Prompted Self-Explanations Improve Learning in Statistics but Not Retention

Robert S. Ryan, PhD[1] and James A. Koppenhofer, BS[1]

Author Note

The participants in the research reported here were treated in accordance with the ethical principles of the American Psychological Association and the regulations governing treatment of human research participants of Kutztown University and of the United States.

Abstract

Background: College students often do not retain what they learn in Statistics in order to apply it in Experimental Psychology. Self-explanation, that is, elaborating on what one is trying to learn by asking questions, making inferences, etc., improves learning (Chi et al., 1989) and may improve retention.

Objective: The purpose of this study was to determine whether self-explanation was superior to students' usual study methods specifically for learning some basic concepts in statistics, and, if so, if it was similarly useful for retention a semester after the initial learning.

Method: We used 199 college students as participants in a randomized, between participant, two-part experiment examining the effects of training by prompting self-explanations as a potential solution to this applied problem.

Results: The self-explanations that we elicited improved initial learning and were superior to students' usual study methods, but did not benefit retention.

Conclusions: Future research on improving the quality of the self-explanations and training with spaced retrieval practice, in order to benefit retention, is suggested.

Teaching Implication: Self-explanation should be implemented for teaching statistics in order to benefit initial learning. However, teachers should explore other methods to accomplish retention.

Keywords: college teaching, teaching statistics, learning statistics, long-term retention

Prompted Self-Explanations Improve Learning in Statistics but Not Retention

Better problem solvers produce comments, called self-explanations, while they study (Chi et al., 1989). These self-explanations contain elaborations, such as connecting the material to be learned to previous knowledge, making inferences, and generating questions and predictions of their answers, thus producing better learning (Chi et al., 1994). However, previous research on instructional methods to improve learning has shown that different instructional principles have benefits for different learning processes (Koedinger et al., 2012). For example, prompted self-explanations increase understanding (Chi et al., 1994) whereas spacing of practice improves memory and fluency (Cepeda et al., 2009). Thus, it is important to take such differences into consideration for purposes of choosing an instructional method for a particular application. The experiment reported here examined whether self-explaining can be usefully applied for the initial acquisition of concepts in statistics and retention of them a full academic semester after they were learned. The statistics materials used were two of the most basic concepts taught in introductory statistics for the behavioral sciences, as will be explained further below.

Possible Benefits of Self-Explanation for Learning Statistics

When choosing an instructional method for a particular application, one concern could be the domain of knowledge. For example, some of the early research that showed the benefits of self-explaining was done in the domains of eighth grade science (Chi et al., 1994) or physics problem solving (Chi & VanLehn, 1991). Therefore, the results of those studies may not apply to the domain of undergraduate statistics. Similarly, more recent studies, such as Talley and Scherer (2013) and Margulieux and Catrambone (2019) showed benefits of self-explaining, but they were in the domains of physiological psychology and learning subgoals in programming, respectively.

Additionally, none of those previous studies examined the effect of self-explaining on retention after a delay.

On the other hand, Leppink et al. (2012) examined the effectiveness of self-explaining and studying worked examples for learning one of the most difficult concepts in the domain of statistics, the central limit theorem. They found that the instructional method interacted with the students' amount of prior knowledge. That finding is important for purposes of improving students' initial acquisition of such material, but it still does not address how long students retained the knowledge.

The Practical Issue of Retention

Retention is an important concern for educational applications. For example, undergraduate psychology majors typically conduct a research project in their experimental psychology or research methods class in which they gather data, and then need to use the appropriate inferential statistics procedure to analyze their data. The students learn how to conduct the procedures in a statistics class. Importantly, they also learn how to determine which statistical procedure is appropriate for analyzing the data depending on how the data was collected. Accordingly, one of the most basic concepts in statistics that they need to learn are the associations between features of a research design and the appropriate procedure for that design. Two such associations are, first, whether to use a *t*-test or an ANOVA depending upon whether there are only two conditions, or more than two, and second, whether to use an independent measures or a repeated measures procedure depending upon whether the conditions are between participants or within participants. However, because such students often take their statistics class at least a semester before the research methods class, they need to retain the knowledge

described above for at least one semester. Unfortunately, based on the teaching experiences of the first author and of his colleagues, they often fail to do so.

Constraints on When Self-Explanation Aids Learning

A recent review of the constraints on when eliciting self-explanations aids learning (Rittle-Johnson & Loehr, 2017) provides some insight into whether that instructional method might be useful for helping statistics students to initially acquire and then retain the basic concepts they need for their experimental psychology class. The review proposed four constraints that should be taken into account when considering whether and how to use an instructional method that involves self-explanation.

First, self-explaining is best for domains such as math and science that are guided by general principles (Aleven & Koedinger, 2002). This is especially true where the applications of principles are consistent, as opposed to where there are many exceptions, such as in English grammar (Wylie et al., 2010, 2011). Statistical tests are mathematical procedures for analyzing data from experimental designs used in psychological science, as well as other sciences. Also, for learning which statistical procedure to use based on the features of a particular research design, the associations are consistent. These facts argue in favor of using self-explaining as an instructional method to learn the associations. However, they do not necessarily address whether self-explaining would be useful for retention.

Second, self-explaining one's own answers can reduce learning if the answers are incorrect. For example, in one study children were prompted to explain their predictions about forecasting earthquakes or about the progress of an ocean voyage. The predictions were often incorrect. They were later poorer at making evidence-based claims than control participants who did not explain (Kuhn & Katz, 2009). However, explaining *why* correct solutions are correct, and

contrasting those solutions to ones that are incorrect, along with explaining why they are incorrect, is useful (de Bruin et al., 2007; Howie & Vicente, 1998; Huk & Ludwigs, 2009; McEldoon et al., 2013; Pillow et al., 2002; Rittle-Johnson, 2006). These findings suggest that in order for self-explaining about which statistical procedure to use for a particular research design to be useful for learning, the participants should be sufficiently trained to self-explain so that their explanations are likely to be correct. However, once again, that does not necessarily mean that those correct explanations would be useful for retention.

The third constraint is in regard to the type of explanation prompt. Prompts can focus learners on one type of information, but possibly at the expense of some other type. For example, in studies of using self-explaining to learn about probability problems or tax-law problems, prompts to explain why some particular step of a procedure is performed focused learners on domain principles, but they resulted in poorer transfer (Berthold & Renkl, 2009; Berthold et al., 2011). In statistics the appropriate statistical procedure to use for analyzing the data from a particular research design is associated with a specific feature of that design. Therefore, transfer of knowledge from one example of a particular design to another would not be an issue as long as it was obvious that both examples contained the same relevant feature. The findings cited above suggest that self-explanation prompts should foster the acquisition of the correct associations if they (a) ask the student to explain why the procedure is appropriate, (b) ask them to do so specifically in terms of the relevant feature, and (c) provide the relevant feature in the prompt. However, retention would still be another question.

The fourth constraint is that it is important to take into consideration the condition to which self-explaining is being compared, both in terms of the time on task and the type of instructional method. In Chi et al. (1994) an attempt to control for time on task was made by

having the control participants re-read the same text that the experimental participants explained. However, the attempt was not successful. The experimental participants spent, on average, 2 hours and 5 minutes, whereas the control participants spent, on average, 1 hour and 6 minutes. However, other studies that did control time on task did show benefits of self-explaining (e.g., Atkinson et al., 2003; Bielaczyc et al., 1995; de Bruin et al., 2007; De Koning et al., 2011). This was usually done by having the control participants engage in their own self-selected study method for the same amount of time as the self-explainers. These findings remind us that, regardless of how it is done, our examination of the benefits of self-explanation, specifically for learning and retaining the associations between features of a research design and the appropriate statistical test, must use a control condition that equalizes time on task.

In terms of comparison to another instructional method, instructional explanations, that is explanations produced by experts, were often equally effective as self-explanation (Cho & Jonassen, 2012; Crowley & Siegler, 1999; de Koning et al., 2010; Rittle-Johnson et al., 2015; Tenenbaum et al., 2008). Also, a meta-analysis of six such studies found no difference in benefits between self-explanation and expert instructional explanations (Wittwer & Renkl, 2010). In the present experiment, self-explanation during studying eight examples was compared to re-stating information in the same studied examples. This was done in order to compare self-explanation to a study method that is similar to study methods that students often use, such as re-reading or highlighting, and to equate time on task.

Self-Explanation and Retention

Few studies in the previous literature examined the use of self-explanation in statistics. There were some studies that did examine the effects of self-explaining and reported that they examined retention. However, one study used the quality of participants' self-explanation as a

measure of retention. That study examined how variations in participants' game playing behavior in a computer game-based learning environment impacted their performance while playing the game, as well as at posttest and after a 1-week retention delay. Participants whose behavior was more persistent (deterministic or controlled), as opposed to more random, produced better self-explanations after the delay (Snow et al., 2016).

Other studies did not use the kind of retention delay in which we were interested in our study. For example, Chi (2018) examined learning HTML computer skills with either self-explaining using a screen shot, self-explaining using a screen cast, self-explaining with no visualization, or no self-explaining. All of the self-explaining conditions produced better retention than the no self-explaining condition. However, the retention test occurred immediately after the manipulation.

De Koning et al. (2011) examined learning about the circulatory system from an animation. They reported that self-explanations enhanced immediate acquisition if the animation was cued by highlighting its relevant parts, but that it did not enhance retention. However, they measured immediate acquisition with an inference test followed by measuring retention with a transfer test. Therefore, the only retention delay was the amount of time for the participants to complete the inference test.

Some studies examined the effects of self-explanation on retention after delays of 1 to 3 weeks, although not in the domain of statistics. For example, pilots who self-explained the reasons for safety procedures were more compliant with those procedures a week later than pilots who did not self-explain them (Molesworth et al., 2011). Mathan (2004) examined participants’ learning of absolute versus relative cell referencing in spreadsheets using a computerized tutor.

Participants who used an intelligent novice model to explain why errors occurred produced better retention 8 days after the initial training began.

Hsu et al. (2016) compared the effects of solitary self-explaining, collaborative self-explaining, and no self-explaining on learning of scientific concepts through computer games. In general, high engagement participants performed better than low engagement participants. On a retention test after a 3-week delay, high engagement was beneficial for both the solitary self-explaining participants and the collaborative self-explaining participants compared to both the control participants and to the low engagement participants.

More importantly for purposes of our examination of a longer retention delay, there was one study that examined retention after a delay similar to our full semester delay, although, again, it was not in the domain of statistics. After training participants and testing them in neurology topics over a 4-week period, Larsen et al. (2013) tested their retention 6 months later. Self-explaining was beneficial for retention when compared to studying a review sheet without self-explaining, although there were also benefits of repeated testing.

The Present Experiment and Hypothesis

Our review of the literature showed that the practical problem of semester-to-semester retention, specifically of basic statistics concepts, by the use of self-explaining has not been addressed. However, there was some evidence that self-explaining might have been somewhat useful for retention over shorter intervals in other domains (e.g., Hsu et al., 2016; Mathan, 2004; Molesworth et al., 2011). Also, the self-explanation review of Rittle-Johnson and Loehr (2017) suggests why self-explaining might be helpful for learning and retention of basic concepts in statistics.

Based on the aforementioned considerations, we conducted a pre-registered, high powered, experiment in which we first trained naive participants in the basic statistics concepts mentioned above. Then they studied eight examples in which they were prompted to either self-explain or to re-state information in the examples. To measure initial acquisition of the concepts they took a pre-test before the training and an immediate posttest after it. A semester later we recruited them to return for a final retention test. Our primary confirmatory research question was whether prompted self-explanation would result in better semester to semester retention of statistics concepts than a restating control condition, given that both conditions result in initial learning.

Method

Participants

Participants were recruited from the Psychology Department's human participant pool. Because of the applied nature of the experiment, the important demographic characteristic of the participants was that they were college undergraduates. Thus, our sample was representative of the students to which we would seek to apply our findings.

We invited the participants to participate in the initial training for the experiment in partial fulfillment of the research participation requirement for their Introductory Psychology class. They were offered a $15 payment for returning for the retention test. The goal, according to the pre-registered plan, was to train enough participants to enable retention testing of 200 participants, 100 in each condition. Using 100 participants per condition, a two independent samples *t*-test between the retention scores of the two conditions has a power of .94 to detect a minimum effect of $d = 0.50$.

We ran 364 participants through the training procedure and were able to recruit a total of 98 control participants and 101 experimental participants to return for the retention test. We decided to do the analysis with the extra participant in the experimental condition, and the shortage of two in the control condition, on the justification that doing so would result in only a very slightly unbalanced design, and it would allow us to come as close as possible to the full 200 required for the full power we were seeking.

Materials

The materials consisted of a pretest, some initial training, eight study examples, a posttest, and a retention test. We constructed three versions of the test and counterbalanced all six possible orderings of the versions of the test with the time of test between participants.

The tests had five examples that described a research situation. One of the test examples with its questions appears in Appendix A on the Open Science Framework (OSF; Ryan & Koppenhofer, 2021). The initial training was designed to acquaint the naive participants with the meaning of terms used in research such as "participants" and "conditions" (see Appendix B on the OSF; Ryan & Koppenhofer, 2021).

The eight study examples were provided in writing on a response sheet. The participants opened an electronic version of the response sheet on a computer in order to work with the prompts embedded in it. The procedure section below will describe in greater detail how the examples were used. Each example described a research situation. The example contained information about the research situation that could be used to notice the features that determined the correct statistical procedure. For example, the information made it possible to notice whether there were only two conditions or more than two conditions, and whether the conditions were between participants or within participants. Also, the example provided the name of the correct

statistical procedure to use (see Appendix C on the OSF for one example as it appeared to the self-explain participants and to the restate participants; Ryan & Koppenhofer, 2021). As will be described below, the experimenter used a script to run one participant at a time through their procedure.

Procedure

The six possible orderings of the versions of the tests were crossed with the two conditions, self-explaining and restating, to form 12 conditions with 25 participants in each condition. We used https://www.random.org/ to assign participant numbers randomly to the first 300 participants (the same procedure was used a second time when we needed to run more than 300 participants through the initial training). Then the experimenters ran the participants in participant number order allowing the advance randomization to determine the condition.

Participants were run one at a time. The experimenter followed a written scripted procedure. After informed consent was obtained, the participant was given a brief explanation of what would occur for the experiment, including that they would later be contacted to recruit them to return for the retention test, for which they would be paid $15.

All of the response sheets were prepared in advance and stored on the OSF (Ryan & Koppenhofer, 2021). The response sheet contained the pretest, the eight study examples, the posttest, and the retention test for one participant. After the brief explanation, the experimenter opened the appropriate response sheet for the participant. The first page collected the contact information of the participant in order to enable recruiting them for the retention test (all those first pages were later discarded).

Pretest

After the participants completed the first page they stopped on a blank page while the experimenter gave them instructions for the pre-test. They were instructed to answer all questions (by highlighting their choice) even if they were only guessing and to do the examples in order, not looking ahead or looking back. Each page of the pretest had one example and its two questions. They were not timed.

Initial Training

After the pre-test, the participant was given the initial training. The instructions to the participant were, “This initial training will familiarize you with the concepts and the terms associated with them. Please read it, make sure you understand it, and ask me any questions you have." The most common question was asking for either a verbal re-iteration of the difference between a between participants versus a within participants design, or asking for examples to illustrate the difference. Once the experimenter was confident that the participant understood the initial training, they moved on to the study examples.

Study Examples

For the study examples, the instructions to the participant made them aware that there would be eight examples, and what they and the experimenter would do for each example. For the first four examples, the instructions to the participant were:

> "For the first four examples, I'll read the sentences of the example, I'll read the question, and I'll read what would be a good answer to the question. Also, sometimes I'll make a comment about the answer. You can just follow along," as shown in the folder for Materials for data collection on the OSF (Ryan & Koppenhofer, 2021).

For the next two examples the instructions to the participant were:

> "For the next two examples, I'd like you to read the sentences of the example. You can just read them to yourself. Then read the question to yourself. Then, after the question I'll ask you to type in what you believe would be a good answer to the question, according to the way we've been doing this. And I'll guide you along for these two examples." For the last two examples the instructions to the participant were, "There are just two examples left. Numbers 7 and 8. You have the idea now. I'll let you do these on your own. Just let me know when you're finished," as shown in the folder for Materials for data collection on the OSF (Ryan & Koppenhofer, 2021).

For the experimental participants, the questions prompted them to explain why a particular statistical procedure (such as t-test) was the correct procedure to associate with the research design (in this case a good explanation would be because the design had only two conditions). They were instructed to use their prior knowledge from the initial training to answer the question. This was done to ensure that their answers were consistent with the definition of self-explanation from previous literature. Thus, they kept the initial training material in front of them. They were encouraged to refer to it to ensure that their explanations were correct.

For the control participants, the questions prompted them to re-state something that they had just read, but without explaining. Thus, they did not have the initial training material in front of them when they did their restating. This method of studying was used as the control to be as similar as possible to the methods, such as highlighting and re-reading, that students are known to typically use. The script that the experimenter used contained the examples. The instructions to *the experimenter* were:

> “You read to the subject. The subject follows along. Be sure to read exactly what is in the script with no ad-libbing (except, do not read the label at the top of the

example). Read slowly, and with expression. Read a phrase at a time, providing emphasis where it will contribute to keeping the subject engaged, and helping the subject understand.

The subjects will see exactly what you see in this script, except for the italicized parts. Therefore, when you come to the italicized parts, you need to pause, and make eye contact with the subject, so that they know you are instructing them in how to do their task," as shown in the folder for Materials for data collection on the OSF (Ryan & Koppenhofer, 2021).

Appendix D on the OSF (Ryan & Koppenhofer, 2021) shows one page of the script for each condition to illustrate how the experimenter instructed the subjects to give good answers to the prompts.

Posttest

After the study examples, the participants took the posttest. The instructions for the posttest were the same as for the pretest except that the participants were encouraged to try to use what they learned from the examples to give the correct answers. Then they were reminded that they would be contacted later to return for the retention test, for which they would be paid $15.

Retention Test

At the beginning of the semester after the participants were initially trained, they were contacted to recruit them to return for the retention test. They were run individually. The experimenter verified that they were the same person who participated in the training, and that, since the training, they had not been exposed, either in a class or by discussions with other students, to the statistics concepts on which they were to be tested. One participant's retention

test data was not able to be used because of such previous exposure. Further details about the procedure are available in the scripts on the OSF (Ryan & Koppenhofer, 2021).

Design

The dependent variable was the participant's scores on the tests. The independent variables were the training condition, which was between participants, and the time of test. The order of the versions of the test was a counterbalancing variable between participants.

Results

We obtained usable data for the pretest, training, and posttest from 364 participants (self-explain, n = 178; restate, n = 186). Of those, we also obtained usable data from the retention test from 199 participants (self-explain, n = 101; restate, n = 98). All of the data was thoroughly checked for accuracy.

All of the analyses reported in this paper were conducted using the R statistical analysis program. All of the analyses of variance (ANOVAs) were conducted using the aov command in the psych package. All of the contrasts were conducted using the lm command in the multcomp package. The scripts for the pre-planned analyses are all available on the Open Science Framework in a file folder for the confirmatory analyses (Ryan & Koppenhofer, 2021). The scripts for any of the exploratory analyses are also available on the Open Science Framework in a file folder for the exploratory analyses.

Acquisition

A preliminary analysis found no main effect for order $F(5, 187) = 1.85$, $p = .105$. Of the three possible interactions with order, only one, the one with time, was significant, $F(10, 374) =$

1.99, $p = .0336$, $\eta^2 = .05$ [1]. Figure 1 shows the six means of the test scores as a function of test time and training condition along with their 95% confidence intervals. An omnibus test confirmed that there were significant differences among those six means, $F(5, 591) = 45.43$, $p < .001$, $\eta^2 = .28$.

There was a significant interaction between time and condition, $F(2, 394) = 12.46$, $p < .001$, $\eta^2 = .06$. Also, across the three test times, the self-explanation condition ($M = 72.2\%$, $SD = 26.4$, $n = 303$, 95% CI [69.24, 75.19]) performed slightly better than the restate control condition ($M = 67.9\%$, $SD = 23.0$, $n = 294$, 95% CI [65.30, 70.55]). The 4.3% difference was statistically significant, although not necessarily meaningfully large, $F(1, 197) = 5.42$, $p = .021$, $\eta^2 = .03$, 95% CI [0.31, 8.26]. There was also a significant, and large, main effect of time, $F(2, 394) = 105.36$, $p < .001$, $\eta^2 = .35$. A pre-planned linear contrast between pretest and posttest, collapsing over conditions, showed a significant 27% improvement, $t(394) = 12.55$, $p < .001$, $d = 0.89$ [2], 95% CI [23.02, 31.55].

In the experimental condition there was a significant, and large, increase from pretest to posttest of 36%, $t(394) = 12.10$, $p < .001$. $d = 1.20$, 95% CI [30.19, 42.08]. In the restate control condition, there was only a significant 18% increase, $t(394) = 5.99$, $p < .001$, $d = 0.61$, 95% CI [12.13, 24.20]. An exploratory contrast (not included in the pre-registered analysis plan) showed that the 16% difference at posttest between the experimental condition and the control condition was statistically significant, and reasonably meaningfully large, $t(591) = 5.37$, $p < .001$, $d = 0.76$, 95% CI [10.27, 22.07].

---

1 All eta squares reported in this paper are partial eta squares. They 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. Eta squares of .01, .06, and .14 are considered small, medium, and large, respectively, retrieved from https://www.spss-tutorials.com, 01/03/22.

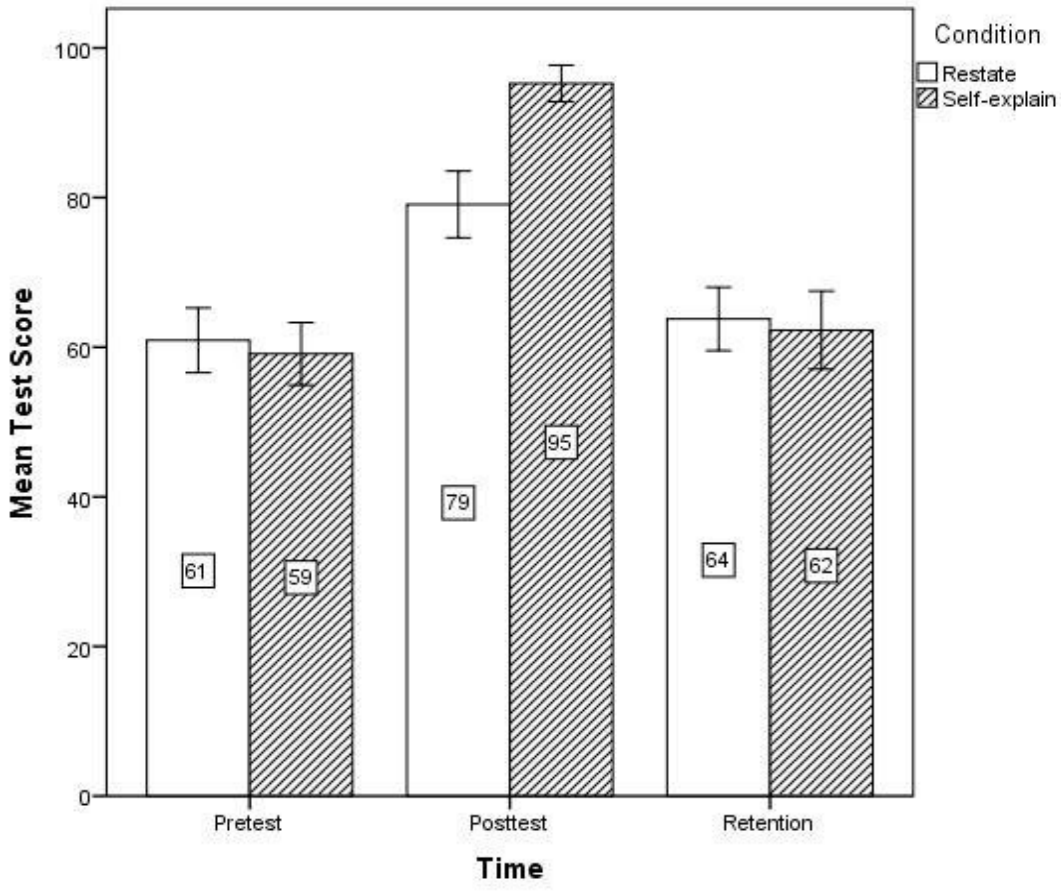
2 A Cohen's *d* of 0.2, 0.5, and 0.8 is considered small, medium, and large, respectively, retrieved from https://www.simplypsychology.ord, 01/03/22.

Retention

A pre-planned contrast just between the posttest and retention test, collapsed across conditions, showed that performance dropped a significant 24%, $t(394) = 11.16$, $p < .001$, $d = 0.79$, 95% CI [20.0, 28.5]. Exploratory contrasts showed that in the experimental condition there was a significant 33% drop, $t(394) = 11.04$, $p < .001$, $d = 1.10$, 95% CI [27.03, 38.91]. In the restate control condition, there was a significant 15% drop, $t(394) = 5.05$, $p < .001$, $d = 0.51$, 95% CI [9.27, 21.34].

Figure 1

Mean Test Performance as a Function of Training Condition and Time of Test



Note. Error bars are 95% confidence intervals.

Pre-Registered Confirmatory Hypothesis

The prediction that the self-explanation condition would be better for retention was not confirmed in that the control condition performed slightly better on the retention test. The contrast between the conditions showed that the less than 2% difference was not significant, $t(591) = 0.50$, $p = .619$, $d = -0.07$, 95% CI [-7.40, 4.40]. Therefore, even though the experimental condition had been beneficial for initial acquisition, there was no evidence that either condition was any more beneficial for retention than the other (or at least it was plausible that there was no real effect greater than $d = 0.50$, given that we had approximately .94 power to detect such an effect).

Exploratory Examination of Quality of Self-Explanations

To measure the quality of the self-explanations, the second author divided each of the participant's explanations on just the last two of the study examples into three portions worth one point each. Participants received one point each for defining a term correctly, for setting up a self-explanatory answer, and for giving the correct information to support that self-explanation. Appendix E on the OSF (Ryan & Koppenhofer, 2021) provides a short illustration of the coding scheme. The coded data, and a more complete description of the coding scheme, including an example from a participant's response, is available in the open materials on the OSF in the folder for results (Ryan & Koppenhofer, 2021).

The second author also categorized the data based on whether it was in regard to questions about the number of conditions in the example (which we believed would be an easier concept) or about whether the conditions were within or between (a harder concept). We had previously determined that there was a difference in difficulty between those two kinds of questions in the following way. In an exploratory analysis we had examined any differences in

performance on the two levels of difficulty of questions by conducting a three-factor mixed between subjects and within subjects analysis of variance on test scores. It compared the performance of the two conditions as a between subjects factor, using the three test times as one within subjects factor and question difficulty as a second within subjects factor. The results for condition and time were the same as in the analyses reported above. There was a condition by difficulty interaction which we do not report because it was likely due to a ceiling effect. However, in support of our intuition about the difference between the types of questions, we found that across conditions and times, performance on the questions that we considered easier, (M = 77.88%, SD = 29.7, n = 597, 95% CI [74.51, 79.26]) was better than the performance on the questions that we considered harder, (M = 63.32%, SD = 33.8, n = 597, 95% CI [60.61, 66.03]) by about 14%, $F(1, 591) = 70.62$, $p < .001$, $\eta^2 = .11$, 95% CI [10.39, 16.74].

Table 1 shows the correlations between the quality and the test scores broken down by the difficulty of the concept and their 95% confidence intervals. Quality was more strongly related to performance on the harder concepts, and more so for the posttest than the retention test. The scatterplots (available in the open materials) showed that a small number of cases may have been driving the correlations. An examination of the individual scores that contributed to the correlation between quality of self-explanation and overall accuracy on the retention test showed that most of the 101 self-explanation quality scores were 10, 11, or 12, with only 6 scores that were either 6, 7, 8, or 9 (none were below 6). Those 6 low scores were associated with exceptionally low retention scores. Similarly, it was also the case that most of the 178 quality of self-explanation scores that were correlated with the posttest scores were 10, 11, or 12 with only 15 scores that were either 5, 6, 7, 8, or 9 (none were below 5). Those 15 low scores were not associated with especially low retention scores. Nevertheless, we still considered them

suspect as possible outliers. Even though any of those extreme scores may turn out to be reproducible, rather than random outliers, for full transparency Table 2 shows the correlations with those scores removed. Once removed, the relationships weakened, but the pattern remained.

Table 1

Correlations Between Quality of Self-Explanations and Test Scores for Easier and Harder Concepts, Including All Scores

	Type of Question		
Test	Easier	Harder	Total
Posttest	-.062 (178)	.334 (178) ***	.261 (178) ***
	[-.207, .086]	[.197, .459]	[.118, .393]
Retention	.097 (101)	.175 (101)	.204 (101) *
	[-.101, .287]	[-.021, .358]	[.009, .384]

Note. n's in parentheses, 95% CI's in brackets

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 2

Correlations Between Quality of Self-Explanations and Test Scores for Easier and Harder Concepts, Excluding Potential Outliers

	Type of Question		
Test	Easier	Harder	Total
Posttest	-.025 (163)	.258 (163) ***	.214 (163) **
	[-.178, .129]	[.109, .396]	[.062, .356]
Retention	.084 (95)	.128 (95)	.101 (95)
	[-.119, .281]	[-.075, .321]	[-.102, .297]

Note. n's in parentheses, 95% CI's in brackets

* $p < .05$, ** $p < .01$, *** $p < .001$

Discussion

Self-explanation resulted in better initial acquisition of statistics concepts than a restating control condition. However, there was very little evidence of a beneficial effect of self-explaining on retention over a semester delay. Our pre-registered analysis plan provided evidence from a high-powered test that these particular prompted self-explanations were not beneficial for retention.

On the other hand, regarding the possible effect of the quality of the self-explanations, in our additional exploratory analyses we did find a weak, positive correlation between the quality of the self-explanations and the overall scores on the retention test, across the easier and harder concepts. However, that correlation may have been misleading, being due to only a few scores that might be considered outliers. Regarding the possibility of a benefit of the quality of the self-explanations for immediate acquisition, the correlation between the quality of the self-explanations and the overall scores on the posttest did not suffer from the problem of outliers. However, the correlation, although statistically significant, was low. Despite the problems with the correlations, there was a consistent pattern that the correlations were stronger between the quality of the self-explanations and the scores for the harder concepts than for the easier concepts, on both the immediate acquisition and the retention tests. That raises the possibility that investigating ways to improve the quality of the self-explanations might uncover benefits for retention as well as for immediate acquisition.

These results suggest that just trying to self-explain, even if the self-explanations are not expressed particularly well, is beneficial for initial acquisition and that expressing them better might not be much more beneficial. Regarding retention, these results suggest that the kinds of

self-explanations that we elicited with our prompts are not beneficial. However, the possibility that better expressed self-explanations might be beneficial, although not strongly supported, was not ruled out.

Benefit for Initial Acquisition

Although the lack of a benefit for semester-to-semester retention was disappointing, even finding a benefit for initial acquisition was promising, given its size, and given the difficulty of learning statistics for undergraduate students. There could be many reasons why learning the concepts involved in statistics is so hard for students. The term “conditions” in an experimental design is used in a much more specific way than in everyday language. The specifics have to do with controlling the possible unwanted effects of extraneous variables. All of those ideas would be difficult for students to relate to their everyday experiences. Therefore, it is encouraging that self-explanation at least had beneficial effects for initial acquisition.

Lack of Retention

Possible reasons for the lack of retention include that there might not have been enough time devoted to the learning tasks, that a learning task other than self-explanation might have been more suitable for retention, and that retention might require a level of motivation during the learning tasks that was not induced by participating in an experiment rather than in an actual class. A class would also reiterate the concepts over the weeks of the course.

This research was aimed at the application of identifying specific instructional methods to improve the retention of what was learned by undergraduate college students in their introductory statistics class so that it could be readily available for use in an experimental psychology class a semester later. Of course, the introductory statistics class is typically a 15-week college level course. And in that course students learn everything from basic descriptive

statistics and research designs up to the logic of hypothesis testing and which test to use for various research designs. Given that our experiment only trained our participants for about an hour, we focused on only giving them some basic terms and definitions for research designs and hypothesis testing procedures, and the correct associations between them. Nevertheless, even for what seemed to us to be a quite simple learning task, retention may have required connecting that new and sometimes unfamiliar material to whatever the students already knew. And that may be something that requires more, at least in terms of time, and possibly in terms of the nature of the task, than what they could accomplish with the training tasks we gave them.

Self-explanation may not have been the best candidate for an instructional principle to best facilitate retention. As pointed out in the introduction, prompted self-explanations increase understanding (Chi et al., 1994) whereas spacing of practice improves memory and fluency (Cepeda et al., 2009). In fact, Ebersbach and Barzagar Nazari (2020) found a benefit of spaced practice specifically for retention and transfer of knowledge in statistics. However, their retention interval was only five weeks, rather than from one semester to the next.

We chose self-explanation as an instructional principle to investigate for purposes of improving retention of statistics concepts for two reasons. First, we found that the specific question we wished to address had been neglected in the previous literature. Second, we did find some evidence that self-explaining might have been somewhat useful for retention over shorter intervals in other domains (e.g., Hsu et al., 2016; Mathan, 2004; Molesworth et al., 2011), and that the nature of statistics might suggest that self-explaining might be helpful for learning and retention of basic concepts in that domain (Rittle-Johnson & Loehr, 2017). In addition, according to Koedinger, et al. (2012), the instructional principle of prompted self-explanation is at the top of a hierarchy of simpler to more complex principles. This contributed to our intuition

that it might be a good candidate for our purposes, given that retention is a more ambitious goal than immediate acquisition. However, Koedinger, et al. also point out that prompted self-explanation affects the learning processes of understanding and sense making, whereas memory and fluency processes are affected by spacing and testing, which are at the bottom of that hierarchy. Thus, even though spacing and testing are at the bottom of the hierarchy, because of the learning processes they affect, and in the light of the findings by Ebersbach and Barzagar Nazari (2020), perhaps they should be further investigated.

Motivation is generally believed to be an important factor in college students' learning and retention. In our experiment we used a $15 payment as a motivator for students to return after a semester delay to take the retention test. However, that payment was only contingent upon returning and taking the test, not upon doing well on it. To capitalize on motivation to do well would require devising a classroom-based experiment, rather than a laboratory experiment. That remains for future research.

Limitations of the Present Study

We wanted to ensure that our experimental manipulation was consistent with previous studies on self-explanation. Therefore, we wanted to be sure that the experimental participants responded to our prompts by using prior information, thus enabling them to not only give a correct answer, but also to connect their answer to a reason for it. Because the participants were naive to statistics, and our training was brief, we reasoned that they needed to have the training materials, which explicitly stated the reason for each answer, in front of them when they responded to the prompts. However, that resulted in a confound, in that the experimental participants had access to that aspect of the training materials for a longer time than the control participants. Therefore, it is possible that such extra access, rather than self-explaining per se,

produced the higher initial acquisition. Furthermore, even though this procedure ensured that the experimental participants gave a response that was a true self-explanation, it also meant that they did not necessarily retrieve the information for that explanation from their memory. In future research, if participants are trained sufficiently to enable them to remember the information and retrieve it when they respond to the prompt, that would remove the confound. Such increased training might also foster better retention.

When the participants in the restatement group worked through examples 5 and 6, they sometimes made mistakes. Those examples were guided practice, so if they did make a mistake the experimenter corrected them. However, that was not done on examples 7 and 8, which the participants did entirely on their own. We did not record any data on the quality of the control group's restatements during the study portion. However, an interested reader can go to the open materials on the OSF (Ryan & Koppenhofer, 2021) and examine the participants' response sheets to extract that data.

Although all the participants took the retention test one full semester after they were trained, the amount of time that elapsed varied. For some participants the interval was over a winter break, whereas for others it was over the summer—a longer time. Also, the week of the semester in which a participant took the tests could vary by 15 weeks. We did not include those varying intervals in our analysis. However, the dates on which we ran the tests were mostly recorded on the response sheet, and they were always recorded on the session log. Those materials are available on the OSF (Ryan & Koppenhofer, 2021).

Future Directions

This research was intended to determine whether a particular instructional method, self-explanation, can improve retention of statistics concepts. However, it was not intended as just

basic research to answer that question, but rather, it was aimed at the applied goal of actually improving retention. Therefore, a next logical step would be to use the possible reasons for the lack of retention to guide future research.

One possible approach would be to conduct future research using actual classrooms. Spreading the training over an entire semester, instead of just the hour or so that occurred in the laboratory experiment, would drastically increase the time devoted to the learning tasks and using many actual classrooms would potentially enable a very large sample size. It would also facilitate using spaced retrieval practice and multiple testing as potentially beneficial instructional methods. Finally, the need to receive a good grade in the class would provide more motivation than that provided in our experiment.

Another drawback of the current research was that the to-be-learned material was very limited. It only involved making associations between a few basic types of research situations and appropriate statistical procedures. A classroom setting would afford testing the effects of the instructional methods on a much more realistic set of to-be-learned materials, including more complex concepts, such as the nature of a sampling distribution and the logic of hypothesis testing. Using such materials would afford more richly interconnected learning. Combining more complex learning material with a manipulation that included the instructional principles of spacing and practicing retrieval might foster better retention. Also, conducting such research in multiple classrooms might afford testing retention by measuring performance in an Experimental Psychology class a semester after the Statistics class. Such a method of testing retention would complement our applied goal of establishing ways of improving actual classroom performance in one of the most difficult areas in the Psychology major.

Conclusion

This study adds to the large body of evidence regarding the potential benefits of self-explaining. However, it also shows that the selection of an instructional activity needs to be guided by empirical findings about the effectiveness of that activity in the particular situation for which it is intended (Koedinger et al., 2012). Specifically, we examined the goal of fostering not only the initial acquisition of knowledge of what statistical procedure to use for a given research design, but also retaining that knowledge. Our evidence suggests that, in this situation, self-explanation is useful for initial acquisition, but for retention, it needs to be either supplemented or replaced by a different instructional activity.

References

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, *26*(2), 147–179. https://doi.org/10.1016/S0364-0213(02)00061-7

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, *95*(4), 774–783. https://doi.org/10.1037/0022-0663.95.4.774

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology*, *101*(1), 70–87. https://doi.org/10.1037/a0013247

Berthold, K., Röder, H., Knörzer, D., Kessler, W., & Renkl, A. (2011). The double-edged effects of explanation prompts. *Computers in Human Behavior*, *27*(1), 69–75. https://doi.org/10.1016/j.chb.2010.05.025

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition & Instruction*, *13*(2), 221–252. https://doi.org/10.1207/s1532690xci1302_3

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications.

Experimental Psychology, *56*(4), 236–246. https://doi.org/http://dx.doi.org/10.1027/1618-3169.56.4.236

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1

Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

Chi, M. T., & VanLehn, K. A. (1991). The content of physics self-explanations. *Journal of the Learning Sciences*, *1*(1), 69–105. https://doi.org/10.1207/s15327809jls0101_4

Chi, T.-Y. (2018). *Computer skill acquisition and retention: The effects of computer-aided self-explanation*. ProQuest Information & Learning. (2017-23162-237).

Cho, Y. H., & Jonassen, D. H. (2012). Learning by self-explaining causal diagrams in high-school biology. *Asia Pacific Education Review*, *13*(1), 171–184. http://dx.doi.org/10.1007/s12564-011-9187-4

Crowley, K., & Siegler, R. S. (1999). Explanation and generalization in young children's strategy learning. *Child Development*, *70*(2), 304. https://dx.doi.org/10.1111/1467-8624.00023

de Bruin, A. B. H., Rikers, R. M. J. P., & Schmidt, H. G. (2007). The effect of self-explanation and prediction on the development of principled understanding of chess in novices. *Contemporary Educational Psychology*, *32*(2), 188–205. https://doi.org/10.1016/j.cedpsych.2006.01.001

de Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2010). Learning by generating vs. receiving instructional explanations: Two approaches to enhance attention cueing in animations. *Computers & Education*, *55*(2), 681–691. http://dx.doi.org/10.1016/j.compedu.2010.02.027

De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2011). Improved effectiveness of cueing by self-explanations when learning from a complex animation. *Applied Cognitive Psychology*, *25*(2), 183–194. https://doi.org/10.1002/acp.1661

Ebersbach, M, & Barzagar Nazari, K. (2020). Implementing distributed practice in statistics courses: Benefits for retention and transfer. *Journal of Applied Research in Memory and Cognition, 9*(4), 532-541. http://dx.doi.org/10.1016/j.jarmac.2020.08.014

Howie, D. E., & Vicente, K. J. (1998). Making the most of ecological interface design: The role of self-explanation. *International Journal of Human-Computer Studies*, *49*(5), 651–674. https://doi.org/10.1006/ijhc.1998.0207

Hsu, C.-Y., Tsai, C.-C., & Wang, H.-Y. (2016). Exploring the effects of integrating self-explanation into a multi-user game on the acquisition of scientific concepts. *Interactive Learning Environments*, *24*(4), 844–858. http://dx.doi.org/10.1080/10494820.2014.926276

Huk, T., & Ludwigs, S. (2009). Combining cognitive and affective support in order to promote learning. *Learning and Instruction*, *19*(6), 495–505. https://doi.org/10.1016/j.learninstruc.2008.09.001

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. https://doi.org/10.1111/j.1551-6709.2012.01245.x

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, *103*(3), 386–394. https://doi.org/10.1016/j.jecp.2009.03.003

Larsen, D. P., Butler, A. C., & Roediger, H. L. I. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, *47*(7), 674–682. https://doi.org/10.1111/medu.12141

Leppink, J., Broers, N. J., Imbos, T., van der Vleuten, C. P. M., & Berger, M. P. F. (2012). Self-explanation in the domain of statistics: An expertise reversal effect. *Higher Education: The International Journal of Higher Education and Educational Planning*, *63*(6), 771–785. https://doi.org/10.1007/s10734-011-9476-1

Margulieux, L. E., & Catrambone, R. (2019). Finding the best types of guidance for constructing self-explanations of subgoals in programming. *Journal of the Learning Sciences*, *28*(1), 108–151. http://dx.doi.org/10.1080/10508406.2018.1491852

Mathan, S. (2004). Recasting the feedback debate: Benefits of tutoring error detection and correction skills [ProQuest Information & Learning]. In *Dissertation Abstracts International: Section B: The Sciences and Engineering* (Vol. 64, Issue 12–B, p. 6350).

McEldoon, K. L., Durkin, K. L., & Rittle-Johnson, B. (2013). Is self-explanation worth the time? A comparison to additional practice. *British Journal of Educational Psychology*, *83*(4), 615–632. http://dx.doi.org/10.1111/j.2044-8279.2012.02083.x

Molesworth, B. R. C., Bennett, L., & Kehoe, E. J. (2011). Promoting learning, memory, and transfer in a time-constrained, high hazard environment. *Accident Analysis and Prevention*, *43*(3), 932–938. https://doi.org/10.1016/j.aap.2010.11.016

Pillow, B. H., Mash, C., Aloian, S., & Hill, V. (2002). Facilitating children's understanding of misinterpretation: Explanatory efforts and improvements in perspective taking. *Journal of Genetic Psychology*, *163*(2), 133. https://doi.org/10.1080/00221320209598673

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1–15. http://dx.doi.org/10.1111/j.1467-8624.2006.00852.x

Rittle-Johnson, B., Fyfe, E. R., Loehr, A. M., & Miller, M. R. (2015). Beyond numeracy in preschool: Adding patterns to the equation. *Early Childhood Research Quarterly*, *31*, 101–112. https://doi.org/10.1016/j.ecresq.2015.01.005

Rittle-Johnson, B., & Loehr, A. M. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review*, *24*(5), 1501–1510. https://doi.org/10.3758/s13423-016-1079-5

Ryan, R. S., & Koppenhofer, J. A. (2021, September 11). Statistics (SE) Fall 2016. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/H8SUT

Shen, C.-Y., & O'Neil, H. (2006). *The Effectiveness of Worked Examples in a Game-Based Learning Environment*. Online Submission.

Snow, E. L., Likens, A. D., Allen, L. K., & McNamara, D. S. (2016). Taking control: Stealth assessment of deterministic behaviors within a game-based system. *International Journal*

of Artificial Intelligence in Education, *26*(4), 1011–1032. http://dx.doi.org/10.1007/s40593-015-0085-5

Talley, C. P., & Scherer, S. (2013). The enhanced flipped classroom: Increasing academic performance with student-recorded lectures and practice testing in a “flipped” STEM course. *Journal of Negro Education*, *82*(3), 339–347. https://doi.org/10.7709/jnegroeducation.82.3.0339

Tenenbaum, H. R., Alfieri, L., Brooks, P. J., & Dunne, G. (2008). The effects of explanatory conversations on children’s emotion understanding. *British Journal of Developmental Psychology*, *26*(2), 249–263. https://doi.org/10.1348/026151007X231057

Wittwer, J., & Renkl, A. (2010). How effective are instructional explanations in example-based learning? A meta-analytic review. *Part of a Special Issue: Cognitive Load Theory: Advances in Research on Worked Examples, Animations, and Cognitive Load Measurement; Review Article*, *22*(4), 393–409. https://doi.org/10.1007/s10648-010-9136-5

Wylie, R., Koedinger, K., & Mitamura, T. (2010). Analogies, explanations, and practice: Examining how task types affect second language grammar learning. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems* (pp. 214–223). https://doi.org/10.1007/978-3-642-13388-6_26

Wylie, R., Sheng, M., Mitamura, T., & Koedinger, K. R. (2011). Effects of adaptive prompted self-explanation on robust learning of second language grammar. In G. Biswas, S. Bull, J.

Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education* (pp. 588–590). https://doi.org/10.1007/978-3-642-21869-9_110