The Pennsylvania State University

The Graduate School

College of Education

A COMPARISON OF THE EFFECTIVENESS OF WORKED EXAMPLES AND PRODUCTIVE FAILURE IN LEARNING PROCEDURAL AND CONCEPTUAL KNOWLEDGE RELATED TO STATISTICS

A Dissertation in

Educational Psychology

by

Michael A. Cook

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2017

The dissertation of Michael A. Cook was reviewed and approved* by the following:

Peggy N. Van Meter
Associate Professor of Education (Educational Psychology)
Dissertation Adviser
Committee Chair

Rayne A. Sperling
Associate Professor of Education (Educational Psychology)

Jonna M. Kulikowich
Professor of Education (Educational Psychology)

Roy B. Clariana
Department Head of Learning and Performance Systems
Professor of Education

David X. Lee
Department Head of Educational Psychology, Counseling, and Special Education
Professor of Education

*Signatures are on file in the Graduate School

Abstract

In statistics instruction, strategies that promote initial student success in instruction, such as worked examples, are frequently used. There is evidence that difficulties at the beginning of the instructional process can significantly hinder student learning, so instruction should be designed to minimize the chances of students failing (i.e., Kirschner, Sweller, & Clark, 2006). However, other instructional paradigms, such as productive failure, have shown evidence supporting the idea that initial difficulty may actually be helpful in the learning process. Furthermore, the use of productive failure in statistics instruction has shown evidence of greater gains for students on conceptual knowledge measures, with no loss of performance on procedural knowledge measures. However, both worked examples and productive failure have usually been compared to direct instruction or some other variation of a control condition of instruction. There is little available research that has directly compared these strategies, either in statistics instruction or other content areas. This study examines the interaction between instruction type (worked examples and productive failure) and knowledge learners acquire (procedural and conceptual knowledge) in the instruction of basic statistics. This interaction is examined at both immediate and delayed posttests. Repeated measures ANOVAs were conducted to address these research questions. This study contributes to the literature by examining whether an instruction by knowledge interaction occurs in statistics instruction on both immediate and delayed posttests, which has not been investigated in previous research.

Table of Contents

Table of Tables

Table of Figures

Chapter 1

Introduction

The content area of statistics is a domain in which students have historically experienced a great deal of difficulty in learning, especially as novices (Garfield & Ahlgren, 1988). These difficulties experienced by students are well-documented, and include issues such as: only "crunching numbers" and not understanding how or why the computations they are performing work (Lavigne, Salkind, & Yan, 2008); an inability to transfer a familiar procedure or concept to a new context (Chervany, Collier, Fienberg, & Johnson, 1977); a tendency to treat statistical knowledge as disparate "bits" of information and not as hierarchically structured knowledge (Lovett & Greenhouse, 2000); and a general inability to differentiate between different types of problems (Yan & Lavigne, 2014). One reason for these difficulties may be related to the structure of instruction in statistics classrooms. In general, statistics textbooks are structured in a manner that emphasizes practice on similar procedures, but does not emphasize conditional knowledge related to when to use these procedures (Chance & Garfield, 2002; Mayer, Sims, & Tajika, 1995). As a result, students in these classes often have trouble selecting when to use an appropriate statistical analysis when encountering a novel problem, even if they have already had experience in using the correct statistical procedure.

While there has been much research documenting the difficulties that both college and high school students have when learning statistics, the research that attempts to explain *why* these difficulties occur is much less developed. Garfield & Ahlgren (1988) observed that there has been very little research that has attempted to explain why these various difficulties occur for students when learning statistics. While they made this observation almost 30 years ago, this observation is still very valid today. The structure of statistics instruction is one proposed reason

for these continuing difficulties (Mayer et al., 1995). Students in these classes are typically asked to practice the same type of calculation related to the topic of instruction (i.e., compute a Pearson correlation between two variables). While this helps to build accuracy and fluency with the procedure, it is difficult for the student to apply this procedure outside of the topic of instruction. Thus, students only apply the concept to the same types of problems during instruction that only focuses on a given analytic procedure. Students generally have far less practice with being given a scenario different from what they studied during instruction and deciding which analytic procedure would be appropriate to use (Yan & Lavigne, 2014). As a result, learners are generally far less able to make correct decisions regarding which analytic procedure to choose when given a novel scenario that is not in the predetermined time frame for covering a specific topic. For example, a learner may be able to compute a correlation between two variables while completing a unit of instruction on correlation. However, if this same student was given a scenario at a later time, outside of this specific instruction, and was asked to determine which analysis would be most appropriate for the scenario and to perform this analysis, it is much less likely that the student would correctly choose to perform a correlation, much less compute the correlation correctly. In short, the application of an appropriate analysis to a scenario is a task students are much less likely to complete correctly (Quilici & Mayer, 2002).

Types of Knowledge

The shortcomings detailed in the statistics instruction literature may be examined in terms of the types of knowledge that are typically emphasized in this instruction. Statistics instruction generally focuses on procedural knowledge (Chance & Garfield, 2002; Lavigne et al., 2008). Procedural knowledge can be defined as "knowledge one has of certain processes or routines" (Anderson, 1983). Statistics instruction may also focus on conceptual knowledge, which is

defined as being comprised of content and discourse knowledge on a topic, as well as how this knowledge operates and under what conditions the knowledge can be applied (Alexander, Schallert, & Hare, 1991). Conditional knowledge is also relevant to statistics instruction. Conditional knowledge, which is defined as a subset of conceptual knowledge related to when and how knowledge should be applied (Alexander & Judy, 1988), is relevant to statistics instruction because it refers to students' ability to correctly choose a procedure to perform when given a problem to solve.

Statistics instruction, with its emphasis on computations (Mayer et al., 1995), focuses heavily on procedural knowledge acquisition. However, conceptual knowledge acquisition is not focused on as much in typical statistics instruction. This is evidenced by students' general inability to connect various statistical concepts (Lovett & Greenhouse, 2000) and to understand how or why the procedures they learn work (Kempthorne, 1980). Further, these conceptual knowledge deficits lead to conditional knowledge deficits (as conditional knowledge is subsumed under conceptual knowledge), resulting in students' inability to choose an appropriate computation to perform when given a statistics problem (Quilici & Mayer, 1996; Yan & Lavigne, 2014). Further, these conceptual knowledge deficits are likely to preclude learners from acquiring expertise with statistical knowledge (Broers, 2009). Expertise in statistical knowledge is important, as experts structure knowledge differently than do novices, allowing for activities such as metacognitive monitoring and strategy selection to occur in addition to the utilization of content-specific problem solving strategies (Baxter, Elder, & Glaser, 1996; Chi, Glaser & Far, 1998).

Research on statistics instruction has extensively documented the difficulties that novice learners experience in statistics classrooms. When these difficulties are examined in terms of

different types of knowledge, it becomes clear that statistics instruction focuses mainly on procedural knowledge acquisition, at the expense of conceptual knowledge acquisition. As a result, students learn how to perform a given computation related to a topic of instruction, but students may not understand how to transfer this knowledge to novel situations. In addition, conceptual knowledge deficits lead to conditional knowledge deficits, as evidenced by students' inability to choose a correct procedure when given a new problem to solve. These conceptual knowledge deficits also make it less likely for a novice learner to eventually achieve expertise with a statistical knowledge. Thus, statistics instruction should incorporate strategies and techniques that promote not only procedural, but also conceptual knowledge acquisition.

Effective Manipulations of Statistics Instruction

Productive failure. One type of instructional manipulation that has been found to be advantageous over direct instruction is productive failure (PF). Productive failure refers to a technique in which learners encounter a novel, open-ended problem at the beginning of instruction that they probably do not know how to solve (Kapur, 2008; 2012). When learners are prompted to solve this novel problem, they are asked to produce as many different solutions as they can. These solutions will probably be incorrect and incomplete, to varying degrees. However, the generative process of creating solutions to unique problems is theorized to help prepare students to learn new content (Schwartz & Bransford, 1998). Further, by generating unique solutions to novel problems, learners may be more likely to understand conceptual knowledge related to the novel problem they encountered, and may also be able to apply that knowledge to other problems more effectively than students who were taught through direct instruction only.

Worked examples. Worked examples are another instructional manipulation that has been found to be very effective at teaching a variety of different types of topics, including various mathematical and statistical content areas (e.g., Atkinson, Derry, Renkl, & Wortham, 2000; Ward & Sweller, 1990). While there is not a precise definition of what worked examples should look like, they are generally thought of as "expert" solutions that novice learners may use to assist in the learning process (Atkinson et al., 2000). A worked example will always contain a statement of the problem, along with a step by step solution for a given problem type. The structure of worked examples is such that they can be used most effectively in domains that contain knowledge that is both declarative (i.e., individual steps) and procedural. Worked examples draw learners' attention to both individual steps and the overall structure of a given problem type. However, use of worked examples in instruction does not guarantee that students will learn more effectively when compared to students administered direct instruction only. The body of research on mathematics has shown that there are several features of worked examples that may be manipulated, and thus may moderate the effectiveness of these examples (Atkinson et al., 2000).

One significant criticism of the use of worked examples, has been that their use, while fostering learning of procedural knowledge in learners, does not necessarily help learners acquire related conceptual knowledge (Aleven & Koedinger, 2002). For example, a worked example may help a student learn how to compute variance and standard deviation, but may not be effective at teaching students how to apply these concepts in different situations other than the teaching example (i.e., variability is a measure of consistency, how variance is related to the distribution of data). Further, a worked example may not address the issue of determining when

it is appropriate to use a specific statistical procedure (i.e., when would it be appropriate to compute a standard deviation?).

Worked examples and productive failure are two instructional techniques that are used in statistics instruction. Many instructional techniques, such as worked examples, are theorized to work because they guide the student at the beginning of the instructional process and prevent students from encountering difficulty. Succeeding at the beginning of instruction is thought to be very important, because it increases the likelihood that learning will occur (e.g., Kirschner et al., 2006). However, there may also be some learning-related benefits to students initially encountering difficulty during instruction (Van Lehn, Siler, Murray, Yamauchi, & Baggett, 2003). In addition, productive failure and worked examples can be compared on the basis of the types of knowledge they are designed to help students learn. Productive failure is designed to be more effective in conceptual knowledge acquisition, while worked examples are designed to be more effective in procedural knowledge acquisition. These two instructional techniques use different methods (early success vs. early difficulty) and are used to teach different types of knowledge in statistics (conceptual vs. procedural), but up to now, there has been no research that has directly compared productive failure to worked examples.

The purpose of this study is to test methods to enhance learning of statistics content by comparing the effectiveness of two commonly tested instructional manipulations on the learning of procedural and conceptual knowledge in introductory statistics instruction. Worked examples and productive failure have both been shown to be effective methods of instruction in mathematics, albeit with different types of knowledge, but they have never been tested against each other in an experiment. Furthermore, worked examples are designed to foster student success early in the learning process, while productive failure emphasizes learning from initial

difficulties in the learning process. Thus, this study allows for a direct comparison of how these types of instruction affect learning of conceptual and procedural knowledge in statistics instruction.

Chapter 2

Review of Literature

Statistics instruction has generally been characterized as being effecting in promoting procedural knowledge acquisition, while not being as effective in promoting conceptual knowledge acquisition (Yan & Lavigne, 2014). The unequal focus on these two types of knowledge in instruction is a main cause of the various problems that have been noted regarding statistics instruction (Garfield & Ahlgren, 1988). To address these shortcomings, instructional techniques that focus on conceptual knowledge acquisition may be necessary. The use of productive failure prompts may be one such technique.

Productive Failure

One type of instructional manipulation that may be effective in teaching statistical content is productive failure (PF). Productive failure refers to an instructional process that requires students to generate possible solutions to a novel type of problem, followed by direct instruction on the topic of interest. These generated solutions will be of varying degrees of correctness and completeness, which is to be expected. This is different from students completing practice problems, as practice problems are administered after instruction, while a productive failure prompt is administered before instruction.

Productive failure can be traced back to other research paradigms which, while not being called "productive failure", still contained many similar components, including a focus on preparing students to learn and seeking opportunities for learning to occur (i.e., Bransford & Schwartz, 1999; Schwartz & Bransford, 1998). These instructional paradigms certainly have constructivist origins, as a significant component of these instructional processes requires students to develop their own explanations or theories in regard to the material being taught.

Likewise, PF also has a distinct constructivist influence, which is evident in the generative student activities that begin a typical PF lesson. However, it is important to note that PF does not refer to aimless "discovery learning" where the students learn new concepts and ideas with little to no guidance from a teacher. The teacher actually plays a very significant role in PF instruction, as the direct instruction that follows the generative activity is important for students to connect their generated solutions an expert's (teacher's) explanations.

Theoretical foundations of productive failure. Development of a student's opportunity to learn is one of the main components of PF and can be found in pre-PF research. One example of this type of instruction is impasse-driven learning (Van Lehn, 1999; Van Lehn et al., 2003). Van Lehn et al. operationally defined an impasse for a student as occurring when "…a student gets stuck, detects an error, or does an action correctly but expresses uncertainty about it " (p. 220). In impasse-driven learning, it is these points that provide the greatest opportunity for learning to occur. Specifically, Van Lehn et al. believe that these impasses motivate the learner to take a greater role in constructing knowledge. This may occur by the student attempting to access more prior knowledge, use objects that are in close proximity (such as a textbook or computer), or ask a more knowledgeable person. They hypothesized that, if a student encounters one (or more) of these impasses during learning, then the student's level of understanding will be higher than if none of these impasses were encountered. However, it is important to note that, just because impasses occur, student understanding and learning will not automatically increase. If an impasse is not resolved, then learning may not occur and the student may become frustrated in the process. If a student encounters an unresolvable impasse, then some sort of explanation needs to be provided to help the student move past this obstacle (Chi, 1996; Renkl, 1997).

While impasse-driven learning is not identical to PF, there are similarities worth noting. An impasse in learning occurs when a student is progressing through learning smoothly and then encounters difficulty, while PF uses initial difficulty to prepare the student to learn. While the timing of the impasse or obstacle differs, overcoming the impasse is a necessary condition of learning. An impasse requires a student to draw on, or activate, prior knowledge in an attempt to resolve the impasse. PF simply moves the impasse before instruction on a new topic begins, as prior knowledge activation may serve as an effective organizing activity for a student. Of course, the role of the teacher is important in regards to either type of instruction: if a student reaches an impasse in learning, then a student will usually need prompting or other assistance from a teacher to resolve the impasse. In both scenarios, the teacher plays an important role in moving the student past the impasse.

Productive failure also has origins in an instructional process called *preparation for future learning* (PFL) (Schwartz & Bransford, 1998). In PFL, a generative strategy is used in advance of direct instruction for students. Specifically, contrasting cases was used as the generative strategy in Schwartz and Bransford's PFL study, as the presentation of contrasting cases was theorized to help students differentiate between dissimilar cases, which in turn gave students the opportunity to sufficiently differentiate prior knowledge, allowing the students to more deeply and usefully connect subsequent direct instructional materials with relevant prior knowledge. Schwartz and Bransford tested the effectiveness of a number of different combinations of analyzing contrasting cases, receiving lectures, and summarizing book chapters. In the third of a series of three experiments, students who analyzed a set of contrasting cases (generation task), followed by a lecture on the material, performed best on predictive tasks (making predictions regarding what would happen in hypothetical psychological experiments)

based on the material they had just learned. Interestingly, these students did not perform as well on a different assessment which tested how many of the "target ideas" they mentioned after reading the instructional text. They hypothesized that the analysis of contrasting cases helped to differentiate between the knowledge structures they needed to use to understand new material (Schwartz & Bransford, 1998, p. 504). In other words, the generation task helped "prepare" students to learn new material more effectively and deeply than simply reading a book chapter, even if the immediate recall results did not show this. However, the generation task by itself was not enough to ensure student learning, as evidenced by the significantly lower posttest scores for students who analyzed two sets of contrasting cases but received no direction instruction on the content. In this context, a generative task worked most effectively when followed by some sort of direct instruction, such as a lecture. Further, there was evidence of deeper conceptual learning fostered by the PFL strategy; students made more accurate predictions, which requires not just knowledge of facts and procedures, but also synthesis of these types of knowledge. Differentiation of contrasting cases may also be building conditional knowledge, as this activity helps students generate unique ways to differentiate between knowledge structures, which will subsequently allow students to access the correct knowledge structure to apply to a future scenario.

Another type of instructional process that is related to PF is called *inventing to prepare for learning* (IPL) (Schwartz & Martin, 2004). IPL builds on the use of analyzing contrasting cases that was used in the PFL learning paradigm. IPL focuses on the role of student invention and generation *before* direct instruction on a given topic. Schwartz and Martin explicitly focus on the role of these invention activities in preparing students to learn from direct instruction that will occur at another point during instruction. Specifically, they believe that appropriate

invention activities can help prepare students to learn how to solve problems not necessarily in the immediate future, but on more long-term assessments. Further, they hypothesized that students that learn through the IPL paradigm (by using an invention strategy prior to direct instruction) would be better prepared to learn a novel statistical concept introduced on a posttest than students who did not use an invention strategy. They conducted two experiments that focused on the use of IPL strategies to teach basic statistical concepts (measures of central tendency) to 9th grade students. There were two main independent variables of interest in this study: the use of IPL versus only direct instruction, and the use of a worked example on a posttest as a resource to help students learn a novel statistical concept and solve a problem related to this novel concept.

Schwartz and Martin found that learners who experienced IPL instruction were in fact better prepared to learn about measures of central tendency, as evidenced by one particular result of interest. Students who were in the IPL condition and received the embedded worked example resource in the posttest performed significantly better on the novel posttest item related to a new transfer topic (normalized scores) than did students in all other groups. Students who were in the IPL condition but did not receive the embedded worked example did not perform as well on the transfer item as did IPL learners who did receive the embedded worked example. Similarly, both groups that did not receive IPL learning did not perform as well on the transfer posttest item, regardless of whether the embedded worked example was included. The difference between the IPL/embedded worked example effect and the other groups on the transfer item was striking; students in the IPL/embedded worked example group answered the transfer item correctly more than twice as often as students in the other groups. This result can be interpreted as evidence of what Schwartz and Martin refer to as a "double transfer" effect. In this specific context, double

transfer refers to how students transfer prior knowledge to influence what they learn, and then what is learned is transferred to the ability to solve a new problem. IPL focuses on the first part of this double transfer; if invention activities help students connect more prior knowledge to what they learn, then they should make more connections between the target content material and prior knowledge, which in turn should help the students apply the knowledge more effectively to subsequent problems.

Development of productive failure. The next step in the development of the productive failure research paradigm was Kapur's (2008) research on the role of ill-structured versus well-structured problems in the teaching of Newtonian kinematics. Kapur designed a study in which students all took the same pre-test, but were then assigned to one of two different instructional conditions. Students in the first condition worked in groups together to solve two well-structured kinematics problems, while students in the second condition worked in groups together to solve two ill-structured kinematics problems. The problems were the same in both groups, but were manipulated to make one set of problems more well-structured, and the other set less structured. Both groups then took the same two posttests; one posttest contained only well-structured problems, and the other posttest contained only ill-structured problems. Predictably, students in the ill-structured problem group encountered more obstacles to learning; the conversations in this group tended to be more "complex, chaotic, and divergent" (p. 409) than conversations that occurred in the well-structured problem group. This quality of conversation carried over to performance on the two questions in instruction: groups that worked on the well-structured problems produced qualitatively better solutions than did students who worked on the ill-structured problems. However, Kapur originally hypothesized that the struggles caused by working on ill-structured problems, when followed by work on well-structured problems (the

well-structured problem posttest, in this scenario), would result in greater gains on the ill-structured problems. Not only did students in the ill-structured condition perform better on the ill-structured problem posttest, but they also performed better on the well-structured problem posttest than did students who received well-structured problems during instruction.

These results raised questions about the role of structure in instructional design. Kapur noted that there has been a large amount of research that essentially asks what a learner can accomplish on an ill-structured problem that will usually be beyond the current abilities of the learner. By their very nature, ill-structured problems are thought to require certain supports and other instructional manipulations that help ensure that the learner does not experience excessive difficulty. Worked examples would certainly be an example of a technique that is used to help students solve novel problems that may be unfamiliar or beyond their abilities. However, by focusing only on ways to minimize the chances of student failure, Kapur argued that researchers are ignoring the possible benefits of allowing a student to experience difficulty at first when trying to solve these types of problems. This is a logical next step when considering the ideas of Van Lehn et al.'s impasse-driven learning, as well as the PFL and IPL research paradigms. Kapur's research combined these paradigms into a new paradigm called productive failure. His first study gave evidence of a positive effect of productive failure; students who were assigned to ill-structured problems very clearly had more difficulties trying to solve these questions than did students solving the same problems, but organized in a more structured manner. However, this initial difficulty led to students performing better not just on a posttest of similar ill-structured problems, but also on a posttest of well-structured items. The better performance on ill-structured problems was an expected finding, since working on ill-structured problems previously may prepare learners to work on similar problems later, but the improved

performance on the well-structured questions was surprising. Kapur hypothesized problem representation, nature of assumptions, and contextual parameters as ways in which students given ill-structured problems gave better solutions that did students who were given well-structured problems. All three of these factors suggest that student work on ill-structured problems was not just increasing student learning on both subsequent ill and well-structured problems, but are also getting at something beyond procedural knowledge; specifically, conceptual and conditional knowledge related to the topics being taught.

Subsequent research has shown how productive failure can be used as an effective instructional strategy in the teaching of mathematics and statistics. Kapur (2012) tested the use of productive failure versus more commonly used direct instruction in the instruction of concepts related to central tendency and variance to 9th grade students in Singapore. Students in the productive failure condition were given a novel question and instructed to work in groups to develop as many possible solutions to the question. After they worked in groups, the teacher gathered the class together and briefly discussed the solutions that the students had created, followed by a period of direct instruction that focused on the concepts and formulas related to central tendency and variance. Students in the direct instruction condition received a lecture from the teacher on the important formulas and concepts, followed by discussion of an example. This was followed by individual practice by students. Both groups of students then took the same posttest, which consisted of three problems testing procedural knowledge, two problems testing conceptual understanding, and one transfer question regarding the use of normalized scores. Students in the productive failure group performed significantly better on the conceptual and transfer measures than did students in the direct instruction group, while procedural knowledge performance was not statistically different. Thus, the process of having students

struggle initially with a novel problem did not hurt their ability to perform the procedure, but led to gains in conceptual knowledge acquisition, as well as a better ability to transfer knowledge to novel scenarios.

In a subsequent study, Kapur (2014) replicated the results of his 2012 study and also tested a condition she called vicarious failure (VF). Vicarious failure is similar to productive failure except that the activity in which students developed solutions to a unique problem was replaced by a phase in which students evaluated the effectiveness of different solutions to the problem. In this study, Kapur used the student-generated solutions from her previous study to use as the solutions evaluated in the vicarious failure condition. Vicarious failure reduces the amount of mental effort that a learner must use by eliminating the generative aspect and replacing it with an evaluative activity that should require fewer cognitive resources. On a posttest containing procedural and conceptual knowledge questions, learners in all three conditions performed at approximately the same level on the procedural knowledge problems, but there were statistically significant differences on conceptual knowledge performance. Specifically, both productive failure and vicarious failure students performed significantly better than direct instruction students on conceptual questions. However, productive failure students scored significantly higher than vicarious failure students on these questions. From these results, it appears that while vicarious failure may be more effective than direct instruction, it is still not as effective as productive failure. It appears as though the generative nature of the a productive failure activity is an important part of what helps students more effectively learn conceptual knowledge, as well as to transfer this knowledge to other scenarios.

Design features of productive failure instruction. There are some design features that should be considered when designing productive failure activities. These features were

discussed in Kapur and Bielaczyc's (2012) study that found a similar positive effect of productive failure in the teaching of ratios when compared to direct instruction. Specifically, they discussed three core principles that should be followed when designing productive failure instruction. The first principle is the creation of a problem-solving context in which a complex problem is worked on, but the problem is not so challenging as to frustrate the learner. The problem should also require the student to activate and use prior knowledge in the creation of multiple representations and solution methods (RSMs). The second principle is the opportunity for learners to explain and elaborate on content. The third principle is the provision of opportunities for learners to compare and contrast failed RSMs, as well as to develop the formula or solution to a problem.

A problem that is used in productive failure instruction must be complex enough to challenge learners; if it is not challenging enough, then learners will simply not experience difficulty. The problem should be ill-structured; well-structured problems can usually be easily solved with one RSM, while ill-structured problems can afford the creation of many different solutions (Greeno, Smith, & Moore, 1993). The complex, unstructured nature of the problem should require students to rely on prior knowledge, as well as to make and justify assumptions that are used in the creation of RSMs (Jonassen, 2000). However, the problem used for productive failure should not be so difficult as to frustrate the learner; thus, careful calibration of the difficulty level of this problem is necessary to ensure success from this activity (Kapur & Bielaczyc, 2012).

Once learners have used their prior knowledge to make assumptions and justifications that eventually become RSMs, there needs to be an opportunity for various RSMs to be compared to each other. This comparison should allow students an opportunity to see why some

RSMs are more or less effective and appropriate than others (diSessa & Sherin, 2000). While this is occurring, the teacher should also discuss the procedural methods of solving the problem, usually through algebra or some other procedure. This last step is important; the productive failure phase of instruction should always be followed by a teacher-led discussion that includes direct instruction on the instructional topic, as this allows students the opportunity to compare various generated RSMs to the established method or procedure used for solving a particular type of problem. It is this combination of generation and comparison that has been theorized to maximize the instructional benefits of productive failure.

In sum, productive failure extends the research paradigms of impasse-driven learning, invention for preparing for learning (IPL), and preparation for future learning (PFL) by using a novel question that students will probably not know how to solve correctly and asking students to generate as many possible solutions as they can. The novel question represents an impasse, and the generation of solutions functions much the same way as the analysis of contrasting cases and other related activities did in IPL and PFL instruction. When compared to direct instruction, students receiving productive failure perform better on conceptual knowledge and transfer questions while performing at approximately the same level on procedural questions. There are several different mechanisms that are theorized to make productive failure especially effective with the learning of conceptual knowledge, including the activation and differentiation of prior knowledge, attention to critical features, explanation and elaboration of these features, and organization and assembly into representations and solution methods (Kapur, 2010; Kapur & Bielaczyc, 2012). The generative strategy that forms the basis for the productive failure strategy has been theorized to be the key to making the aforementioned mechanisms work, as other

alternative forms of failure that do not contain a generative activity, such as vicarious failure, do not have the same positive effect on the learning of conceptual knowledge.

While productive failure has been used to teach both procedural and conceptual knowledge in statistics, and has been theorized to be particularly effective in the teaching of conceptual knowledge, there are some possible drawbacks. One significant drawback may be the amount of mental effort that students use to respond to a productive failure prompt. The generation of solutions to a novel problem is likely to be a cognitively taxing process for students. Higher amounts of mental effort have been theorized to be detrimental to student learning (Paas & van Merrienboer, 1994), so other instructional techniques in statistics education have focused on reducing the amount of mental effort that students use in the learning process. One of the most commonly used techniques is worked examples.

Worked Examples and Cognitive Load Theory

Much of the research on the development and use of worked examples has been framed through Cognitive Load Theory. Cognitive load can be thought of as the mental effort that is required of a learner to process new information (Sweller, van Merrienboer, & Paas, 1998). Cognitive Load Theory (CLT) is based on the idea that there is a limited capacity for working memory, which has been theorized to be seven bits of information, plus or minus two bits (Miller, 1956). Thus, if more working memory is required to attend to and process information, then a higher level of cognitive load is imposed on the learner. According to CLT, schematic construction is analogous to the learning of procedures. A schema can be defined as a way in which bits of information can be organized according to how they should be used (e.g., Chi, Glaser, & Rees, 1982). In short, schematic construction occurs when bits of information are attended to and processed in working memory and then either combined into a new knowledge

structure or assimilated into an already existing structure. When discussing learning of procedural knowledge, the individual steps can be thought of as individual facts, or declarative knowledge (Alexander et al., 1991). Procedural knowledge, then, would be knowledge of how these steps are put together and used. Thus, a learner's acquisition of a new procedure is essentially the construction of a new schema or the modification of an existing schema. It is important to note that while declarative, procedural, and conditional knowledge are discussed heavily in cognitive load and worked example literature, conceptual knowledge is not mentioned quite as often.

The issue of cognitive load is important in regards to issues related to instruction, due to the limited capacity of working memory. If the cognitive load of instructional materials is too high, then learning may be negatively affected. Thus, according to CLT, instruction should be designed to make overall levels of cognitive load as low as possible. However, not all types of cognitive load are the same. There are three types of cognitive load: intrinsic, extraneous, and germane. When designing instruction, consideration of each of the three types of cognitive load is necessary.

Intrinsic cognitive load refers to cognitive load imposed by the nature of the materials being learned. For example, learning individual elements in chemistry or multiplication tables in math would be tasks that entail low level of intrinsic cognitive load. By contrast, tasks such as performing multivariate calculus or learning grammar in a foreign language would be tasks that have a large amount of intrinsic cognitive load. It is important to note that the intrinsic cognitive load of a concept or procedure generally cannot be manipulated or altered very much. As Sweller et al. (1998) point out, mathematical operations and procedures tend to be very high in intrinsic cognitive load, due to the amount of knowledge that must be attended to simultaneously.

Extraneous cognitive load refers to the cognitive load imposed on the learner by the instructional materials and processes. For example, instruction that is unclear about what content a student should be focusing on would be high in extraneous cognitive load, as the search processes the student engages in are extraneous to the content being learned (Roelle, Lehmkuhl, Beyer, & Berthold, 2015). While intrinsic cognitive load is constant, extraneous cognitive load may be adjusted. Ideally, extraneous cognitive load should be adjusted to be as low as possible, especially when teaching subject material that is already high in intrinsic cognitive load, due to the additive nature of this construct. Poorly designed instructional materials may result in high levels of extraneous cognitive load being imposed on the learner, which in turn may make learning much more difficult for the learner.

Germane cognitive load is the third type of cognitive load that may be present in instructional materials. Germane cognitive load refers to load imposed on the learner while the learner is attending to and processing novel information (van Merrienboer, 1997). For example, providing self-explanations to explain how steps of a procedure work together would be an example of germane cognitive load, as the self-explanation forces the learner to attend to the content and construct unique meaning (Renkl & Atkinson, 2003). Germane cognitive load is sometimes difficult to differentiate from intrinsic cognitive load, as both are related to the content being taught to the learner (Kalyuga, 2011). It is important to consider that all three types of cognitive load are not necessarily mutually exclusive; adjustments to instruction may be made that decrease extraneous cognitive load while increasing germane cognitive load

In short, the amount of cognitive load imposed on a learner is an important factor in determining how effectively material may be attended to and processed. If the cognitive load is too high for a given task, then working memory may become overloaded and learning may be

hindered. Cognitive load theory is one of the main motivations behind the use of worked examples in statistics research; presentation of all of the steps of a procedure (or, how an expert would solve a problem) allows the student to focus on individual steps initially, reducing the need to search for appropriate next steps. Productive failure, by contrast, imposes larger amounts of cognitive load than do worked examples, due to the generative and ill-structured nature of the task. Worked examples are used because they are designed to reduce extraneous load and increase germane load, which is associated with more relevant attention to and processing of content

The Development of Worked Examples in Mathematics Instruction

The use of worked examples in instruction originated from research examining the use of practice problems and their effect on student learning. For much of the 1970s and early 1980s, the most commonly used instructional technique was to simply have learners practice problems over and over again until they found a way to solve the problems on their own (e.g., Simon & Chase, 1973). Many researchers during this time concluded that the amount of practice a learner completed was directly related to the achievement level of the learner. However, research in the early 1980s started to examine instructional methods that were different from simply giving learners sets of practice problems. One reason for this had to do with research on differences between how "experts" and "novices" think through the same types of questions. Researchers were finding that experts and novices approached problems differently at the beginning of the problem solving process, and that there were qualitative differences in the processes that experts and novices used in arriving at answers to problems from domains such as physics and mathematics (Chi, Feltovich, & Glaser, 1981; Silver, 1979). In fact, Chi et al. (1981) specifically cited differences in schematic construction for solving certain types of problems as a major

difference between experts and novices. While experts already had constructed an appropriate schema that may be easily followed to solve a question, novices had not yet constructed a corresponding schema for a problem, and had to rely on other processes instead, which were likely more time consuming and less accurate.

The ways in which experts and novices solve similar problems is an important difference, and one of the main reasons why the use of worked examples began to be tested in research. Since experts and novices approach problems differently and attend to different types of details, there must be fundamental principles regarding how experts solve problems that could be taught to novice learners to help them learn more efficiently and effectively. In the novice/expert literature of the early 1980s, the schematic construction for specific types of questions was cited as one of the main differences between expert and novice learners (e.g., Chi et al., 1981). Novices, lacking these pre-existing knowledge structures, would instead use methods such as means-ends analysis, which may be successful, but are considerably more time consuming and inefficient. Sweller & Cooper (1985) found that, when learning basic algebra, schematic construction was a critical element in students' accuracy in solving problems. Generally, students who were more experienced in solving algebra problems also had more effective schemata constructed for different types of algebra problems. Further, the use of worked examples resulted in students needing less time and displaying higher accuracy on algebra problems when compared to students who only received practice problems. Similarly, Cooper & Sweller (1987) found that students who were given worked examples for algebraic transformation problems were more successful on both similar and dissimilar types of problems that required adaptation of previously constructed schemata. Sweller & Cooper's work showed that using worked examples helped students' schematic construction related to solving algebra

problems, as well as improved students' performance in a shorter period of time than would only the use of practice problems.

One of the most widely cited early studies on worked examples is Zhu and Simon's (1987) series of experiments that examined the use of worked examples versus traditional "learning by doing" (learning with practice problems) in several different mathematical contexts. In their main experiment, Zhu and Simon compared test scores and verbal protocols of students who were learning how to factor quadratic expressions from either worked examples or practice problems. In the initial experiment, students in both conditions learned how to factor these types of expressions fairly quickly, although students who were only given practice problems made more errors earlier in the learning process. These experiments were then replicated later in instruction on a variety of different types of content including exponents, geometry, and ratios and fractions. In all of these experiments, the students who received worked examples performed as well or better on posttests, which consisted of five open-response questions on a certain procedure, than did students who only received practice problems. Furthermore, students who were given worked examples generally required less time to learn these different procedures (i.e., factoring algebraic expressions, exponent rules, solving for unknowns in ratios and proportions). It is important to note that, in all of Zhu and Simon's studies, the instructional materials were focused mainly on the learning of mathematical rules, which generally consist of procedural (i.e., the steps necessary to execute a rule) knowledge. Additionally, in their last study, in which Zhu and Simon tested the use of worked examples versus practice problems over entire years of instruction, they found that the class that used worked examples required almost one third less time to learn an entire year's worth of material than did the class that used practice problems. While this result by itself does not necessarily mean that using worked examples

could save large amounts of instructional time, it does show that the use of worked examples may be advantageous to learners both in terms of knowledge acquisition, as well as instructional time needed, especially when the content being taught consists mainly of rules, or procedural knowledge.

From the results of these early studies, it is clear that worked examples had a positive effect on learning in mathematics. However, why were worked examples more effective than traditional practice problems? Why was the presentation of an "expert" solution so helpful to novice learners? Researchers have hypothesized that worked examples present novices with a expert "model" of how to optimally work through a problem. In terms of cognitive load theory, a worked example would be a model of an appropriate and effective schema that an expert would use to solve a given problem. Schematic construction is cognitively taxing, as it requires a learner to take multiple bits of information and categorize them in a way in which the information will be used to solve a particular type of problem (Chi et al., 1982). Because this task may be high in cognitive load for most learners, it is very possible that working memory could become overloaded, which may slow down or interfere with learning. A worked example would help the novice learner categorize and organize information; by doing this, fewer bits of information have to be attended to in working memory simultaneously, and thus, cognitive load for a task may be decreased. The worked example allows learners to not have to construct schemata completely from the bottom up in working memory; rather, they only have to focus on individual elements of a task. This may also allow the learner to achieve automation with the procedure more quickly, which will then allow the learner to use the procedure more efficiently and accurately when solving problems in the future (Sweller et al., 1998).

Worked examples have been shown to be advantageous to only using practice problems when teaching mathematics, but there are some design issues that may moderate the effectiveness of worked examples in instruction. Subsequent research focused on specific design features related to worked examples, and how worked examples should optimally be designed and integrated into instruction of various types of mathematics. It is important to note that since much of this research was performed with the CLT framework in mind, the purpose of much of this research was to reduce the cognitive load imposed by tasks on learners, as decreased cognitive load was hypothesized to be directly related to improved learning outcomes for students. It is also important to consider the types of knowledge that much of the worked example research has focused on. This research mostly focused on content in mathematics that could be classified as rule or property-based; this type of knowledge is generally procedural in nature. Rourke and Sweller (2009) concede that much of the worked example research has been performed in highly structured, algorithmic domains. The type of knowledge that worked examples are shown to be most effective at helping to teach will be an important issue to consider throughout a review of the worked example literature.

Worked examples and multiple representations. One important feature of worked examples in instruction is how the worked examples are constructed. Just because worked examples are incorporated into instruction does not guarantee that students will learn the material more effectively. According to CLT theory, worked examples are effective in mathematics instruction because they help reduce the cognitive load imposed on learners by certain mathematical tasks. However, if a worked example is designed in such a way that it may increase cognitive load, then learners may have more trouble learning the materials. For example, Sweller and various colleagues, in a series of different experiments, found that when

worked examples contained both text and diagrams that the learner had to attend to simultaneously, then learners actually experienced difficulties. This effect is known as the split-attention effect (Ayres & Sweller, 1995; Tarmizi & Sweller, 1988). Tarmizi and Sweller used various types of worked examples to teach geometry to students. Students were either assigned to a standard problem solving condition or a worked example condition where problems were presented in pairs: one worked example, and then one problem to solve. In the worked example condition, learners were required to use information about circles in conjunction with given problems and diagrams. Because learners were required to attend to such a large amount of information simultaneously, the worked examples did not have their expected effect. While most previous research had shown that worked examples were advantageous to using practice problems only, Tarmizi and Sweller found no significant differences on posttests, and actually found that students in the regular problem solving condition took significantly less time to complete the posttest. The researchers concluded that, due to the format and presentation of information, the worked examples did not help facilitate learning. Their study was one of the first to show that not all worked examples are effective, and more specifically showed that the split-attention effect is something that must be considered when incorporating worked examples into instruction.

The process of attending to multiple pieces of information at the same time does not have to be an impediment to learning, though. If designed correctly, examples using information from sources can be helpful to a learner. Specifically, presenting information in different modes (i.e., auditory and visually) may be more effective than simply presenting all information visually. Mousavi, Low, and Sweller (1995) tested this hypothesis in a series of four experiments examining how different types of instruction may help students learn about geometric proofs.

Mousavi et al. used worked examples that showed a geometry diagram and corresponding proof. These diagram/proof pairs were presented in three different ways: visual only (diagram and explanations), visually with an auditory explanation of the proof, and a visual diagram with an explanation presented both visually and aurally. The results showed that there were no differences in posttest performance, but time spent on the posttest was lower for students in the two conditions that used both visual and auditory representations of the proofs.

A couple of caveats exist with the practice of using dual modes with worked examples, though. Jeung, Chandler, and Sweller (1997) found that, when diagrams were especially complex in diagram-explanation pairs in geometry worked examples, then the advantages in time on task found by Mousavi et al. (1995) disappeared. Similarly, if information presented in both modes is redundant, then the positive effects of using dual modes disappeared (redundancy effect; Bobis, Chandler, & Sweller, 1994). Specifically, this redundancy effect may be more pronounced for learners with higher levels of prior knowledge, and may be evidence of the expertise reversal effect (Kalyuga, Chandler, & Sweller, 1998). Thus, information presented in dual modes should not be overly complex in either mode, but information contained in both modes should not be redundant if the use of dual modes in worked examples is to be optimized.

In general, the research on worked examples in mathematics instruction has yielded several important design principles. Worked examples that are structured in ways such that learners do not have to process several pieces of information simultaneously are preferable, as this avoids the split-attention effect, which results in higher levels of cognitive load. If multiple bits of information need to be presented simultaneously, it may be preferable to present information through different modes, such as visually and aurally. This is known as the modality effect, and it may help offset negative effects of split attention. Finally, worked examples tend to

be more effective if the specific tasks and sub-goals are highlighted or otherwise cued for the learner (Catrambone & Holyoak, 1990). This may allow the learner to focus more on the details of a task, and less on searching for an appropriate strategy to complete the task.

Use of multiple worked examples. While a large portion of research on worked examples in mathematics has focused on how to design individual worked examples and incorporate them into instruction, there has also been much research on the nature of the number and type of worked examples that are included. Research on worked examples has shown that not only is the content of worked examples important, but the layout and sequencing of multiple worked examples throughout instruction is important in maximizing their positive instructional effects.

One very important consideration when designing instruction with worked examples is the number of worked examples to include. Reed and Bolstad (1991) used instruction on rate and time computations to construct six different conditions which involved different combinations of single and multiple examples, as well as simple and complex examples. They found that students who were in the condition that used both simple and complex worked examples outperformed students in all other conditions, including students who were exposed to one simple or one complex example. This finding has been replicated in additional studies. (e.g., Brunstein, Betts, & Anderson, 2009; Cooper & Sweller, 1987).

In addition to the number and type of worked examples or completion problems that should be used in instruction, the variability of the types of worked examples is another important consideration in instructional design. Paas and van Merrienboer (1994) tested this effect in an experiment in which they taught geometry to high school students. Specifically, they hypothesized that using varied examples in instruction would be more beneficial for learners

than using the same type of examples. Further, they also hypothesized that using varied worked examples would be advantageous to using varied examples as practice problems. To test these hypotheses, the researchers assigned learners to one of four conditions: low-variability practice problems, low-variability worked examples, high-variability practice problems, and high-variability worked examples. On a posttest measure consisting of transfer tasks, they found that students who received high-variability worked examples performed better on the posttest than did students who received low-variability worked examples.

Another way in which worked examples may vary is the types of details and stories that are used to give context to tasks. Previous researchers have noted that, while expert learners tend to be very skilled at paying attention to structural details of tasks, novice learners tend to focus on surface level details which are not necessarily related to successfully completing a task (Ross, 1989). For this reason, using different stories and contexts for worked examples on the same topic may be advantageous, as this may help learners focus on the important structural details, while not being distracted by surface details. Quilici and Mayer (1996) tested this hypothesis by modifying instruction on basic statistical procedures (chi square tests, t tests, and correlations). They assigned students to one of three groups: worked examples emphasizing structure (different stories and contexts), worked examples emphasizing surface details (same context for all worked examples of a given topic), and no instructional materials. On a card sort posttest measure, students in the structure-emphasizing condition sorted different types of problems based on their structure, while students in the other two conditions sorted in similar ways, attending more to surface features. This result shows that designing a series of worked examples with different contexts and surface stories is an optimal design strategy, as it helps learners focus on the

important structural details of a task while not being distracted by similar surface features that are not relevant to the task.

Overall, research on using multiple worked examples in mathematics instruction has yielded several design strategies that should be considered in order to maximize potential benefits to learners. Variability in terms of the types of worked examples included in instruction may be beneficial to learning. Worked examples may also vary in terms of the stories and contexts used. Use of different stories and contexts in multiple worked examples on a similar topic will help learners attend to important structural features of questions and not be distracted by irrelevant surface features.

Worked Examples: A Limitation?

One recurring theme in the worked examples research is the idea that the use of a worked example provides a model that students can refer to and use when solving new types of problems. By providing this model for students, a worked example is theorized to decrease the amount of cognitive load needed to learn a procedure, as the example provides a structure for the learner to use, thus freeing working memory to attend to the individual steps, and not the overall structure of the procedure. The use of a worked example is also a way to help ensure that students do not encounter excessive difficulty when learning a new topic. By providing a model for students to follow, worked examples are, in part, providing a mechanism that guards against students failing to learn a new topic (Reiser, 2004). Most instructional manipulations in mathematics instruction (and other content areas) focus on the "success" of students learning new topics, and minimizing difficulties (Kapur & Bielaczyc, 2012). Obviously, success when learning a new topic should be the ultimate goal of instruction. However, there may be some learning-related benefits to students initially encountering difficulty when learning a new topic

(DeCaro & Rittle-Johnson, 2012). This initial difficulty has to be followed by success for the student; if the failure persists, then learning will not occur, and the student may also become frustrated. If, though, a student attempts to solve problems or learn new material and encounters initial failure, there may still be opportunities for the learner to benefit.

The worked example research can also be viewed through the type of knowledge that worked examples are generally used with. In most of the worked example research involving mathematics instruction, the worked example instruction is generally compared to direct instruction or direct instruction followed by practice problems. This research tends to focus mainly on content that is procedural in nature, such as algebraic equations (e.g., Cooper & Sweller, 1987; Zhu & Simon, 1985), geometric rules (e.g., Paas & van Merrienboer, 1994; Tarmizi & Sweller, 1998), and statistical computations (e.g, Renkl, 1997; 2002). There is not as much of a focus on conceptual knowledge, and the results are more mixed regarding the effectiveness of worked examples on conceptual knowledge acquisition. Thus, a comparison of worked examples against another type of instruction that is designed for instruction of different types of knowledge, such as conceptual knowledge, may constitute an interesting and useful comparison. The research is fairly consistent regarding the effect of worked examples on learning procedural knowledge when compared to direct instruction and practice. However, a comparison of worked example instruction against another type of instruction on not just procedural knowledge, but also conceptual knowledge related to statistics, may advance the research on statistics instruction, especially in regards to the types of knowledge that different types of instruction are most effect in teaching.

Worked Examples versus Productive Failure: A Comparison

There are certainly some similarities between the use of worked examples and productive failure in statistics instruction, and instruction in general. Both types of instruction are more effective when paired with a form of direct instruction such as a lecture or an expository text. Both types of instruction are designed to allow learners to attend to critical features of a problem, even if the methods are different. By attending to critical features, students may begin to differentiate and classify different types of problems and build conditional knowledge regarding the appropriate procedure to use for a specific type of problem. Prior knowledge plays a role in the effectiveness of both of these strategies, although the exact nature of the effect is somewhat different for worked examples and productive failure.

However, there are also several ways in which the worked example strategy differs greatly from the productive failure strategy. The most obvious difference is when students individually work on problems, and specifically when this work occurs in relation to direct instruction. When using worked examples, direct instruction is usually at the beginning, followed by worked examples, and then practice problems for students. With productive failure, however, the order is changed: students begin with on a novel problem and attempt to generate as many solutions as they can. This activity is then followed by direct instruction, which may also then be followed by practice problems, as in worked examples. Essentially, a main difference is simply the order in which components of instruction are arranged. With worked examples, a new concept or topic is introduced through direct instruction, modeled with worked examples, and then followed by practice, during which students may refer back to worked examples. Productive failure uses a novel problem to introduce a new topic, which is then followed by direct instruction, which may also include a teacher working through examples

relevant to the topic being discussed. The use of worked examples is a way in which instruction is manipulated to ensure that a student does not encounter difficulty when learning a new topic (Kapur, 2008). Productive failure, however, operates on the assumption that a student will encounter difficulty, but that this failure prepares the student to learn in the immediate future. The design of instruction, along with the role of practice, is completely between these two instructional strategies, as worked examples focuses on modeling of procedures and minimizing the possibility of student failure, while productive failure uses practice on a novel problem as a way to bring about initial failure, while may then lead to advances in learning.

The difference between productive failure and worked examples is also very pronounced when considering the types of knowledge that each instructional technique is designed to help students learn. Productive failure has shown evidence of facilitating conceptual knowledge acquisition related to statistical topics such as central tendency and variance more effectively than direct instruction with practice problems (Kapur, 2012; 2014). Worked examples have shown evidence of facilitating procedural knowledge acquisition related to statistical topics such as probability (Renkl, 1997; 2002) and statistical distributions (Catrambone & Holyoak, 1990). Worked examples have been used much more commonly in mathematics instruction, including statistics instruction, than have productive failure prompts. However, the research on statistics instruction has generally cited a lack of conceptual knowledge learning as one of the most important issues in statistics education (Lavigne et al., 2008). If conceptual knowledge learning is an important issue in statistics education, then use of a technique such as productive failure, which is designed to encourage conceptual knowledge acquisition, may be a useful complement to worked examples, which generally are more effective at fostering procedural knowledge acquisition.

Cognitive load is also an issue on which differences may occur between PF and WE. Productive failure activities are generally very high in cognitive load because the task of generating solutions to a novel problem is a cognitively taxing activity. Worked examples, by contrast, have shown evidence of lower levels of cognitive load. Cognitive load theory states that lower levels of cognitive load, especially extraneous load, are associated with better learning outcomes (Sweller et al., 1998). While worked examples are theorized to be effective because they reduce cognitive load, productive failure is theorized to be effective despite the higher cognitive load associated with this task. It is possible that the larger amount of cognitive load associated with productive failure is germane to learning of conceptual knowledge. Thus, the higher levels of cognitive load associated with productive failure may not be problematic in instruction, and may actually be indicative of conceptual knowledge acquisition

Productive failure and worked examples have clear differences in terms of the design of instruction, the types of knowledge they are used for, and subsequently, the amount of cognitive load imposed on the learner. Productive failure has shown evidence of being more effective than direct instruction and practice problems in teaching conceptual knowledge, while being not significantly different in teaching of procedural knowledge. Worked examples have been shown to be effective in promoting procedural knowledge acquisition in statistics, but the evidence of effectiveness in related conceptual knowledge acquisition is less clear. Productive failure generally results in a higher amount of cognitive load being imposed on a learner than does worked examples, but this increased cognitive load may not be problematic. The literature on statistics instruction cites a lack of conceptual and conditional knowledge as one of the most pervasive issues in the field (i.e., Garfield & Ahlgren, 1988; Yan & Lavigne, 2014). Worked examples are effective at teaching procedural knowledge related to statistics, but productive

failure may be more effective in teaching conceptual knowledge related to statistics. Consideration of the interaction between types of instruction and knowledge to be learned in statistics may be useful in addressing the documented shortcomings in current statistics education.

Purpose of Study

Statistics instruction has been well-documented as being very difficult for novice learners (Garfield & Ahlgren, 1988). Productive failure and worked examples have both shown evidence of being effective in statistics instruction, but these two instructional techniques have not yet been tested against each other. More importantly, productive failure and worked examples are designed to target different types of knowledge (conceptual and procedural knowledge, respectively). The purpose of this study, then, is to compare the effectiveness of productive failure (PF) against well-designed worked-example instruction (WE), as well as a control condition, which will consist of participants reading an expository text. This will allow for a direct comparison of the effectiveness of worked examples and productive failure on both procedural and conceptual knowledge questions on the same statistics topic, using the same instructional materials, in the same population.

In addition, this study compares the effects of worked examples against productive failure on delayed posttest performance. There is little productive failure research than examines the effectiveness of this approach on a delayed posttest, but the nature of the PF activity, as well as research from instructional paradigms related to PF (i.e., Schwartz & Bransford, 1998) suggest that PF may be more effective on a delayed posttest than other forms of instruction. Worked examples have shown some evidence of effectiveness in terms of delayed posttest performance, but the amount of the literature regarding this topic is much smaller, and the effect

is not as clear. It is expected that there will be some sort of drop in performance from immediate to delayed posttest, but an examination of the interaction between instructional condition and immediate and delayed posttest performance may have important implications to both instruction and research. Thus, in addition to comparing the effectiveness of worked examples against productive failure on conceptual and procedural knowledge problems, there is also a comparison of the effects of these two instructional strategies on immediate and delayed posttests. This allows for an examination of immediate versus delayed effects of these instructional strategies on knowledge acquisition.

Cognitive load is also measured in this study. Students who engage in productive failure consistently report higher levels of mental effort than do students who are administered other types of instruction. However, this increased cognitive load may not be detrimental to learning, and may in fact be helpful in building conceptual and conditional knowledge in learners. Thus, cognitive load is measured not just during the instructional manipulation (worked examples versus productive failure versus direct instruction), but also during posttests. This allows for comparisons of cognitive load to be made not just during the productive failure or use of worked examples, but also how these activities influence cognitive load later in instruction.

Research Questions and Hypotheses

Research Question #1: Are the effects of instructional condition on type of knowledge students learn uniform across time?

Hypothesis: PF is theorized to be effective in large part because of prior knowledge activation, allowing for new content to be processed in ways that allow for more meaningful processing and encoding in long term memory. Thus, it is hypothesized that there is an interaction between the effect of instructional condition on procedural and conceptual knowledge and time of test (immediate vs. delayed). Specifically, the effect of the productive failure prompt may be more noticeable on the delayed posttest compared to the immediate posttest, as the productive failure activity should result in more retention of material than does worked examples or an expository text, which do not promote prior knowledge activation as much. However, since PF is theorized to be more effective in conceptual knowledge acquisition, while WEs are theorized to be more effective in procedural knowledge acquisition, the effect might not be uniform across both types of knowledge.

Research Question #2a: Is there an interaction between instructional condition and the type of knowledge students learn?

Research Question #2b: What is the effect of instructional condition on student learning of conceptual and procedural knowledge?

Hypothesis: Because productive failure and worked examples focus on different types of knowledge, then the hypothesis is that there should be an interaction between type of instruction and acquisition of different types of knowledge. Specifically, Productive Failure and Worked Example (PF and WE, respectively) participants will outperform control condition participants on both measures of conceptual and procedural knowledge. However, PF conditions participants

will outperform WE condition participants on measures of conceptual knowledge, while performing approximately the same on measures of procedural knowledge.

Research Question #3: Is there an interaction between instructional condition and performance on immediate and delayed posttests?

Hypothesis: PF is theorized to be effective due to the generative nature of the task, as well as the prior knowledge activation that the activity is designed to promote. These factors are related to more meaningful processing and encoding, which in turn may manifest in the form of better performance on a delayed posttest than participants receiving other instruction. Thus, it is hypothesized that there is an interaction between instructional condition and performance on immediate and delayed posttests. Specifically, PF condition participants are expected to perform better on the delayed posttest than on the immediate posttest, but this difference may not be evident (or may actually be reversed), on the immediate posttest.

Research Question #4a: Is there an interaction between instructional condition and the amount of mental effort students perceive they use?

Research Question #4b: What is the effect of instruction type on the amount of mental effort students perceive they use after different activities?

Hypothesis: PF condition participants will, on average, perceive they use more mental effort than WE and control condition participants, especially after the initial activity. Specifically, because PF requires a learner to create novel solutions to a problem that a learner has probably never encountered before, perceived mental effort may be especially high in this initial activity, before dropping later in instruction. While this may seem problematic, if PF participants are outperforming participants in both other conditions on measures of conceptual knowledge and

performing equally as well on measures of procedural knowledge as learners in the worked example condition, then this increased mental effort may actually be advantageous for the learner, which is contrary to much of the worked example research and cognitive load theory, in general. However, the difficulties that participants encounter while completing the productive failure activity should be addressed by direct instruction, so the discrepancy in cognitive load after the initial activity should decrease on the posttests.

Chapter 3

Methods

Participants

The participants for this study were drawn from an undergraduate biology course at a large university in the northeastern United States. Participants were assigned to one of three conditions, Productive Failure (PF), Worked Example (WE), or Control (C) in this study. Given the nature of the course participants were drawn from, it was assumed that most students' prior knowledge of statistics ranged from none to intermediate, as some of these students already had taken a statistics class in college.

In all, 313 students participated in the study and completed the pretest, immediate posttest, and delayed posttest. However, only 260 of the students completed all of the parts of the study and followed directions for giving informed consent for participating in the study. Of these 260 students, 83 students were assigned to the worked example condition, 89 students were assigned to the productive failure condition, and 88 students were assigned to the control condition. In terms of gender, 68.4% of the overall sample was female, and the average participant age was 19.42 years. In terms of ethnicity, 67.4% of the participants identified themselves as White, followed by Asian (15.3%), Hispanic (4.5%), and Black (4.2%). The average standardized math test scores were 629.64 for SAT Math scores and 28.39 for ACT Math scores. The average college GPA of participants was 3.31.

Design of Study

Participants were randomly assigned to one of the three instructional conditions. Participants in the PF condition worked on an open-ended problem related to correlation and asked to generate as many unique solutions to the problem as they could. This was followed by

a period of direct instruction. Participants in the WE condition studied a worked example on correlation and then completed a practice problem, which was then followed by direct instruction. The control condition read an expository text on correlation, which was then followed by direct instruction. For all three conditions, an immediate posttest followed direct instruction. A week later, participants in all three conditions were administered a delayed posttest. Cognitive load was measured after the instructional manipulation (worked example, productive failure prompt, or expository text) and after both immediate and delayed posttests.

Materials

Pretest/prior knowledge measure. A topic-specific pretest was administered to students. This pretest contained eight multiple choice items. These items assessed basic factual and conceptual knowledge related to correlation, including basic definitions, possible values of correlations, and determination of magnitude and direction of correlations from scatterplots. This pretest was similar to pretests that have been used in previous studies in both the worked example and productive failure literature (i.e., Grosse & Renkl, 2007; Kapur, 2014). However, it is important to point out that, even though this pretest was similar to others used in similar research, this pretest was measuring topic knowledge, and not necessarily domain knowledge. Domain knowledge may be defined as knowledge broadly related to a subject area (i.e., statistics) (Alexander, Kulikowich, & Schulze, 1994), while topic knowledge may be defined as knowledge related specifically to one concept within a domain (Alexander et al., 1991). Thus, this pretest measured prior correlation (topic) knowledge, but not prior domain (statistics) knowledge. One point was awarded for each correct answer and zero points for each incorrect answer, which resulted in a range of scores between zero and eight. Appendix A contains the complete pretest that was administered to all participants in this study.

Instructional materials. All participants received a form of instruction related to the topic of correlation. PF condition participants received an open-ended problem on the topic of correlation and were instructed to generate as many solutions as they could to the problem, consistent with prompts used in previous PF research (i.e., Kapur, 2012; 2014). The PF prompt is shown in Appendix B. WE condition participants received a worked example that was designed in accordance with the principles of optimal worked example design (i.e., Atkinson et al., 2000). The worked example contained completely worked out calculations for each step of the procedure, along with a brief description of what each step in the procedure was achieving. Introduction of new steps were indicated by bold-faced and/or underlined text, as were important values and calculations, as has been recommended by previous researchers (Yan & Lavigne, 2014). In addition, a table was included in some steps to help guide the learner through the calculations. Following the worked example, learners were given another problem to attempt, as is consistent with the worked example followed by practice problem ordering in typical worked example studies (i.e., Mwangi & Sweller, 1998). Appendix C shows the worked example and practice problem that was presented to WE condition participants. Control condition participants were given a text that discussed the concept of correlation and the formula for calculating Pearson's *r*, but this text did not include any worked examples; rather, this was an expository text that simply described this content including: the definition of correlation, magnitude and direction of correlation, possible values of a correlation, and the formula for computing a Pearson correlation.

Posttest. The posttest contained seven items; three items assessing procedural knowledge and four items assessing conceptual knowledge. The complete correlation posttest is shown in Appendix D. Two of the open-response procedural questions required students to calculate the

correlation between two variables, given raw data of a small number of observations. There was one closed-response procedural question in which students were presented the four steps of calculating a correlation coefficient, as well as one extra step. Students were required to choose the correct steps and put them in the correct order. There were two different forms of the posttest. These forms consisted of the same items in different orders. The same two forms of the posttest were also used as the delayed posttest.

Measurement of cognitive load. Cognitive load was measured at the end of three activities, resulting in three measurements of cognitive load for each participant. The measure consisted of one question asking learners to rate the level of perceived mental effort they used on the task they just completed, similar to the measure developed by Paas (1992). Ratings ranged from one to nine, with lower ratings indicating lower levels of perceived mental effort and higher ratings indicating higher levels of perceived mental effort. Measurements of cognitive load were taken after the initial activity (worked example, productive failure prompt, or expository text), after the immediate posttest, and after the delayed posttest. This is preferable to only measuring cognitive load once at the end of instruction, as using only a single cognitive load measure at the end of instruction may be less reliable because the learner only remembers mental effort used on the last activity, or only remembers the task requiring the most mental effort (Schmeck, Opfermann, van Gog, Paas, & Leutner, 2015). Multiple observations of cognitive load allowed for more accurate assessments of average or overall mental effort expended by learners.

Procedures

The experimenter visited the class to recruit participants. Students were told that the experiment covered statistics content, which was not a focus of the course, but that they would receive extra credit for their participation in the study. Informed consent was administered

through a course management system; participants had to give consent for their data to be used in this study. Demographic data were also collected by administering a survey via the same course management system.

All activities in this study were conducted in experimental sessions held in university classrooms. Participants signed up for these sessions using an online sign-up system. Participants were required to sign up for one session each from two different groups of sessions: the first group of sessions contained the instructional manipulation, direct instruction, and immediate posttest. The second group of sessions contained the delayed posttest, and was administered the week following the first group of sessions. Participants were assigned to conditions based on the first session that they signed up for. Of the six sessions offered in the first group of sessions, two sessions each were for PF, WE, and control conditions.

Participants in all three conditions completed the instructional manipulation first. Thus, PF condition participants received the open-response question, WE condition participants received the worked example and practice problem, and control condition participants received the expository text on correlation. Participants in all three conditions were given 10 minutes to complete the corresponding activity.

After the instructional manipulation, all participants were then administered a 20 minute period of direct instruction. This direct instruction consisted of a lecture-style format in which the topics discussed in the instructional manipulation were explained in more detail. To control for instructor effects, the experimenter lectured on the concept of correlation and the formula for computing correlation to all conditions. The same slides and examples were used across all three conditions, and the experimenter used a script to ensure that the presentations were as similar as

possible to each other. The instruction was designed in a way to mimic the instruction that would occur in an introductory statistics classroom, albeit in a compressed time frame.

Participants in all three conditions were then administered the immediate posttest. A maximum of 30 minutes was given for participants to complete this posttest. For participants in all three conditions, the instructional manipulation, direct instruction, and immediate posttest comprised the first experimental sessions and took a total of one hour to complete.

The following week, participants in all three conditions were administered the delayed posttest. As with the immediate posttest, participants were given 30 minutes to complete this measure.

Scoring Rubric for Posttests

Each posttest item was worth four points, with partial-credit available for all items. For open-response procedural items, use of the correct formula but with computational errors resulted in partial credit being awarded, with more errors resulting in fewer points being awarded. The closed-response procedural item was simply scored by counting the number of steps correctly placed in the right order for calculating a correlation coefficient. For conceptual items, the rubric varied, depending on the question. Table 1 describes the rubric that was used for scoring each of the correlation posttest items. The same posttest was administered to all three groups.

Table 1

Scoring rubric for correlation posttest items

Type of item	Scoring rules	Procedural/Conceptual
Open-ended procedural questions (Items 1 and 4 on posttest)	1 point for correct summations 1 point for correct numerator (substitution and operations) 1 point for correct denominator (substitution and operations) 1 point for correct computation and answer 1 point TOTAL if all calculations are wrong but there is some indication of correct formula use	Procedural
Scenario questions (Items 2 and 5)	1 point for correct choice 1 point for mentioning positive or negative 2 points for applying the relationship correctly	Conceptual
Scatterplot question (Item 3)	2 points for choosing correct scatterplot 1 point for correct explanation 1 point for us of some form of the word “line” or “linear”	Conceptual
Interpretation of a correlation coefficient (Item 6)	1 point for direction (mention positive or negative) 1 point for magnitude 2 points for explanation	Conceptual

Interrater Agreement of Posttest Scores

All immediate and delayed posttest responses were scored by two raters. These raters were the experimenter and a graduate student from the educational psychology program at the university where the study was conducted. Both raters worked together to create the scoring rubric above. Both raters rated all responses, and a random 20% of the posttests were chosen to calculate agreement. All responses from the posttests selected were used to calculate interrater

agreement. Interrater agreement was calculated by dividing the number of scores the two raters disagreed on by the total number of scores compared between raters. Interrater agreement was 86.3%; of the 13.7% of scores on which raters disagreed, the vast majority of disagreements were one-point differences in score. Disagreements of more than one point where resolved by discussion between the raters.

Chapter 4

Results

Prior Knowledge Across Conditions

The average score on the eight item multiple choice correlation pretest was 5.69, with a standard deviation of 1.72. There were no differences of mean pretest score between the three conditions, $F(2, 255) = .051$, $p = .95$. Thus, prior knowledge was not different across the three conditions. Cronbach's alpha for the pretest was .54, which is a low, but still acceptable level of reliability. Correlations between pretest scores and posttest scores were generally quite low, with correlations between pretest and conceptual knowledge scores of .25 and .29 on immediate and delayed posttests, respectively, and correlations between pretest and procedural knowledge scores of .252 and .34 on immediate and delayed posttests, respectively.

Assumptions

Mixed ANOVAs were performed to investigate each of the research questions. Assumptions of mixed ANOVAs include normality, independence of observations, homogeneity of variance and covariance, and sphericity. The independence of observations assumption is met by the nature of data collection; observations from different students can be assumed to be independent. Shapiro-Wilk tests were performed to assess normality of each of the dependent variables. Even though these tests indicated significant deviation from normality in terms of skewness, none of the skewness values were larger in magnitude than 1. Homogeneity of variance was met for the most part; the procedural knowledge scores on the immediate posttest did not meet this assumption, but repeated measures ANOVA is robust in cases where sample sizes are approximately equal across groups and the ratios of group variances are not significantly different from 1 (Cohen, 2013). Homogeneity of covariance was tested using Box's

M; for some analyses, the p-value for the test was less than .05 but greater than .01. While this result would suggest that the homogeneity of covariance assumption has been violated, Tabachnick & Fidell (2007) state that, due to the high sensitivity of the Box's M test, a mixed ANOVA design is still robust with approximately equal sample sizes and a p-value of at least .001 for Box's M test. The sphericity assumption was met for all analyses, using Mauchly's W.

Effect of Instruction on Learning Across Time and Type of Knowledge

A 3 (instructional condition) x 2 (type of knowledge) x 2 (immediate and delayed posttest) mixed ANOVA was performed to test for a three way interaction between instructional condition, type of knowledge, and time of posttest, as described in research question #1. This three way interaction was found to not be statistically significant, $F(2, 238) = 1.397$, $p = .249$. Thus, the effect of instruction was not the same across time and type of knowledge. Figures 1 and 2 plot the average performance of participants across groups and posttests, in terms of percentages of points obtained, on procedural and conceptual knowledge items, respectively. Table 2 gives the means and standard deviations of percentages of points obtained by participants across all instructional conditions on procedural and conceptual knowledge items at both posttests.

Figure 1. Average procedural knowledge performance on two posttests (in percentage of points earned)

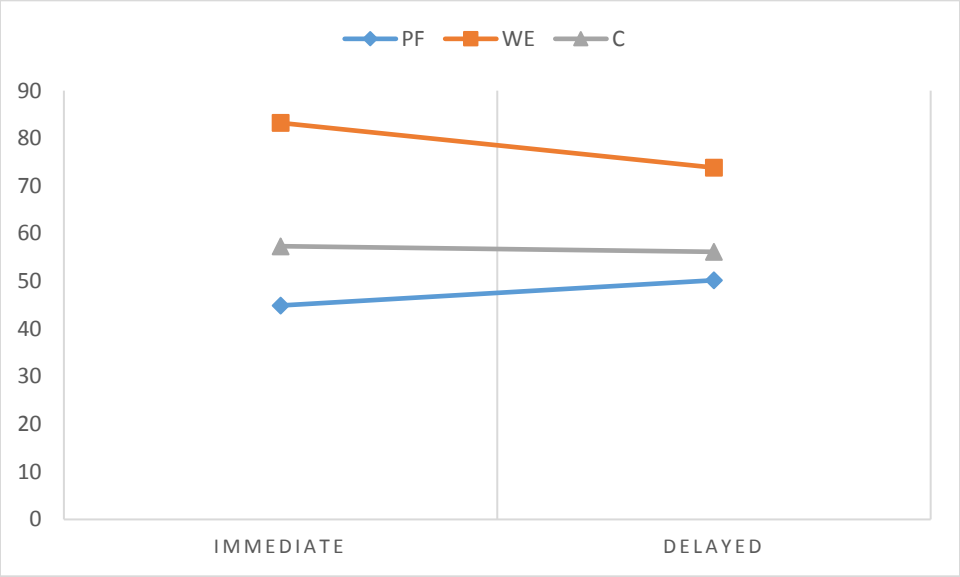


Figure 2. Average conceptual knowledge performance on two posttests (in percentage of points earned)

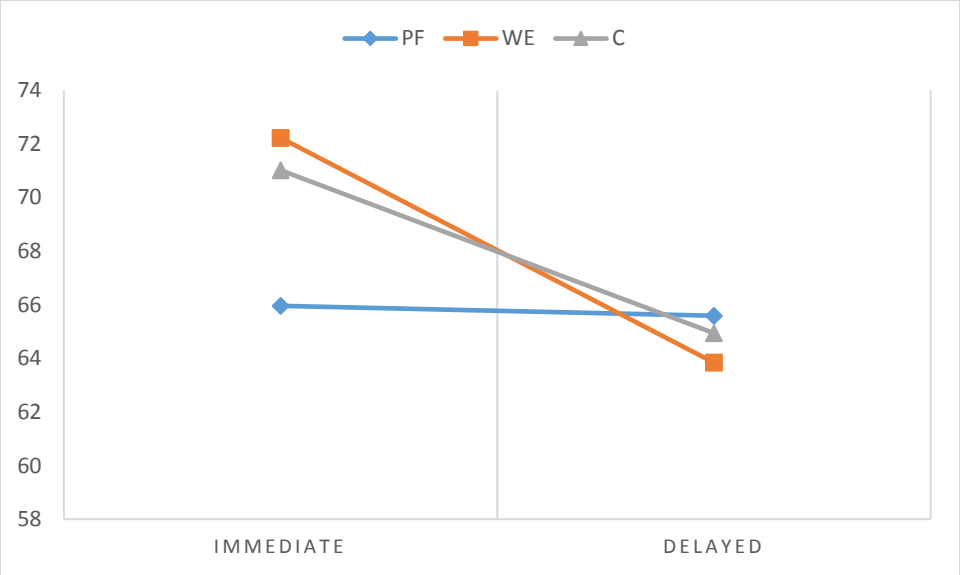


Table 2

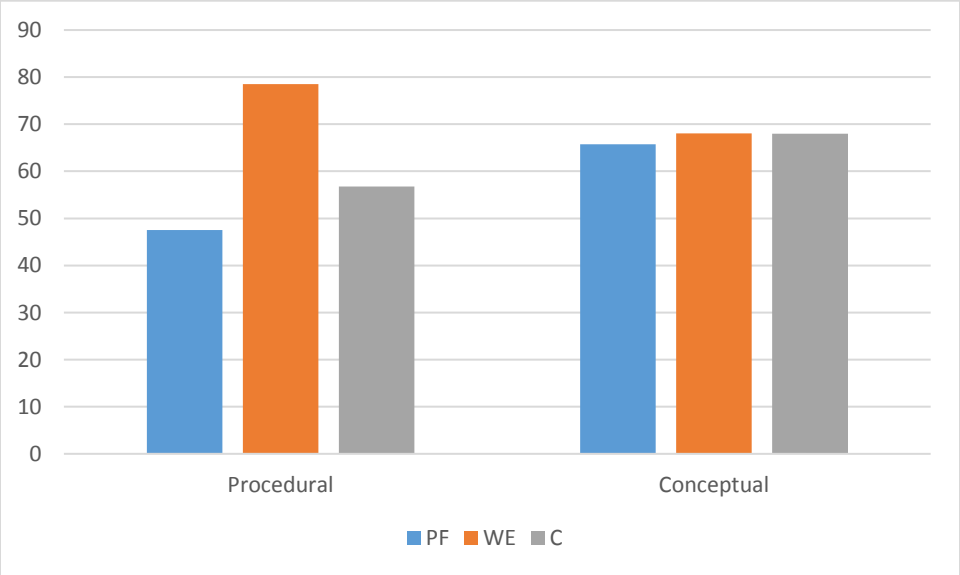
Average performance, in percentages of points (SD), across conditions and posttests

	Procedural (Immediate)	Procedural (Delayed)	Conceptual (Immediate)	Conceptual (Delayed
PF	44.90 (32.46)	50.20 (29.35)	65.96 (16.91)	65.59 (16.82)
WE	83.23 (19.13)	73.84 (27.92)	72.23 (16.22)	63.84 (15.17)
C	57.35 (34.67)	56.17 (31.43)	71.02 (12.26)	64.94 (12.75)

Effect of Instruction on Learning

The 3 x 2 x 2 mixed ANOVA found a statistically significant interaction between instructional condition and type of knowledge, $F(2, 238) = 25.507$, $p < .001$, $\eta^2 = .174$. Thus, the effect of instruction was different across types of knowledge. To further test for simple effects, the 3 (condition) x 2 (type of knowledge) interaction was broken down by type of knowledge, resulting in two one-way ANOVAs, one each on procedural and conceptual knowledge. Procedural and conceptual knowledge scores, which were percentages of points obtained, were collapsed across both posttests. Figure 3 plots average percentages of points obtained on procedural and conceptual knowledge problems by condition, collapsed across both posttests.

Figure 3. Average performance on procedural and conceptual knowledge items by condition (in percentage of points earned)



The first ANOVA, on procedural knowledge scores, showed a statistically significant effect of instruction condition, $F(2, 238) = 27.857$, $p < .001$, $\eta^2 = .190$. Tukey post-hoc testing showed the WE condition participants scored significantly better on procedural knowledge questions than did PF and control condition participants. Control condition participants

performed better on procedural knowledge questions than did PF condition participants, but this different was not quite statistically significant (p = .081).

The second ANOVA was identical to the first analysis, with the only change being percentage of points obtained on conceptual knowledge items as the dependent variable. This analysis showed no statistically significant effect of condition, F(2, 238) = .727, p = .485, $\eta^2 = .006$. The order of the instructional conditions was the same as with procedural knowledge, but none of the differences came close to approaching statistical significance. Means and standard deviations for total procedural and conceptual knowledge scores across the three conditions are found in Table 3.

Table 3

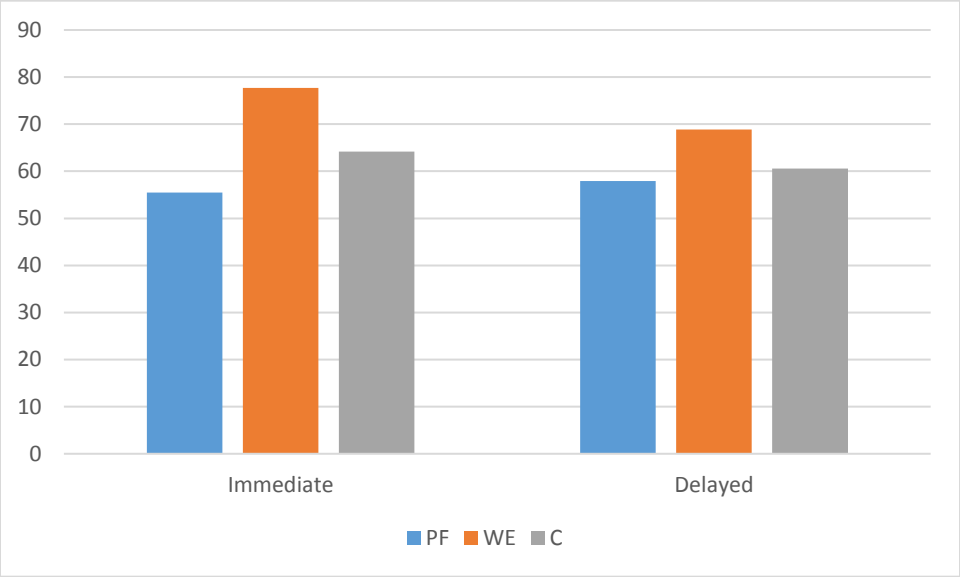
Average percent scores (SD) for procedural and conceptual knowledge measures across posttests

	Mean procedural score (SD)	Mean conceptual score (SD)
PF	47.55 (28.82)	65.77 (14.21)
WE	78.53 (20.94)	68.03 (14.04)
Control	56.76 (30.62)	67.98 (12.96)

Effect of Instruction on Learning: Immediate vs. Delayed Posttest

The 3 x 2 x 2 mixed ANOVA found a statistically significant interaction between instructional condition and time of posttest (immediate vs. delayed), F(2, 238) = 12.316, p < .001, $\eta^2 = .094$. Thus, the effect of instruction was different across time. Figure 4 plots the average percentage of points obtained on procedural and conceptual knowledge items across all three conditions. WE condition participants performed best on both posttests, followed by the control and PF conditions, but the gaps between the conditions became much smaller at the delayed posttest.

Figure 4. Performance by condition on both posttests (in percentage of points earned)



To further examine simple effects, two one-factor ANOVAs using instructional condition as the independent variable were performed on immediate and delayed posttest percentages, as type of knowledge was collapsed for these analyses. For the first analysis, which used immediate posttest percentages as the dependent variable, a statistically significant effect of instructional condition was found, $F(2, 238) = 29.701$, $p < .001$, $\eta^2 = .200$. Tukey post-hoc testing showed that WE condition participants scored significantly higher on the immediate posttest than did control and PF condition participants, and control condition participants scored significantly higher than did PF condition participants.

The second analysis was similar to the previous analysis, with delayed posttest percentage becoming the dependent variable. A statistically significant effect of instructional condition was found, $F(2, 238) = 8.088$, $p < .001$, $\eta^2 = .064$. Tukey post-hoc testing showed that WE condition participants again scored significantly better than did control and PF condition participants. However, the difference between the control and PF conditions was not significant. Moreover, the magnitude of the differences between the conditions became smaller, and the

effect size of condition on delayed posttest scores was less than one third the effect size on immediate posttest scores. Table 4 shows the average total procedural and conceptual knowledge percentages of points obtained for each of the three conditions.

Table 4

Average percent scores (SD) for immediate and delayed posttests

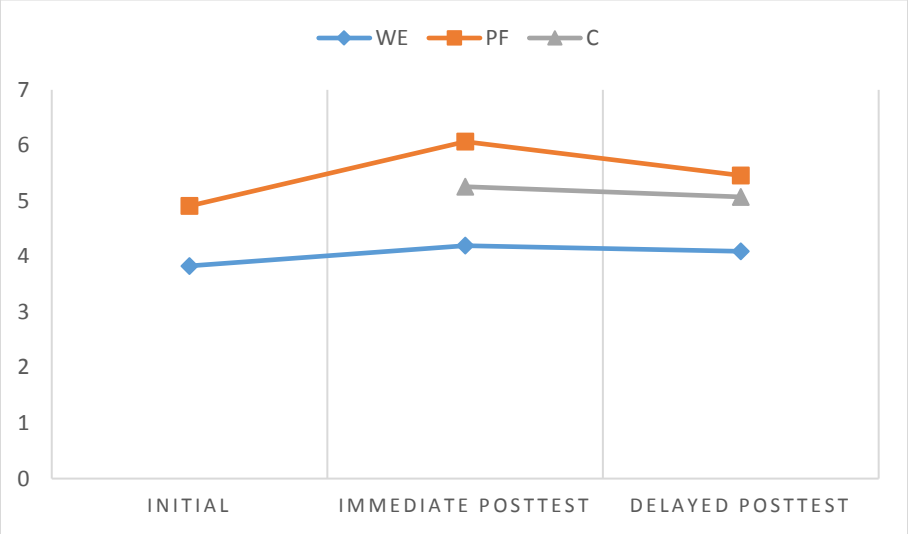
	Immediate Posttest	Delayed Posttest
PF	55.45 (19.77)	57.89 (17.91)
WE	77.73 (14.01)	68.84 (18.21)
Control	64.19 (21.36)	60.55 (18.03)

Effect of Instruction on Cognitive Load

A mixed ANOVA was performed to test for an interaction between cognitive load measurements and instructional condition. Instructional condition was the independent variable, and cognitive load reported at three different times was the repeated measure. Because of an error, no cognitive load measurements were available for the control condition after the initial activity, so there were only two levels of instructional condition. A statistically significant interaction between instructional condition and cognitive load was found, $F(2, 127) = 7.397$, $p = .001$, $\eta^2 = .104$. Students in the worked example condition reported less cognitive load than did students in the productive failure condition, but this gap became much smaller after the delayed posttest. A second mixed ANOVA was also performed to examine the effect of condition on cognitive load measurements after each of the two posttests, as cognitive load measurements were available for all three conditions. There was not a statistically significant interaction between condition and cognitive load reported after both posttests, $F(2, 197) = 2.667$, $p = .072$, $\eta^2 = .026$. Figure 5 shows the mean cognitive load reported in each condition after the initial activity and both posttests. Across the three activities, the order of conditions in terms of

cognitive load remained the same, although the load reported by PF condition participants dropped noticeably more quickly between the immediate and delayed posttest.

Figure 5. Cognitive load after initial activity and both posttests



An independent t-test and two one-way ANOVAs were performed to further test the simple effects of instructional condition on cognitive load at each report of cognitive load in the experiment. The independent t-test on cognitive load after the initial activity, which was used due to the lack of cognitive load data for the control condition, showed that WE students reported significantly less cognitive load than were PF students, $t(160) = -4.418$, $p < .001$. Table 5 shows the means and standard deviations for self-reported cognitive load at each of the three observations.

Table 5

Means and standard deviations of self-reported cognitive load after three activities

	Cognitive Load 1 (after initial activity)	Cognitive Load 2 (after immediate posttest)	Cognitive Load 3 (after delayed posttest)
Worked Example	3.83 (1.58)	4.19 (1.64)	4.09 (1.78)
Productive Failure	4.91 (1.53)	6.07 (1.67)	5.46 (1.74)
Control	n/a	5.26 (1.68)	5.07 (1.83)

A one-factor ANOVA was performed to test the effect of instructional condition after both the immediate and delayed posttests. After the immediate posttest, a statistically significant effect of instructional condition was found, $F(2, 236) = 24.312$, $p < .001$, $\eta^2 = .173$. Tukey post-hoc tests showed that WE participants reported significantly less cognitive load on the immediate posttest than did PF and control condition participants. In addition, participants in the control condition reported significantly less cognitive load than did PF condition participants. Thus, PF condition participants reported significantly higher levels of cognitive load than did participants in the other two conditions.

The one-factor ANOVA that was performed on cognitive load measurements after the delayed posttest showed a statistically significant effect of instructional condition on cognitive load, $F(2, 212) = 11.562$, $p < .001$, $\eta^2 = .098$. Tukey post-hoc tests showed that WE condition participants reported significantly less cognitive load on the delayed posttest than did PF and control condition participants. Participants in the PF and control conditions did not report significantly different levels of cognitive load. The differences between the two groups became smaller in magnitude, even though the order conditions in terms of increasing cognitive load remained the same.

Chapter 5

Study 2

Introduction

This study was designed to be similar to the main study, but take place embedded within an introductory statistics classroom, instead of sampling from a biology course. Embedding the experiment within an introductory statistics classroom was done to ground the study within typical instruction, as well as to more easily assess transfer of knowledge in a realistic learning environment. Embedding the study within a statistics class was also theorized to maximize participation in the study, as participation was designed not to take place outside of scheduled class meetings, and the content in the study was directly related to the course content. However, this theory did not work as well in practice, as recruitment and retention in the study proved to be far more difficult than expected. The sample size was such that a three condition design would have had such little statistical power that effects would have been nearly impossible to detect. Thus, a modified design was used in this study.

In addition, the research questions for this study were slightly different than those used in the main study. This is due to the fact that the delayed posttest was replaced with a transfer posttest on simple regression in this study, as effect of instructional condition on transfer was a research question of interest in the initial design of the study.

Research Questions and Hypotheses

Research Question #1: Are the effects of instructional condition on type of knowledge students learn uniform across correlation and simple regression (transfer) posttests?

Hypothesis: PF is theorized to be effective in large part because of prior knowledge activation, allowing for new content to be processed in ways that allow for more meaningful processing and encoding in long term memory. Thus, it is hypothesized that there is an interaction between the effect of instructional condition on procedural and conceptual knowledge and the type of test (immediate vs. transfer). Specifically, the effect of the productive failure prompt may be more noticeable on the transfer posttest, compared to the immediate posttest, as the productive failure activity should result in more acquisition of conceptual and conditional knowledge than do worked examples or an expository text, which do not promote prior knowledge activation and connection, and thus, may not promote acquisition of conceptual and conditional knowledge, as much. However, since PF is theorized to be more effective in conceptual knowledge acquisition, while WEs are theorized to be more effective in procedural knowledge acquisition, the effect might not be uniform across both types of knowledge, even on a transfer posttest.

Research Question #2a: Is there an interaction between instructional condition and the type of knowledge students learn?

Research Question #2b: What is the effect of instructional condition on student learning of conceptual and procedural knowledge?

Hypothesis: Because productive failure and worked examples focus on different types of knowledge, then the hypothesis is that there should be an interaction between type of instruction and acquisition of different types of knowledge. Specifically, PF conditions participants should outperform WE condition participants on measures of conceptual knowledge, while performing approximately the same on measures of procedural knowledge.

Research Question #3: Is there an interaction between instructional condition and performance on immediate and transfer posttests?

Hypothesis: PF is theorized to be effective due to the generative nature of the task, as well as the prior knowledge activation that the activity is designed to promote. These factors are related to more meaningful processing and encoding, which in turn may manifest in the form of better performance on a transfer posttest than the performance of participants receiving other instruction. Thus, it is hypothesized that there is an interaction between instructional condition and performance on immediate and transfer posttests. Specifically, PF condition participants are expected to perform better on the transfer posttest than on the immediate posttest, but this difference may not be evident (or may actually be reversed), on the immediate posttest.

Research Question #4a: Is there an interaction between instructional condition and the amount of mental effort students perceive they use?

Research Question #4b: What is the effect of instruction type on the amount of mental effort students perceive they use after different activities?

Hypothesis: PF condition participants will, on average, perceive they use more mental effort than WE condition participants, especially after the initial activity. Specifically, because PF requires a learner to create novel solutions to a problem that a learner has probably never encountered before, perceived mental effort may be especially high in this initial activity, before dropping later in instruction. While this may seem problematic, if PF participants are outperforming participants in the WE condition on measures of conceptual knowledge while performing equally as well on measures of procedural knowledge, then this increased mental effort may actually be advantageous for the learner, which is contrary to much of the worked example research and cognitive load theory, in general. However, the difficulties that participants encounter while completing the productive failure activity should be addressed by direct instruction, so the discrepancy in cognitive load after the initial activity should decrease on the posttests.

Method

Sample

The participants for this study were drawn from an undergraduate introductory statistics course at a large university in the northeastern United States. Participants were assigned to one of two conditions in this study. Given the nature of the course participants were drawn from, it was assumed that most of the students had little to no prior knowledge of the topic of correlation.

28 participants initially signed up for the study, although attrition between posttests lowered the sample size to 25 participants that completed all activities and measures. Of these 25 participants, 11 were assigned to the PF condition and 14 were assigned to the WE condition. The average age of the sample was 19.62 years of age. Nearly the entire sample (96%) was female. The sample was mostly White (88%), while 4% each was Black and Hispanic, and one participant declined to identify ethnicity. Average SAT Math score was 563, while the average ACT Math score was 26. Average GPA of the sample was 3.48.

Design of Study

The design of the study was similar to that used in the main study, but with a few key differences. One important difference was that participants were only assigned to one of two different groups (PF or WE). Based on pilot testing, only about 30-40 students were expected to participate in the study, which would have resulted in low statistical power for the proposed statistical analyses. Thus, the control condition was dropped from this study.

Participants were assigned to one of the two instructional conditions. Participants in the PF condition worked on an open-ended problem and were instructed to generate as many unique solutions to the problem as they could. This was followed by a period of direct instruction. Participants in the WE condition studied a worked example and then completed a practice

problem, which was then followed by direct instruction. For both conditions, an immediate posttest followed direct instruction. Three weeks later, participants in both conditions were administered a simple regression posttest, after instruction on simple linear regression in the course was completed. Cognitive load was measured after the instructional manipulation (worked example or productive failure prompt) and after both immediate and simple regression (transfer) posttests.

Materials

Pretest. The pretest was the same eight item pretest that was used in the main study.

Instructional materials. The instructional materials were the same productive failure prompt and worked example with practice problem combinations that were used in the main study.

Immediate posttest. The immediate correlation posttest was the same as the immediate posttest used in the main study.

Simple regression posttest. The simple regression posttest consisted of six open-response items; two procedural knowledge questions and four conceptual knowledge questions. The procedural knowledge questions required participants to compute the slope and intercept of regression equations, given means, standard deviations, and the correlation between two variables. The conceptual knowledge questions required students to interpret parts of regression equations, determine whether simple linear regression is appropriate for a scenario, and explain why one predictor would be a better choice for a regression model than another predictor. The simple regression posttest is found in Appendix E. There were two different forms of this measure.

Measurement of cognitive load. Cognitive load was measured in the same manner as it was measured in the main study. Students were asked to rate the amount of mental effort they used after the first task (worked example/productive failure), after the correlation posttest, and after the regression posttest. The same question adapted from Paas (1992) was used.

Procedures

The experimenter visited the class to recruit participants. Students were told that the experiment covered content from their course and that it required no extra time outside of regularly scheduled class sessions. Students that participated in the study received extra credit. Informed consent was administered through a course management system; participants had to give consent for their data to be used in this study, as in the main study. Demographic data were also collected by administering a survey via the same course management system.

All activities in this study were conducted in experimental sessions held in university classrooms. Participants signed up for these sessions using an online sign-up system. These sessions took place during regularly scheduled class meetings, so no extra time outside of class was required: these sessions contained the instructional manipulation, direct instruction, and immediate posttest. The regression posttest was administered later in the course, also during a regularly scheduled class meeting, when the instructor finished lecturing on simple regression.

Participants in both conditions completed the instructional manipulation first. Thus, PF condition participants received the open-response question and WE condition participants received the worked example and practice problem. Participants in both conditions were given 10 minutes to complete the corresponding activity.

After the instructional manipulation, all participants were then administered direct instruction. This instruction was performed by the usual instructor of the statistics class students

were sampled from, and covered the same content as was covered in the direct instruction from the main study. Because the two experimental sessions were offered during class but in different locations, a video of this instruction was made and shown to both conditions. The video was constructed to be as similar as possible to the direct instruction used in the main study in terms of content, explanation, and length. The video was approximately 18 minutes long, and was shown on a projection screen in the classrooms where data collection was being conducted. After the video instruction, students were given 30 minutes to complete the immediate posttest, which was the same as the immediate posttest that was used in the main study. In all, these three activities took approximately one hour for participants in both conditions to complete.

Three weeks later, when the instructor of the course completed her lectures on simple linear regression, participants in both conditions were administered the simple regression posttest. This posttest was also administered within a class session in one location, as it was not necessary to separate participants by condition for this assessment. Participants were given 25 minutes to complete the simple regression posttest.

Scoring of Posttests

Scoring of the immediate posttest was conducted using the same rubric as was used in the main study. Scoring of the regression posttest was performed in a similar manner. Each item was worth four points. For the procedural items, partial credit was given in a similar manner to the procedural correlation items. Use of the correct formula, but with errors in computation resulted in partial credit, with more errors resulting in fewer points awarded. For the two conceptual knowledge questions that required students to interpret regression coefficients, two points were given for identifying the correct part of the regression equation, and two points were given for the correct interpretation (i.e., "for every one point increase in X"…). For the question

that required participants to choose one of two possible predictors for a dependent variable, two points were given for choosing the correct variable, and two points were given for the explanation of why a variable was chosen. For the other conceptual question, which asked students whether a linear regression model would be appropriate based on a series of listed data points, two points were awarded for making the correct decision and two points were awarded for the explanation. Because each item was out of a total of four points, total scores on the posttest could range from 0-24

Interrater Agreement of Posttest Scores

All correlation and simple regression posttest responses were scored by two raters. These raters were the experimenter and a graduate student from the educational psychology department at the university where the study was conducted. Both raters worked together to create the rubrics for both posttests. Both raters rated all responses, and a random 20% of the responses were chosen to calculate agreement. Interrater agreement was calculated by dividing the number of scores the two raters disagreed on by the total number of scores. Interrater agreement was 91.7%; of the 8.3% of scores on which raters disagreed, nearly all of the disagreements were one point. Disagreements of more than one point where resolved by discussion between the raters.

Results

Differences in Prior Knowledge Across Conditions

The average score on the eight item multiple choice correlation pretest was 4.09 with a standard deviation of 1.84. The Cronbach's alpha for this test was .57, which is a low, but acceptable, level of reliability. There were no significant differences in mean prettest performance between the two groups, $t(30) = .546$, $p = .59$.

Assumptions/Descriptive Statistics

Mixed ANOVAs were performed to investigate each of the research questions. Assumptions of mixed ANOVAs include normality, independence of observations, homogeneity of variance and covariance, and sphericity. Shapiro-Wilk tests were performed to assess normality of each of the dependent variables. These tests showed that the assumption of normality was met for these data. The observations from different students met the independence of observations assumption. The homogeneity of variance and covariance were met, along with the sphericity assumption.

The average total score on the regression posttest was 7.15 with a standard deviation of 4.40. Cronbach's alpha for this posttest was .74, which is an acceptable value for reliability. A floor effect was evident, with an average score of just over seven points out of a possible 24. There was also a large amount of variation in posttest scores, as evidenced by the standard deviation of over 4.5 points.

Effect of Instruction on Learning Across Two Posttests

A 2 (instructional condition) x 2 (type of knowledge) x 2 (immediate and delayed posttest) mixed ANOVA was performed to test for a three way interaction between instructional condition, type of knowledge, and type of posttest, as described in research question #1. This three way interaction was not found to be statistically significant, $F(1, 23) = 0.65$, $p = .43$. Thus, the effect of instructional condition across type of knowledge and type of posttest was not statistically significant. Figures 6 and 7 plot the average performance of participants across groups and posttests on procedural and conceptual knowledge items, respectively. Table 6 gives the means and standard deviations of percentages of points obtained by participants across both instructional conditions on procedural and conceptual knowledge items at both posttests.

Figure 6. Average procedural knowledge performance on two posttests (in percentage of points earned)

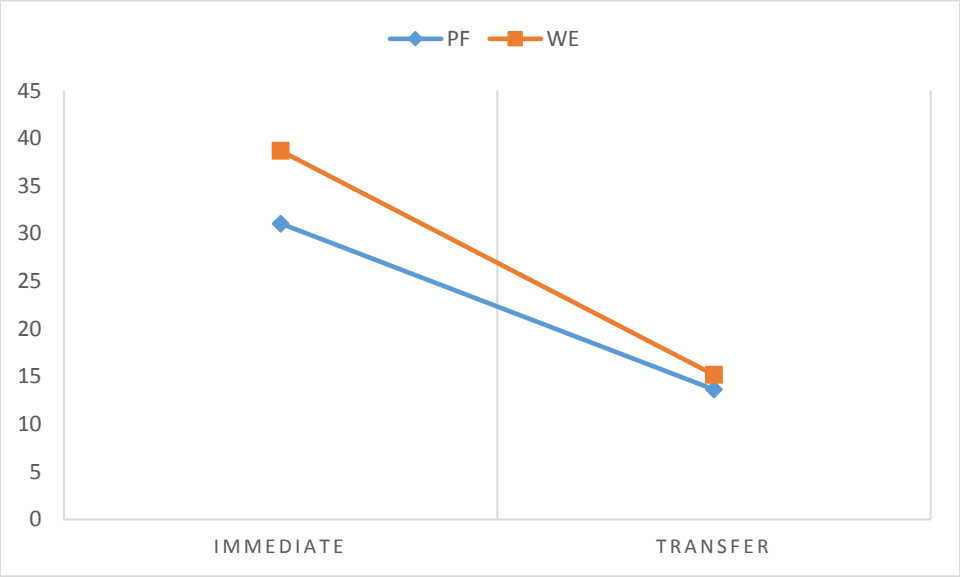


Figure 7. Average conceptual knowledge performance on two posttests (in percentage of points earned)

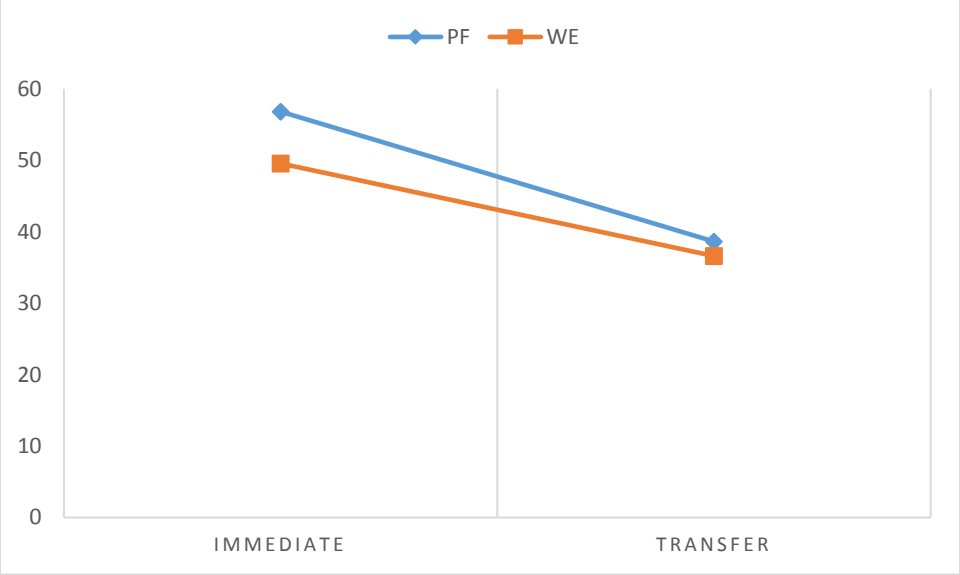


Table 6

Average procedural and conceptual knowledge scores, in percentages (SD): correlation posttest

	Procedural (Immediate)	Procedural (Transfer)	Conceptual (Immediate)	Conceptual (Transfer)
PF	31.06 (28.16)	13.63 (23.35)	56.81 (15.68)	38.63 (20.50)
WE	38.69 (27.88)	15.18 (26.48)	49.55 (14.18)	36.60 (18.68)

Effect of Instruction on Learning

The 2 x 2 x 2 mixed ANOVA did not find a statistically significant interaction between instructional condition and type of knowledge, F(1, 23) = 2.06, p = .17, $\eta^2 = .08$. Thus, the effect of instructional condition was not different across type of knowledge. Follow-up simple effects tests were not performed, as there was no interaction between instructional condition and type of knowledge. The main effect of instructional condition was also not significant, F(1, 23) = .0001, p = .99. Figure 8 plots average performance on procedural and conceptual knowledge problems by condition, collapsed across both posttests. Table 7 displays the means and standard deviations for procedural and conceptual knowledge scores by condition.

Figure 8. Performance on procedural and conceptual knowledge items by condition (in percentage of points earned).

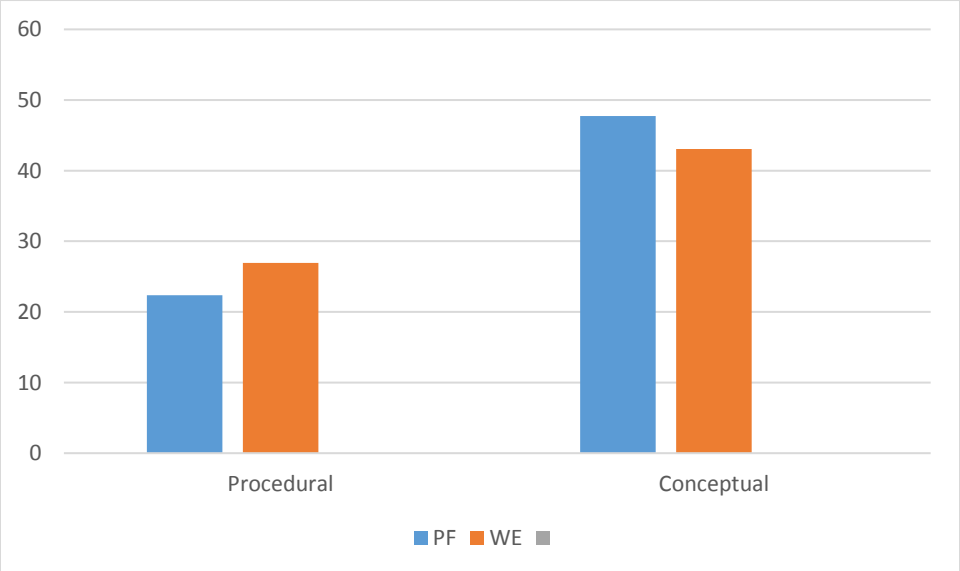


Table 7

Average percent scores (SD) for procedural and conceptual knowledge measures across posttests

	Mean procedural score (SD)	Mean conceptual score (SD)
PF	22.35 (20.09)	47.73 (15.25)
WE	26.93 (21.39)	43.08 (10.70)

Effect of Instruction on Learning: Immediate vs. Transfer Posttest

The 2 x 2 x 2 mixed ANOVA did not find a statistically significant interaction between instructional condition and type of posttest (immediate vs. transfer), $F(1, 23) = .002$, $p = .963$, $\eta^2 = .0001$. Thus, the effect of instructional condition was not different across immediate and delayed posttests. Figure 9 plots the average total procedural and conceptual knowledge scores of all three conditions. WE condition participants performed better on the immediate posttest, but PF condition participants performed better on the transfer posttest. However, these gaps were very small on both tests, even though the conditions switched order between posttests.

Figure 9. Performance by condition on both posttests (in percentage of points earned)

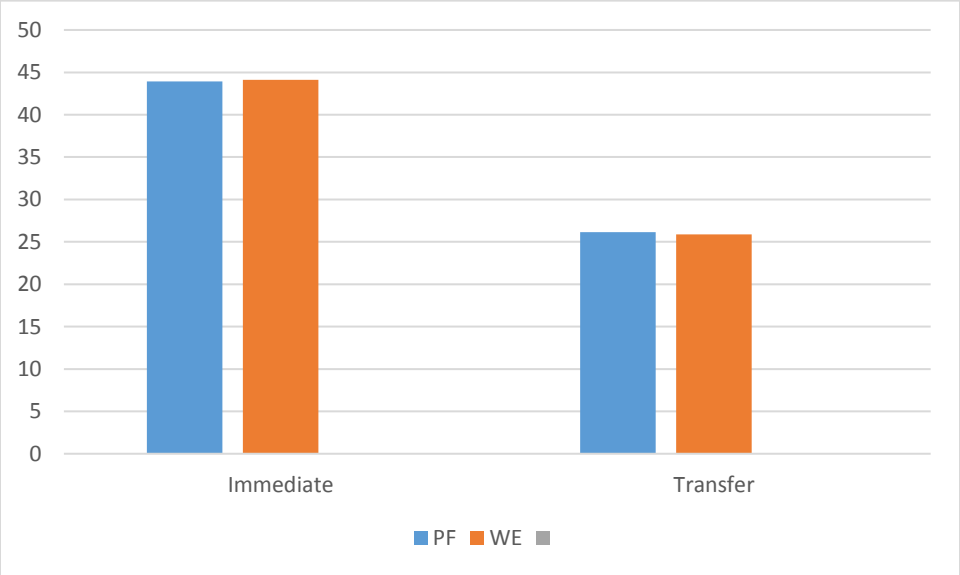


Table 8

Average percent scores (SD) for immediate and transfer posttests

	Immediate Posttest	Transfer Posttest
PF	43.94 (18.75)	26.14 (20.69)
WE	44.12 (17.82)	25.89 (19.28)

Effect of Instruction on Cognitive Load

A mixed ANOVA was performed to test for an interaction between cognitive load measurements and activity. This test did not find a statistically significant interaction between cognitive load and activity, F(2, 17) = .373, p = .694, $\eta^2 = .042$. Thus, instructional condition did not have significantly different effects on cognitive load across the three time points. Figure 10 plots the average reported cognitive load for both conditions at each of the three time points. Because no significant interaction between instructional condition and cognitive load over the three time points was found, no further simple effects tests were performed. Table 9 displays the means and standard deviations of cognitive load reports at each of the three time points in the study.

Figure 10. Cognitive load after three activities

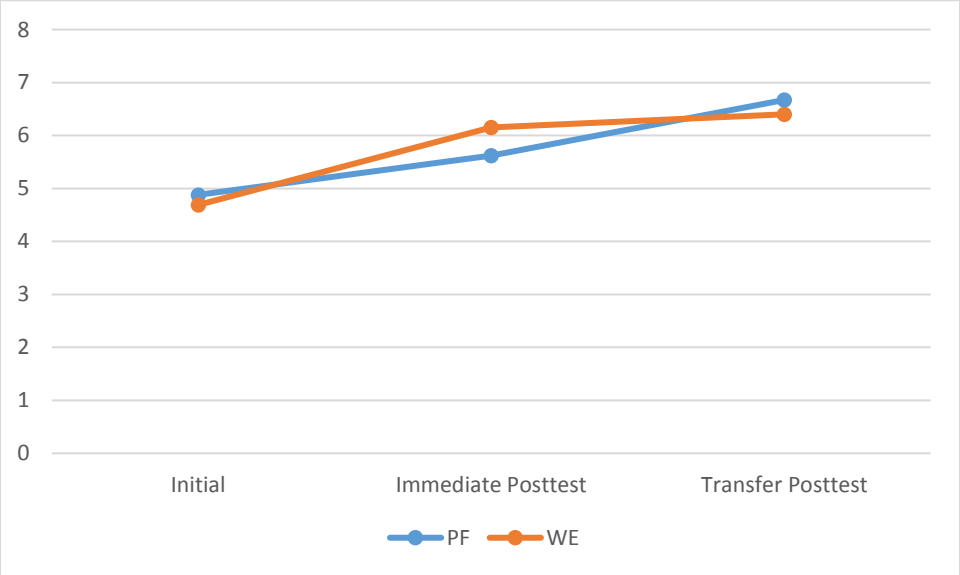


Table 9

Means and standard deviations of self-reported cognitive load after three activities

	Cognitive Load 1 (after initial activity)	Cognitive Load 2 (after immediate posttest)	Cognitive Load 3 (after transfer posttest)
Worked Example	4.88 (1.76)	5.62 (2.90)	6.67 (1.83)
Productive Failure	4.69 (1.46)	6.15 (1.723)	6.4 (1.63)

Chapter 6

Discussion

The purpose of this study was to examine the effects of two different instructional techniques used in basic statistics instruction on learning of procedural and conceptual knowledge related to the topic of correlation, on both immediate and delayed posttests. The hypotheses predicted an interaction between instruction type and knowledge; specifically, that productive failure would be more effective than worked examples in conceptual knowledge instruction, while not being different from worked examples in procedural knowledge instruction. The results of the study only partially supported these hypotheses. While interactions were found between instruction type and knowledge acquisition, the simple effects were not as hypothesized. Rather, the results showed that PF condition participants performed significantly lower on procedural knowledge questions than did participants in the other two conditions. PF condition participants did not show any differences on conceptual knowledge questions compared to participants in the other two conditions. These results were consistent across both the immediate and delayed posttests. Higher amounts of cognitive load were initially reported by PF condition participants, and this discrepancy in cognitive load between the PF condition and the other two conditions did not become smaller across subsequent activities. Further, the increased cognitive load reported by PF condition participants did not appear to result in more germane processing, as evidenced by the lack of significantly higher scores for these participants on any measures.

One important conclusion from this study is that WE condition participants performed significantly better on procedural knowledge items than did PF and control condition participants. This is a result that is consistent with much of the worked example literature, which

generally has found that worked examples are more effective than other types of instruction in promoting learning of procedural mathematical and statistical knowledge (i.e., Sweller & Cooper, 1985; Renkl, 1997). In addition, students in the worked example group reported significantly lower levels of cognitive load, after both instruction and immediate and delayed posttest, which is again consistent with much of the literature regarding worked examples and their effect on student-reported cognitive load (i.e., Gerjets et al., 2004; Paas, 1992). The positive effects of productive failure that have been reported in previous research (i.e., Kapur, 2012; 2014) were not replicated in this study. PF condition participants scored significantly lower than both WE and control condition participants on procedural knowledge measures. Productive failure was more effective with conceptual knowledge than with procedural knowledge, consistent with previous research, but productive failure did not show evidence of being more effective than worked examples on conceptual knowledge problems.

In addition, PF condition participants also reported significantly higher levels of cognitive load both after completing the open-ended generative question and after completing both posttests. While proponents of productive failure would argue this is not problematic because the increased cognitive load may be an indicator of activation and attendance to prior knowledge, the posttest results do not give evidence that this increased cognitive load was germane to learning. Rather, these results tend to give evidence more consistent with researchers such as Kirschner et al. (2004); in their experiment, unguided learning led to worse learning outcomes when compared to worked examples and simply reading a text to prepare for instruction. In terms of cognitive load theory, the productive failure prompt appeared to impose more extraneous than germane cognitive load on learners, as the PF prompt actually appeared to hinder learning of the procedure to calculate a correlation coefficient.

In contrast to the stark differences between conditions on procedural knowledge and cognitive load measures, there were no significant differences between any of the conditions on conceptual knowledge measures. While the order of conditions in terms of performance was still the same (WE, control, PF), the means and standard deviations of all three conditions were very similar. Thus, less definitive conclusions may be made regarding the effectiveness of these different types of activities on the learning of conceptual knowledge. While productive failure seemed to have a more positive effect on conceptual knowledge than on procedural knowledge, PF condition participants did not have significantly higher scores on conceptual knowledge measures than did WE and control condition participants. In fact, the PF condition participants still had the lowest average performance on conceptual knowledge measures, although this difference did not come close to approaching statistical significance.

Performance on the delayed posttest generally followed the same patterns as did performance on the immediate posttest; WE condition participants performed significantly better on procedural knowledge questions than did PF and control condition participants, but there were no significant differences on conceptual knowledge measures. Two notable patterns emerged from the data, though. One is that the differences between conditions on procedural knowledge became smaller on the delayed posttest. On the immediate posttest, WE condition participants scored significantly higher on procedural knowledge than the control and PF condition participants, but control group participants almost scored significantly higher than PF condition participants. On the delayed posttest, the gap between WE and the other two conditions shrank, even though the difference was still significant, and the difference between PF and the control condition nearly disappeared. For conceptual knowledge, while there were no significant differences between conditions on either posttest, PF condition participants actually performed

best, on average, on conceptual knowledge questions on the delayed posttest, while WE condition participants performed worst.

Overall, WE condition participants showed the largest decreases in performance between immediate and delayed posttest. PF condition participants, though, experienced almost no drop in scores, on average, from immediate to delayed posttest. Worked examples clearly appeared to be more effective than the other two conditions at helping students acquire procedural knowledge, even with the drop in performance from immediate to delayed posttest. However, the closing of the gap between PF and WE conditions on the posttest does give evidence of possible benefits of using PF instruction. Specifically, WE may be more effective for immediate recall, while PF's advantages may not be evident until delayed measures of recall are administered.

While there was clear evidence regarding procedural knowledge acquisition, the evidence regarding conceptual knowledge acquisition was much less clear. Worked examples, while being found to be effective at helping students learn procedural knowledge in statistics, do not have as clear of a relationship with the learning of conceptual knowledge in statistics. The evidence is much more tenuous of a positive relationship between use of worked examples and conceptual knowledge acquisition. The results of this study do not provide much evidence to make this relationship any clearer. While the WE condition did have the highest average conceptual knowledge score of the three conditions on the immediate posttest, this difference did not approach statistical significance. Further, the WE condition actually had the lowest average score on conceptual knowledge items on the delayed posttests, although the average score was not significantly different from the average scores of the other two conditions. Because participants in all three conditions performed approximately the same on conceptual knowledge

measures, it is difficult to draw many conclusions about the types of instructional activities or that may lead to increases in the learning of conceptual knowledge in statistics, although productive failure appeared to work slightly better on the delayed posttest.

One important consideration in interpreting these results is the content that was used for instruction. Worked examples have been used in the instruction of many different mathematical topics such as algebra (Cooper & Sweller, 1987), geometry, (Mousavi et al., 1995), and statistics (Catrambone & Holyoak, 1990; Paas, 1992). Productive failure has been used in the instruction of algebra (Kapur & Bielaczyc, 2012) and basic statistics (Kapur, 2014), but generally has not been used in as wide a variety of different mathematical and statistcal topics as have worked examples. The statistical concepts that have generally been used for either worked examples or productive failure have been probability (i.e., Paas, 1992; Renkl, 2002) and measures of central tendency, such as mean and variance (Kapur, 2014). Correlation was chosen as the basic statistical topic for instruction in this study because it had not been used before in previous worked example or productive failure research, and because correlation was a topic that could lead very logically to other topics that could require students to transfer their knowledge of correlation to solve other questions, like those related to simple linear regression.

The nature of the topic of correlation is important to consider because the procedural and conceptual knowledge associated with correlation is a bit different from that associated with other statistical topics on which worked examples and productive failure have been tested. One important difference is the structure of procedural knowledge, which mostly consists of how to calculate a correlation coefficient. The procedure to calculate a correlation is quite intensive, relatively speaking, especially when compared to computing probabilities, basic statistics and other procedural knowledge assessed in other studies. There are several different, yet equivalent,

computational procedures that can be used to compute a Pearson correlation coefficient, but even the simplest of these procedures requires several sets of computations that, while not necessarily being difficult, may be tedious, especially for novice learners. The method that was used here (and is outlined in the worked example) is a method that is commonly used is a very straightforward procedure of calculating a Pearson correlation. Still, though, this procedure is much more complex than most calculations students in introductory statistics classes would generally be accustomed to performing. The variance calculation is probably closest to the correlation calculation in terms of complexity and number of steps, but even the variance calculation is comparatively simpler (subtract the mean from each observation, square each difference, sum the squared differences and divide by sample size minus one). Because of the relative complexity of the correlation calculation procedure, the productive failure prompt may have been more ill-suited for this content than for instruction on other basic statistical and algebraic topics.

Limitations of the Study

There were several limitations to this study that may have affected the results that were found. One limitation is the population of interest and the population that was used in the study. Originally, the population of interest for this study was novice statistics students who were enrolled in an undergraduate statistics course at a large university. To obtain participants form this population, an undergraduate introductory statistics course offered by the College of Education at this university was to be sampled from. This class is offered twice a year and has an enrollment of approximately 80 students each semester. The original intent was to embed the study within the course of the class; the first round of data collection (introductory activity, direct instruction, and posttest) was to occur as students were beginning to learn about correlation in

class, and then a transfer posttest on simple linear regression was to be administered to students at the end of the regression unit in class. Thus, the design of this experiment would allow for the study of the effects of instruction type on the learning of procedural and conceptual knowledge not just on a chosen topic, in this case correlation, but also on simple linear regression, which usually follows correlation in the course of instruction. In other words, this study design would have allowed for the examination of the effect of these instructional activities on a course of instruction, and not just one specific topic in isolation, in the classroom environment. However, sampling did not go according to plan, and in both pilot testing and this study, only very small sample sizes were obtained. Specifically, 48 students were sampled the first time, but a significant portion of these students never took the regression posttest, so there was only an effective sample size of 22 students. Similarly, I was only able to obtain a sample of 28 students from the second round of data collection, and of these 28, only 25 completed the regression posttest. This made the administration of three conditions untenable, and even with two conditions and no control group, I had very little statistical power. This led to a suggestion to try sampling from a different class, which worked very well and yielded a comparatively large sample size, but also raised other potential issues.

The new class that was sampled from was an intermediate biology class. As such, the population shifted slightly from only novice statistics learners to a mix of novice and intermediate statistics learners. While there was still large variation in terms of topic-specific prior knowledge, in general, this sample had higher levels of prior knowledge than did the previous samples drawn from the introductory statistics class. This change in prior knowledge may have affected how worked examples and productive failure prepared students for instruction; both of these activities have generally been used to prepare novice learners to learn

new material. Since 59.4% of the participants in the study had already taken an introductory statistics course, then they would not be "new" to the content. In theory, this could have led to the expertise reversal effect with the worked example condition (Kalyuga et al., 2003), even though the results showed WE condition participants performed significantly better on procedural tasks. The expertise reversal effect was unlikely, though, because even though some students rated themselves as "experts" on correlation and performed well on the prior knowledge pretest, they clearly did not have expert-level prior knowledge of correlation. Similarly, students who had sufficient prior knowledge may not engage with the productive failure prompt either, as they may not feel the need to generate solutions to a problem they already know how to solve (almost an expertise reversal effect for productive failure). Again, though, this effect was not evident in the present study, as only a negligible percentage of PF condition participants did not attempt to generate any solutions.

Another limitation of using a different sample was that the biology class did not allow for the integration of the experiment within the class. Thus, while most of the original research questions could still be addressed, the question of how the use of the different instruction types would affect learning of other topics later in a course could not be addressed. This was replaced by a question addressing how these instructional activities affected performance on procedural and conceptual knowledge problems as measured by a delayed posttest. While this is a still an important question that addresses an issue relevant gaps in both PF and WE research, it is slightly different from the original focus of the study.

When comparing the main study with the smaller study performed here, there are few major differences in the results of the studies. WE condition participants performed better on procedural knowledge items in both studies, although there was no statistical significance in the

smaller study. There were no significant differences in performance on conceptual questions, although PF condition participants in the smaller study had a higher average score on conceptual knowledge items than did WE condition participants (no statistical significance due to low power). There were also no significant differences in cognitive load in the smaller study, but PF condition participants generally used higher amounts of cognitive load and WE condition participants generally used lower amounts of cognitive load in the main study. Essentially, the lack of statistical power in the small study resulted in most of the differences in results between studies, with the possible exception of PF condition participants' performance on conceptual questions. With more power, this test in the smaller study may have indicated that productive failure students scored significantly higher on conceptual knowledge items. This finding could provide tentative support for the theory that productive failure can be effective when integrated into classroom instruction, especially for conceptual knowledge acquisition.

The instructional time allocated to instruction was an additional limitation of this study. This was a limitation regardless of which sample was used. The original intent of using an introductory statistics class was motivated in part by the opportunity to include more instruction to students in addition to the instruction administered in the study. However, by using a class session, time constraints limited the amount of time that could be dedicated to direct instruction in the study. Specifically, with the pretest taking 5 minutes, the PF/WE/control activity taking about 10 minutes, and the posttest taking about 25 minutes on average, that only left about 20 minutes for direct instruction out of the hour-long experimental sessions that were scheduled. 20 minutes is a very short amount of time to teach concepts related to correlation, as well as the procedure for calculating a correlation, so this short amount of instructional time may have weakened the effects of the different conditions, especially productive failure, which requires a

thorough period of direct instruction following the activity to maximize effectiveness. Even in the introductory statistics class, which ran in a 75 minute time period, there was still only about 25-30 minutes for direct instruction, which is an improvement, but still a short time period for teaching a topic like correlation effectively. For example, Kapur (2014) used two hours for the productive failure activity and direct instruction on measures of central tendency. A longer period of instructional time would have been desirable, but was unfortunately not feasible for this study.

Implications for Future Research

Research that examines how productive failure affects student performance on transfer tasks would be an important extension of this research. The original study was designed to study performance on transfer tasks in statistics, but because of the small sample size obtained, these analyses did not have enough statistical power, resulting in the main study that used immediate and delayed posttests on the same topic instead, due to logistical constraints. There has been plenty of research on worked examples that show varying degrees of effectiveness in terms of their effect on student performance on transfer tasks, but very little of this type of research has been performed with productive failure. Kapur (2014) found that students who completed productive failure activities performed better on transfer tasks than did students who were only administered practice and direct instruction. However, Kapur's transfer tasks consisted of only three multiple choice items that required students to apply the procedures and concepts related to variance and standard deviation to the concept of normalization. Similarly, in an earlier (2012) study, Kapur used one procedural item to assess the transfer of knowledge from standard deviations to normalization and comparisons of different distributions. Three multiple choice items out of 19 items and one item out of six open-response items may not give incontrovertible

evidence that productive failure is effective at helping students transfer knowledge to novel tasks. Further research that examines the effect of productive failure on transfer tasks in statistics, especially in comparison to other activities that are commonly used in statistics instruction, would help give a clearer picture of the effect of productive failure to foster students' transfer of knowledge to novel problems.

The type of content that is used for productive failure instruction is another factor that could influence future research. Previous research has generally only looked at PF's effects on basic statistics and algebra concepts. The present study examined the effect of productive failure instruction on acquisition of knowledge related to correlation, and the results diverged from the results of Kapur and others in their research on productive failure. The nature of knowledge related to correlation, especially procedural knowledge, may be more complex than the types of procedural knowledge used in other productive failure studies; Kapur (2014) acknowledged a ceiling effect on procedural knowledge questions due to the relatively straightforward nature of the calculations. The results of this study did not show the same positive effects, so there may be an interaction with the type of mathematical and statistical knowledge being taught and the effectiveness of productive failure. Future research should examine different topics in statistics to examine which types of knowledge productive failure works best for in terms of instruction, i.e, whether the instructional-fit hypothesis (Nokes, Hausmann, Van Lehn, & Gershman, 2011) applies to PF and WE usage in statistics instruction.

In general, additional research should be performed to examine how and under what types of conditions and for which types of knowledge productive failure instruction works. The results of the present study are not meant to suggest that productive failure cannot be an effective organizing introductory instructional activity for students. Rather, these results suggest

ambiguity as to how and when productive failure may be used effectively, and this speaks in part to the comparatively small amount of literature examining productive failure use. By contrast, there is a large volume of research that shows the types of knowledge that worked examples may be effective in teaching, as well as theories explaining how worked examples are effective. More research on productive failure should be done to examine when and how it can be most effectively incorporated into statistics instruction.

References

Aleven, V.A., & Koedinger, K.R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26,* 147-179. doi: 10.1207/s15516709cog2602_1

Alexander, P.A., & Judy, J.E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research, 58,* 375-404. doi: 10.3102/00346543058004375

Alexander, P.A., Kulikowich, J.M., & Schulze, S.K. (1994). How subject-matter knowledge affects recall and interest. *American Educational Research Journal, 31,* 313-337. doi: 10.3102/00028312031002313

Alexander, P.A., Schallert, D.L., & Hare, V.C. (1991). Coming to terms: how researchers in learning and literacy talk about knowledge. *Review of Educational Research, 61,* 315-343. doi: 10.3102/00346543061003315

Anderson, J.R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press

Atkinson, R.K., Derry, S.J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70,* 181-214. doi: 10.3102/00346543070002181

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

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R.E.

Mayer (Ed.) *Cambridge Handbook of Multimedia Learning.* Cambridge, England: Cambridge Academic Press.'

Baxter, G.P., Elder, A.D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31,* 133-140. doi: 10.1207/s15326985ep3102_5

Bobis, J., Sweller, J., & Cooper. M. (1994). Demands imposed on primary-school students by geometric models. *Contemporary Educational Psychology, 19,* 108-117. doi: 10.1006/ceps.1994.1010

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24,* 61-100. doi:10.3102/0091732X024001061

Broers, N.J. (2009). Using propositions for the assessment of structural knowledge. *Journal of Statistics Education, 17,* 19-38.

Brunstein, A., Betts, S., & Anderson, J.R. (2009). Practice enables successful learning under minimal. *Journal of Educational Psychology, 101,* 790-802. doi: 10.1037/a0016656

Catrambone, R., & Holyoak, K.J. (1990). Learning and subgoals and methods for solving probability problems. *Memory & Cognition, 18,* 593-603. doi: 10.3758/BF03197102

Chance, B. L., & Garfield, J. B. (2002). New approaches to gathering data on student learning for research in statistics education. *Statistics Education Research Journal, 1,* 38-41.

Chervany, N.L., Collier, R.D., Fienberg, S., & Johnson, P. (1977). A framework for the development of measurement instruments for evaluating the introductory statistics course. *American Statistician, 31,* 17-23. doi: 10.1080/00031305.1977.10479186

Chi, M.T. (1996). Constructing self-explanations and scaffolded explanations in tutoring.

Applied Cognitive Psychology, 10, S33-S49. doi: 10.1002/(SICI)1099-0720(199611)10:7%3C33::AID-ACP436%3E3.3.CO;2-5

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

Chi, M.T., DeLeeuw, N., Chiu, M.H., & LaVancher, C.(1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439-477. doi:10.1207/s15516709cog1302_1

Chi, M.T., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152. doi: 10.1207/s15516709cog0502_2

Chi, M.T., Glaser, R., & Far, M.J. (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum

Chi, M.T., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.) *Advances in the psychology of human intelligence.* (pp. 7-75). Hillsdale, NJ: Erlbaum

Cohen, B.H. (2013). *Explaining Psychological Statistics*. Hoboken, NJ: John Wiley & Sons.

Cooper, G.A. & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem solving transfer. *Journal of Educational Psychology, 79,* 347-362. doi: 10.1037/0022-0663.79.4.347

Decaro, M.S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of Experimental Child Psychology, 113,* 552-568. doi: 10.1016/j.jecp.2012.06.009

diSessa, A.A., & Sherin, B.L. (2000). Meta-representation: An introduction. *Journal of Mathematical Behavior, 19,* 385-398. doi: 10.1016/S0732-3123(01)00051-7

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*, 44-63. doi: 10.2307/749110

Gerjets, P., Scheiter, K., & Catrambone, R. (2004). Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures. *Instructional Science, 32(1-2),* 33-58. doi: 0.1023/B:TRUC.0000021809.10236.71

Gerjets, P., Scheiter, K., & Catrambone, R. (2006). Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learning and Instruction, 16,* 104-121. doi:10.1016/j.learninstruc.2006.02.007

Greeno, J.G., Smith, D.L., & Moore, J.L. (1993). Transfer of situated learning. In D.K Detterman & R.J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction.* (pp. 99-167). Norwood, NJ: Ablex.

Grosse, C.S., & Renkl, A. (2007). Finding and fixing errors in worked examples: can this foster Learning outcomes? *Learning and Instruction, 17,* 612-634.

Jeung ,H., Chandler, P., & Sweller, H. (1997). The role of visual indicators in dual sensory Mode instruction. *Educational Psychology, 17,* 329-433. doi:10.1080/0144341970170307

Jonassen, D.H., (2000). Towards a design theory of problem solving. *Educational Technology, Research, & Development, 48,* 63-85. doi: 10.1007/BF02300500

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review, 23,* 1-19. doi: 10.1007/s10648-010-9150-7

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). Expertise reversal effect. *Educational*

Psychologist, 38, 23-31. doi: 10.1207/S15326985EP3801_4

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors, 40,* 1-17. doi: 10.1518/001872098779480587

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26,* 379-424. doi: 10.1080/07370000802212669

Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science, 38,* 523-550. doi: 10.1007/s11251-009-9093-x

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40,* 651-672. doi: 10.1007/s11251-012-9209-6

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38,* 1008-1022. doi: 10.1111/cogs.12107

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *The Journal of the Learning Sciences, 21,* 45-83. doi: 10.1080/10508406.2011.591717

Kempthorne, O. (1980). The teaching of statistics: Content versus form. *American Statistician, 34,* 17–21. doi: 10.1080/00031305.1980.10482704

Kirschner, P.A., Sweller, J., & Clark, R.E. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist, 41,* 75-86. doi: 10.1207/s15326985ep4102_1

Lavigne, N.C., Salkind, S.J., & Yan, J. (2008). Exploring college students' mental representations of inferential statistics. *Journal of Mathematical Behavior, 27,* 11-32. doi: 10.1016/j.jmathb.2007.10.003

Lovett, M.C., & Greenhouse, J.B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician, 54,* 196-206. doi: 10.1080/00031305.2000.10474545

Mayer, R.E., Sims, V., & Tajika, H. (1995). A comparison of how textbooks teach mathematical

problem solving in Japan and the United States. *American Educational Research Journal, 40,* 257-265. doi: 10.2307/1163438

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychology Review, 63,* 81-97. doi: 10.1037/h0043158

Mousavi, S.Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and oral presentation modes. *Journal of Educational Psychology, 87,* 319-334. doi: 10.1037/0022-0663.87.2.319

Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: the effect of example format and generating self-explanations. *Cognition & Instruction, 16,* 173-199. doi: 10.1207/s1532690xci1602_2

Nokes, T. J., Hausmann, R. G., M., Vanlehn, K., & Gershman, S. (2011). Testing the instructional fit hypothesis: The case of self-explanation prompts. *Instructional Science, 39,* 645-666. doi: 10.1007/s11251-010-9151-4

Paas, F.G.W.C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive load approach. *Journal of Educational Psychology, 84,* 429-434. doi: 10.1037/0022-0663.84.4.429

Paas F.G.W.C., & van Merrienboer, J.J.G. (1994). Variability of worked examples and transfer of geometrical problem solving skills: a cognitive load approach. *Journal of Educational Psychology, 86,* 122-133. doi: 10.1037/0022-0663.86.1.122

Quilici, J.W., & Mayer, R.E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88,* 144-161. doi: 10.1037/0022-0663.88.1.144

Quilici, J.W., & Mayer, R.E. (2002). Teaching students to recognize structural similarities

between statistics word problems. *Applied Cognitive Psychology, 16, 325-342.* doi: 10.1002/acp.796

Reed, S.K., & Bolstad, C.A. (1991). Use of examples and procedures in problem solving. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17,* 753-766.

Reiser, B.J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences, 13,* 423-451. doi: 10.1207/s15327809jls1303_2

Renkl, A. (1997). Learning from worked-out examples: a study on individual differences. *Cognitive Science, 21,* 1-29. doi: 10.1207/s15516709cog2101_1

Renkl, A. (2002). Worked-out examples: instructional explanations support learning by self-explanations. *Learning and Instruction, 12,* 529-556. doi:10.1016/S0959-4752(01)00030-5

Renkl, A. (2005). The worked-out example principle in multimedia learning. In R.E. Mayer (Ed.) *Cambridge Handbook of Multimedia Learning.* Cambridge, England: Cambridge Academic Press.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skills acquisition: A cognitive load perspective. *Educational Psychologist, 38,* 15–22. doi: 10.1207/S15326985EP3801_3

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23,* 90-108. doi: 10.1006/ceps.1997.0959

Roelle, J., Lehmkuhl, N., Beyer, M., & Berthold, K. (2015). The role of specificity, targeted learning activities, and prior knowledge for the effects of relevance instructions. *Journal of Educational Psychology, 107*, 705-723. doi: 10.1037/edu0000010

Ross, B.H. (1989). Remindings in learning and instruction. In S. Vosnaiadou & A. Rotony (Eds.), *Similarity and Analogical Reasoning* (pp. 438-469). Cambridge, MA: Cambridge University Press.

Rourke, A., & Sweller, J. (2009). The worked-example effect using ill-defined problems: learning to recognize designers' styles. *Learning and Instruction, 19,* 185-199. doi: 10.1016/j.learninstruc.2008.03.006

Schmeck, A., Opfermann, M., van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: differences between immediate and delayed ratings. *Instruction Science: An International Journal of the Learning Sciences, 43*, 93-114. doi: 10.1007/s11251-014-9328-3

Schwartz, D.L, & Bransford, J.D. (1998). A time for telling. *Cognition and Instruction, 16,* 475-522. doi: 10.1207/s1532690xci1604_4

Schwartz, D.L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficacy of encouraging original student production in statistics instruction. *Cognition and Instruction, 22,* 129-184. doi: 10.1207/s1532690xci2202_1

Silver, E.A. (1979). Student perceptions of relatedness among mathematical verbal problems. *Journal for Research in Mathematics Education, 10,* 195-210. doi: 10.2307/748807

Simon, H.A., & Chase, W.G., (1973). Skill in chess. *American Scientist, 61,* 394-403.

Sweller, J., (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science, 12,* 257-285. doi: 10.1207/s15516709cog1202_4

Sweller, J., & Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59-89. doi:10.1207/s1532690xci0201_3

Sweller, J., van Merrienboer, J.J.G., & Paas, F.G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251-296. doi:10.1023/A:1022193728205

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Tarmizi, R.A., & Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology, 80,* 424-436. doi: 10.1037/0022-0663.80.4.424

Van Lehn, K. (1999). Rule learning events in the acquisition of a complex skill: an evalution of cascade. *The Journal of the Learning Sciences, 8,* 71-125. doi: 10.1207/s15327809jls0801_3

Van Lehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggestt, W.B., (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21,* 209-249. doi: 10.1207/S1532690XCI2103_01

Van Merrienboer, J.J.G. (1990). Strategies for programming instruction in high school: program completion vs. program generation. *Journal of Educational Computing Research, 6,* 265-287. doi: 10.2190/4NK5-17L7-TWQV-1EHL

van Merrienboer, J.J.G. (1997). *Training complex cognitive skills: A four-component instructional design model for technical training.* Englewood Cliffs, NJ: Educational Technology Publications

Ward M.., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and*

Instruction, 7, 1-39. doi: 10.1207/s1532690xci0701_1

Yan, J., & Lavigne, N.C. (2014). Promoting college students' problem understanding using schema-emphasizing worked examples. *The Journal of Experimental Education, 82*, 74-102. doi: 10.1080/00220973.2012.745466

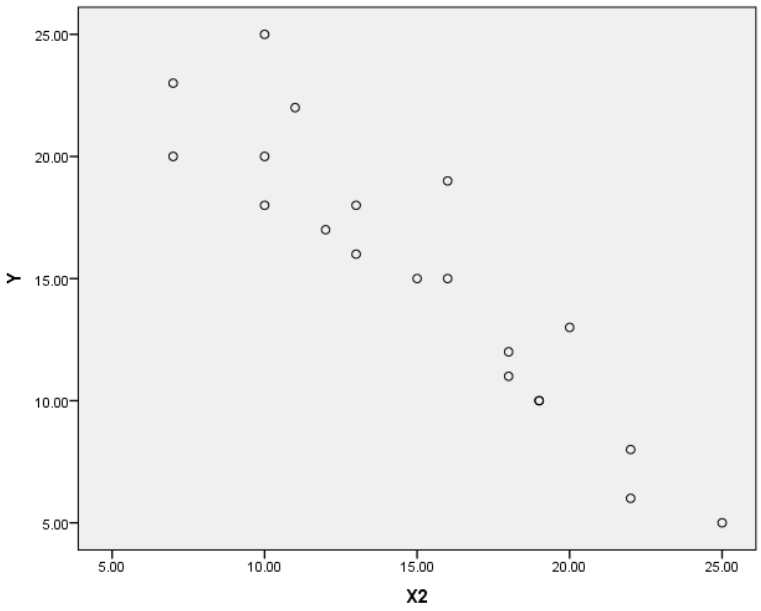
Zhu, X., & Simon, H.A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*, 137-166. doi: 10.1207/s1532690xci0403_1

Appendix A: Correlation Pretest

#1. Which of the following choices correctly orders the correlations from *smallest* to *largest* in terms of **magnitude?**

a. -.75, -.10, +.35, +.65
b. +.65, +.35, -.10, -.75
c. -.75, +.65, +.35, -.10
d. -.10, +.35, +.65, -.75

#2. What is most likely the observed correlation between the two variables shown in the scatterplot below?



a. -0.75
b. -0.25
c. 0.25
d. 0.75

#3. If there is NO relationship between two variables, what would the observed correlation between these two variables be?

a. -1
b. -0.5
c. 0
d. 0.5
e. 1

#4. If there is a strong, negative correlation observed between two variables X and Y, which of the following statements is true?

a. All values of X must positive, while all values of Y must be negative
b. Larger values of X are associated with larger values of Y
c. There is no association between values of X and values of Y
d. Larger values of X are associated with smaller values of Y

#5. Which of the following values could NOT be a correlation coefficient?

a. -1.0
b. 0
c. .57
d. 1.3

#6. A **correlation** between two variables refers to

a. the difference between the ranges of the two variables
b. the relationship between two variables
c. how changes in one variable cause a change in the other variable
d. the difference between the means of the two variables

#7. **Covariance** may be described as

a. how two variables change together
b. the combined range of two variables
c. how similar the means of two variables are
d. the combining of two variables

#8. Chelsea, who works at a car dealership, was asked by her boss, Peggy, to determine the relationship between number of cars sold in a week, the amount of money spent on advertising, and the average outdoor temperature. Would Chelsea be able to compute a correlation that describes the relationship between these three variables?

a. Yes, because a correlation can measure the relationship between any number of variables
b. Yes, because all of the variables are continuous
c. No, because the variables are on different scales of measurement
d. No, because a correlation can only measure a relationship between two variables.

Appendix B: Prompt for Productive Failure Condition

In soccer, elite goal scorers are valued very highly, often commanding some of the top salaries in the sport. Mike, a general manager of a soccer club, is considering paying a lot of money for one of these players. However, he wants to determine how strong the relationship is between having a player who scores a lot of goals and the number of games a team wins. Mike has obtained the following data:

Player	Goals	Team Wins
Chris	10	16
Leo	12	18
Jamie	9	15
Luis	7	14
Wayne	12	17

Help Mike by developing as many methods as you can of measuring the relationship between the number of goals the leading scorer scores and the number of wins for that player's team. Please show all formulas and calculations on the paper

Appendix C: Worked Example and Practice Problem

In soccer, elite goal scorers are valued very highly, often commanding some of the top salaries in the sport. Mike, a general manager of a soccer club, is considering paying a lot of money for one of these players. However, he wants to determine how strong the relationship is between having a player who scores a lot of goals and the number of games a team wins. Mike has obtained the following data:

Player	Goals (*X*)	Team Wins (*Y*)
Chris	10	16
Leo	12	18
Jamie	9	15
Luis	7	14
Wayne	12	17

Step 1: Find the values of X^2, Y^2, and XY for all observations:

Player	Goals (*X*)	Team Wins (*Y*)	**X^2**	**Y^2**	**XY**
Chris	10	16	**100**	**256**	**160**
Leo	12	18	**144**	**324**	**216**
Jamie	9	15	**81**	**225**	**135**
Luis	7	14	**49**	**196**	**98**
Wayne	12	17	**144**	**289**	**204**

Step 2: Compute the sums of X, Y, X^2, Y^2, and XY:

Player	Goals (*X*)	Team Wins (*Y*)	X^2	Y^2	XY
Chris	10	16	100	256	160
Leo	12	18	144	324	216
Jamie	9	15	81	225	135
Luis	7	14	49	196	98
Wayne	12	17	144	289	204
Sum	$\sum \boldsymbol{X = 50}$	$\sum \boldsymbol{Y = 80}$	$\sum \boldsymbol{X^2 = 518}$	$\sum \boldsymbol{Y^2 = 1290}$	$\sum \boldsymbol{XY = 813}$

Remember, **Σ** means "the sum of all…". For example, $\sum X$ denotes the sum of all Xs, which is 10 + 12 + 9 + 7 + 12 = 50.

Step 3: Compute SS_X and SS_Y:

$$SS_X = \sum X^2 - \frac{(\sum X)^2}{n} = 518 - \frac{50^2}{5} = 518 - \frac{2500}{5} = 518 - 500 = \mathbf{18}$$

$$SS_Y = \sum Y^2 - \frac{(\sum Y)^2}{n} = 1290 - \frac{80^2}{5} = 1290 - \frac{6400}{5} = 1290 - 1280 = \mathbf{10}$$

Step 4: Compute the sum of cross products ($\sum(X - \bar{X})(Y - \bar{Y})$):

$$\sum(X - \bar{X})(Y - \bar{Y}) = \sum XY - \left[\frac{(\sum X)(\sum Y)}{n}\right] = 813 - \frac{(50)(80)}{5} = 813 - 800 = \mathbf{13}$$

Step 5: Substitute the values from Steps 3 and 4 into the computational equation for a correlation coefficient. The result will be the correlation between X and Y:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{(SS_X)(SS_Y)}} = \frac{13}{\sqrt{(18)(10)}} = \frac{13}{\sqrt{180}} = \frac{13}{13.416} \approx \mathbf{.969}$$

The observed correlation between X and Y is **+.969**.

Now try a problem on your own! Feel free to refer back to the previous example at any time while working on the following question.

Lana is a 5th grade teacher who has administered two exams to her class: a reading comprehension exam and a spelling exam. Here are the scores on those two exams from 5 of her students:

Comprehension Score	Spelling Score
8	9
6	7
7	5
5	6
4	8

What is the correlation between comprehension scores and spelling scores? Be sure to show all of your steps!

Appendix D: Correlation Posttest

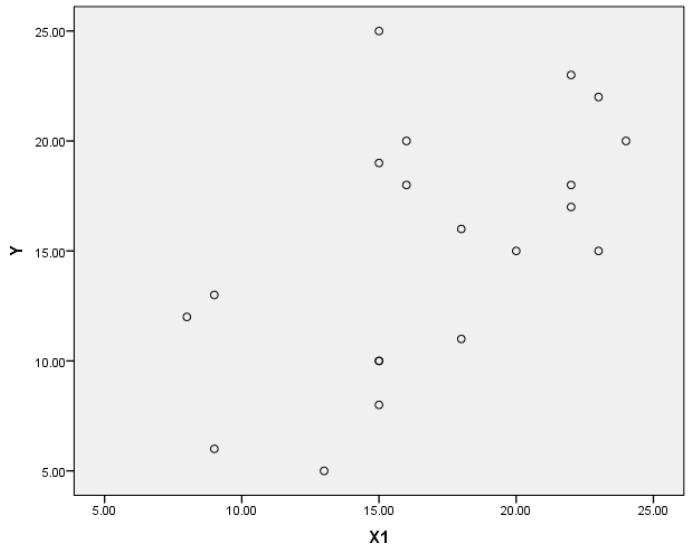
#1. Consider the following scenario:
Matt is a fitness instructor, and he is looking for variables that may be associated with the time it takes individuals to run one mile. He hypothesizes that hours of sleep per night may be positively related with the time it takes individuals to run a mile, as people who are more rested may be more prepared to run the following day. He collected data from 8 people. These data are displayed in the table below:

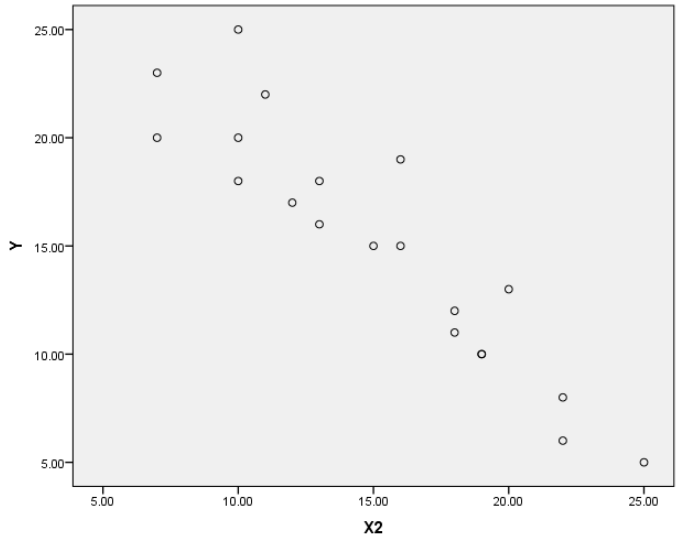
Time to run a mile (in minutes) (X)	Average hours of sleep per night (in hours) (Y)
5.5	8
4.8	7
6.2	8
5.2	6
7.3	8
7.5	6
6.5	7
5.0	6

What is the observed correlation between time it takes a person to run a mile and the average hours of sleep a person gets per night? Be sure to show all of the steps and calculations that you use.

#2. Mallory is looking to buy a new car. She drives a lot, but is looking for a car that is inexpensive to run. Fuel efficiency is one of the factors that she is taking into account when choosing a new car. Do you think that Mallory should buy a car with *higher* or *lower* fuel efficiency? Be sure to consider the relationship between operating expenses and fuel efficiency when answering this question.

#3. Consider the following scatterplots of Y with X1 and Y with X2, respectively:





Which scatterplot indicates a *stronger* correlation? Why?

#4. Consider the following pairs of scores on two different tests:

	Reading Comprehension (X)	Vocabulary (Y)
Becky	8	9
Tom	7	9
Carol	6	8
Pam	8	7
Trudy	5	7

Compute the correlation between reading comprehension scores and vocabulary scores. Please be sure to show all steps!

#5. Lindsay operates a large ski resort in New England. Because her business is very dependent on weather, she pays close attention to changes in weather predictions. Do you think that Lindsay is hoping for *larger* or *smaller* amounts of snow this winter? In your justification, please be sure to describe what you believe the relationship is between amount of snowfall and ski resort business.

#6. Maria works at a car dealership. Her boss believes that there is a correlation between the number of sick days an employee takes and the number of cars the employee sells. Maria collected data and found a correlation of -.65 between these two variables. How could Maria describe the meaning of this correlation to her boss?

#7. The steps for computing a correlation coefficient (along with one extra step) are listed below. Please correctly choose the four steps for computing a correlation coefficient and list them in the correct order.

Compute the sums of X, Y, X^2, Y^2, and XY
Calculate the range of both X and Y
Compute X^2, Y^2, and XY for all observations
Divide the sum of the cross products by the square root of SS_X and SS_Y
Compute SS_X, SS_Y, and the sum of cross products

Appendix E: Regression Posttest

#1. Mike wants to construct a simple regression equation to predict the number of goals a hockey player scores over the course of a season by using the average distance a hockey player skates during a game. He speculates that hockey players that skate longer distances tend to score more goals. The average number of goals scored is 16 with a standard deviation of 3.5, and the average distance skated during a game is 4.5 miles with a standard deviation of 1.2 miles. The correlation between average distance skated and goals scored is + 0.4.

a. Calculate the slope for the regression equation.
b. Calculate the intercept.
c. Write the regression equation.

#2. Jamie ran a regression analysis that used minutes of exercise per day to predict the number of push-ups an individual could perform in two minutes. She obtained the following regression equation:

$$\hat{y} = 20 + .95x$$

a. How should Jamie describe the effect of exercise on the predicted number of push-ups performed?

b. Can Jamie make any inferences about how well this regression line fits the data she collected? Why or why not?

#3. Consider the following data on the following variables: age and distance a person can run.

Age (in years)	Distance (in miles)
10	1.5
20	4.8
30	6.1
40	6.5
50	5.6
60	4.7
70	2.3

Based on the relationship between age and distance a person can run, would it be appropriate to use age to predict distance a person can run using simple linear regression? Why or why not?

#4. Marta would like to use student motivation data to predict reading test scores (maximum 10 points). The average student motivation score was 3.5 with a standard deviation of 0.7, and the average reading test score was 8.3, with a standard deviation of 1.2. The correlation between motivation scores and reading test scores is +0.55.

a. Calculate the slope for the regression equation.
b. Calculate the intercept
c. Write the regression equation.

#5. Max would like to use a simple regression equation to predict the number of scarves he can sell. Max has the following information: the price of each of the scarves that he sells, and the number of pictures of each type of scarf he has posted on his Instagram page that he uses for advertising. He found that the correlation between scarves sold and price was -.55, and the correlation between scarves sold and Instagram pictures of a scarf was +.3. Which predictor would provide Max with the more accurate prediction of scarves sold? Why?

#6. Cheryl, a reading teacher, is examining how different variables affect students' scores on a reading comprehension test. In one analysis, she used scores on a spelling test to predict reading comprehension scores. Here is the regression equation produced from this analysis:

$$\hat{y} = -3.5 + 0.8x$$

If Cheryl's principal asked her what she expected to happen to reading comprehension scores as spelling test scores increased, what should she say?

Curriculum Vita – Michael Cook
mac397@psu.edu

EDUCATION

Doctor of Philosophy, Educational Psychology
Penn State University, University Park, PA
2013 to 2017

Master of Science, Educational Psychology
Penn State University, University Park, PA
2009-2013

Instructor, EDPSY 406, Pennsylvania State University, University Park
Spring 2012, Fall 2013, Spring 2014-Spring 2015 (different version of course for Spring 2014-Spring 2015), Spring 2017

Instructor, EDPSY 10, Pennsylvania State University, University Park
Fall 2012

Statistical Consultant, Data Learning Center, University Libraries, University Park, 2016-2017

Research Team Member, ADHD/Vocabulary Grant Team, University Park, 2012-2015

Research Team Member, Intelligent Tutoring of the Structure Strategy (ITSS) Team, University Park, 2010 - 2012

Publications

Morgan., P.L., Farkas, G., Hillemeier, M., Mattison, R., Li, H., & **Cook, M**. (2015). Minorities are Disproportionately Under-represented in Special Education: Longitudinal Evidence Across Five Disability Conditions. *Educational Researcher, 44,* 278-292.

Morgan, P.L., Li, H., **Cook, M**., Farkas, G., Hillemeier, M.M., & Lin, Y (2015). Which Kindergarten Children are at Greatest Risk for Attention-deficit/Hyperactivity and Conduct Disorder Symptomatology as Adolescents? *School Psychology Quarterly, Advance online publication.*

Wijekumar, K., Meyer, B.J.F., Lei, P., Lin, Y., Johnson, L.A., Shurmatz, K., Spielvogel, J., Ray, M.N., & **Cook, M**. (2014). Improving reading comprehension for 5th grade readers in rural and suburban schools using web-based intelligent tutoring systems. *Journal of Research on Educational Effectiveness, 7,* 331-357

Lin, Y., Morgan, P. L., Hillemeier, **M., Cook**, Maczuga, S., & Farkas, G. (2013). Reading, mathematics, and behavioral difficulties interrelate: Evidence from a cross-lagged panel design and population-based sample of US upper elementary students. *Behavioral Disorders, 38,* 212-227.

Suen, H.K., & **Cook, M.** (2011). Validity of assessment results and uses. In James A. Banks (ed.) *Encyclopedia of diversity in education*. Thousand Oaks, CA: Sage