statistically significant result, whereas researcher B might not because researcher A might have a
   a. higher average for his treated group.
   b. smaller sample.
   c. larger treatment effect.
   d. larger standard error.

3. Suppose two researchers each conducted an independent measures t test. If both had the same sized samples and the same size treatment effect, but the samples that one of them used had more variability than the samples that other used, then the researcher with the more variable samples would see
   a. a larger value for the t statistic.
   b. the same size standard error.
   c. a smaller value for the t statistic.
   d. a smaller standard error.

4. If everything else is equal, larger samples will
   a. be less likely to lead to a significant result.
   b. be more likely to lead to a significant result.
   c. be more likely to lead to a larger treatment effect.
   d. be more likely to lead to a smaller treatment effect.

Appendix D (cont.)

5. In a repeated measures t test, the average of the difference scores is
   a. always zero.
   b. always greater than zero.
   c. the standard error.
   d. the treatment effect.

6. The repeated measures t test eliminates variability due to individual differences because
   a. the treatment effect is very small.
   b. the standard error is very small.
   c. only the difference scores are used in calculating the test.
   d. the actual before and after scores are used in calculating the test.

7. In a repeated measures t test, what problem is created when some subjects have much different scores than other subjects before the treatment?
   a. It results in a small value for the t statistic.
   b. It results in a very large standard error.
   c. It results in a small treatment effect.
   d. It doesn't matter because only the change from before to after the treatment is used to calculate the test.

8. In the repeated measures t test, removing variability due to individual differences results in a smaller standard error if what is also true?
   a. The sample is small.
   b. The before and after scores are positively correlated.
   c. The treatment effect is large.
   d. Removing the variability due to individual differences always results in a smaller standard error.