# Why Correlation Doesn't Imply Causation: Improving Undergraduates Understanding of Research Design

by

**Ciara Louise Willett**

BS, St. Mary's College of Maryland, 2014

MS, Seton Hall University, 2017

MS, University of Pittsburgh, 2021

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

**Ciara Louise Willett**

It was defended on

August 24, 2022

and approved by

Kelly Goedert, Professor, Psychology, Seton Hall University

Timothy Nokes-Malach, Professor, Psychology

Christian Schunn, Professor, Psychology

Dissertation Director: Benjamin Rottman, Associate Professor, Psychology

**Why Correlation Doesn't Imply Causation: Improving Undergraduate Students' Understanding of Research Design**

Ciara Louise Willett, PhD

University of Pittsburgh, 2022

Understanding when it is appropriate to make causal inferences from a statistical result is a fundamental skill for science literacy. Prior research has concentrated on erroneous causal judgments about observational studies, but there is little research on whether people understand that experiments provide stronger justification for causal claims. Our study tested the efficacy of an intervention at improving students' ability to discriminate between correlation and causation. Students were taught how to use causal diagrams to illustrate possible explanations for a statistical relation in an experiment versus an observational study. To evaluate the intervention's efficacy, intro psych (Experiments 1-3) and research methods (Experiment 1) students decided whether to make causal inferences about hypothetical observational studies and experiments. In Experiment 1, we tested multiple methods of instruction to see which worked best. Intro psych students learned more when they completed practice problems that involved generating self-explanations, whereas research methods students learned more from making analogical comparisons or reading worked examples. Critically, we found that students struggled with identifying the study design, which is the first step in correlation-causation discrimination. In Experiment 2, we added instructions to the Self Explanation intervention about how to identify observational studies versus experiments. Our modifications did not improve this skill nor students' ability to discriminate between correlation and causation. The most successful intervention was in Experiment 3, which explicitly pointed out that people often make errors when evaluating evidence from observational studies and repeated

the importance of considering study design when making causal judgments. A second goal of Experiment 3 was to test the influence of students' expectations about the direction of the statistical relationship on their evaluation of the evidence. Students made more causal inferences about study outcomes that were in the same direction as their prior beliefs than outcomes they thought were implausible. After the intervention, students still used their prior beliefs to decide whether to make a causal judgment, but they also more strongly considered the study design in their evaluation of evidence. In general, our intervention improved students' understanding of causality, but its efficacy may also depend on their prior knowledge.

## Table of Contents

**Preface** ..... **xi**

**1.0 Introduction** ..... **1**

**1.1 Using Causal Diagrams to Improve Correlation-Causation Discrimination** ..... **3**

**1.2 Different Methods of Instruction for Improving Correlation-Causation Discrimination** ..... **7**

**1.2.1 Worked Examples** ..... **9**

**1.2.2 Analogical Comparison** ..... **11**

**1.2.3 Self Explanation** ..... **12**

**1.2.4 Active versus Passive Learning** ..... **14**

**1.3 Testing Research Design Abilities** ..... **15**

**1.4 Influence of Prior Beliefs** ..... **16**

**1.5 Summary of Three Experiments** ..... **19**

**2.0 Experiment 1** ..... **21**

**2.1 Methods** ..... **21**

**2.1.1 General Procedure** ..... **22**

**2.1.2 Pre-Test Measures** ..... **23**

**2.1.3 Intervention** ..... **26**

**2.1.4 Post-Test Measures** ..... **30**

**2.2 Results** ..... **32**

**2.2.1 Ability to Correctly Identify the Study Design from Vignettes** ..... **32**

**2.2.2 Correlation-Causation Discrimination for the Vignettes** ..... **35**

**2.2.3 Pre-Test Explanations of "Correlation Doesn't Equal/Imply Causation" ...42**

**2.2.4 Pre-Test Explanations and Correlation-Causation Discrimination .....42**

**2.2.5 Causal Diagrams Task.....44**

**2.3 Discussion ..... 47**

**3.0 Experiment 2 ..... 50**

**3.1 Methods ..... 51**

**3.1.1 Participants, Attention Check, and Exclusion Criteria.....51**

**3.1.2 Design and Intervention .....52**

**3.1.3 Pre-Test and Post-Test Measures .....52**

**3.2 Results..... 54**

**3.2.1 Ability to Correctly Identify the Study Design from Vignettes .....54**

**3.2.2 Correlation-Causation Discrimination for the Vignettes .....56**

**3.2.3 Design a Study Task.....60**

**3.2.4 Causal Structures Task .....62**

**3.3 Discussion ..... 65**

**4.0 Experiment 3 ..... 67**

**4.1 Methods ..... 71**

**4.1.1 Prior Belief Manipulation .....71**

**4.1.2 Intervention .....73**

**4.2 Results..... 74**

**4.2.1 Ability to Correctly Identify the Study Design from Vignettes .....74**

**4.2.2 Correlation-Causation Discrimination for the Vignettes .....77**

**4.2.2.1 Efficacy of the Intervention on Correlation-Causation Discrimination ..... 77**

**4.2.2.2 Effects of Prior Beliefs on Correlation-Causation Discrimination ... 78**

**4.3 Discussion ..... 84**

**5.0 Discussion ..... 86**

**5.1 Lessons Learned About Improving Correlation-Causation Discrimination ..... 87**

**5.2 Undergraduate Students' Ability to Design their Own Studies ..... 92**

**5.3 Implications for Theories about Prior Beliefs ..... 95**

**5.4 Implications for Theories about Instructional Techniques ..... 97**

**5.5 Future Directions for Improving the Intervention and Tailoring it to Different Populations ..... 99**

**5.6 Conclusions ..... 102**

**Appendix A Results for Interpreting Statistical Tests ..... 104**

**Bibliography ..... 106**

## List of Tables

**Table 1 Mixed Effects Models Testing Believed Study Design Discrimination ..... 35**

**Table 2 Mixed Effects Models Testing for Correlation-Causation Discrimination..... 40**

**Table 3 Mixed Effects Models Comparing the Three Interventions..... 41**

**Table 4 Explanation Accuracy as a Predictor of Pre-Test Correlation-Causation Discrimination ..... 44**

**Table 5 Testing for Differences in Endorsement of Causal Diagrams..... 47**

**Table 6 Mixed Effects Model Testing Believed Study Design Discrimination ..... 56**

**Table 7 Mixed Effects Model Testing for Correlation-Causation Discrimination by Actual Design ..... 59**

**Table 8 Mixed Effects Model Testing for Correlation-Causation Discrimination by Believed Design ..... 60**

**Table 9 Mixed Effects Model Testing Probability of Designing an Experiment..... 62**

**Table 10 Testing for Differences in Endorsement of Causal Diagrams..... 64**

**Table 11 Mixed Effects Models Testing Believed Study Design Discrimination ..... 76**

**Table 12 Mixed Effects Model Testing for Correlation-Causation Discrimination by Actual Design ..... 82**

**Table 13 Mixed Effects Model Testing for Correlation-Causation Discrimination by Believed Design ..... 83**

**Appendix Table 1 Testing for Improvement in Interpreting Statistical Tests..... 105**

## List of Figures

**Figure 1 Examples of using causal diagrams to illustrate possible explanations for a correlation ..... 5**

**Figure 2 Procedural Timeline of Experiment 1 ..... 23**

**Figure 3 Examples from Causal Diagram Tutorial ..... 27**

**Figure 4 Portion of Reverse Causality Section in the Practice Problems ..... 29**

**Figure 5 Feedback in the Reverse Causality Section for the Experiment Problem ..... 30**

**Figure 6 Three Possible Explanations in the Causal Diagrams Task ..... 31**

**Figure 7 Believed Study Design Discrimination ..... 34**

**Figure 8 Correlation Causation Discrimination by Actual Study Design ..... 38**

**Figure 9 Correlation Causation Discrimination by Believed Study Design ..... 39**

**Figure 10 Endorsement in Causal Diagrams Task ..... 46**

**Figure 11 Believed Study Design Discrimination ..... 55**

**Figure 12 Correlation Causation Discrimination in Study 2 ..... 58**

**Figure 13 Participants' Study Designs ..... 62**

**Figure 14 Endorsement in Causal Diagrams Task ..... 64**

**Figure 15 Believed Study Design Discrimination ..... 76**

**Figure 16 Correlation Causation Discrimination by Study Design ..... 80**

**Figure 17 Correlation Causation Discrimination by Study Design and Evidence Congruency ..... 81**

## Preface

This work is dedicated in memory of Eli Harris Roth, who was so proud of his Ella's, put out our (my) fires, and showed us how to be grownups. E, you are endlessly loved and dearly missed.

I am immensely grateful to have so many wonderful people that have helped me get to this point in my academic journey. Thank you all for being in my corner, I am indebted to you all.

First, thank you to the people who directly helped with the (many) words you see on the following pages. My mentor Benjamin Rottman, who helped me grow into a better writer, scientist, and teacher. Thank you for your encouragement and guidance over the past five years, I truly could not have done this without you.

Thank you to my committee members, Kelly Goedert, Timothy Nokes-Malach, and Christian Schunn, who provided kind and insightful feedback. Kelly was my first mentor, who introduced me to the world of causal learning, and it was an absolute gift to bookend my graduate student chapter with her on my committee. Tim supported a year of this work with an ITM grant, which I am so thankful for as it provided me with the time and space to devote to this research.

Thank you to the Basic and Applied Cognition lab at the University of Michigan, Priti Shah, Audrey Michal, and Colleen Seifert, whose earlier work inspired the direction of the current research. Thank you for your kindness in collaboration, your sharing of insights and materials.

Thank you to the many research assistants who spent hours of their time coding on this project – Danielle Nebril, Brianna Hale, London Claridy, Ashley Harbaugh, Rabia Khan, and Kayla Grutkowski. Our meetings are some of my fondest memories of my dissertation, and I am so grateful for your help; I wish you all the best in your future endeavors.

And now, I would like to thank all the people in my life who have cheered me on and have gotten me to this point. To my family, thank you for everything. Mom, you are brilliant, caring, resilient, inspire me endlessly, and remind me that I can do hard things. Dad, you taught me how to write, never give up, and the joy of making people laugh. Allison, you light up every room and make the hardest days feel like the best days. Jen, I am forever moved by your strength.

Lindsey, Emma, and Cara, my Ella's – thank you for always picking up the phone, for every piece of advice and much-needed laugh or distraction, for proofreading my emails, and for being there to binge every episode of reality tv that has ever been created. I could not have accomplished any of this without you three by my side.

Nana, the Tansey's, the Waddington's, Adam, Asher, Joanna, Stephanie, Vicki, Leo, the Riley's – your love and support crosses oceans and state lines. You are the best family I could have ever asked for and I love you so very much.

Evan, Evie, and Chloe – thank you for every FaceTime and voice memo that brought me sunshine on cloudy and rainy Pittsburgh days. I need my own pair of duck boots, Duck Boots.

My Maryland family – who have spent the past five years calling me "Dr Willay" and joking that they're going to show up with signs saying "D-fence" – how did I get so blessed to have friends like you for three decades?! Thank you for being my biggest cheerleaders, for always checking in, and all the impromptu encouragement when I needed it most.

Petra, Lorraine, Josh, Orma, Brett, Kevin, Ewan, Stef, Jamie, Liam, Heather, Gaby, Kole, Griff, Alex, Josh, Bill, Lizzy, Lauren, Becca, Grace, Gran, and everyone else who made my time here so special – thank you for all the family meals that nourished me, the walks that grounded me, the jumping grids that gave me confidence, and the 530 am social hours (aka swim practice) that kept me going. Thank you for making Pittsburgh home.

Kevin, Cory, Zac, Yiwen, and Sara – I am lucky to have had lab mates that I call close friends. Thank you for the group chats, the never-ending support (see also the willingness to share your code and wisdom!), the lunch spots we got addicted to (and then sick of), and even all the GeoGuessr games I never won. Oh and just an FYI, this is the year I'm winning the league.

And finally, to the best roommate and friend, Brett, thank you for filling our little apartment with the most joy, candles, and food, and for making sure that I ate my vegetables and Lili did not.

## 1.0 Introduction

"Correlation does not imply causation" is a rule introduced in science classes and frequently repeated throughout research methods and statistics textbooks (e.g., Leary, 2012). Its emphasis is well justified – the ability to discriminate between correlation and causation is a key component of scientific reasoning. This idea, however, is not one from modern science but instead has more ancient origins; the Latin phrase "cum hoc ergo propter hoc" means "with this, therefore because of this", and refers to how the mere co-occurrence of two events does not provide evidence that one caused the other. However, despite the longevity of this concept, people often make causal inferences about two co-occurring events without sufficient evidence (Bleske-Rechek et al., 2015; Meijer, 2007). More recently, educational interventions have been designed to improve understanding that a statistical relation on its own does not imply causation; for example, by teaching students how to generate alternative explanations for why two events may have co-occurred (Seifert et al., 2022). Across three studies, we developed and tested an intervention with the goal of improving college students' correlation-causation discrimination abilities and their understanding of when a statistical relation does and does not warrant a causal claim.

Developing a better understanding of why correlation does not equal/imply causation is imperative for both scientists and the public. For scientists, a key goal of research is uncovering causal relations, but certain criteria must be met to make a causal claim. In the 19$^{th}$ century, philosopher John Stuart Mill suggested three necessary conditions for causality: 1) covariation between the cause and effect, 2) temporal precedence such that the cause precedes the effect, and 3) ruling out alternative possibilities for covariation, such as a common cause or confounded

relationship (Mill, 1872). Thus, covariation is necessary but insufficient for making causal inferences.

Understanding this concept is a core tenet of scientific literacy. For scientists, uncovering causal relations is a key goal of research, so it is imperative that they know how to make appropriate conclusions based on their methodology and results. In the American Psychological Association's (APA) list of comprehensive learning goals for undergraduate psychology majors, this skill is a core aspect of the "scientific inquiry and critical thinking" goal (Halonen et al., 2013). Specifically, upon completing their major, students should know to "limit cause-effect claims to research strategies that appropriately rule out alternative explanations". However, even well-trained scientists make erroneous causal claims based on correlational studies (Haber et al., 2018; Han et al., 2022; Parra et al., 2021; Robinson et al., 2007).

The incorrect use of causal language to describe correlational findings is also prevalent in media articles (Adams et al., 2019; Bratton et al., 2020; Cofield et al., 2010; Haber et al., 2018). One study found that 49% of health news articles made causal claims about correlations, with 22% of articles making erroneous causal claims directly in the news headlines (Haneef et al., 2015). Thus, it becomes the responsibility of the public to evaluate the validity of these causal claims. However, correlation-causation discrimination is quite challenging, and people often make causal claims from correlational findings. For example, a sample of Midwestern adults were randomly assigned to read a vignette that described either an observational study or an experiment, and they were equally likely to make causal claims about correlational studies as they were for experiments (Bleske-Rechek et al., 2015). Given that it is highly difficult to retroactively correct misinformation (Lewandowsky et al., 2012), it is imperative to uncover methods of improving

correlation-causation discrimination so that people can adequately evaluate the validity of causal claims upon first encountering them.

In this paper, we conducted three studies to answer four main questions. First, does a causal diagrams intervention improve correlation-causation discrimination in undergraduate students? Second, what are the ideal methods of instruction for such an intervention? Third, can the intervention also improve related research methodology skills, like the ability to design observational studies and experiments? And fourth, can the intervention lessen the influence of students' prior beliefs on their assessments of whether a statistical relationship is causal or not?

### 1.1 Using Causal Diagrams to Improve Correlation-Causation Discrimination

A comprehensive understanding of "why correlation does not imply causation" involves knowing when it is versus when it is not appropriate to make a causal claim about a statistical relation. Much of the prior research in this domain has focused on studying the tendency for people to make causal claims about observational studies (Michal, Seifert, et al., 2021; Seifert et al., 2022), which is typically framed as a bias in reasoning (Halonen et al., 2013). However, these studies can only address reasoning about scenarios in which causal claims for statistical relations are unjustified, rather than in contrast to scenarios in which causal claims are justified. Because experiments can rule out alternative explanations for a statistical relation, people should make stronger causal claims about experiments than observational studies; we call this "correlation-causation discrimination".

Correlation-causation discrimination requires a foundational knowledge of research design. One must be able to identify the correct study design to determine whether a causal claim

is warranted or not. In an observational study, there are many possible explanations for a statistical relationship. Consider the following example in which a group of researchers collected survey data and found that having yellow teeth and lung cancer are correlated. If we made the claim that "yellow teeth causes lung cancer", we would be ignoring alternative possible explanations for the correlation. One possibility is a reverse causality explanation in which lung cancer causes yellow teeth instead of the other way around. A more likely possibility is a common cause/confound explanation, in which smoking cigarettes is the cause of both lung cancer and yellow teeth. If the smoking variable is not controlled for, it can look like there is a causal relationship between yellow teeth and lung cancer when there is none.

In an experiment, however, random assignment to conditions rules out alternative explanations for a statistical relation. This means we can be more confident that a simple cause-effect explanation is how two variables are related. Random assignment to conditions rules out reverse causality by ensuring that the possible cause precedes the effect in time. Experiments also rule out possible confounds by eliminating the possibility of systemic differences between conditions at pretest, assuming random assignment worked properly. Thus, because there are fewer alternative explanations for a statistical relation in an experiment versus an observational study, there is more justification for a causal claim in an experiment.

We can use causal diagrams, sometimes called directed acyclic graphs (DAGs) to represent the possible explanations for a statistical relation. In an observational study, all three diagrams in Figure 1 (simple cause-effect, reverse causality, common cause/confound) are possible explanations for a statistical relation. In an experiment, random assignment rules out the reverse causality (Figure 1B) explanation and reduces the likelihood of a common cause (Figure 1C) explanation.

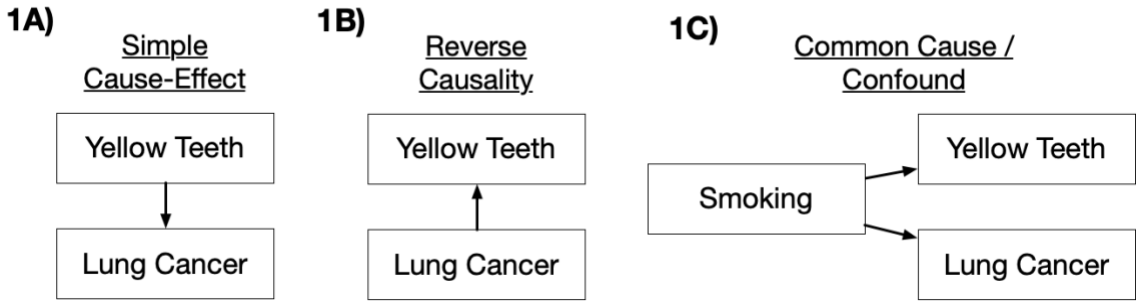


**Figure 1 Examples of using causal diagrams to illustrate possible explanations for a correlation**

Causal diagrams are visual representations of complex systems and have previously been used as tools for science communication and generating testable hypotheses like potential confounds or reverse causality explanations for a statistical relation (Pearl, 1995; Pearl & Mackenzie, 2018). In education, causal diagrams have been successfully used as a tool for improving scientific reasoning skills. For example, students' analytical skills can improve by drawing diagrams to represent the causal structure of an argument (Harrell, 2012). Other studies have shown that students can better understand complex causal systems when also provided with illustrations of the structures in diagrams versus text descriptions of the systems alone (McCrudden et al., 2007; van Loon et al., 2014).

At a high level, causal diagrams appear to facilitate learning by providing a simple concrete illustration of a multivariate causal system. They also enable "cognitive offloading", or a reduction in the amount of effort needed to parse complex relationships (Cheng et al., 2001). Thus, causal diagrams may help students to process that there are many alternative explanations for a statistical relation in an observational study (including multivariate relationships), that there are fewer explanations for a statistical relation in an experiment, and that this means we can make stronger causal claims about experiments than observational studies. In the current paper, we used causal diagrams as a tool to improve students' correlation-causation discrimination.

Seifert et al. (2022) recently showed that teaching undergraduate students how to use causal diagrams can reduce the number of causal claims they make about observational studies. During the intervention, students were given an example of an observational study and an explanation of why it is problematic to draw causal conclusions from the correlational findings. Next, they completed practice problems in which they came up with alternative explanations for the statistical relation and read explanations of how to use causal diagrams to evaluate the possibility of reverse causality and common cause explanations. To test for learning during the intervention, students were given a task that involved deciding whether to make a causal judgment about a headline that merely described a correlation. The number of causal judgments about correlational findings reduced from pre to post, which suggests that the intervention was successful.

Additionally, students got better at generating alternative explanations using causal diagrams for the statistical relation at posttest compared to pre (Seifert et al., 2022). Prior work has shown that teaching students how to generate their own alternative explanations for a statistical relation can be quite difficult (Sibulkin & Butler, 2019). We contend that understanding there are more alternative explanations for a statistical relation in an observational study compared to an experiment is a critical component for improving correlation-causation discrimination. Thus, the findings from Seifert et al. (2022) provide preliminary evidence that causal diagrams may be an effective tool at helping students learn the concepts necessary for improving correlation-causation discrimination. However, testing for improved correlation-causation discrimination requires asking participants to make judgments about both observational studies and experiments.

In the current study, we expanded upon the findings of Seifert et al. (2022) by developing and testing our own intervention to improve correlation-causation discrimination. Specifically, we used causal diagrams to teach students not only about the possibility of alternative explanations

for a statistical relation in an observational study, but also how those alternative explanations can be ruled out in an experiment. We then tested for improvement in correlation-causation discrimination by having participants decide whether to make causal judgments about both hypothetical observational studies and experiments.

### 1.2 Different Methods of Instruction for Improving Correlation-Causation Discrimination

Taking a foundational course in research methods in psychology is associated with better scientific reasoning (VanderStoep & Shaughnessy, 1997) and critical thinking skills (Bensley et al., 2010; Penningroth et al., 2007) in general. However, undergraduate students struggle with understanding the difference between correlation and causation even after taking research methods (Meijer, 2007) or other STEM-related courses (Bleske-Rechek et al., 2015; List et al., 2022; Owens, 2018). Given that a key learning goal for psychology majors is to understand when it is and is not appropriate to make a causal claim (Halonen et al., 2013), we wanted to test different methods of instruction to see if one method was best for teaching students about correlation-causation discrimination.

In the current study, we first presented students with foundational knowledge about why correlation does not imply causation, including information about how to use causal structures to illustrate alternative explanations for a statistical relation that are possible in observational studies but not experiments. Students then completed two practice problems that reinforced these concepts by asking them to consider different alternative explanations (mechanism, reverse causality, confound) in the context of a hypothetical observational study and a hypothetical experiment.

Practice, which involves completing a repeated series of problems or tasks, is a well-established instructional technique that helps students learn the procedural steps of problem solving (see Richey & Nokes-Malach, 2015 for a review). Thus, providing students in the current study with a series of practice problems may facilitate learning and improve correlation-causation discrimination.

Undergraduate psychology students' understanding of various concepts and skills in research methods can greatly improve with practice, like knowing how to interpret and critique scientific articles (Kershaw et al., 2018) and discriminate between causal versus correlational language in writing (Mueller & Coon, 2013). However, practice may not be sufficient for improving some critical research methods skills. In Sibulkin and Butler (2019), undergraduate research methods in psychology students were given examples of reverse causality explanations for statistical relations in hypothetical observational studies and completed practice problems throughout the rest of the semester, in which they generated their own reverse causality explanations for novel examples. The accuracy of students' explanations improved from the first practice problem (39%) to the fourth and final practice problem (56%), but performance was still quite poor at the end of the study because 44% of students were unable to generate reverse causality explanations after repeated instruction and practice.

In addition to practice, there are other methods of instruction that utilize alternative techniques to facilitate learning during problem solving or task completion. In Experiment 1, we compared three methods of instruction during practice: Worked Examples (the learner sees a problem set and the solutions to the problems), Analogical Comparison (the learner answers a problem set by comparing two example problems), and Self Explanation (the learner answers a problem set and generates an explanation for their answers). We chose these methods because each

has been widely researched in the literature and has its own unique benefits and limitations (see Richey & Nokes-Malach, 2015 for a review). Thus, each method could plausibly help students learn more about correlation-causation discrimination.

### 1.2.1 Worked Examples

Worked Examples involve providing students with example practice problems and solutions that illustrate how to solve the problem (Kalyuga et al., 2001; Renkl, 2014; Richey & Nokes-Malach, 2013; Sweller & Cooper, 1985). Providing students with the correct solution reduces their cognitive load while learning; for example, participants spend less time and effort on trial-and-error to solve the problem (Owen & Sweller, 1985). The reduction in cognitive load improves students' memory for the declarative features of problems like the procedural steps (Sweller, 1988), which means they can later apply this information to novel problems. Compared to Analogical Comparison and Self Explanation, Worked Examples are particularly good for students acquiring declarative knowledge and conceptual knowledge that is necessary for problem solving. However, there is less evidence that Worked Examples help students identify misconceptions in their own knowledge, which is a key strength of the other two methods (Richey & Nokes-Malach, 2015).

There are a few limitations regarding the effectiveness of Worked Example instruction. One constraint is that students must have enough content knowledge to support learning from Worked Examples. For example, undergraduate students needed a foundational understanding of relevant laws and theorems in order to benefit from Worked Examples of physics problems (Renkl, 2014). On the other hand, providing too much information in Worked Examples can lead to less learning compared to Worked Examples that involve less instructional explanations (Richey &

Nokes-Malach, 2013). When students are given too much information, this can stop them from spontaneously engaging with the material (e.g., generating their own explanations) beyond passively reading the provided problem solutions. If spontaneous engagement is critical for learning, then providing students with too much information in Worked Examples will impede that process. In contrast, Self Explanation and Analogical Comparison instruction encourage more engagement with the material by design, which may yield greater learning.

The efficacy of Worked Examples may depend on the prior knowledge and understanding of the learner. If the student already has substantial knowledge of the subject, Worked Examples may not be effective. For example, mechanical trade apprentices with less experience learned more from Worked Examples instead of practice problems without solutions, whereas those with more experience learned more from practice problems without solutions because the Worked Examples contained redundant information (Kalyuga et al., 2001). Instead, Worked Examples may have a bigger impact if people have less prior knowledge or understanding about the material (Cooper & Sweller, 1987). The potential influence of prior knowledge is particularly relevant to the current study, because we compared the efficacy of the three instructional techniques with an intro psych and a research methods sample. Because intro psych students have less prior knowledge about research methodology, they may benefit more from Worked Example instruction. However, we expect that very few (if any) students would have prior knowledge about using causal diagrams to illustrate alternative explanations for a statistical relation. Therefore, the research methods sample may also be naïve to the information and benefit from Worked Example instruction.

### 1.2.2 Analogical Comparison

Analogical Comparison involves students comparing example problems that share similar features (Gentner, 1983; Gentner et al., 2003; Gick & Holyoak, 1983). The version of Analogical Comparison that is most like what we used in the current study, is when participants are asked to compare or contrast problem features by making analogies between examples provided by the researcher. Because the examples are provided, the learner does not have to worry about memory retrieval and can instead focus on identifying similar features across problems, which they can apply to novel scenarios. According to Gentner's (1983) structural mapping-theory, for successful problem solving of novel scenarios, the learner must first uncover similar structural features across examples at a deep level and then map those features onto similar features in the novel problem (see also Novick & Holyoak, 1991).

Whereas Worked Examples teach the learner about surface-level problem features, Analogical Comparison instruction can improve understanding of common structural features across problems (Blanchette & Dunbar, 2000; Cummins, 1992). However, with Analogical Comparison, people sometimes exhibit bias for making analogies about superficial similarities between examples rather than deeper structural similarities between examples. When students only learn about surface-level features from analogical comparisons, then they may only be able to solve problems that share superficial features and not problems that share structural features (Holyoak & Koh, 1987).

One solution is to specifically structure Analogical Comparison instruction so that the problems highlight or provide hints about structural similarities in the examples. Having students answer questions that directly ask about similarities between structural features in examples can improve performance for novel problems, compared to if students are simply given the examples

in succession without any questions that prompt analogical comparison (Gentner et al., 2003). This means that students may require support or guidance in the development of connections for Analogical Comparison to be effective, because students may not spontaneously engage in analogical comparisons on their own. For example, Gick and Holyoak (1983) found that using diagrams to illustrate problem solutions was most effective when the diagrams were also supplemented with Analogical Comparison instruction, versus no analogical comparisons.

In the current paper, Analogical Comparison instruction may improve correlation-causation discrimination because one of our goals is to teach students about the causal structures that underlie statistical relations in observational studies and experiments. Analogical Comparison is particularly effective at facilitating learning about similarities in structural features across examples (Gentner, 1983). First, causal diagrams are inherently structures, and because Analogical Comparison is good for learning about structural features in particular, this method may be ideal for learning during the intervention. Second, Analogical Comparison may help students learn about another problem feature that is critical for correlation-causation discrimination, the design of the study; students can compare study designs across different examples and apply this information when making inferences about causation.

### 1.2.3 Self Explanation

Self Explanation involves students generating their own explanations of the material, such as summarizing key concepts in a reading (Hausmann & VanLehn, 2007). Although some prior research has studied the effects of Self Explanations that are spontaneously generated by the learner (e.g., Chi et al., 1989), we explicitly asked participants questions that required generating Self Explanations during the intervention. Our Self Explanation instruction was similar to the

version used in Seifert et al. (2022), in which participants were told to generate their own examples of alternative explanations for a statistical relation and draw diagrams to represent those relations, rather than the participants spontaneously coming up with these examples on their own.

One proposed mechanism of learning via Self Explanation is that this method provides opportunities for students to identify gaps in their own knowledge or understanding (Chi, 2013). Therefore, even if students generate incorrect explanations, they can still learn from Self Explanation instruction. For example, a group of undergraduate psychology and biology students completed an intervention in which they learned and practiced generating Self Explanations (e.g., monitoring their understanding of the material, explaining what happened in their own words, making predictions about what would happen next in the text, etc.) to improve their reading comprehension for science texts (McNamara, 2004). At posttest, students who completed the intervention – even those who generated poor Self Explanations – had better reading comprehension than those who did not complete the intervention.

Only Seifert et al. (2022) have specifically studied the effects of incorporating causal diagrams into Self Explanation instruction. After completing Self Explanation problems that included drawing diagrams to represent alternative explanations for a statistical relation, students were less likely to make causal claims about observational studies compared to before instruction. Other studies have shown that supplementing Self Explanation with visual representations of information, aside from causal diagrams, can facilitate learning. Students who were prompted to generate their own visual explanations of complex mechanical and chemical systems demonstrated a better understanding of those systems than students who were prompted to generate purely verbal explanations (Bobek & Tversky, 2016). Another study found that Self Explanation practice was most beneficial for learning about new content if Self Explanation was preceded by instructional

material presented in a diagrams format (e.g., drawings of the human circulatory system) instead of a text format (Ainsworth & Loizou, 2003). Thus, incorporating visual illustrations with Self Explanation practice, like teaching students how to represent different causal relations in diagrams and then providing them with the opportunity to practice generating their own examples of diagrams and explanations for the relations, may improve correlation-causation discrimination.

### 1.2.4 Active versus Passive Learning

All three interventions differ in terms of whether they involve active (Analogical Comparison or Self Explanation) or passive (Worked Example) learning. Educational interventions that involve active learning have been associated with improving critical thinking in relation to psychology (Penningroth et al., 2007, cf. McLean & Miller, 2010), understanding of research methodology (Kreher et al., 2021; LaCosse et al., 2017), and reducing erroneous pseudoscientific and paranormal beliefs (Lawson & Brown, 2018; McLean & Miller, 2010). Therefore, active learning in the Analogical Comparison and Self Explanation practice conditions may be particularly beneficial at facilitating learning about correlation-causation discrimination. On the other hand, passive learning in the Worked Example condition could be more beneficial given that most students will be novices with causal diagrams and therefore might benefit from step-by-step instructions and examples. In sum, there are plausible reasons for why any of the three techniques could improve correlation-causation discrimination.

### 1.3 Testing Research Design Abilities

The first step of correlation-causation discrimination, and of evaluating whether a study outcome warrants a causal claim, is to identify the design of the study. People should be more confident in making a causal claim about an experiment than about an observational study. Thus, understanding the critical difference between the two study designs – that random assignment is used in experiments but not observational studies – is foundational knowledge for correlation-causation discrimination. This is also foundational knowledge for the ability to properly design studies, which is another research methods skill that the APA pinpoints as a learning outcome for undergraduate psychology students to master before graduating (Halonen et al., 2013).

In the developmental literature, many studies have looked at the acquisition of one critical skill for designing experiments – whether the child understands how to control for variables in a multivariate study (Chen & Klahr, 1999; Klahr et al., 2011; Klahr & Nigam, 2004; Kuhn et al., 2000; Kuhn & Dean Jr, 2005). For example, if a researcher is designing a study to test the effect of a potential cause (e.g., caffeine intake as a cause of sleep quality), they should control for possible confounds by holding alternative causes (e.g., stress) constant. Thus, these studies have tested whether elementary children have the foundational knowledge for designing research that is robust to critiques, like alternative explanations for the statistical elation. In a different domain, the biology literature, researchers have tested undergraduate students' knowledge of experimental design and their ability to design experiments (Brownell et al., 2014; Shanks et al., 2017; Sirum & Humburg, 2011). For example, Shi et al. (2011) found that biology majors struggled with understanding the purpose of a control group in an experiment and also with recognizing experiments from observational studies.

Across these domains, there seems to be a substantial focus on whether people can design experiments or not. None of the prior research, however, has specifically tested whether students can design both experiments and observational studies when prompted. Because random assignment to conditions is not always possible in research, especially in the field of psychology, it is important for students to learn how to design both types of studies. Knowing how to design an experiment instead of an observational study, or vice versa, requires students to understand that experiments use random assignment to conditions, but observational studies do not. Similarly, this understanding is also necessary for correlation-causation discrimination because the first step in deciding whether to make a causal conclusion should be to determine the study design. In Experiment 2, we tested whether our intervention also improved students' abilities to design their own observational studies and experiments. The key goal of our intervention was to improve correlation-causation discrimination; because the same foundational knowledge is required for designing an observational study versus an experiment, it is possible that the causal diagrams intervention might also improve students' research design abilities.

### 1.4 Influence of Prior Beliefs

A third direction of the current paper was to test the role of prior beliefs in correlation-causation discrimination. When one is learning about a study, for example when reading the news, or hearing about scientific findings in a course, they likely have expectations about how the results of that study will turn out. In turn, these prior expectations may influence whether they accept the conclusions from the study; for example, if the results are incongruent with their expectations, they may be less likely to accept the findings. Michal et al. (2021b) gave participants examples of

studies that tested hypothetical educational interventions (e.g., studying in a messy versus tidy classroom. When participants were asked to evaluate the scientific evidence in the study, they were more likely to make judgments based on whether the study findings were plausible or not rather than evaluating the quality of the study evidence.

In the field of causal learning, several studies have shown that prior beliefs can influence people's judgments about causal relationships in a variety of ways (Alloy & Tabachnik, 1984; Fugelsang & Thompson, 2000, 2001; Garcia-Retamero et al., 2009). For example, prior beliefs can affect whether someone believes that a variable is the cause or the effect in a relationship (White, 1995). Additionally, people are more likely to discount or rule out a possible cause of an outcome (e.g., allergic reactions, car accidents) if they think that another cause is more believable (Fugelsang & Thompson, 2001). However, there is little research about how prior beliefs might affect correlation-causation discrimination.

One study by Michal and colleagues (2021a) had participants read hypothetical media article vignettes about observational studies. The vignettes were pretested to ensure that half of the vignettes were belief-congruent in the sense that most people thought the result was plausible (e.g., social media use is negatively correlated with well-being), and the other half were belief-incongruent in that most people thought the result was implausible (e.g., time spent on homework is negatively correlated with grades). When the correlational findings were belief-congruent, and participants thought the statistical relation was plausible, they were more likely to justify a causal claim. Additionally, participants generated fewer alternative explanations for belief-congruent vignettes; although there were alternative explanations for a statistical relation in both the belief-congruent and belief-incongruent cases, participants generated more alternative explanations for the result when they thought it was implausible.

These findings are consistent with a dual-process account of reasoning. When people reason about information that already aligns with their expectations, their reasoning can be biased or they may rely on heuristics to make judgments (Evans & Curtis-Holmes, 2005; Nickerson, 1998). For example, when people have prior knowledge about a topic, they tend to engage in positive testing strategy and seek out evidence that confirms those beliefs rather than searching for information that would invalidate them (Goedert et al., 2014). On the other hand, when people reason about belief-incongruent information, they engage in more analytical reasoning and are more critical of the study findings (Koehler, 1996; Kunda, 1990; Lord et al., 1979; Shah et al., 2017; Thompson & Evans, 2012). If people have strong prior beliefs, and they have a good reason to have those prior beliefs, it is unlikely that they will immediately accept evidence that disconfirms those expectations. Instead, people may be more motivated to generate alternative explanations for evidence that they find to be implausible or disagree with, as in Michal et al. (2021a).

In the current research we expanded upon the literature on prior beliefs in two ways. First, we tested the influence of prior beliefs on whether people made causal judgments about observational studies and experiments, rather than only asking participants about observational studies. In Michal et al. (2021a)., participants made causal claims about observational studies, and they made even stronger causal claims about belief-congruent observational studies than belief-incongruent ones. However, we do not know how participants' expectations about the statistical relationship might affect their judgments about experiments.

One possibility is that the influence of prior beliefs will be the same for experiments as for observational studies; that participants will make stronger causal judgments when the evidence is consistent with their prior beliefs and make weaker causal judgments when the evidence is belief

incongruent. Another possibility is that prior beliefs have less of an impact when participants are deciding whether to make causal judgments about experiments than observational studies, because experiments provide stronger evidence for causality. This could have implications for correlation-causation discrimination; for example, if prior beliefs do not affect judgments about experiments, but judgments for evidence-congruent observational studies are more causal than for evidence-incongruent observational studies, then people would have better correlation-causation discrimination for evidence-incongruent scenarios. A third possibility is that prior beliefs have a greater impact when people make evaluate the findings from experiments than observational studies, which would also affect the extent of correlation-causation discrimination.

The second way that we expanded on the existing research on prior beliefs is that we tested whether our intervention decreases the influence of prior beliefs on correlation-causation discrimination. Our intervention encourages students to think more critically about study design when making causal judgments; they learn how to generate alternative explanations for a correlation in an observational study, and how experiments rule out these alternative explanations through random assignment. If the intervention works, after the intervention, when students are asked to decide whether a study provides sufficient evidence for causality, they may rely less on whether the evidence is congruent with their prior beliefs and more on information about study design.

### 1.5 Summary of Three Experiments

Across three experiments, we tested the efficacy of interventions aimed at improving correlation-causation discrimination (i.e., making stronger causal claims for experiments than

observational studies) among undergraduate students (Experiments 1-3). A main theme across these interventions was teaching students how to use causal diagrams to illustrate alternative explanations for statistical relations and that they should make stronger causal claims about experiments because there are fewer possible explanations for a statistical relation than in an observational study. In Experiment 1, we tested whether different methods of instruction – Worked Example, Analogical Comparison, Self Explanation – were more successful at improving correlation-causation discrimination. In Experiment 2, we tested whether the intervention would also improve students' ability to design both observational studies and experiments, which is another critical research methods skill. Additionally, we tested whether they could identify the causal structures that were possible alternative explanations for a statistical relation in their hypothetical studies, based on the type of study they were prompted to design. Finally, in Experiment 3, we tested the effects of prior beliefs on correlation-causation discrimination and whether the intervention would reduce bias due to prior beliefs.

## 2.0 Experiment 1

The primary goal of Experiment 1 was to test whether certain methods of instruction (Worked Example, Analogical Comparison, Self-Explanation) led to greater improvement in participants' ability to discriminate between correlation and causation. Specifically, whether they understand that they can make stronger causal inferences about statistical relationships in an experiment versus an observational study. We used several measures to assess students' correlation-causation discrimination abilities, the efficacy of the intervention, and students' prior knowledge of why correlation does not imply causation.

### 2.1 Methods

A total of 602 participants completed the study. 265 participants completed the study to fulfill a requirement for an Introduction to Psychology course. 337 participants were recruited from a Research Methods in Psychology course; two instructors assigned the survey as a homework assignment. Students were encouraged to email the study administrator if they did not wish their data to be used in analyses ($N$ = 1). An additional 2 participants from the intro psych sample were excluded from analyses because they did not show effort on any of the pre-test or post-test qualitative measures (e.g., wrote "I don't know" for every question). Ultimately, 599 participants (263 in intro psych, 336 in research methods) were included in analyses.

### 2.1.1 General Procedure

Participants completed the entire study online in Qualtrics. Participants were randomly assigned to one of the three intervention conditions: Analogical Comparison ($N$ = 86 in intro psych; 114 in research methods), Self-Explanation ($N$ = 89 in intro psych; 110 in research methods) and Worked Example ($N$ = 88 in intro psych; 112 in research methods).

The general procedure for Study 1 is outlined in Figure 2. The study involved a pre-test, intervention, and post-test. The intervention comprised two parts. The first part which we call the 'tutorial' was the same for all three conditions that talks about the difference between observational studies vs. experiments and introduces causal diagrams for as potential explanations for a statistical relation. The second part of the intervention involved further learning and practice, and was different for the three conditions (worked example, analogical comparison, or self-explanation). The full texts for the interventions and tutorial are available at https://osf.io/eug96.

Students were encouraged to take the entire study in one sitting. However, some students (38%) opened the assignment and returned to it on a later day. Of the students who completed the assignment in one day, the median completion time was 69 minutes in intro psych and 65 minutes in research methods, though this could include time not working on the study.

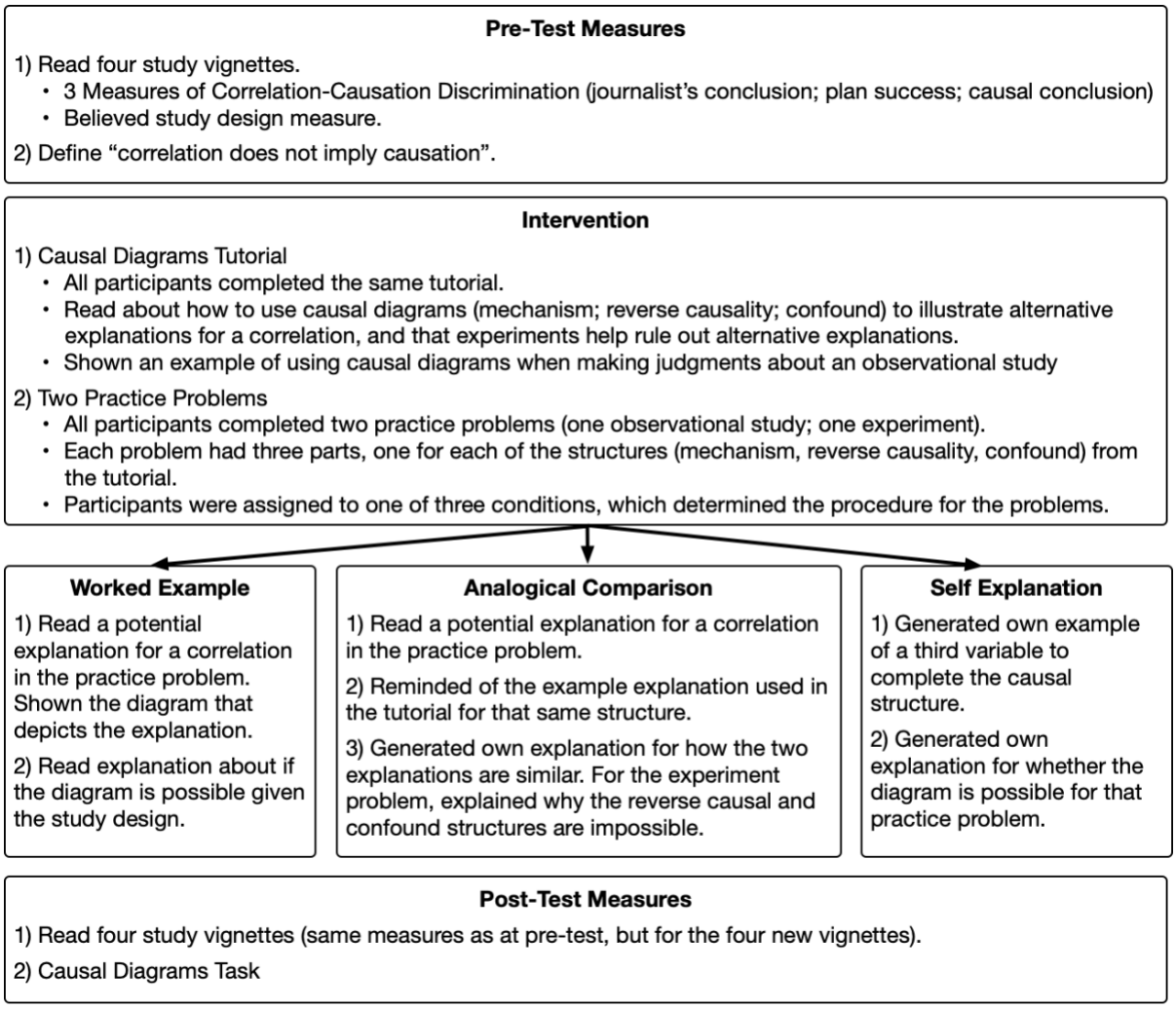


**Figure 2 Procedural Timeline of Experiment 1**

### 2.1.2 Pre-Test Measures

Participants did a task designed to test if they can discriminate correlation from causation. In particular, we tested whether they know to only make causal claims for statistical relations in experiments and not observational studies. Furthermore, we tested whether they know that an intervention proposed to act on the independent variable of an experiment should indeed have an influence on the dependent variable, whereas interventions proposed to act on the independent

variable of an observational study are not justified because the independent variable may not cause the dependent variable.

This task involved reading about four imaginary research studies. Two of the studies were experiments and the other two were observational studies, though participants needed to figure out which type of study it was on their own. For each study they answered four questions. The first three tested for correlation-causation discrimination (journalist's conclusion, plan success, and causal conclusion). The fourth asked if they believe that the study was an experiment or observational study; accurately identifying the study type is necessary for correct correlation-causation discrimination.

Participants were told to "imagine that you are reading the newspaper and a journalist has written a series of articles on recently published studies". Each article was presented as a one-paragraph vignette about a study that found a statistical relation between two variables. All but one of the vignettes was adapted from Michal et al. (2019) or other materials provided by their lab. We converted half of the vignettes so that they described experiments, not observational studies. An example vignette about an observational study was as follows:

> "Researchers at the Sleep Research Society have found that people who are tired spend more money on food purchases. Participants reported the average number of hours that they slept each night and then submitted their grocery and restaurant bills to the researchers at the end of every week for analysis. Participants who reported sleeping less than five hours a night spent more money on food. So, it follows that getting more sleep will reduce the amount of money someone spends on food."

On the same page as each vignette, participants answered four questions. Three of these questions probed correlation-causation discrimination[1]. First, each vignette ended with a

[1] The first two questions (journalist's conclusion and plan success) were from Michal et al. (2019).

journalist's conclusion proposing an intervention, and the participant judged the **journalist's conclusion**, "To what extent do you think that the study findings support the journalist's conclusion that [getting more sleep will reduce the amount of money someone spends on food]?", rated on a 7-point scale (1 = *the finding does not support the journalist's conclusion;* 4 = *unsure*; 7 = *the finding strongly supports the journalist's conclusion*).

Second, the participant made a **plan success** judgment, in which they evaluated the likelihood of a successful intervention that was based on the study's findings: "After reading about this study, [Jane decides to get at least six hours of sleep before going to the grocery store to reduce her spending on food]. How likely is it that [Jane's] plan will work?" (1 = *not at all likely for the plan to work;* 4 = *unsure*; 7 = *extremely likely for the plan to work*). They were also asked to "Explain your reasoning. Why do you think that [Jane's] plan is likely/unlikely to work?". We only asked participants to explain their reasoning in Experiment 1, and do not analyze these results in the current work.

Whereas the above two questions focused on a proposed intervention, the third question more directly asked about whether the study supports a **causal conclusion** with the following question: "Do you think that this study shows that [getting more sleep] causes [people to spend less money on food]?"

Third, the participants made a judgment about the **believed study design** by selecting whether they thought the study was an observational study or an experiment. This question is not about correlation-causation discrimination per se; however, the first step in deciding whether to make a causal inference from a study or not involves first whether the design is an observational study or an experiment. Incorrect identification of the study design could prevent participants from drawing appropriate causal conclusions.

After participants read the four studies and answered the above questions, they were asked to explain "What does the phrase 'Correlation does not equal/imply causation' mean to you?". Although this phrase is quite common, our hypothesis was that most students do not fully understand why this is true. This qualitative measure is a different way of assessing correlation-causation discrimination because it is an open-ended opportunity for students to articulate their prior knowledge and depth of understanding about the concept. Additionally, the amount of prior knowledge that participants have about correlation vs. causation could affect the extent to which the intervention improves correlation-causation discrimination. This was only asked at pre, not post.[2]

### 2.1.3 Intervention

There were two parts in the intervention: a causal diagrams tutorial (see OSF for exact text) and two practice problems to elaborate the concepts. All three conditions saw the same tutorial. At the beginning of the tutorial, participants read a one-paragraph summary of a blog post written for the New York Times, "Walkable Neighborhoods Cut Obesity and Diabetes Rates" (Bakalar, 2016). We chose this example because the headline used a causal claim about findings from an observational study. After reading the summary paragraph, participants answered the five measures from the correlation-causation discrimination task.

---

[2] At pre-test and post-test, participants were also asked to apply the concept of correlation-causation discrimination to interpreting the results of statistical tests. Only research methods students completed this measure because we thought intro psych students would not have enough prior statistical knowledge. For conciseness and because this measure was less central to our main goals, the methods and results for this measure are presented in Appendix A.

On the following page, there was a detailed explanation about how the author made a causal claim about an observational study and that causal claims should only be made about experiments. At the end of the explanation, they were given an abbreviated summary: "If it is an experiment/randomized control trial, then we can conclude that the independent variable caused the dependent variable. If it is an observational study (or otherwise known as "correlational" study), then we can conclude that two variables are related, but we don't know how."

The last two pages of the tutorial introduced causal diagrams as a method for illustrating different causal explanations for a statistical relation in an observational study (Figure 3); participants were told that "all three of these explanations can explain why there is a negative correlation". At the end of the tutorial, participants completed a "knowledge check" where they had to identify the correct name for the three main causal structures (mechanism, reverse causality, common cause/confound). They saw the correct answers on the following page.

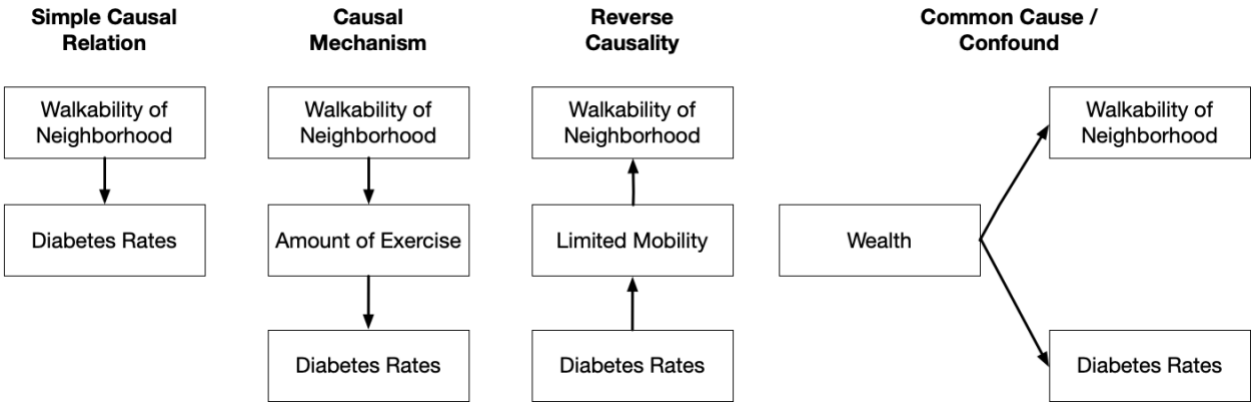


**Figure 3 Examples from Causal Diagram Tutorial**

In the second part of the intervention, participants were randomized to three conditions: Worked Example, Analogical Comparison and Self Explanation. The three conditions are similar in that they all included one practice problem about an observational study (a positive relation between watching TV and feeling tired during the day) and one about an experiment (positive relation between text message reminders and medication adherence).

For both the observational study and experiment practice problems, participants read a one-paragraph vignette in the style of a media report. They made a judgment about the believed study design and were given the correct answer on the following page. The rest of the procedure involved three sections, one for each causal diagram (mechanism, reverse causality, confound); the goal was to teach them that all three causal structures are possible for an observational study but only the mechanism causal structure is possible for an experiment. At the beginning of the reverse causality and common cause sections, participants were asked "Given that this is an [observational study/experiment], do you have to worry about [reverse causality/common cause]?" and were then given the correct answer.

The rest of the procedure for the sections differed across the three methods of instruction, all of which had the potential to improve correlation-causation discrimination. To highlight some of the similarities and differences between the three conditions, both the experiment and observational study passages for the reverse causality section is shown in Figure 4. In the **Worked Example** condition, we gave participants an example of a causal structure and an explanation for whether that structure was a possible explanation for a statistical relation in the practice problem. In the **Analogical Comparison** condition, participants were given the same example but no explanation. Instead, participants had to compare or contrast the practice problem example with the analogous causal structure from the tutorial (Figure 4). In the **Self Explanation** condition, participants came up with their own example of a causal structure and had to explain how the causal structure was or was not possible for the hypothetical study. Due to a programming error, most participants were not asked to explain how the confound structure was impossible for the experiment; after the error was identified, we added the question for the last 53 participants in the intro psych sample.

Because it is possible that participants in the Analogical Comparison and Self Explanation conditions would give incorrect responses, they were given feedback through an example of a correct response afterwards (see Figure 5 for an example). This example of a correct response was meant to be as similar as possible to the worked example condition while also answering the analogical comparison and self-explanation questions.

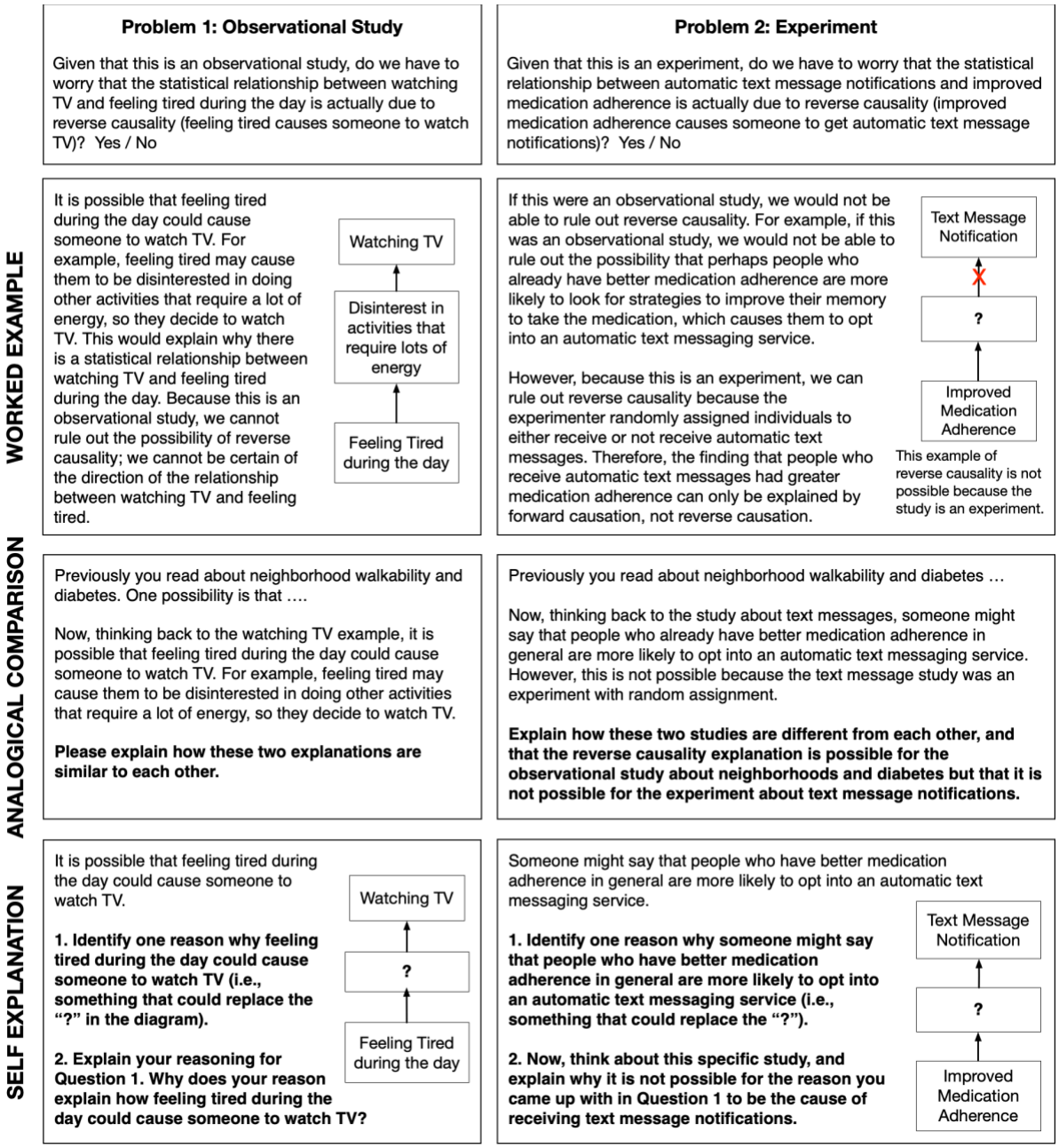


**Figure 4 Portion of Reverse Causality Section in the Practice Problems**

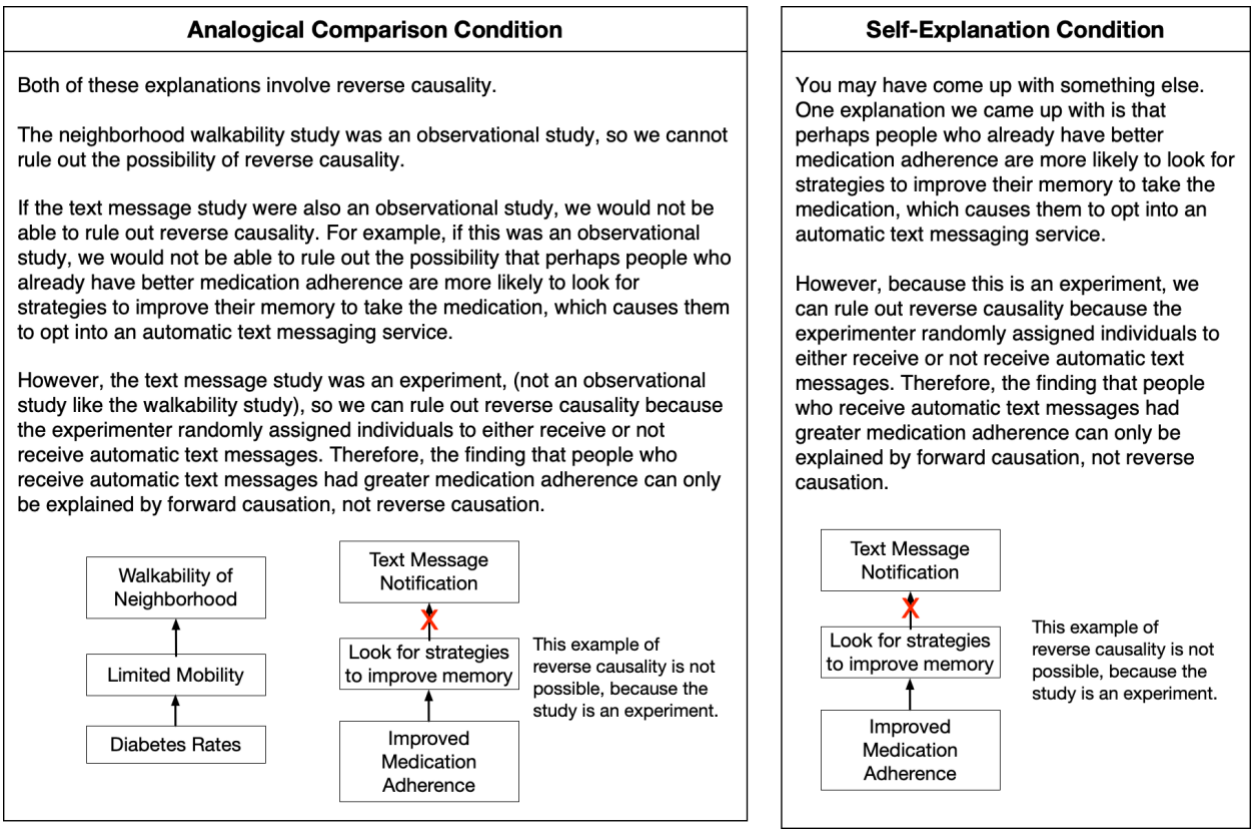


**Figure 5 Feedback in the Reverse Causality Section for the Experiment Problem**

### 2.1.4 Post-Test Measures

First, participants did the same tasks that they did in the pre-test, but with four different vignettes. We counterbalanced the order, so that half of participants saw one block of four at pre-test and the other half saw the block at post-test.

Next, participants completed a novel task that we call the 'causal diagrams' task. The goal of this task was to test whether participants would endorse all three causal structures (mechanism,

reverse causality, and confound) as explanations for an observational study, but only endorse the mechanism structure as an explanation for an experiment.[3]

Participants were told to imagine that they surveyed a group of undergraduate students and found "students who meditated at least 10 minutes per day reported greater life satisfaction than students who did not meditate at all". Next, they were shown three diagrams (Figure 6) and were asked "from this finding, which of the following relationships are possible?". Because the study was observational, they should endorse all three diagrams as possible explanations; however, they were not told it was observational and had to figure that out on their own.

Next, participants read an example of an experiment: "100 undergraduate students are randomly assigned to either meditate for at least ten minutes per day or not meditate at all. After three weeks, the researchers survey the undergraduate students and find that the students who meditated at least 10 minutes per day reported greater life satisfaction than the students who did not meditate at all." Participants answered the causal diagrams question; they should only endorse the causal mechanism as a possible explanation because random assignment to conditions rules out the alternative explanations.

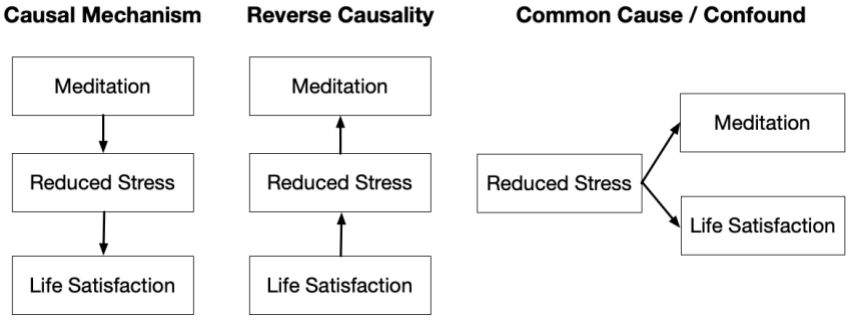


**Figure 6 Three Possible Explanations in the Causal Diagrams Task**

[3] In addition to the experiment and observational study, we asked participants to design their own experiments and they completed the 'causal diagrams' task for their own designed study. The task ended up being too easy and most participants successfully designed their own experiments (instead of observational studies). Because most participants were at ceiling, we do not discuss this measure or results further and designed a better task for Experiment 2.

## 2.2 Results

Registered analyses, anonymized datasets, and analysis scripts are available on OSF. Any deviations from the registration are noted in the manuscript. For mixed-effects models, we followed Barr et al.'s (2013) suggestions for how to simplify random slopes if a model does not converge. The most complex model that converged is reported for each analysis – detailed notes for model convergence are available in the analysis scripts on OSF. The intro psych and research methods samples were analyzed separately, with one exception, noted below. To test for possible differences between the two sections of research methods, we conducted a mixed-effects regression for each of the three correlation-causation measures with section × actual design (within subjects: observational versus experiment) as a fixed effect, actual design as a by-subject random slope, and vignette as a random intercept. Because there were no main effects of class section or any section × design interactions, the two research methods sections were analyzed together.

### 2.2.1 Ability to Correctly Identify the Study Design from Vignettes[4]

When deciding whether to infer causality from a study, the first step should be to determine whether it is an observational/correlational study or an experiment/randomized control trial. In the vignettes, the key difference was whether the term "randomly assigned" was used or not. To test whether participants could discriminate between the two study designs at pretest, and to test for learning during the intervention, we ran a regression with an interaction between actual design and block. We used contrast coding for actual design (-0.5 = observational study; 0.5 = experiment)

[4] These analyses were not included in the registered analysis plan.

and dummy coding for block (0 = pre; 1 = post). The dependent measure was whether participants said that the vignette was an observational study or an experiment (Model 1 in Table 1). The most complex model that converged included actual design as a by-subject random slope and vignette as a random intercept. For all analyses presented in this paper, we only included the most important predictors in the tables and text, meaning those that aligned with our registered hypotheses; our registered hypotheses and full set of results are available on OSF.

The effect of actual design was significant for both samples; because of contrast coding, this means that participants could discriminate between observational studies and experiments at pretest. Participants were more likely to say that the study was an experiment if it was an actual experiment (filled shapes in Figure 7) as opposed to an observational study (white shapes). In both classes, the actual design × block interactions were significant, which means that participants' study design discrimination was better at posttest than at pretest.

To test whether the improvement in study design discrimination differed across the three interventions, we tested for a three-way interaction (Model 2 in Table 1). The most complex model that converged included a by-subject random intercept and slope for actual design, and a by-vignette random intercept. Only one of six interactions was significant; in the intro psych sample, improvement in study design discrimination was greater in the Worked Example condition (circles in Figure 7) than in the Analogical Comparison condition (triangles).

In sum, participants could discriminate between study designs at pretest and got better at posttest. Thus, participants learned about study design discrimination during the intervention. In the intro psych class, the Worked Example condition showed more improvement than the Analogical Comparison condition; however, this was the only difference across the interventions for the two samples. Although study design discrimination generally improved across the

interventions, it was still imperfect at posttest; this could have critical implications for correlation-causation discrimination.

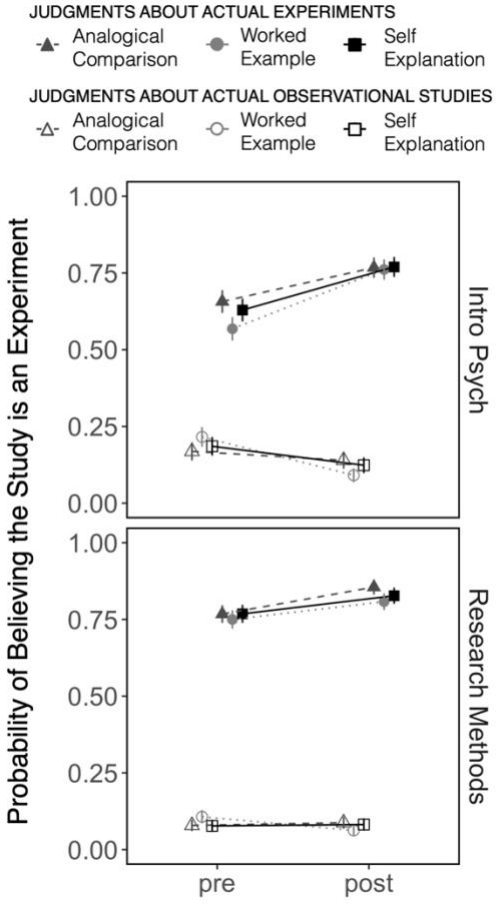


**Figure 7 Believed Study Design Discrimination**

**Table 1 Mixed Effects Models Testing Believed Study Design Discrimination**

| | Intro Psych | | | Research Methods | | |
|---|---|---|---|---|---|---|
| Predictor | Odds Ratio | 95% CI | *p* | Odds Ratio | 95% CI | *p* |
| Model 1: Block × Actual Design | | | | | | |
| Block (B) | 1.12 | 0.88, 1.42 | .350 | 1.21 | 0.93, 1.57 | .152 |
| Actual Design (A) | 11.93 | 6.91, 20.61 | **<.001** | 94.60 | 40.79, 219.41 | **<.001** |
| Block × Actual Design | 4.76 | 2.93, 7.72 | **<.001** | 2.02 | 1.20, 3.40 | **.008** |
| Model 2: Block × Actual Design × Intervention | | | | | | |
| B × A × I (WE vs AC) | 3.26 | 1.10, 9.66 | **.033** | 1.58 | 0.49, 5.05 | .441 |
| B × A × I (SE vs AC) | 1.52 | 0.52, 4.40 | .440 | 0.89 | 0.28, 2.88 | .848 |
| B × A × I (SE vs WE) | 0.47 | 0.16, 1.37 | .166 | 0.56 | 0.17, 1.82 | .339 |

Note. Boldface = $p < .05$. I = Intervention; AC = Analogical Comparison; WE = Worked Example; SE = Self Explanation. Results for effects and interactions not listed here are on OSF.

### 2.2.2 Correlation-Causation Discrimination for the Vignettes

Participants made three correlation-causation judgments for each vignette: support for the journalist's conclusion, likelihood of a successful plan, and support for a causal conclusion. We transformed the judgments to be on a scale of -3 to +3; participants should make more positive judgments (more causal) for experiments and more negative judgments (less causal) for observational studies.

Our original analysis plan tested for correlation-causation discrimination by comparing judgments for actual observational studies versus actual experiments. However, this analysis is unable to fully detect correlation-causation discrimination abilities. Consider a hypothetical participant who sometimes confuses a vignette describing an experiment for an observational study, or vice versa, but who correctly makes positive judgments for studies that they believe to be experiments and negative judgments for studies that they believe to be observational. An

analysis that tests judgments based on actual design tests if participants correctly identify the study type and correctly makes correlation-causation discrimination judgments; but if they get one part correct and the other incorrect, this analysis will not be able to detect their partial knowledge. Thus, we decided to analyze correlation-causation discrimination in two ways; based on the actual study design and the participants' believed study design.

To test for correlation-causation discrimination at pretest and learning during the intervention, we conducted separate mixed-effects regressions for each measure. Model 1 tested for an actual design × block interaction and Model 2 tested for a believed design × block interaction (Table 2). The most complex model that converged included a by-subject slope and random intercept for design × block, and a by-vignette random intercept. There were significant effects of actual design and believed design for both samples across all three measures; at pretest, participants made more positive judgments for actual experiments than actual observational studies and for believed experiments than believed observational studies (Figure 8 and Figure 9). In both samples, there is evidence that correlation-causation discrimination improved according to believed study design, as seen by five out of six significant Believed Study Design × Block interactions in Table 2. However, there is little evidence of improved correlation-causation discrimination according to actual study design; none of the Actual Study Design × Block interactions were significant though some were close. In sum, correlation-causation discrimination did improve after the interventions, but only partially, because complete correlation-causation discrimination also requires accurate study design discrimination.

To test for differences in learning across the three interventions, we tested the three-way interactions in Table 3. These regressions test whether the amount of correlation-causation discrimination increased more in one model versus another with pairwise comparisons. The

maximal model converged for almost all regressions and included a by-subject slope and random intercept for design × block, and a by-vignette random intercept. The Self Explanation (SE) condition was the best intervention for the intro psych sample. Regarding the actual design analyses, SE was better than Worked Example (WE) for all three measures, and better than Analogical Comparison (AC) on the causal conclusion measure; all the other 5 comparisons were not significant. For believed design, there were no differences for any of the 9 comparisons.

However, for the research methods sample, SE was the worst for improving correlation-causation discrimination. It was significantly worse than WE on the causal conclusion measure for the actual design analysis, and significantly worse than AC for the journalist's conclusion and causal conclusion measures for the believed design analysis; all the other 15 comparisons were not significant.

In sum, participants exhibited significant correlation-causation discrimination at pretest, and this ability improved after the intervention but only partially; correlation-causation discrimination improved based on believed but not actual study design. The Self Explanation condition produced the best learning of the three conditions for the intro psych class but the worst for the research methods class. That said, for the most part the three conditions all performed fairly similarly.

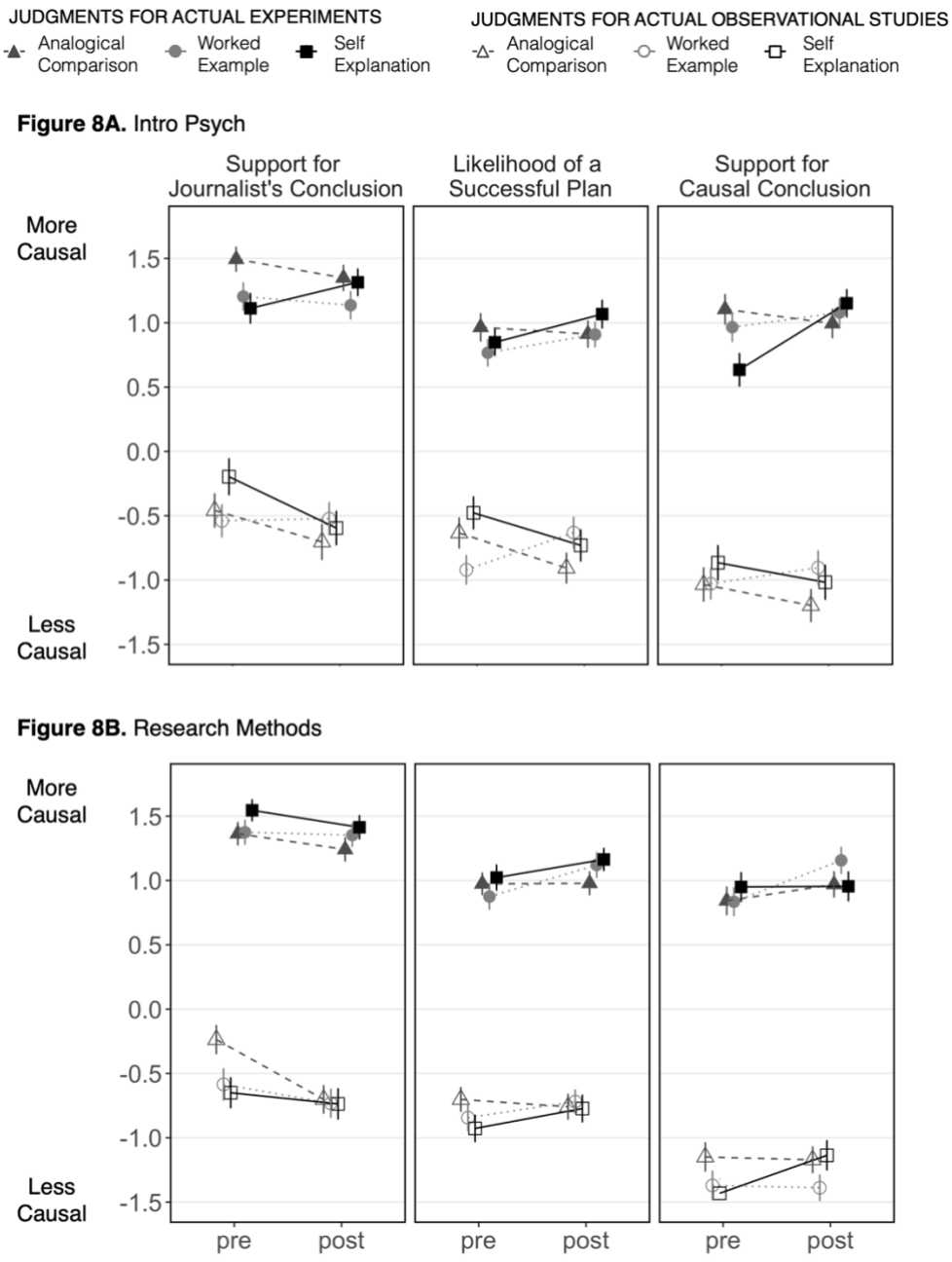


**Figure 8 Correlation Causation Discrimination by Actual Study Design**

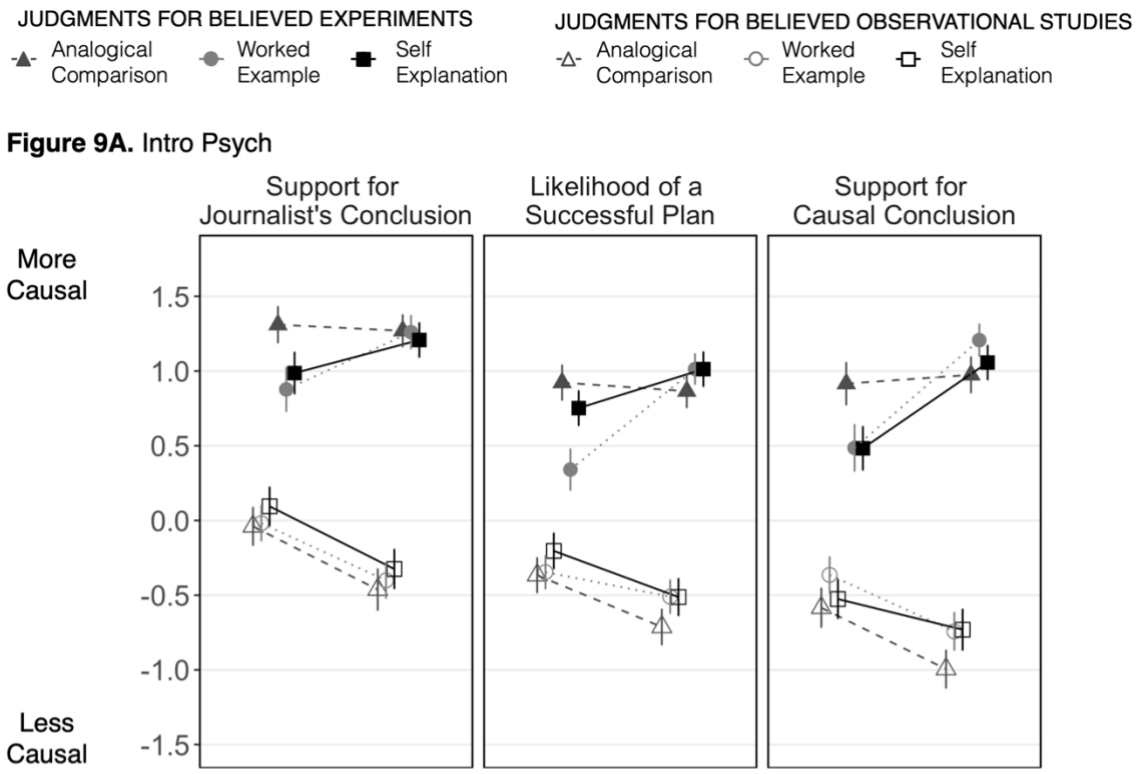


**Figure 9 Correlation Causation Discrimination by Believed Study Design**

**Table 2 Mixed Effects Models Testing for Correlation-Causation Discrimination**

| | Intro Psych | | | Research Methods | | |
|---|---|---|---|---|---|---|
| Dependent Measure | *β* | 95% CI | *p* | *β* | 95% CI | *p* |
| Actual Study Design from Model 1 [Actual Study Design × Block] | | | | | | |
| Journalist's Conclusion | 0.90 | 0.73, 1.08 | **<.001** | 1.03 | 0.78, 1.29 | **<.001** |
| Plan Success | 0.89 | -0.03, 0.24 | **<.001** | 1.04 | 0.69, 1.38 | **<.001** |
| Causal Conclusion | 0.96 | 0.75, 1.16 | **<.001** | 1.12 | 0.87, 1.36 | **<.001** |
| Interaction from Model 1 [Actual Study Design × Block] | | | | | | |
| Journalist's Conclusion | 0.11 | -0.02, 0.24 | .092 | 0.08 | -0.03, 0.19 | .159 |
| Plan Success | 0.11 | -0.03, 0.24 | .113 | 0.03 | -0.08, 0.14 | .552 |
| Causal Conclusion | 0.12 | 0.00, 0.25 | .055 | 0.04 | -0.07, 0.14 | .506 |
| Believed Study Design from Model 2 [Believed Study Design × Block] | | | | | | |
| Journalist's Conclusion | 0.27 | 0.15, 0.40 | **<.001** | 0.44 | 0.34, 0.55 | **<.001** |
| Plan Success | 0.27 | 0.16, 0.39 | **<.001** | 0.45 | 0.34, 0.56 | **<.001** |
| Causal Conclusion | 0.26 | 0.14, 0.38 | **<.001** | 0.62 | 0.51, 0.74 | **<.001** |
| Interaction from Model 2 [Believed Study Design × Block] | | | | | | |
| Journalist's Conclusion | 0.19 | 0.05, 0.33 | **.007** | 0.13 | 0.02, 0.24 | **.023** |
| Plan Success | 0.26 | 0.12, 0.39 | **<.001** | 0.14 | 0.03, 0.25 | **.015** |
| Causal Conclusion | 0.20 | 0.06, 0.34 | **.007** | 0.05 | -0.07, 0.16 | .421 |

Note. Boldface = $p < .05$. Positive $\beta$ for main effect of design is evidence of correlation-causation discrimination. Positive $\beta$ for design × block interaction is evidence of learning. The effect of block is not included in the table because it tests for bias (not correlation-causation discrimination) and none of the effects were significant; see OSF for full results.

**Table 3 Mixed Effects Models Comparing the Three Interventions**

| | Intro Psych | | | Research Methods | | |
|---|---|---|---|---|---|---|
| Dependent Measure | *β* | 95% CI | *p* | *β* | 95% CI | *p* |
| Model 3: Actual Design × Block × Intervention (WE vs AC) | | | | | | |
| Journalist's Conclusion | -0.08 | -0.41, 0.25 | .623 | -0.12 | -0.39, 0.15 | .380 |
| Plan Success | -0.22 | -0.56, 0.13 | .218 | 0.03 | -0.24, 0.31 | .807 |
| Causal Conclusion | -0.01 | -0.32, 0.29 | .929 | 0.09 | -0.17, 0.36 | .492 |
| Model 3: Actual Design × Block × Intervention (SE vs AC) | | | | | | |
| Journalist's Conclusion | 0.28 | -0.05, 0.61 | .094 | -0.20 | -0.47, 0.08 | .161 |
| Plan Success | 0.15 | -0.20, 0.49 | .403 | -0.04 | -0.32, 0.23 | .761 |
| Causal Conclusion | 0.33 | 0.02, 0.63 | **.037** | -0.22 | -0.49, 0.05 | .111 |
| Model 3: Actual Design × Block × Intervention (SE vs WE) | | | | | | |
| Journalist's Conclusion | 0.36 | 0.04, 0.69 | **.029** | -0.07 | -0.35, 0.20 | .599 |
| Plan Success | 0.36 | 0.02, 0.70 | **.037** | -0.08 | -0.35, 0.20 | .586 |
| Causal Conclusion | 0.34 | 0.03, 0.64 | **.029** | -0.31 | -0.58, -0.04 | **.025** |
| Model 4: Believed Design × Block × Intervention (WE vs AC) | | | | | | |
| Journalist's Conclusion | 0.06 | -0.28, 0.40 | .741 | -0.15 | -0.42, 0.12 | .279 |
| Plan Success | 0.13 | -0.23, 0.48 | .484 | 0.01 | -0.26, 0.28 | .939 |
| Causal Conclusion | 0.17 | -0.17, 0.51 | .327 | -0.07 | -0.35, 0.21 | .605 |
| Model 4: Believed Design × Block × Intervention (SE vs AC) | | | | | | |
| Journalist's Conclusion | 0.06 | -0.28, 0.40 | .725 | -0.31 | -0.58, -0.04 | **.027** |
| Plan Success | 0.02 | -0.33, 0.37 | .916 | -0.03 | -0.29, 0.24 | .850 |
| Causal Conclusion | 0.08 | -0.26, 0.42 | .639 | -0.30 | -0.58, -0.02 | **.035** |
| Model 4: Believed Design × Block × Intervention (SE vs WE) | | | | | | |
| Journalist's Conclusion | <0.01 | -0.34, 0.34 | .985 | -0.16 | -0.44, 0.12 | .257 |
| Plan Success | -0.11 | -0.46, 0.24 | .549 | -0.03 | -0.31, 0.24 | .804 |
| Causal Conclusion | -0.09 | -0.42, 0.25 | .606 | -0.23 | -0.51, 0.05 | .113 |

Note. Boldface = $p < .05$. SE = Self Explanation; WE = Worked Example; AC = Analogical Comparison. Positive correlation coefficients mean that correlation-causation discrimination improved more for the condition listed first than second. Results for effects and interactions not listed here are on OSF.

### 2.2.3 Pre-Test Explanations of "Correlation Doesn't Equal/Imply Causation"

The goal of the intervention was to use causal diagrams to teach undergraduate students *why* correlation does not imply causation. Before the intervention, participants were asked to explain the phrase "correlation doesn't equal/imply causation" – the purpose of this measure was to gauge participants' prior knowledge. We identified two themes from a subset of participants' responses and two coders classified the remaining responses (92% agreement; $\kappa = 0.75$). Three responses were not coded due to lack of effort by the participant. The two main themes identified were 1) that there are alternative possible explanations for a correlation (e.g., mentioning the possibility of a coincidence, reverse causality, confound), and 2) that an experiment (or "controlled study" or manipulation of the IV/random assignment to conditions) is needed to make a causal claim.

83% of intro psych participants and 76% of research methods participants mentioned neither theme in their responses; these participants simply re-stated the phrase in their own words. 15% of intro psych participants and 18% of research methods participants only mentioned Theme 1. Less than 1% of intro psych participants and 4% of research methods students only mentioned Theme 2. Very few mentioned both themes. In sum, even if participants know correlation does not equal causation, they struggle to explain why. Because few participants mentioned Theme 2 or both themes, we treated their explanations as binary (i.e., any theme or not) in subsequent analyses.

### 2.2.4 Pre-Test Explanations and Correlation-Causation Discrimination

We hypothesized that at pretest, the quality of participants' explanations for "correlation doesn't equal/imply causation" would be correlated with their correlation-causation discrimination

abilities. Recall that at pretest, students had some correlation-causation discrimination because they made stronger causal judgments for the experiment vignettes than the observational study vignettes. However, their pretest explanations were quite poor in both classes. To assess whether pretest explanation quality was predictive of pretest correlation-causation discrimination, we ran mixed-effects regressions for the three correlation-causation measures and separately tested for actual design × explanation accuracy (Model 1) and believed design × explanation accuracy (Model 2) interactions. The maximal model converged in all cases, with by-subject random slopes and random intercepts for study design, and a by-vignette random intercept (Table 4).

In the research methods class, three of the six design × explanation accuracy interactions were significant; believed design × explanation accuracy was significant for the causal conclusion and journalist's conclusion measures, and actual design × explanation accuracy was significant for the causal conclusion measure. In all three cases, at pretest, research methods students with better quality explanations had better correlation-causation discrimination. This suggests that at least to some extent, the free-response explanations are capturing similar knowledge as the correlation-causation discrimination measures.

However, only half of the interactions were significant for the research methods sample, and the journalist's conclusion measure was only barely significant ($p = .047$). Furthermore, none of the interactions were significant for the intro psych class, meaning that explanation accuracy was not a strong predictor of pretest correlation-causation discrimination. This aligns with the fact that intro psych students had especially poor pretest explanation accuracy, despite exhibiting correlation-causation discrimination when making judgments about the vignettes at pretest. Altogether, these results suggest that the free response measure of "What does correlation does not

equal/imply causation mean to you?" may be more sensitive than the correlation-causation discrimination measures for capturing gaps in participants' understanding.

**Table 4 Explanation Accuracy as a Predictor of Pre-Test Correlation-Causation Discrimination**

| | Intro Psych | | | Research Methods | | |
|---|---|---|---|---|---|---|
| Predictor | *β* | 95% CI | *p* | *β* | 95% CI | *p* |
| Interaction from Model 1 [Actual Study Design × Explanation Accuracy] | | | | | | |
| Journalist's Conclusion | 0.09 | -0.22, 0.40 | .557 | 0.18 | -0.05, 0.42 | .129 |
| Plan Success | <0.01 | -0.32, 0.33 | .991 | 0.15 | -0.08, 0.37 | .193 |
| Causal Conclusion | 0.11 | -0.18, 0.15 | .460 | 0.27 | 0.05, 0.50 | **.017** |
| Interaction from Model 2 [Believed Study Design × Explanation Accuracy] | | | | | | |
| Journalist's Conclusion | 0.14 | -0.19, 0.47 | .398 | 0.23 | 0.00, 0.47 | **.047** |
| Plan Success | 0.02 | -0.30, 0.33 | .919 | 0.18 | -0.05, 0.42 | .125 |
| Causal Conclusion | 0.18 | -0.13, 0.48 | .261 | 0.40 | 0.16, 0.64 | **.001** |

Note. Positive *β* means better correlation-causation discrimination for participants who provided an accurate explanation. Results for effects and interactions not listed here are on OSF.

### 2.2.5 Causal Diagrams Task

In the intervention, participants were taught to use causal diagrams to represent alternative explanations for a correlation. If the intervention was successful, they should know that there are more possible explanations for a correlation in an observational study (mechanism, reverse causality, confound) than in an experiment (mechanism). To test this, we analyzed participants' posttest endorsements of the three causal structures as possible explanations for a hypothetical observational study versus a hypothetical experiment. (These questions were only asked at posttest, not pre.) For each causal structure, we ran a mixed-effects logistic regression that tested for an interaction between study design prompt and intervention. The dependent variable was whether the participant endorsed the structure as a possible explanation. The results for the two

samples were quite similar; for conciseness, we present the results of both samples together in Table 5 and Figure 10 (see OSF for separate analyses). The most complex model that converged included a by-subject random intercept.

For the reverse causality and confound diagrams, participants should endorse both structures for the observational study but not the experiment, and this is what was found; the effect of study design was significant for both structures. At the same time, the proportion of endorsement of reverse causality and a confound was less than 50% for observational studies, which means that participants still do not have a good grasp that these structures can explain a statistical relation in an observational study.

In contrast to the reverse causality and confound structures, participants should endorse a mechanism as an explanation for both study designs. However, there was a significant effect of study design; participants were more likely to endorse the mechanism structure for the experiment than for the observational study. This is inconsistent with what they were taught in the intervention, but it is not too concerning because the rate of endorsement of the mechanism structure was high for both experiments and observational studies. None of the interactions were significant, so there is no evidence that one condition worked any better than another.

In sum, though participants had some understanding of different possible causal explanations for experiments vs. observational studies, many participants still did not understand that a confound or reverse causality can explain the results of an observational study.

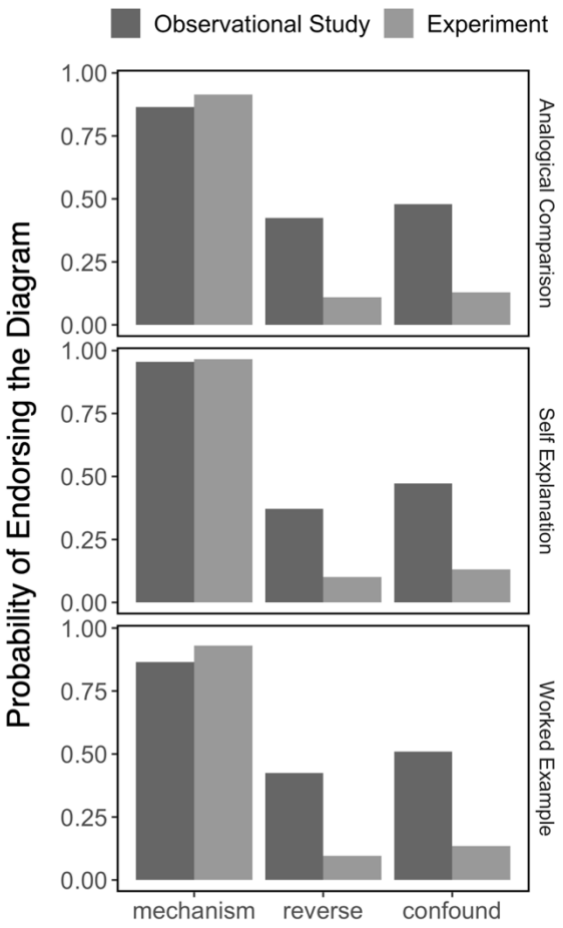


**Figure 10 Endorsement in Causal Diagrams Task**

**Table 5 Testing for Differences in Endorsement of Causal Diagrams**

| Dependent Measure | Odds Ratio | 95% CI | *p* |
|---|---|---|---|
| Study Design | | | |
| Model 1: Mechanism | 5.30 | 1.50, 18.72 | **.010** |
| Model 2: Reverse | 0.17 | 0.10, 0.29 | **<.001** |
| Model 3: Confound | 0.16 | 0.10, 0.27 | **<.001** |
| Study Design × Intervention (AC vs WE) | | | |
| Model 1: Mechanism | 1.78 | 0.30, 10.56 | .524 |
| Model 2: Reverse | 0.85 | 0.40, 1.82 | .673 |
| Model 3: Confound | 0.93 | 0.46, 1.86 | .829 |
| Study Design × Intervention (AC vs SE) | | | |
| Model 1: Mechanism | 0.38 | 0.05, 3.04 | .363 |
| Model 2: Reverse | 1.13 | 0.53, 2.40 | .753 |
| Model 3: Confound | 1.04 | 0.51, 2.10 | .919 |
| Study Design × Intervention (SE vs WE) | | | |
| Model 1: Mechanism | 0.21 | 0.03, 1.80 | .156 |
| Model 2: Reverse | 1.33 | 0.61, 2.88 | .471 |
| Model 3: Confound | 1.12 | 0.56, 2.26 | .751 |

Note. Boldface = $p < .05$. SE = Self Explanation; WE = Worked Example; AC = Analogical Comparison. Intervention was not a significant predictor of endorsement in any of the analyses ($p > .05$), see OSF for the full results.

## 2.3 Discussion

The main goal of Experiment 1 was to test the efficacy of different interventions aimed at improving correlation-causation discrimination abilities. There were a few key findings.

First, both intro psych and research methods were able to discriminate between correlation and causation at pretest. Although prior research has shown that people tend to make causal

judgments about observational studies, which is typically framed as a bias in reasoning (Halonen et al., 2013), we found that students knew to make more causal judgments about experiments than observational studies. Regarding scientific reasoning abilities, this result establishes that even before the intervention, students at least somewhat understand that correlation does not imply causation. On the other hand, most students were unable to adequately explain "why correlation does not imply/equal causation" at pretest. Thus, this qualitative measure may be more sensitive at detecting students' understanding and suggests that there is room for improvement in students' understanding. Additionally, even though correlation-causation discrimination for the vignettes was present at pretest, there was also room for improvement; participants could learn to make even more causal judgments for experiments and less causal judgments for observational studies.

Second, for the vignettes, students somewhat struggled with identifying whether the study design was an observational study or an experiment. At pretest, both intro psych and research methods students failed to identify experiments as experiments in at least 25% of cases. Because accurate identification of study design is critical for deciding whether to make causal inferences about statistical relations, inaccurate study design identification means students cannot have a complete understanding of correlation-causation discrimination. Study identification improved at posttest in both classes, but it was still imperfect, particularly for identification of experiments. This had critical implications for learning; collapsing across the three interventions, correlation-causation discrimination only improved from pre to post based on participants' beliefs about study design rather than the actual design of the study. Thus, participants only partially learned about correlation-causation discrimination; they learned to make more causal judgments about studies they think are experiments and less causal judgments about studies they think are observational

studies. To have a complete understanding, they must also learn how to identify the correct study design so that they can be more accurate in their correlation-causation discrimination.

Third, participants' correlation-causation discrimination did get better from pre to posttest, but only based on participants' beliefs about study design, if collapsing the analyses across the three interventions. However, even though there was some improvement in correlation-causation discrimination, at posttest, participants made causal claims about actual or believed observational studies in 22-25% of cases and failed to make causal claims about actual or believed experiments in 25% of cases. Thus, there was room for improvement in correlation-causation discrimination based on both participants' beliefs about study design and the actual design of the study. One possible reason why correlation-causation discrimination was not very strong, in addition to participants having difficulty with correctly identifying study designs, is that students did not fully understand a key lesson in the tutorial. After the intervention, less than half of participants endorsed the reverse causality and confound structures as possible reasons for a statistical relation in an observational study, which was a critical message in the tutorial. In Experiment 2, we improved the tutorial in several ways to further improve posttest correlation-causation discrimination and study design discrimination.

Fourth, we found that different methods of instruction were successful for the two classes. In the intro psych class, participants learned the most about correlation-causation discrimination in the Self Explanation condition. In the research methods class, participants learned more in the Analogical Comparison and Worked Example conditions compared to the Self Explanation condition. Thus, considering prior knowledge or abilities may be key in determining best methods of instruction for teaching about why "correlation doesn't imply causation".

## 3.0 Experiment 2

In Experiment 2, we made a number of modifications to improve the Self Explanation intervention, test learning with additional dependent measures, and compare the Self Explanation intervention against a control with no intervention.

First, in Experiment 2 and 3 we focused on the Self Explanation condition. In Experiment 1 there was mixed evidence about which intervention was best depending on the dependent measure, analysis, and sample. All three of the interventions were quite similar – they had the same tutorial and participants did the same practice problems. Thus, for Experiments 2 and 3 we chose to focus on one condition, making improvements to that condition, and testing new dependent measures. In Experiments 2 and 3 we only had access to an Intro Psych sample, and we chose to focus on the Self Explanation condition because it was the most promising for Intro Psych in Experiment 1. We now refer to the Self Explanation condition as the intervention.

Second, Experiment 1 compared three interventions, but did not compare them to a control condition. In Experiment 2 we compared a single intervention (Self Explanation) to a control to isolate the influence of learning due to the intervention vs. learning merely due to repeated practice with the vignettes and dependent measures.

Third, we made several changes to improve the intervention. In Experiment 1, participants had difficulty discriminating between observational studies and experiments, which is fundamental for correlation-causation discrimination. Participants also did not fully understand that they should endorse reverse causality and confounds as possible explanations for a correlation in an observational study. We made changes to directly address both of these.

Fourth, we added a third set of vignettes called ‘implicit experiments’, that describe experiments without explicitly mentioning random assignment. In actual media articles and scientific abstracts, random assignment is often implied rather than being explicitly stated, so it is important to assess whether students recognize such cases as experiments.

Fifth, in Experiment 2 we tested if the intervention would also help students be able to design experiments and observational studies, by asking them to design both at pre and at post. This is similar to being able to identify if a study is experimental or observational, but also requires more original thinking.

## 3.1 Methods

### 3.1.1 Participants, Attention Check, and Exclusion Criteria

A total of 399 participants completed Experiment 2 as part of a requirement for an Introductory to Psychology course. We added an attention check measure, the “sports participation” question from Oppenheimer et al. (2009), at the beginning of the study. Most participants (98%) passed the attention check on the first attempt. If they did not pass after five attempts, they could still complete the study but were excluded from analyses ($N$ = 3); this was registered on OSF.

### 3.1.2 Design and Intervention

Participants were randomly assigned to either the Self Explanation Intervention or the Control condition. The intervention consisted of both the tutorial and Self Explanation practice problems from Experiment 1, with some modifications. In the tutorial, we made two major changes. First, we added instructions for how to discriminate between observational studies and experiments, including how to discriminate study designs when random assignment is only implied. Second, we added an example to show how random assignment in an experiment rules out alternative explanations (reverse causality, confounds) for a statistical relation. In the practice problems, we broke up some of the questions into multiple parts so that participants were more likely to answer each part of the question.

The control condition did not contain the tutorial or the Self Explanation Intervention; participants simply did the pre-test and then the post-test questions immediately afterwards. Since this experiment was given in an educational setting, participants in the Control condition received the tutorial and practice problems after the posttest measures as a form of instruction for the class.

### 3.1.3 Pre-Test and Post-Test Measures

At pretest and posttest, participants completed the same three tasks. First, they read and made judgments about six randomly ordered vignettes – two observational studies, two explicit experiments, and two implicit experiments. We used the same vignettes for the observational studies and explicit experiments from Experiment 1 but removed any references to sample size in the vignettes. We counterbalanced the two blocks of six vignettes to pre versus post. After reading each vignette, participants answered the three correlation-causation discrimination measures from

Experiment 1 and whether they thought the study was an "Observational/Correlational Study" or an "Experiment."[5] We did not include the qualitative "explain your reasoning" response for the plan success measure.

Second, participants designed two observational studies and two experiments at pretest and posttest. They were told to imagine themselves as a researcher who was testing a hypothesis (e.g., "people who consume caffeine will have disrupted sleep") and were instructed, "In 1-2 sentences, design an [observational study (not an experiment) / experiment (not an observational study)] that tests this hypothesis". We counterbalanced the scenarios so that half of participants were told to design an observational study for a particular hypothesis, and the other half were told to design an experiment.

Third, participants completed the causal structure task, which was embedded within the design an experiment task. After designing a study, participants were asked "Imagine that you ran your proposed [observational study/experiment] and you found a statistical relation … which of the following diagrams could possibly explain this statistical relation?" If participants were told to design an observational study, they should endorse all three structures as a possible explanation for a correlation; if participants were told to design an experiment, they should only endorse the mechanism structure as a possible explanation.

[5] In Study 1 it is possible that when asked to identify the study design participants answered in a superficial way by noticing that some vignettes included the words "random assignment" and then selecting the "Experiment / Randomized Controlled Trial" option. In Experiments 2 and 3 we removed "Randomized Controlled Trial" and simply called it an "Experiment" to avoid this issue.

## 3.2 Results

Before analyzing the data, we decided to use two additional exclusion criteria that were not registered on OSF, aside from the attention check which was. The study was designed so that participants could complete it within 40-50 minutes; the median time of completion was 49 minutes. We excluded participants from analyses if they took less than 20 minutes total ($N$ = 11), because this timeframe seemed indicative of very low effort. After reading participants' responses in the "design a study" task, we also dropped participants who demonstrated insufficient effort on more than one item in the same block ($N$ = 9) or did not provide enough information to code their responses for more than two items in the same block ($N$ = 9). The final analyses included 367 participants, with 181 in the Self Explanation condition and 186 in the Control condition.

### 3.2.1 Ability to Correctly Identify the Study Design from Vignettes

We conducted similar analyses as in Experiment 1 to test whether participants were able to identify the study design ('believed study design') from the actual design in the vignettes. In one mixed effects model, we tested for a three-way interaction between block, actual design, and intervention. However, we ran separate analyses to compare discrimination for implicit experiments versus observational studies and explicit experiments versus observational studies (Table 6).

The most complex model that converged included a by-subject random slope and random intercept, and a by-vignette random intercept. There was a significant effect of actual design for both analyses. This means that at pretest (due to the contrast coding), participants discriminated between observational studies (triangles in Figure 11) and explicit experiments (squares), and also

between observational studies and implicit experiments (circles). There were also significant interactions between actual design and block, which means that participants learned to distinguish observational studies from both implicit and explicit experiments better at post than pre. As is obvious from Figure 11, participants learned to better identify experiments; their ability to identify observational studies was already quite good at pre and did not change.

For both analyses, there was no significant three-way interaction, meaning that the amount of learning was the same in the intervention and control conditions. This suggests that our modification to the intervention to teach students how to identify observational studies versus experiments unfortunately did not work. Still, participants got better at identifying study types in both conditions, suggesting that practice even without feedback was beneficial.

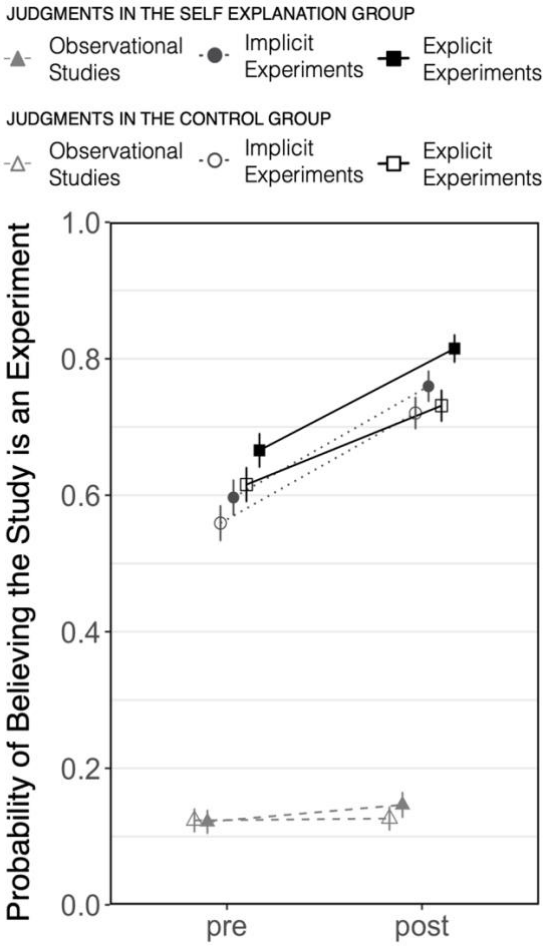


**Figure 11 Believed Study Design Discrimination**

**Table 6 Mixed Effects Model Testing Believed Study Design Discrimination**

| | Explicit Experiment vs. Observational Study | | | Implicit Experiment vs. Observational Study | | |
|---|---|---|---|---|---|---|
| Predictor | Odds Ratio | 95% CI | *p* | Odds Ratio | 95% CI | *p* |
| Block | 0.67 | 0.56, 0.80 | **<.001** | 0.66 | 0.55, 0.79 | **<.001** |
| Actual Design | 0.09 | 0.06, 0.15 | **<.001** | 0.11 | 0.09, 0.15 | **<.001** |
| Block × Actual | 0.58 | 0.40, 0.84 | **.004** | 0.52 | 0.36, 0.73 | **<.001** |
| B × A × Intervention | 1.03 | 0.51, 2.09 | .933 | 1.26 | 0.62, 2.55 | .523 |

Note. Boldface = $p < .05$. B = Block (pre versus post); A = Actual Design (Explicit Experiment versus Observational or Implicit Experiment versus Observational). Results for effects and interactions not listed here are on OSF.

### 3.2.2 Correlation-Causation Discrimination for the Vignettes

We conducted similar analyses as in Experiment 1 to test correlation-causation discrimination (Figure 12). For discrimination based on the actual design, we compared observational studies vs. explicit experiments and observational studies vs. implicit experiments (Table 7). For discrimination based on the believed design, we compared the studies that participants thought were observational vs. those that they thought were experiments, regardless of whether they were actually observational studies, explicit experiments, or implicit experiments (Table 8).

Both models tested for a three-way interaction to assess if learning improved more in the intervention than control. The most complex model that converged included a by-subject random slope and intercept for the interaction between block and design, and a by-vignette random intercept. There were significant effects of actual design (Table 7) and believed design (Table 8) across all three measures and analyses. This means that at pretest, participants could discriminate

between observational studies and both types of experiments based both on the actual design and believed design.

The actual design × block interactions, which tested for improved discrimination based on actual design, was only significant for one of the six cases: only comparing implicit experiments versus observational studies and only for the journalist's conclusion measure (Table 7). This was mainly driven by the journalist's conclusion judgments for implicit experiments becoming more causal (circles in Figure 12A). Overall, given that only one of these six interactions was significant, it means that there still was not considerable learning to discriminate experiments vs. observational studies.

For the analyses of learning based on believed design, there were significant interactions with block for two out of the three measures: journalist's conclusion and plan success (Table 8). This appears to be driven both by participants learning to make more causal judgments for believed experiments (squares in Figure 12B) and less causal judgments for believed observational studies (triangles).

None of the three-way interactions, which tested whether learning improved more in the intervention than control group, were significant for discrimination based on the actual design. This fits with the fact that there was little improvement in the actual design in the two-way interactions. For believed study design, the three-way interaction was significant for the journalist's conclusion measure. Participants in the Self Explanation intervention condition (black squares in Figure 12B) learned to make more causal judgments for believed experiments but not for the control (white squares). The other two three-way interactions were not significant.

Altogether, these results suggest that the intervention improved correlation-causation discrimination a bit (for the journalist's conclusion for believed study design), and there was some

learning due to practice across both the intervention and control in two other cases (journalist's conclusion for actual study design for implicit experiments versus observational studies and plan success for believed study design). Still, there is considerable room for improvement.

**Figure 12A.** Judgments by Actual Study Design

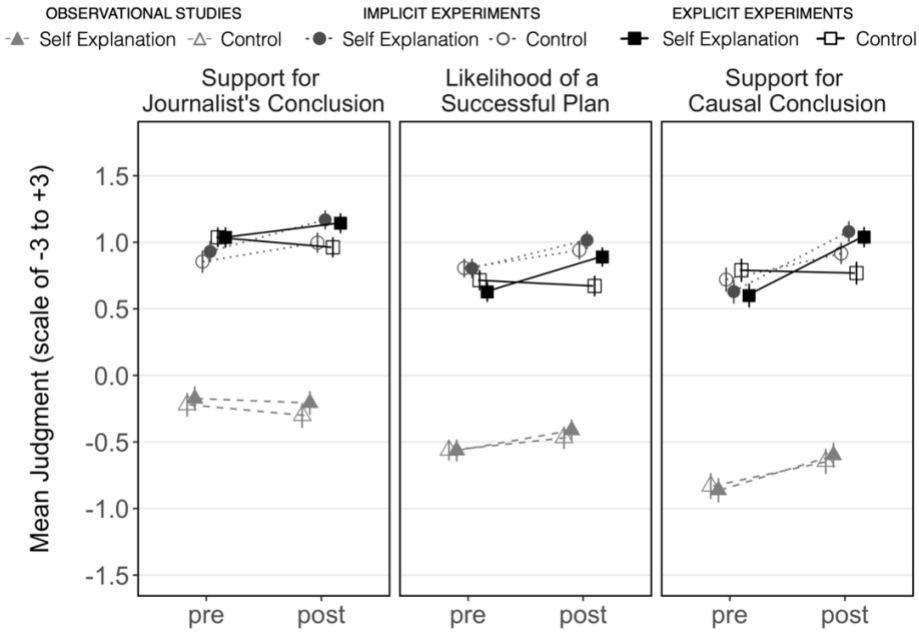


**Figure 12B.** Judgments by Believed Study Design

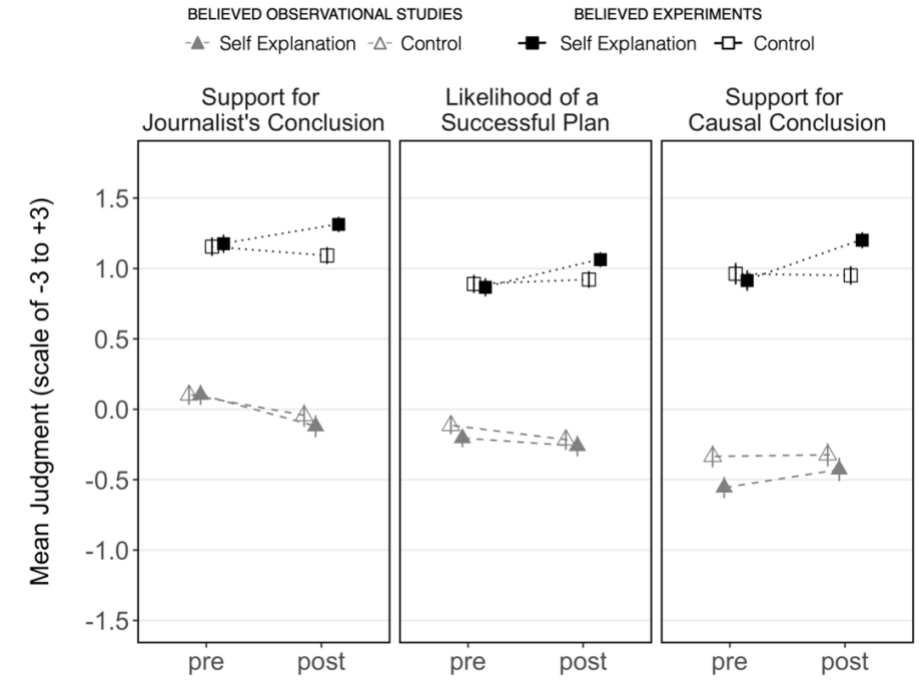


**Figure 12 Correlation Causation Discrimination in Study 2**

**Table 7 Mixed Effects Model Testing for Correlation-Causation Discrimination by Actual Design**

| | Explicit Experiment vs. Observational Study | | | Implicit Experiment vs. Observational Study | | |
|---|---|---|---|---|---|---|
| Dependent Measure | *β* | 95% CI | *p* | *β* | 95% CI | *p* |
| Actual Study Design from Model [Actual Study Design × Block × Intervention] | | | | | | |
| Journalist's Conclusion | 0.70 | 0.46, 0.94 | **<.001** | 0.63 | 0.39, 0.87 | **<.001** |
| Plan Success | 0.75 | 0.38, 1.12 | **<.001** | 0.84 | 0.54, 1.15 | **<.001** |
| Causal Conclusion | 0.79 | 0.52, 1.06 | **<.001** | 0.79 | 0.54, 1.03 | **<.001** |
| Actual Study Design × Block | | | | | | |
| Journalist's Conclusion | 0.05 | -0.06, 0.17 | .341 | 0.14 | 0.03, 0.26 | **.016** |
| Plan Success | 0.01 | -0.11, 0.13 | .879 | 0.02 | -0.09, 0.14 | .701 |
| Causal Conclusion | 0.01 | -0.10, 0.11 | .918 | 0.05 | -0.06, 0.16 | .352 |
| Actual Study Design × Block × Intervention | | | | | | |
| Journalist's Conclusion | 0.07 | -0.16, 0.29 | .560 | -0.03 | -0.25,0.20 | .829 |
| Plan Success | 0.13 | -0.10, 0.37 | .277 | -0.05 | -0.29, 0.19 | .680 |
| Causal Conclusion | 0.17 | -0.04, 0.38 | .119 | 0.09 | -0.13, 0.31 | .422 |

Note. Boldface = $p < .05$. Positive $\beta$ for main effect of design is correlation-causation discrimination. Positive $\beta$ for design × block interaction is learning. Results for effects and interactions not listed here are on OSF.

**Table 8 Mixed Effects Model Testing for Correlation-Causation Discrimination by Believed Design**

| Dependent Measure | $\beta$ | 95% CI | $p$ |
|---|---|---|---|
| Believed Study Design from Model [D × B × Intervention] | | | |
| Journalist's Conclusion | 0.44 | 0.35, 0.52 | **<.001** |
| Plan Success | 0.40 | 0.32, 0.48 | **<.001** |
| Causal Conclusion | 0.53 | 0.45, 0.62 | **<.001** |
| Believed Study Design × Block | | | |
| Journalist's Conclusion | 0.11 | 0.02, 0.21 | **.018** |
| Plan Success | 0.10 | 0.00, 0.19 | **.044** |
| Causal Conclusion | -0.02 | -0.11, 0.07 | .711 |
| Believed Study Design × Block × Intervention | | | |
| Journalist's Conclusion | 0.24 | 0.05, 0.43 | **.012** |
| Plan Success | 0.11 | -0.04, 0.15 | .221 |
| Causal Conclusion | 0.14 | -0.04, 0.32 | .136 |

Note. Boldface = $p < .05$. Positive $\beta$ for main effect of design (D) is correlation-causation discrimination. Positive $\beta$ for design × block (B) interaction is learning. Results for effects and interactions not listed here are on OSF.

### 3.2.3 Design a Study Task

At pretest and posttest, participants were prompted to design two observational studies and two experiments, to test if participants got better at designing both sorts of studies. A total of 2871 responses were coded as either an experiment or an observational study (92% agreement; $\kappa = 0.79$), after dropping 65 items due to lack of effort ($N = 4$) or not enough information to code ($N = 61$). A response was coded as an experiment if there was implied or explicit random assignment and there was more than one participant in each group.

We ran a regression testing for a three-way interaction between design prompt, block, and intervention, using our standard contrast codes. The dependent measure was whether they designed

an experiment or not[6] (Model 1 in Table 9). The most complex model that converged included a by-subject random intercept. The effect of design prompt was significant, which means that at pretest, they were more likely to design an experiment when told to design an experiment (squares in Figure 13) versus when told to design an observational study (triangles).

There was no significant interaction between design prompt and block, which tested for improvement in design abilities at posttest, and there was no three-way interaction. However, this is likely because participants were already able to design observational studies when prompted to do so at pre so could not exhibit learning. To address this, we tested for a block × intervention interaction using the subset of cases in which participants were prompted to design an experiment (squares in Figure 13).[7] The most complex model that converged included a by-subject random slope and intercept for block. There was a significant block × intervention interaction; participants in the Self Explanation condition got considerably better at designing experiments at posttest, but there was no change in the Control condition (odds ratio = 7.37, 95% CI [1.77, 30.63], $p$ = .006).

In sum, participants could discriminate between designing an observational study and an experiment at pretest; they were particularly adept at designing observational studies. Additionally, participants in the intervention learned how to better design experiments at posttest; however, they clearly still have much room to improve.

[6] In the registration, we used accuracy (designing an experiment vs. observational study when asked to) as the dependent variable. After visualizing the data, we believed it was more straightforward and better aligned with other analyses to present our findings using probability of designing an experiment instead.

[7] This analysis was not in the preregistration.

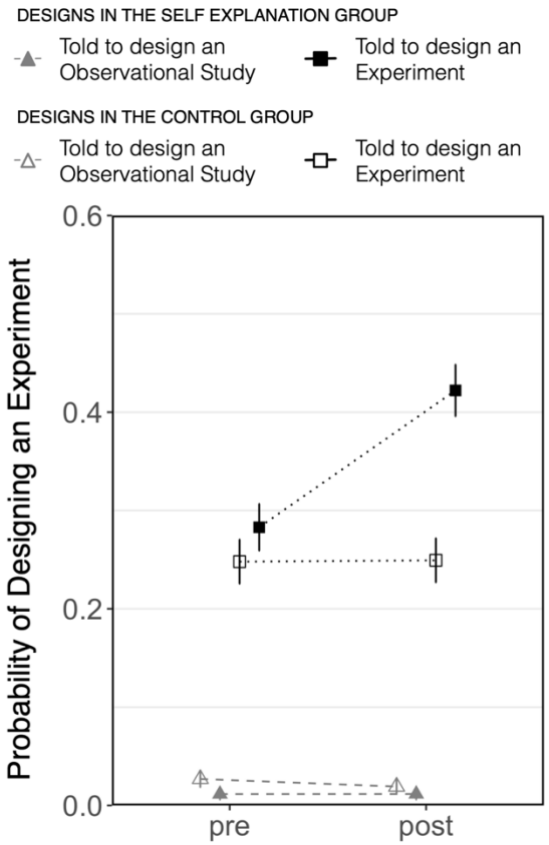


**Figure 13 Participants' Study Designs**

**Table 9 Mixed Effects Model Testing Probability of Designing an Experiment**

| Predictor | Odds Ratio | 95% CI | *p* |
|---|---|---|---|
| Block (B) | 1.20 | 0.74, 1.93 | **.461** |
| Design Prompt (D) | 71.43 | 34.88, 146.25 | **<.001** |
| Block × Design | 2.28 | 0.88, 5.91 | .091 |
| B × D × Intervention | 1.94 | 0.29, 13.05 | .495 |

Note. Boldface = $p < .05$. Results for effects and interactions not listed here are on OSF.

### 3.2.4 Causal Structures Task

After designing a study, participants were asked whether to endorse three causal structures (mechanism, reverse causality, confound) as potential explanations for a significant statistical

result. In the intervention, participants were taught that they should endorse all three structures for an observational study and only the mechanism structure for an experiment.

We ran separate analyses for the three causal structures, using our standard contrast coding. The regression tested for a three-way interaction between design prompt, block, and intervention, with endorsement of the structure (possible versus not possible explanation) as the dependent variable. The most complex model that converged included a by-subject random slope and random intercept for design prompt, and a by-scenario random intercept (Table 10). There were significant effects of design prompt for the reverse causality and confound structures but not the mechanism structure, which means that at pretest, there was some understanding that a mechanism explanation is equally possible for both designs, but reverse causality and confound explanations are more likely in observational studies (triangles in Figure 14) and not in experiments (squares).

The design prompt × block interactions were significant for the reverse causality and confound structures, but not the mechanism structure, suggesting learning. Furthermore, both interactions were qualified by significant three-way interactions. In the Self Explanation condition, at posttest, participants were more likely to endorse reverse causality and confound explanations for observational studies (grey triangles in Figure 14) and less likely to endorse these structures for experiments (black squares). In the Control condition, the change in posttest judgments for observational studies (white triangles) and experiments (white squares) was either nonexistent or much smaller than in the Self Explanation condition.

In sum, our changes to the intervention to explain why random assignment can rule out alternative explanations in an experiment but not observational studies were successful, as participants in the Self Explanation condition improved. Still, there is additional room for

improvement; in particular, posttest endorsement for reverse causality could be higher for observational studies and endorsement for a confound could be lower for the experiments.

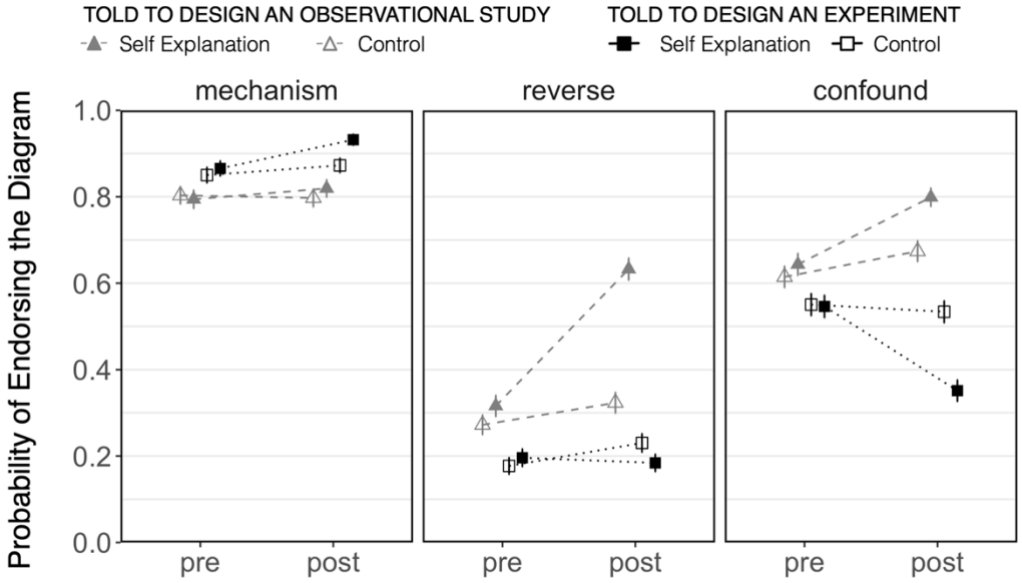


**Figure 14 Endorsement in Causal Diagrams Task**

**Table 10 Testing for Differences in Endorsement of Causal Diagrams**

| Structure | Odds Ratio | 95% CI | *p* |
|---|---|---|---|
| Design Prompt in [Design Prompt × Block × Intervention] | | | |
| Mechanism | 0.73 | 0.34, 1.56 | .413 |
| Reverse | 0.45 | 0.34, 0.60 | **<.001** |
| Confound | 0.63 | 0.50, 0.80 | **<.001** |
| Design Prompt × Block | | | |
| Mechanism | 1.57 | 0.92, 2.67 | .098 |
| Reverse | 0.40 | 0.27, 0.60 | **<.001** |
| Confound | 0.29 | 0.21, 0.42 | **<.001** |
| Design Prompt × Block × Intervention | | | |
| Mechanism | 1.20 | 0.41, 3.46 | .742 |
| Reverse | 0.13 | 0.06, 0.29 | **<.001** |
| Confound | 0.21 | 0.10, 0.42 | **<.001** |

Note. Boldface = $p < .05$. Results for effects and interactions not listed here are on OSF.

## 3.3 Discussion

In Experiment 2, we made several modifications to improve the intervention and test other dependent measures. Some of these changes were more successful than others; there was learning across multiple measures in the Self Explanation condition, but not always due to the intervention.

During the intervention, participants successfully learned which causal structures are possible explanations for a correlation in an observational study versus an experiment. This was a key lesson in the tutorial; we believed that this understanding is fundamental to improving other measures of discrimination. The success of the intervention at teaching this point suggests that the improvements to the tutorial after Experiment 1 were worthwhile, because participants did not perform as well in Experiment 1.

Additionally, the intervention successfully improved participants' ability to design experiments. Participants in both conditions were quite good at designing observational studies at pretest, but only participants the Self Explanation condition improved at designing experiments. Designing the correct type of study, like identifying the correct study design in a vignette, requires understanding that experiments use random assignment whereas observational studies do not. In Experiment 2, we added instructions about how to differentiate between the two designs. This likely helped participants when designing experiments at posttest. However, it had no effect on participants' ability to discriminate between study designs when reading vignettes, because both groups showed the same amount of improvement at posttest.

We hoped that the changes to the tutorial would also improve correlation-causation discrimination, which was quite weak in Experiment 1. However, only one of the three-way interactions was significant in Experiment 2. And like in Experiment 1, the difference in judgments for observational studies versus experiments was not very strong. There was room for

improvement, and our changes to the intervention did not seem to help much. In Experiment 3, we tested whether the intervention would be more successful under different circumstances; specifically, hypothetical scenarios that were consistent or inconsistent with participants' prior beliefs.

## 4.0 Experiment 3

In Experiment 2, there was no evidence that the intervention helped participants learn how to identify whether the study design was an observational study or an experiment, and there was minimal evidence that the intervention helped participants improve at discriminating between correlation and causation. In Experiment 3, we made more improvements to the tutorial, and tested the impact of participants' prior beliefs on correlation-causation discrimination and the effectiveness of the intervention.

First, we modified our approach for some of the instructional material from the tutorial. One difference between the intervention in Experiments 1 and 2 and the intervention in Seifert et al. (2022) is that they framed making causal judgments about observational studies as a bias, which they called "causal theory error". Although we told students about people's tendency to erroneously make causal judgments about observational studies, we did not name this error, nor did we tell them that it was important to avoid making this error; we assumed that the latter was implied because the content of the intervention involved learning when people should and should not make causal claims based on the design of a study. In Experiment 3, we modified the intervention to be more like the one in Seifert et al. (2022), which was successful at reducing causal theory error. We added text to the intervention that defined causal theory error and explained that participants should avoid making this error. This modification could potentially motivate participants to learn about discriminating between correlation and causation, or to apply this principle to novel scenarios.

In addition, we added more explicit instructions to the Experiment 3 intervention that outlined the exact steps one should take to reduce the likelihood of causal theory error in reasoning.

In Experiment 2, we did not want to be too heavy-handed and over explain each concept; we wanted students, at least to some extent, to engage in independent critical reasoning about the content and attempt to make connections between concepts. Research on Worked Examples, which is the instructional technique most similar to the tutorial portion of the intervention, has found that providing too much information in instructional explanations can prevent the learner from spontaneously engaging with the material (Richey & Nokes-Malach, 2013). Our reasoning for a less "heavy-handed" approach is consistent with this finding. However, because there was little to no improvement in correlation-causation discrimination and study design identification following the intervention in Experiment 2, this approach was not very successful. In Experiment 3, we changed this strategy by including itemized steps for reducing the likelihood of causal theory error: 1) identifying the design of the study, and then 2) using the study design to make appropriate inferences about possible explanations for a statistical relation, which should yield stronger causal judgments for experiments than observational studies.

Second, we added a manipulation to test the effects of participants' prior expectations about the direction of a statistical relationship on their ability to discriminate between correlation and causation and whether the Self Explanation intervention was successful. Imagine a scenario in which someone believes that taking short breaks during the workday will increase productivity. Subsequently they read about a study, either an experiment or an observational study, which found that taking short breaks during the workday actually led to a decrease in productivity. How might the fact that the reader's prior beliefs are incongruent with the study findings influence whether they believe the study provides strong evidence of a negative causal relationship? For one, if the reader thinks that the study findings are implausible because they are belief-incongruent, they may be less likely to endorse conclusions that short breaks during the workday decrease productivity.

Second, they may be less likely to think the study's findings support that taking short breaks during the workday causes a change in productivity. Third, they may be less likely to believe that if someone makes a change in their life (stop taking short breaks to improve productivity), that the change will be successful.

On the one hand, such hypotheses make sense from the perspective that if it one has good reasons to hold a belief, a single study likely should not completely overturn their perspective. On the other hand, some of the measures we have been studying, such as "To what extent do you think that the study findings support the journalist's conclusion?" and "Do you think that this study shows that [taking short breaks during the workday] causes [a decrease in productivity]?" can be answered in light of what the study demonstrates about causality rather than necessitating a change in one's worldview based on the findings.

When learning about a study, at least for studies that are pertinent to one's life, we believe that people will very often have prior beliefs about the likely outcomes of the study. Thus, we think it is important to understand how prior beliefs affect the ability to assess the causal support that a study provides. Individuals are likely to think more critically about study findings that they believe are implausible or do not fit with their expectations about the outcome (Evans & Curtis-Holmes, 2005; Lord et al., 1979; Nickerson, 1998). When people think that only one outcome is plausible, and they hear about an observational study in which the results are incongruent with that outcome, they will generate more alternative explanations for that belief-incongruent finding than they would for a belief-congruent finding (Michal et al., 2021a). Furthermore, when asked to evaluate causal conclusions made by journalists about observational studies, participants were less likely to endorse causal conclusions that were incongruent with their prior beliefs. If there are more alternative explanations for a statistical relation, it makes sense to think that there is more evidence

for a stronger causal relation; at the same time, if one does not believe a causal relationship is possible because that would not match their prior beliefs, they may consider alternative reasons why two variables are statistically related.

However, we do not know how participants' prior beliefs will affect the evaluation of evidence from experiments. One possibility is that prior expectations about a study outcome will have the same effect when people are deciding whether to make causal inferences for both observational studies and experiments. Another possibility is that prior beliefs will have less of an impact when making causal judgments about experiments. If someone understands that experiments have strong internal validity, and then learns about the results of an experiment in which the outcome is incongruent with the person's prior beliefs, they may be more willing to update their prior beliefs based on that new information because the evidence comes from an experiment.

In Experiment 3, participants read and made judgments about observational studies and experiments that were either congruent or incongruent with their prior beliefs. We manipulated evidence congruency by having participants pick a statement that most closely aligned with their expectations about a statistical relationship (e.g., whether taking short breaks during the workday increases or decreases productivity) and then randomly assigning them to read a vignette with text that was either congruent or incongruent with their prior expectations. Manipulating evidence congruency across participants – rather than having the same vignettes be belief-congruent or incongruent for each participant – solves two issues. First, because we ask each participant about their expectations, and modify the vignettes based on that choice, we can be more certain that evidence congruency was successfully manipulated for individual participants and not just on

average. Second, this manipulation reduces the possibility of a confound and the possibility of other differences across vignettes as the reason for any differences in causal judgments.

## 4.1 Methods

A total of 399 participants were recruited from the Intro Psych subject pool at the University of Pittsburgh. All participants passed the attention check; thus, participants were only excluded from analyses if there was an error in data collection ($N$ = 20) or they took less than 15 minutes to complete the study ($N$ = 9), which we believed to be indicative of low effort. The final analyses included 370 participants. In Experiment 3, we removed the "design a study task"; the only pretest and posttest measures were the correlation-causation discrimination vignettes.

### 4.1.1 Prior Belief Manipulation

In Experiment 3, we added a manipulation to assess the effects of participants' prior beliefs on their correlation-causation discrimination, and whether the intervention mitigated any effects of prior beliefs. Thus, the study design for Experiment 3 was a 3 actual design (observational study versus explicit experiment versus implicit experiment, within subjects) × 2 block (pre versus post, within subjects) × 2 intervention (Control versus Self Explanation, between subjects) × 2 evidence-congruency (consistent with prior beliefs versus inconsistent, within subjects). Participants made judgments about a total of twelve vignettes. They read six vignettes at pretest and posttest, one for each combination of actual design and evidence congruency, in a random order. At the end of the vignettes, there was either a positive statistical relation or a negative statistical relation between

the two variables of interest. We assigned half of the vignettes to conclude with evidence that was congruent/consistent with participants' prior beliefs about the statistical relation, and the other half to conclude with evidence that was incongruent/inconsistent.

We manipulated evidence-congruency by asking participants to make a judgment about their prior belief about a statistical relation, and then modifying the vignette so that the evidence was either congruent or incongruent with that belief. Participants made judgments about a total of twelve vignettes; six vignettes presented in a random order at pretest and at posttest, with one vignette for each combination of actual design and evidence congruency. Half of the vignettes were assigned to be congruent with participants' prior beliefs, and the other half were assigned to be incongruent with participants' prior beliefs. People are hesitant to update their prior beliefs after learning about new evidence that is incongruent with those prior beliefs, especially if the topic is central to their ideology or worldviews (Lewandowsky et al., 2012). Because of this, we designed the stimuli so that participants were likely to have expectations about the outcome of the study but the outcome was unlikely to be central to their identity or core values (e.g., "Having more daily screen time (e.g., watching TV, using a tablet) [improves/worsens] children's social skills").

To assess prior beliefs, we first asked participants if they thought there would be either a positive or negative statistical relation between two variables. For example, they were asked whether "Exercising close to bedtime **improves** sleep quality" or "Exercising close to bedtime **worsens** sleep quality". On the following page, participants rated how strongly they agreed with that belief on a scale of 0 (*I don't really have an opinion one way or the other*), 1 (*somewhat agree*), or 2 (*strongly agree*). We used the information from the first question (participants believed there would be a positive or statistical relation) to change the text of the vignette so that the results were either congruent or incongruent with participants' prior beliefs. After reading each

vignette, participants completed the three correlation-causation discrimination measures, and said whether they thought the design was an observational/correlational study or an experiment.

We conducted two rounds of pilot testing with participants from Amazon Mechanical Turk to make sure that the vignettes involved scenarios for which participants believed there would be a positive or statistical relationship. We wanted to rule out, to the greatest extent possible, the likelihood that participants would not have prior expectations about the statistical relationships for the scenarios in our vignettes. During the pilot test, participants selected which one of three statements that most closely aligned with their prior beliefs: a positive statistical relation ("Yoga increases energy levels"), a negative statistical relation ("Yoga decreases energy levels"), or that "I do not have an opinion about these statements one way or the other". In the first round, participants ($N$ = 25) completed this task for 21 different scenarios. In the second round, a new group of participants ($N$ = 24) completed the task for 30 scenarios; we retested the scenarios from the first round that participants had prior beliefs about and tested some new scenarios to replace ones from the first round for which participants had no opinion. After the second round, we identified twelve scenarios in which less than 12% of participants had no opinion; we were satisfied that most participants in our study would have prior beliefs about these vignettes.

### 4.1.2 Intervention

Participants were randomly assigned to complete either the Self Explanation ($N$ = 178) or Control ($N$ = 192) conditions. The Self Explanation intervention was mostly the same as in Experiment 2, aside from some edits to improve clarity. There was one key exception, which was that we added text that introduced the term "causal theory error", the idea that people often make causal claims about observational studies, and a solution for how to reduce causal theory error.

These additions were modeled after Seifert et al. (2022), who found that correlation-causation discrimination improved after teaching students about causal theory error and how to reduce bias in judgments for observational studies.

In our intervention, we introduced "causal theory error" at the end of the tutorial (i.e., immediately before the two practice problems) by saying, "people often make causal claims about findings from observational studies. This is called causal theory error". Then, we explained that there are two steps for reducing causal theory error. First, participants must "figure out whether the study is purely observational or an experiment". Second, they must "consider what causal diagrams are possible based on the study design" and were reminded of the causal structures that are possible for observational studies versus experiments.

## 4.2 Results

### 4.2.1 Ability to Correctly Identify the Study Design from Vignettes

To test if participants could identify the correct study design from the actual design in the vignettes, we conducted similar analyses to Experiment 2[8]. To test whether improvement in study design discrimination differed across the three interventions, we tested for a three-way interaction

[8] In our OSF registration, our analysis plan also included "evidence congruency" as a predictor of study design identification. We decided not to include it for two reasons. First, in hindsight, we do not have strong justification for why prior beliefs about the direction of the statistical relation would affect subjects' judgments about study design. When we included it in the model, evidence congruency had no effect and no interactions with the other variables (all $p$'s > .05; see OSF for results). Second, there were problems with convergence when we included evidence congruency.

between actual design, block, and intervention. The most complex model that converged included a by-subject random intercept and slope for actual design with correlations removed, and by-vignette random intercepts (Table 11).

There was a significant effect of study design for both analyses. This means that at pretest, participants could discriminate between both explicit experiments (squares in Figure 15) and observational studies (triangles), and implicit experiments (circles) and observational studies. There was a significant interaction between actual design and block for the implicit experiment versus observational study analysis, which meant that participants got better at identifying the correct study design at posttest. More importantly, in both analyses, there were significant interactions between actual design, block, and intervention. This means that participants learned more during the intervention than control; participants in the Self Explanation group (filled shapes in Figure 15) were better at distinguishing both implicit and explicit experiments from observational studies at posttest compared to pre, but there was no improvement in the Control group. Learning in the Self Explanation group seems mainly driven by an improved ability to identify implicit experiments and explicit experiments; there was not much improvement at identifying observational studies. In contrast, participants in the Control group actually got worse at identifying observational studies at posttest (triangles in Figure 15).

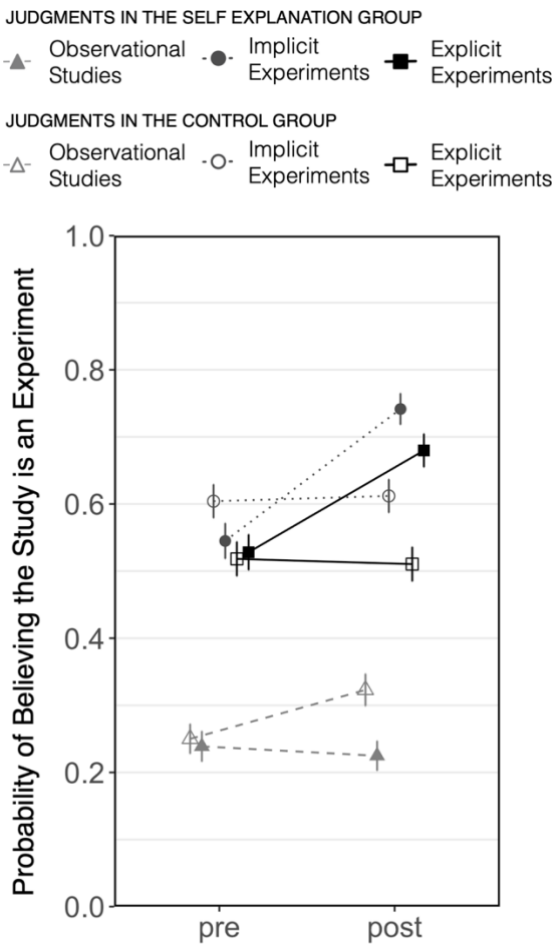


**Figure 15 Believed Study Design Discrimination**

**Table 11 Mixed Effects Models Testing Believed Study Design Discrimination**

| | Explicit Experiment vs. Observational Study | | | Implicit Experiment vs. Observational Study | | |
|---|---|---|---|---|---|---|
| Dependent Measure | *β* | 95% CI | *p* | *β* | 95% CI | *p* |
| Actual Design | 0.26 | 0.16, 0.44 | **<.001** | 0.16 | 0.09, 0.29 | **<.001** |
| Actual Design × Block | 0.86 | 0.63, 1.17 | .330 | 0.65 | 0.46, 0.91 | **.013** |
| A × B × Intervention | 0.27 | 0.14, 0.50 | **<.001** | 0.18 | 0.09, 0.36 | **<.001** |

Note. Boldface = $p < .05$. A = Actual Design; B = Block. Results for effects not listed here are available on OSF.

### 4.2.2 Correlation-Causation Discrimination for the Vignettes

We conducted similar analyses as in Experiment 2, but this time included evidence congruency (consistent or inconsistent with prior beliefs) as an additional predictor. We used binary contrast codes (-0.5 = evidence is incongruent with prior beliefs versus 0.5 = evidence is congruent with prior beliefs).[9] The models tested for a four-way interaction between design, block, evidence congruency, and intervention (Table 12 and Table 13). The most complex model that converged included a by-subject random slope and intercept for the design × block interaction and evidence congruency, and a by-vignette random intercept. Not all possible predictors or interactions are included in Table 12 and Table 13. This decision was made for ease of interpretation purposes because we tested for a complex four-way interaction. We only included the most important predictors in the main paper, meaning those that aligned with our registered hypotheses; our registered hypotheses and full set of results are available on OSF.

#### 4.2.2.1 Efficacy of the Intervention on Correlation-Causation Discrimination

In this section we report findings that appear in Rows 1-3 of Table 12 and Table 13. There were significant effects of actual design and believed design across all three measures and analyses (Row 1 of Table 12 and Table 13). This means that at pretest, participants could discriminate between observational studies (triangles in Figure 16) and both types of experiments (circles and squares in Figure 16). Additionally, almost all the actual design × block and believed design × block interactions were significant, which means that participants got better at discriminating

[9] We also ran analyses using continuous contrast codes (-0.5 = strongly incongruent; -0.25 = somewhat incongruent; 0 = no opinion; 0.25 = somewhat congruent; 0.5 = strongly congruent) that we calculated using participants' judgments about the strength of their prior beliefs. The results were very similar and so we only report the results from binary contrast codes; see OSF for all results.

correlation from causation, both based on the actual design and believed design, and for both explicit and implicit experiments (Row 2 of Table 12 and Table 13). The only exception was that there was no interaction between actual design and block for the plan success measure in either analysis.

Most importantly, the three-way interaction between actual design, block, and intervention tests whether the improvement in correlation-causation discrimination was larger in the Self Explanation condition than Control. The interaction was significant for almost all measures in both analyses (Row 3 of Table 12 and Table 13). As seen in Figure 16, the means for the experiments vs. observational studies got farther apart at post than at pre for the Self Explanation group (filled in shapes). However, in the Control group (white shapes), the means remained fairly flat from pre to post and did not get farther apart. The only exception was that there was no believed design × block × intervention interaction for the plan success measure. For the most part, however, there was clear evidence that the intervention worked.

#### 4.2.2.2 Effects of Prior Beliefs on Correlation-Causation Discrimination

This section reports on the influence of prior belief on correlation-causation discrimination; the results appear in Rows 4 – 8 of Table 12 and Table 13. First, the main effect of prior belief was significant for all three measures and analyses (Row 4 of Table 12 and Table 13). At pretest, judgments were more causal for studies in which the evidence was congruent with participants' prior beliefs (diamonds in Figure 17) than judgments for evidence-incongruent studies (triangles). Specifically, participants' beliefs about the plausibility of the direction of a finding influenced their judgments about whether the results provide causal evidence (supported the journalist's causal conclusion, whether implementing an intervention based on the results would have the outcome implied by the results, and whether the study shows that the independent variable causes the

dependent variable). Furthermore, the regression weights reveal the participant's prior belief about the plausibility of the finding usually had a considerably larger influence on these judgments about evidence for causality than whether the study was actually an experiment or observational study and whether they believed that the study was an experiment or observational study.

Second, Row 5 of Table 12 and Table 13 reports the interaction of evidence congruency, block, and intervention. One of out of the six of these tests was significant for the actual design regression, and two out of three were significant for the believed design regression; all the significant ones were negative. These negative interactions can be interpreted as the effect of prior belief decreasing from pre to post, primarily in the Self Explanation condition. To help interpret this finding, the evidence congruency × block × intervention interactions can be compared to the actual design × block × intervention interactions. The latter were all significant, and positive, evidence of becoming more influenced by the actual study design due to the intervention. In contrast, the former were sometimes significant and negative, evidence of becoming somewhat less influenced by one's prior beliefs due to the intervention. In sum, this is evidence of a desirable form of learning to focus less on one's prior beliefs when making causal claims.

Third, none of the higher-level interactions involving evidence congruency and believed study design were significant (Rows 6-8 of Table 13). And very few of the higher-level interactions involving evidence congruency and actual study design were significant (Rows 6-8 of Table 12), with the following exception: Two of the six evidence congruency × actual study design interactions in Row 6 of Table 12 were significant. This means that at pretest, correlation-causation discrimination was better when the study findings were congruent with participants' beliefs and worse when they were incongruent; the difference in judgments for observational studies and explicit experiments was greater for the evidence-congruent (diamonds in Figure 17A) than the

evidence-incongruent vignettes (triangles) at pretest. This, like the first finding, provides additional evidence that prior beliefs can impact causal claims.

**Figure 16A.** Judgments by Actual Study Design

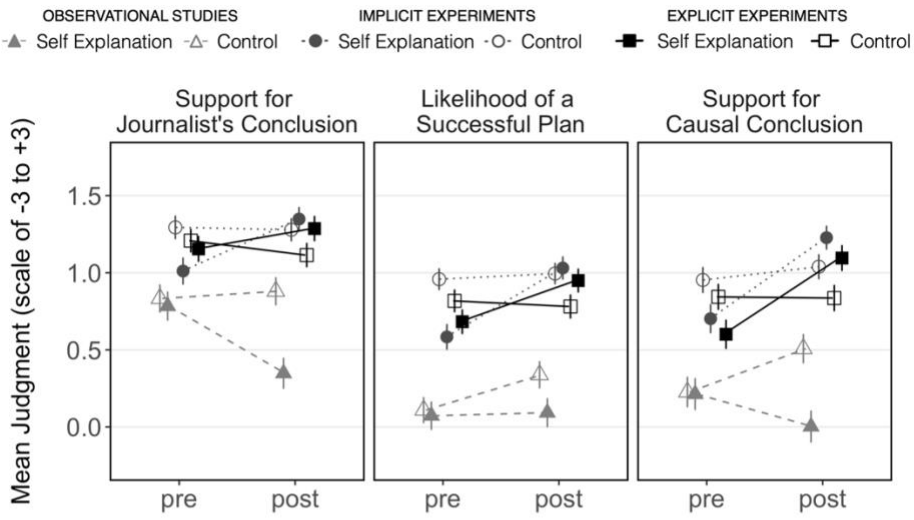


**Figure 16B.** Judgments by Believed Study Design

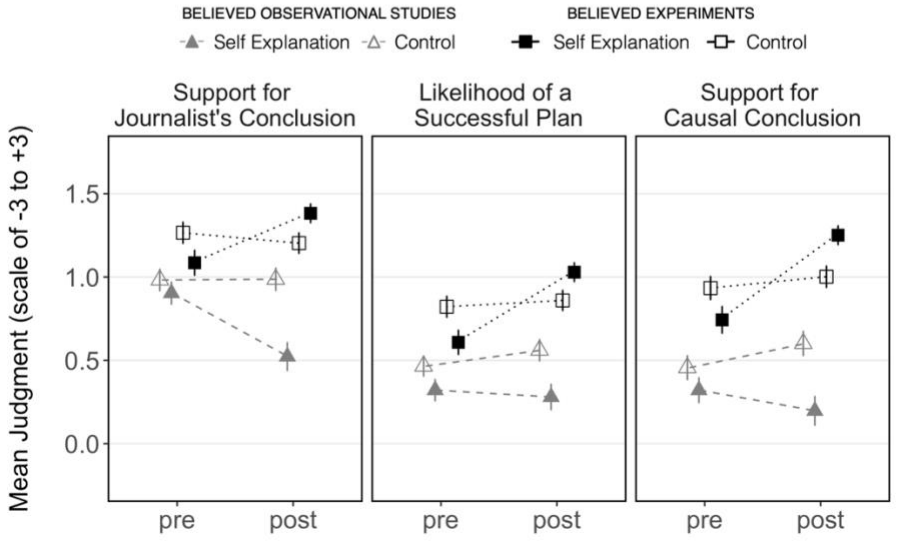


**Figure 16 Correlation Causation Discrimination by Study Design**

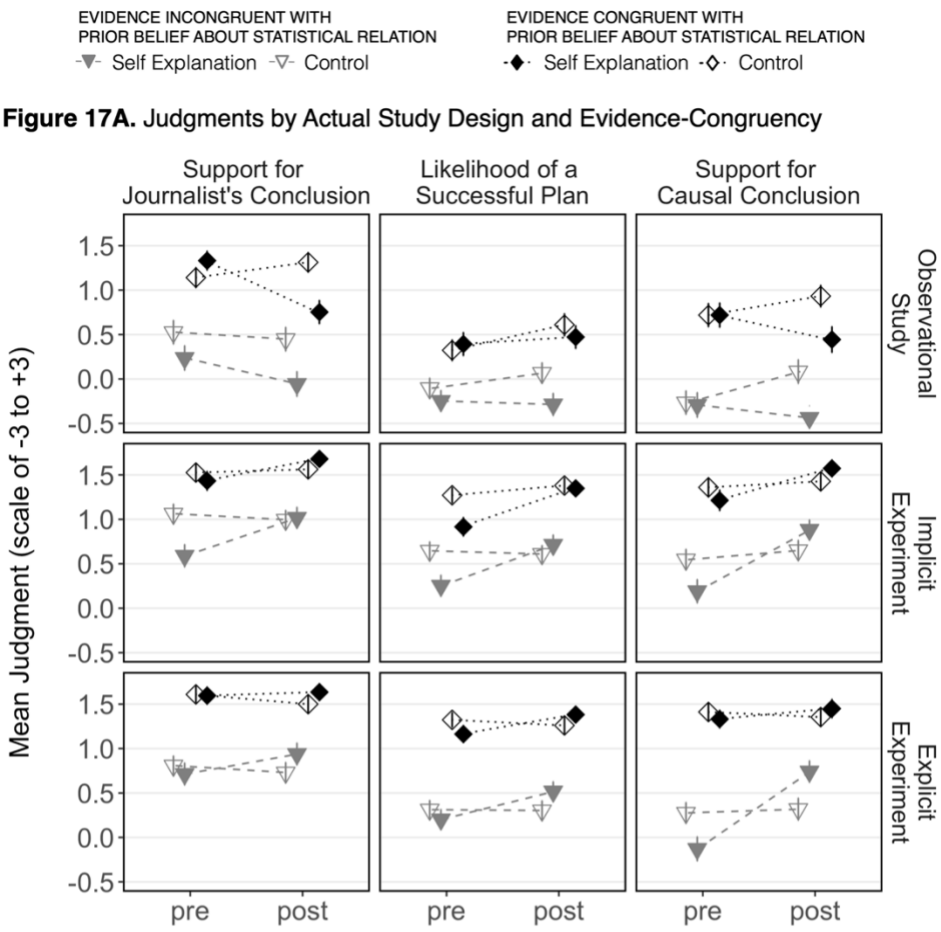


**Figure 17A.** Judgments by Actual Study Design and Evidence-Congruency

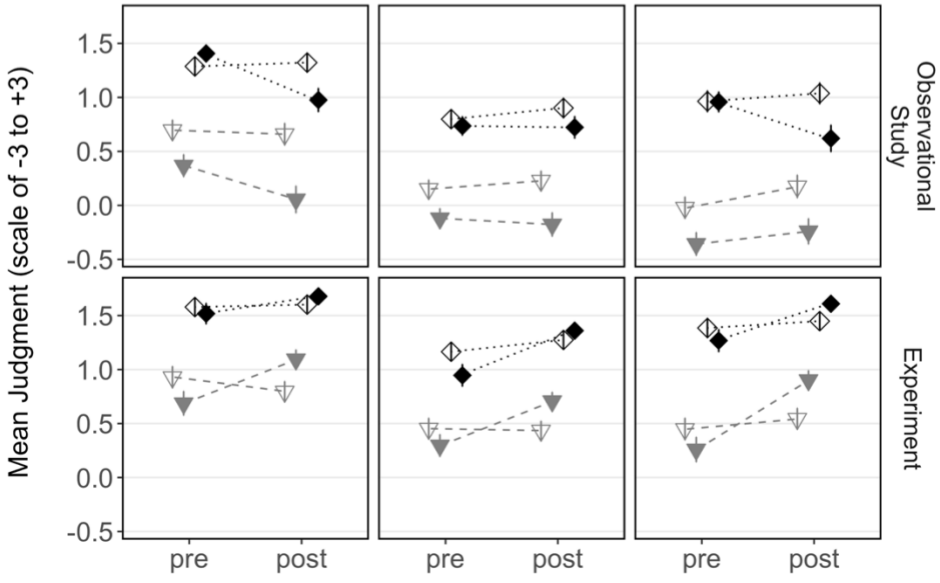


**Figure 17B.** Judgments by Believed Study Design and Evidence Congruency

**Figure 17 Correlation Causation Discrimination by Study Design and Evidence Congruency**

**Table 12 Mixed Effects Model Testing for Correlation-Causation Discrimination by Actual Design**

| | Explicit Experiment vs. Observational Study | | | Implicit Experiment vs. Observational Study | | |
|---|---|---|---|---|---|---|
| Dependent Measure | *β* | 95% CI | *p* | *β* | 95% CI | *p* |
| **1. Actual Study Design** | | | | | | |
| Journalist's Conclusion | 0.23 | 0.06, 0.39 | **.006** | 0.20 | 0.06, 0.34 | **.005** |
| Plan Success | 0.41 | 0.14, 0.69 | **.003** | 0.43 | 0.15, 0.70 | **.002** |
| Causal Conclusion | 0.28 | 0.07, 0.49 | **.009** | 0.33 | 0.17, 0.49 | **<.001** |
| **2. Actual Study Design × Block** | | | | | | |
| Journalist's Conclusion | 0.13 | 0.02, 0.24 | **.020** | 0.22 | 0.11, 0.33 | **<.001** |
| Plan Success | -0.03 | -0.16, 0.10 | .684 | 0.06 | -0.06, 0.18 | .313 |
| Causal Conclusion | 0.11 | 0.01, 0.22 | **.038** | 0.16 | 0.06, 0.27 | **.003** |
| **3. Actual Study Design × Block × Intervention** | | | | | | |
| Journalist's Conclusion | 0.45 | 0.23, 0.67 | **<.001** | 0.51 | 0.29, 0.73 | **<.001** |
| Plan Success | 0.35 | 0.10, 0.61 | **.007** | 0.40 | 0.16, 0.64 | **.001** |
| Causal Conclusion | 0.58 | 0.37, 0.80 | **<.001** | 0.52 | 0.31, 0.74 | **<.001** |
| **4. Evidence Congruency** | | | | | | |
| Journalist's Conclusion | 0.51 | 0.42, 0.59 | **<.001** | 0.45 | 0.37, 0.53 | **<.001** |
| Plan Success | 0.47 | 0.38, 0.55 | **<.001** | 0.37 | 0.29, 0.46 | **<.001** |
| Causal Conclusion | 0.63 | 0.54, 0.71 | **<.001** | 0.53 | 0.45, 0.61 | **<.001** |
| **5. Evidence Congruency × Block × Intervention** | | | | | | |
| Journalist's Conclusion | -0.18 | -0.39, 0.03 | .092 | -0.24 | -0.45, -0.03 | **.026** |
| Plan Success | 0.04 | -0.18, 0.27 | .720 | -0.03 | -0.26, 0.20 | .801 |
| Causal Conclusion | -0.17 | -0.38, 0.04 | .113 | -0.11 | -0.32, 0.09 | .273 |
| **6. Evidence Congruency × Actual Study Design** | | | | | | |
| Journalist's Conclusion | -0.01 | -0.16, 0.14 | .854 | -0.12 | -0.27, 0.02 | .098 |
| Plan Success | 0.29 | 0.13, 0.45 | **<.001** | 0.10 | -0.07, 0.26 | .239 |
| Causal Conclusion | 0.17 | 0.02, 0.32 | **.028** | -0.03 | -0.17, 0.12 | .704 |
| **7. Evidence Congruency × Actual Study Design × Block** | | | | | | |
| Journalist's Conclusion | -0.03 | -0.24, 0.18 | .781 | 0.03 | -0.18, 0.24 | .798 |
| Plan Success | -0.12 | -0.34, 0.11 | .315 | -0.04 | -0.27, 0.19 | .740 |
| Causal Conclusion | -0.16 | -0.37, 0.05 | .137 | -0.02 | -0.23, 0.18 | .823 |
| **8. Evidence Congruency × Actual Design × Block × Intervention** | | | | | | |
| Journalist's Conclusion | 0.25 | -0.17, 0.68 | .241 | 0.14 | -0.27, 0.56 | .501 |
| Plan Success | -0.05 | -0.50, 0.41 | .839 | -0.19 | -0.65, 0.27 | .424 |
| Causal Conclusion | -0.34 | -0.76, 0.09 | .118 | -0.22 | -0.63, 0.19 | .285 |

Note. Boldface = $p < .05$. Results for effects and interactions not listed here are on OSF.

**Table 13 Mixed Effects Model Testing for Correlation-Causation Discrimination by Believed Design**

| Dependent Measure | *β* | 95% CI | *p* |
|---|---|---|---|
| **1. Believed Study Design** | | | |
| Journalist's Conclusion | 0.12 | 0.05, 0.20 | **.001** |
| Plan Success | 0.10 | 0.03, 0.17 | **.005** |
| Causal Conclusion | 0.22 | 0.14, 0.29 | **<.001** |
| **2. Believed Study Design × Block** | | | |
| Journalist's Conclusion | 0.18 | 0.07, 0.28 | **.001** |
| Plan Success | 0.13 | 0.03, 0.22 | **.013** |
| Causal Conclusion | 0.13 | 0.03, 0.23 | **.010** |
| **3. Believed Study Design × Block × Intervention** | | | |
| Journalist's Conclusion | 0.44 | 0.24, 0.65 | **<.001** |
| Plan Success | 0.30 | 0.10, 0.50 | **.003** |
| Causal Conclusion | 0.39 | 0.18, 0.59 | **<.001** |
| **4. Evidence Congruency** | | | |
| Journalist's Conclusion | 0.46 | 0.39, 0.53 | **<.001** |
| Plan Success | 0.45 | 0.38, 0.52 | **<.001** |
| Causal Conclusion | 0.08 | 0.03, 0.13 | **.001** |
| **5. Evidence Congruency × Block × Intervention** | | | |
| Journalist's Conclusion | -0.19 | -0.35, -0.02 | **.026** |
| Plan Success | -0.01 | -0.19, 0.17 | .905 |
| Causal Conclusion | -0.21 | -0.37, -0.05 | **.012** |
| **6. Evidence Congruency × Believed Study Design** | | | |
| Journalist's Conclusion | -0.01 | -0.13, 0.11 | .874 |
| Plan Success | -0.06 | -0.19, 0.07 | .392 |
| Causal Conclusion | -0.06 | -0.18, 0.06 | .339 |
| **7. Evidence Congruency × Believed Study Design × Block** | | | |
| Journalist's Conclusion | -0.07 | -0.24, 0.10 | .400 |
| Plan Success | 0.03 | -0.15, 0.22 | .725 |
| Causal Conclusion | 0.02 | -0.15, 0.19 | .807 |
| **8. Evidence Congruency × Believed Design × Block × Intervention** | | | |
| Journalist's Conclusion | -0.06 | -0.40, 0.28 | .714 |
| Plan Success | -0.14 | -0.51, 0.24 | .469 |
| Causal Conclusion | 0.12 | -0.21, 0.46 | .466 |

Note. Boldface = $p < .05$. Results for effects and interactions not listed here are on OSF.

## 4.3 Discussion

In Experiment 3, we made improvements to the intervention and tested the effects of participants' prior beliefs on correlation-causation discrimination. There were a couple key findings.

First, the improvements we made to the intervention were successful. In the Self Explanation condition, intro psych students became better at identifying whether a vignette described an observational study or an experiment, and at discriminating between correlation and causation. This contrasts with what we found in Experiment 2, in which there was no evidence that the intervention improved students' ability to identify observational studies versus experiments, and there was minimal evidence that the intervention improved correlation-causation discrimination. In the general discussion, we discuss some possible reasons why students may have learned from the intervention in Experiment 3 than from the previous two versions.

Second, we found that participants' beliefs about the plausibility of the study outcome affected their judgments for both observational studies and experiments. The effect of prior beliefs on judgments for observational studies were consistent with Michal et al. (2021a), and we also extended these findings to experiments; for both observational studies and experiments, people are more likely to endorse causal claims about statistical relations when the results are consistent with their expectations. At pretest, prior beliefs had a greater impact on participants' evaluation of the study evidence than the actual design of the study. However, after the intervention, the influence of prior beliefs lessened, at least for some of the analyses. This coincided with improved correlation-causation discrimination based on actual and believed study design in the Self Explanation condition, which suggests that participants may have started placing less weight on

their prior beliefs and more emphasis on the study design when reasoning about causality from study findings.

Additionally, our study is the first to show the impact of prior beliefs on correlation-causation discrimination, because we could directly compare participants' evaluation of experiments and observational studies when the evidence was either congruent or incongruent with their prior beliefs. For the plan likelihood and causal conclusion measures, in which participants judged the likelihood of a successful intervention based on the study findings and how strongly the study supported a causal claim, there was a significant interaction between evidence congruency and actual design. Correlation-causation discrimination based on actual design was better at pretest for the evidence-congruent vignettes than for evidence-incongruent vignettes. This is the first evidence that prior beliefs can have at least some influence on people's ability to discriminate between correlation and causation.

# 5.0 Discussion

Across three studies, we tested an intervention that uses causal diagrams to improve undergraduate students' knowledge of key research methods concepts and skills. There were three key findings.

First, we found that our intervention improved students' understanding across multiple measures. Previous work has studied the extent to which people infer causality from observational studies (Seifert et al., 2022). However, because people were not also asked to evaluate evidence from experiments, which provide more evidence for causation, there is a gap in the current literature about the extent to which people understand when statistical results do and do not imply causation. Our study showed that a well-tailored intervention that teaches students about alternative explanations for a statistical relation improved their ability to discriminate between correlation and causation. Additionally, the intervention improved students' ability to design experiments, a critical research skill, which we tested in a novel task in which students must design the correct type of study (observational or experimental) given a prompt. Another beneficial outcome, which we believe is critical to the two previous outcomes, is that the intervention improved students' ability to identify if a study was an experiment or an observational study.

Second, we tested different methods for teaching students about why correlation does not imply causation. In Experiment 1, we compared three different types of practice (Analogical Comparison, Worked Example, and Self Explanation). Whereas the intro psych students learned the most from Self Explanation practice, research methods students improved the least in this condition. Because students in intro psych classes have less foundational knowledge, these results suggest that prior knowledge may play a role in the effectiveness of certain interventions. Another

consideration is the extent to which students require extensive handholding during learning. Although intro psych students learned the most in the Self Explanation condition, which requires more independent practice than the other two conditions, students learned more when Self Explanation practice was preceded by step-by-step procedural instructions for how to stop making erroneous causal judgments about observational studies.

Third, participants' prior expectations about the likelihood of a finding had a strong influence on their judgments of whether the study should be interpreted as providing causal evidence for both observational studies and experiments. This is an expansion on prior research (Michal, Seifert, et al., 2021), which has only studied the effects of prior beliefs on people's judgments of whether observational studies warrant causal claims. Additionally, we found that correlation-causation discrimination is better when study evidence is congruent with people's prior beliefs. Most importantly, our intervention led to a reduction in the influence of participants' prior beliefs and an increase in the influence of the study design on their judgments of whether the findings support a causal relationship.

Overall, though the intervention could still be improved, our findings provide clear evidence that it helps students learn to discriminate correlation from causation.

### 5.1 Lessons Learned About Improving Correlation-Causation Discrimination

Across three studies, we made iterative changes to the intervention, and across the three studies the intervention appeared to become more successful. During this process, we learned three main lessons about factors that we think are important to include in teaching correlation-causation discrimination. These lessons were mainly from cross-study comparisons rather than perfectly

controlled comparisons within an experiment, so we cannot claim for certain that these were the causal factors that led to success in Experiment 3, but they are our best guesses.

The first lesson we learned was that students may struggle with identifying if a study is an experiment versus an observational study, which can prevent accurate correlation-causation discrimination. The second lesson we learned was to motivate students not to make a "causal theory error". The third lesson we learned was that it seems necessary to be extremely explicit (more explicit than what we initially thought was necessary) about initial steps for having correlation-causation discrimination: 1) identifying the study design, and 2) considering the possibility of alternative explanations for the relation based on the study design.

Prior to conducting any of the studies, we conceptualized correlation-causation discrimination as occurring in two steps. The first step in deciding whether a statistical relationship is evidence of a causal relationship is to identify whether the design is an observational study or an experiment. In the second step, the reader can use that information to decide whether alternative explanations are possible for the statistical relation. If it is an observational study, there are potentially two alternative explanations for a statistical relation, a confound and reverse causality, which means that one should be hesitant to make a causal claim. In contrast, an experiment provides stronger evidence for causality because random assignment reduces the likelihood of a confound, and the temporal precedence inherent in manipulating the independent variable eliminates the possibility of reverse causality. Thus, a causal claim is more warranted for experiments than it is for observational studies.

Our first lesson came from Experiment 1. In Experiment 1, we focused primarily on improving knowledge of the second step in correlation-causation discrimination. We made the intentional decision to mention the difference between observational studies and experiments, but

not to provide extensive details about how to determine the design of a study from a written description. This was because we wanted students to engage in some independent critical thinking and not overexplain each concept covered in the tutorial. Furthermore, we believed this was a task that students would be able to complete without extensive handholding; there was an obvious difference between the vignettes that participants could use to identify the study designs, in which all the experiments used the phrase "random assignment" whereas the observational studies did not. At pretest in Experiment 1, only 60% of intro psych students and 75% of research methods students correctly identified experiments as experiments. In general, they were more likely to say that a study was an experiment if it was an experiment versus if it was an observational study; however, this performance was clearly not at ceiling in either sample. Participants got better at this task with practice, there was still considerable room for improvement on average and the intervention did not help. Furthermore, in Experiment 1, we found that participants' correlation-causation discrimination was related to their beliefs about the study design, which was good news because it was evidence that the students were successfully implementing Step 2, but not Step 1.

Thus, in Experiment 2 we added an explanation about how to look for language that refers to whether there is implicit or explicit random assignment to conditions in the study. Unfortunately, this additional instruction still did not improve students' ability to identify the study design and had minimal effects on correlation-causation discrimination. That said, the intervention was successful at helping students design experiments versus observational studies and understanding which causal structures (mechanism, reverse causality, confound) are possible explanations for a statistical relation in an experiment versus an observational study. Both findings suggest that the intervention was working to some extent, although not necessarily for improving correlation-causation discrimination.

The second and third lessons came from the failure of the intervention in Experiment 2 to improve correlation-causation discrimination, and the eventual success in Experiment 3. In Experiment 3, we decided to make two additional changes to the intervention. First, we looked back at the intervention from Seifert et al.'s (2022) study, which was successful at reducing causal inferences for observational studies. One thing they did was tell participants to avoid making a "causal theory error," and explained that there is a tendency for people to make an error in reasoning where they make causal claims about observational studies. Second, we even more clearly listed the two steps for correlation-causation discrimination, "First, figure out whether the study is purely observational or an experiment … Second, consider what causal diagrams are possible based on the study design". These changes were worthwhile, because after the intervention, students got better at identifying whether the design was an observational study or an experiment, and also at discriminating correlation from causation (making stronger causal judgments for experiments than for observational studies).

Because we made both modifications to the intervention simultaneously, it is possible that either modification, or some combination of the two, is why the intervention was successful in Experiment 3. The second lesson we learned has to do with the first modification – naming the phenomenon of causal theory error during the intervention. In Experiment 3, we taught students about "causal theory error", why it was important to avoid, and the steps they could take to reduce the likelihood of making this error. In this way, we framed causal theory error as an error students could learn to avoid making, and that students can develop greater competency for critical scientific reasoning skills. When students have a greater sense of self-efficacy or feel more confident that they can learn and achieve their goals, they are more motivated to learn compared to students who are less confident in their abilities (Pintrich, 2003; Pintrich & Schunk, 2002). Thus,

in our study, there are two potential sources of motivation in Experiment 3 – that making causal judgments for correlational studies is an error in reasoning (which students would likely want to avoid), and that they are able to learn how to avoid this error. Increased motivation to learn could also explain why we saw improved study design identification in Experiment 3 but not Experiment 2 – although Experiment 2 included very detailed instructions for how to identify observational studies versus experiments from text, we did not include the potentially motivating information about causal theory error.

The third lesson we learned was that to improve correlation-causation discrimination is that it may be necessary to be extremely explicit about the steps the reader must take to reduce causal theory error. Breaking problems down into these kinds of smaller subgoals can facilitate problem solving during learning (Catrambone, 1998), which would explain why explicitly stating that there are two steps to improving correlation-causation discrimination, and these are the order to complete them in, would help students get better at correlation-causation discrimination. During the interventions in Experiment 1 and 2, students learned these subgoals (identifying the study design, considering the possibility of alternative explanations) but we did not provide them with these steps in a numbered list. In Experiment 3, we learned that highlighting these two skills as ordered steps for reducing causal theory error may be critical for improving correlation-causation discrimination.

Though we think that the three lessons mentioned above are likely the most responsible for the success of the intervention in Experiment 3, there are also some other factors that may have made a difference. First, all versions of the interventions in our experiments relied on causal structures to teach about alternative explanations for a correlation, which is similar to Seifert et al. (2022). We suspect that causal diagrams are a useful tool for teaching correlation-causation

discrimination, though we do not know of any studies that have compared it to another form of instruction.

Second, in Experiment 3, we also asked participants to state their expectations about the outcome of a study and included vignettes that were sometimes consistent and sometimes inconsistent with their expectations. It is possible that these procedures may have led participants to be more aware of their prior beliefs and possibly encouraged them to not use their prior beliefs and instead use the study design when making judgments about causality.

We hope other researchers and educators might incorporate these findings when designing interventions that target learning about correlation and causation (Michal, Seifert, et al., 2021), and perhaps other important skills like the ability to discriminate between science and pseudoscience (McLean & Miller, 2010) and critical thinking more generally (Tiruneh et al., 2014), among others.

### 5.2 Undergraduate Students' Ability to Design their Own Studies

In addition to the correlation-causation discrimination measures, we tested a novel task that probed students' ability to design both observational studies and experiments. Prior work, primarily in the biological sciences, has studied students' ability to design robust experiments (Brownell et al., 2014; Shanks et al., 2017; Sirum & Humburg, 2011) but not their ability to design experiments and observational studies. Thus, we do not know much about students' ability to design observational studies, a common paradigm in all domains of research. The APA suggests that undergraduate psychology majors should be able to design both correlational studies and simple between-subjects experiments after taking multiple lower-level psychology courses

(Halonen et al., 2013). Therefore, we thought it was important to assess undergraduate psychology students' ability to design both types of studies, not just experiments.

In the current paper, students read a prompt that described a hypothetical statistical relation; they were randomly assigned to design either an observational study or an experiment to test the outcome. To be successful at this task, students must understand the critical difference between the two designs – that experiments use random assignment to conditions, whereas observational studies do not. This is the same foundational knowledge required for when students identified the design of a hypothetical study after reading a short vignette. Thus, both tasks require the student to be able to discriminate between the two types of study designs. The "design a study" task, however, is a more advanced test of this ability, because it requires more independent and creative thinking for how to measure or manipulate variables.

In Experiment 2, at pretest, participants were quite good at designing observational studies but not experiments. Students successfully designed an observational study almost every time they were asked to at pretest (98%), whereas they were less successful at designing experiments (27%). We can compare this performance to the study design identification task in the vignettes at pretest, in which students correctly identified experiments in 63% of cases and were even more successful at identifying observational study designs in the vignettes (88%). Although both tasks require understanding that only experiments use random assignment, these results show that designing studies is a more challenging application of that principle. However, the intro psych students did show some ability to discriminate between the two designs for the design a study measure at pretest, because they were more likely to design an experiment when told to design an experiment versus an observational study. Still, there was obvious room for improvement.

Critically, we found that our intervention helped students get better at designing experiments. After the intervention, participants in the Self Explanation condition successfully designed an experiment when told to do so in 42% of cases, meaning there was a 15% improvement from pretest to posttest. One thing that we changed in Experiment 2 was that we explained how to identify study designs in vignettes by determining whether random assignment was used or not. This change was unsuccessful for improving students' ability to identify studies from pre-written descriptions, but it may have helped them when designing their own. Additionally, they may have benefitted from seeing our examples of hypothetical observational studies and experiments during the intervention. For each example, we reminded students why the study design was an observational study or an experiment by pointing out how the text either referenced random assignment to conditions or did not.

It is possible that any of these parts – telling students how to identify the design of a study in a text or providing examples of observational studies versus experiments – may have helped students get better at designing their own experiments. However, they ultimately require even more scaffolding or instruction; when prompted to design an experiment after the intervention, most students (58%) could not, and did not include random assignment to conditions in their designs. In Experiment 3, we found that telling students about "causal theory error" and providing the exact steps for reducing causal theory error may have helped improve study design identification from pre to post. Perhaps these changes might also motivate or help students to discriminate between study designs when designing their own experiments. Future studies could test whether this method is successful for this measure as well, which is a critical but challenging research methods skill for undergraduate psychology students to master.

### 5.3 Implications for Theories about Prior Beliefs

In the current study, there were three different measures that we used to probe how students evaluated the findings from observational studies versus experiments – whether there is sufficient support for a journalist to make a causal conclusion based on the study findings, whether the study findings support a causal conclusion about the statistical relationship, and whether implementing changes in one's life to elicit a favorable outcome based on the results of the study would be successful. All three measures were highly influenced by students' prior expectations about the outcome of a study; students drew stronger causal conclusions about vignettes in which the research findings were congruent with their prior expectations about the statistical relationship. For example, if they believed there would be a positive relationship between exercising close to bedtime and sleep quality, they were more likely to justify a causal claim for the results of an observational study or experiment showed a positive relationship between the variables (evidence-congruent) instead of a negative relationship (evidence-incongruent).

Our findings about how prior beliefs influence judgments of causality is consistent with previous research, which shows that people tend to make stronger causal inferences about correlational data that is consistent with their worldview or prior beliefs (Evans & Curtis-Holmes, 2005; Lord et al., 1979; Nickerson, 1998). However, this research has focused on the effects of prior beliefs on judgments about observational studies, whereas people must often draw inferences from experiments as well. Our study expanded upon the previous research to show that students not only weigh prior expectations when evaluating findings from observational studies (Michal, Seifert, et al., 2021), but also when making judgments about the outcome of an experiment.

Furthermore, because we asked participants to evaluate findings from both observational studies and experiments, we were able to directly compare regression weights to assess the

magnitude of the effect that prior beliefs and study design have on participants judgments for observational studies versus experiments. At pretest, people were more likely to rely on their prior beliefs than study design to decide whether they should infer causality from a statistical relation. However, our intervention reduced bias due to prior beliefs at posttest; during the intervention, participants were taught to pay more attention to the design of the study when making causal judgments, and in doing so, seemed to pay less attention to their own prior beliefs.

One of the main theories in the prior belief literature is a dual account of reasoning (Evans & Curtis-Holmes, 2005; Nickerson, 1998). When people encounter a study result that they believe to be implausible and must decide whether that result justifies a causal claim, they will engage in a more analytical reasoning approach and generate alternative reasons for the statistical relation. If the reader can identify alternative explanations for the study outcome, and thereby explain the reason for the implausible result, they may not have to change their perspective about the relationship between the two variables. We found that prior beliefs affected judgments for both observational studies and experiments, which means that they may not have considered study design as a factor in whether such alternative explanations are even possible. However, our intervention helped students in the Self Explanation condition get better at using study design to make judgments about causation, instead of being primarily motivated by their prior beliefs.

The question becomes, how did our intervention reduce reliance on prior beliefs? It is notable that this intervention was able to reduce such a strong influence on reasoning about causes, so how did the intervention motivate them to think more about study design when making causal judgments? One possibility, as we discussed in the previous section about lessons learned, was that students may have been more motivated to learn because of how we framed "causal theory error" as a bias that should be avoided and can be avoided by following a two-step procedure. This

motivation may have also helped students overcome the tendency to be biased by prior expectations about the statistical relationship and incorporate the consideration of study design; this means that students may have been more likely to consider causal claims about belief-incongruent experiments and think more critically about making causal claims for belief-congruent observational studies.

### 5.4 Implications for Theories about Instructional Techniques

One of the main goals of Experiment 1 was to compare methods of instruction as means of improving correlation-causation discrimination and students' ability to recognize whether a study was an experiment or an observational study from text. For intro psych students, the Self Explanation condition seemed to be better for improving correlation-causation discrimination based on actual study design, compared to the Worked Example condition. However, there was little evidence that the interventions differentially affected students' ability to discriminate between believed observational studies and believed experiments or their ability to identify the study designs in vignettes. Thus, the benefits of Self Explanation in the intro psych class were limited to improving correlation-causation discrimination by actual design, and for the most part only found when comparing Self Explanations with Worked Example instruction.

Conversely, the research methods students seemed to learn more from the Worked Example and Analogical Comparison interventions. However, these differences were not always stable. For example, the Analogical Comparison condition seemed to be better than Self Explanation at improving correlation-causation discrimination by believed design, and the Worked Example condition seemed to be better than Self Explanation at improving correlation-causation

discrimination by actual design. In sum, different interventions worked better for the two samples, and within each sample, the effects of the interventions were not always stable across the analyses.

One question was whether some of the differences in performance across the interventions could be explained by theory, and the mechanisms of these instructional methods for facilitating learning. We found more evidence that the Self Explanation intervention helped improve students' understanding in intro psych, compared to the other two conditions. However, we only found these benefits when looking at analyses of correlation-causation discrimination that used actual design and not believed study design as a predictor. At pretest, the difference in intro psych students' judgments for experiments versus observational studies was greater when comparing judgments based on participants' beliefs about design, rather than the actual designs in the vignettes. This meant that correlation-causation discrimination was not as good at pretest if analyzing by actual study design than by believed study design for the intro psych sample. Because Self Explanation instruction helps the learner identify gaps in their own knowledge or understanding (Chi, 2013), the Self Explanation intervention may have led to more learning (compared to the other two conditions) for correlation-causation discrimination by actual study design, because students struggled with this at pretest.

It is also possible that some of these differences are not meaningful, because for the most part, the three instructional methods seemed to work quite similarly. Although we believed there was enough evidence that Self Explanation somewhat improved correlation-causation discrimination for the intro psych sample, the other interactions were too inconsistent across the measures and samples to identify stable patterns that suggest different instructional methods are better for certain tasks. In Experiment 2, we tested an updated verison of the Self Explanation intervention versus a Control condition, and found little evidence of learning during the

intervention. We did, however, find evidence of learning after Experiment 3, which mostly involved edits to the text in the tutorial rather than the practice problems. Recall in Experiment 1, the tutotrial text was the same for the Worked Example, Analogical Comparison, and Self Explanation conditions; the key difference was how participants completed the practice problems. As such, it would be interesting to once again test for different effects of the instructional methods now that we have made successful revisions to the tutorial.

## 5.5 Future Directions for Improving the Intervention and Tailoring it to Different Populations

In Experiment 1, we compared the effectiveness of Self Explanation, Worked Example, and Analogical Comparison interventions at improving correlation-causation discrimination for intro psych students and research methods students. If intro psych students had less prior knowledge at pretest, this could potentially explain why the Self Explanation condition was more effective for this sample in some of the comparison analyses. Admittedly, we should be careful not to make strong generalizations from cross-sample comparisons of prior knowledge. However, there were some fairly clear differences between the two groups, across a few measures.

At pretest, intro psych students had worse pre-test correlation-causation discrimination than research methods students, and they were not as good at identifying the type of study design in vignettes. Intro psych students were also more likely to make causal claims about vignettes that they thought were observational studies (39%) compared to the research methods sample (31%), and they made fewer causal claims about vignettes they thought were experiments (65%) compared to the research methods sample (74%). Thus, at pretest, intro psych students were less

advanced than research methods students at skills like identifying the correct study designs and to some extent, discriminating between correlation and causation. Because there were differences at pretest, this opens the possibility that differences in prior knowledge led to some interventions having a bigger impact than others.

In Experiments 2 and 3, we decided to focus on the Self Explanation intervention, and only collected data from samples of intro psych participants. Although the intervention in Experiment 2 was not successful at improving correlation-causation discrimination, our modifications to the intervention in Experiment 3 seemed effective because students appeared to get better at discriminating between correlation and causation at posttest. An alternative explanation for why there was greater improvement in correlation-causation discrimination after the intervention in Experiment 3, however, is that there was something inherently different about the sample of participants in Experiment 3 (data collected during the Spring 2022 semester) compared to the sample of participants in Experiment 2 (data collected during the Fall 2021 semester).

When we compared participants' pretest performance across studies, we realized that the participants in Experiment 2 may have had more relevant prior knowledge. At pretest, fewer intro psych students made causal claims about observational studies in Experiment 2 (33%) than students in Experiment 3 (53%), but they made an equal number of causal claims about experiments in Experiment 2 (68%) and Experiment 3 (67%). This means that before the intervention, students in Experiment 3 were more likely to make a causal theory error than in Experiment 2. And, if participants in Experiment 3 had more to learn about discriminating between correlation and causation at pretest, this could explain why the intervention helped them improve at posttest. Thus, the intervention we designed for Experiment 3 may be most effective for students who have less prior knowledge.

If the current intervention is only effective for students with a certain amount of prior knowledge, then tailoring the intervention to reach others with more or even less knowledge may help to widen the reach of its impact. In Experiment 3, students seemed to have less prior knowledge or correlation-causation discrimination abilities than the participants in Experiment 2, but they were still able to somewhat discriminate between observational studies and experiments. However, the adults in Bleske-Rechek et al. (2015) had much lower levels of correlation-causation discrimination, and they were just as likely to make causal judgments about observational studies as they were for experiments. In the future, we could potentially test the intervention with groups of varying levels of prior knowledge to see the boundary conditions of effectiveness for the intervention in Experiment 3; perhaps the intervention would also be successful for participants like those in Bleske-Rechek et al. (2015), who were equally willing to make causal judgments about observational studies and experiments. Alternatively, participants with less prior knowledge may require even more foundational knowledge for the intervention to be effective.

One concern is that providing too much redundant information to participants with background knowledge will have no effect on learning. This may be why there was no improvement in Experiment 2; the instruction may have been too heavy-handed for a sample with more advanced levels of prior knowledge. When students are given more instructional information than needed, they may not spontaneously generating their own explanations about the content or engaging with the material more generally, which could impede learning (Richey & Nokes-Malach, 2013). Instead, the amount of instructional explanations should be as minimal as possible, but tailored for more naïve learners who may require more detailed explanations during the preliminary stages of learning (Renkl, 2002).

More advanced learners may benefit from strategies that involve less scaffolding in the sense of heavy-handed instructional explanations. In the future, we plan to test another iteration of the intervention in Experiment 3 that includes more practice problems, but with feedback. In Experiment 2, there was some evidence of learning in both the Control and Self Explanation groups, which means that people got better at discriminating between correlation and causation simply due to repeated practice with the measures. We did not see learning from practice in Experiment 3, presumably because the students did not have enough foundational knowledge to benefit from practicing with more examples. Since practice was effective for the more advanced sample, we want to test whether including practice with feedback during the intervention – with additional vignettes and the same correlation-causation measures – might improve correlation-causation discrimination even further.

## 5.6 Conclusions

Correlation-causation discrimination is a critical skill not only for scientists, but also for the general public. For example, imagine someone who reads about the outcome of a scientific study on Twitter, and learns that people who drink more coffee are more likely to have lung cancer. If the study is observational, but the reader is now convinced that drinking coffee causes lung cancer because the two are statistically related, they may modify future behaviors based on those conclusions. For example, they might stop drinking coffee to reduce the likelihood of lung cancer. However, such an intervention could be ineffective (or in some scenarios even harmful), if the two variables are related for some other reason (e.g., smokers are more likely to drink coffee than non-smokers, and smoking causes lung cancer). Reasoning about such causal relationships can become

even more complicated when that information conflicts with people's prior expectations about the study outcomes, and people may rely more on prior beliefs than considering the methodological design of the study to decide if these variables are causally related or not.

Thus, it is imperative to identify methods for improving scientific literacy, specifically methods to improve correlation-causation discrimination. Across three studies, we tested the efficacy of an intervention that uses a causal diagrams tutorial to improve correlation-causation discrimination. One of the desired learning outcomes was that in the future, participants can apply what they learned to actual studies they read about in the media. Ideally, in the future, they will be able to use that knowledge to recognize when authors make erroneous causal claims about observational studies. However, although our intervention improved correlation-causation discrimination, there remained substantial room for further progress.

Future steps could be to continue making iterative improvements in the hopes of finding an even more effective method of teaching students about why correlation does not imply causation. But our efforts may not be successful, and even if they are successful, it is unlikely that we will be able to reach everyone who is likely to encounter erroneous causal claims about correlational findings, which are rife in media articles (Cofield et al., 2010; Haber et al., 2018; Haneef et al., 2015). Thus, our work also highlights the need for authors to make accurate claims about study findings, and to reduce the demand on the reader. We found that people can learn to discriminate between correlation and causation, even when they encounter information that is congruent or incongruent with their prior expectations. However, there was still much room for improvement after the intervention, and so accounting for biases like "causal theory error" when writing about scientific findings may be necessary for people to draw appropriate conclusions.

## Appendix A Results for Interpreting Statistical Tests

Participants in the research methods sample were answered two questions about interpreting the results of different statistical tests, "Suppose you do a study with an independent variable and a dependent variable and you **run a** [**Pearson correlation/t-test**] to test if they are statistically related. You get a significant result. You should conclude…". For each test, they selected one option from the following: 1) Correlation does not imply causation, so you should conclude that the independent variable **does not** cause the dependent variable; 2) You should conclude that it is possible that the independent variable causes the dependent variable – you would need to know more about the study to make a determination; 3) You should conclude that the independent variable **does** cause the dependent variable; 4) I'm not sure. The order of the two questions was randomized, and they were also asked at the end of the post-test measures.

Neither of the statistical test questions (*t*-test or Pearson's correlation) provided enough information about the design of the study to draw a causal conclusion. In both scenarios, the correct answer was "You should conclude that it is possible that the independent variable causes the dependent variable – you would need to know more about the study to make a determination". These questions require a more nuanced understanding of causality and study design – that regardless of the statistical test, study design is a key factor in interpreting significant results. Previously, we have asked research methods students these questions on exams; even at the end of semester, students seem to struggle with this concept. If participants' performance improves after the intervention, this would be evidence that the intervention leads to far-transfer by means of applying correlation-causation discrimination to making inferences about statistical results.

I coded participants' responses as either correct or incorrect, and conducted a mixed-effects regression (separately for each question) to assess whether accuracy improved after the intervention; I included block × intervention as a fixed effect and a by-subject random slope (Table A3). There was a significant main effect of block for the *t*-test question, which was moderated by a significant block × intervention interaction when comparing the Worked Example and Self Explanation conditions. In the Self Explanation condition, there was no improvement in participants' responses from pre-test (50.91%) to post-test (50.91%). However, there was clear improvement in the Worked Example condition from pre-test (37.50%) to post-test (51.79%). Given that post-test performance was quite similar, the improvement is likely due to pre-test differences across the three groups. These results suggest that while an intervention may help with performance on this far-transfer task, there may be a limit to the benefits.

**Appendix Table 1 Testing for Improvement in Interpreting Statistical Tests**

| | Pearson's Correlation | | | *t*-test | | |
|---|---|---|---|---|---|---|
| Predictor | *b* | 95% CI [LL, UL] | *p* | *b* | 95% CI [LL, UL] | *p* |
| Block | -0.09 | [-0.18, 0.00] | .051 | -0.11 | [-0.20, -0.02] | **.014** |
| Intervention (AC vs WE) | 0.09 | [-0.02, 0.20] | .117 | -0.07 | [-0.18, 0.05] | .247 |
| Intervention (AC vs SE) | 0.08 | [-0.03, 0.19] | .175 | <0.01 | [-0.12, 0.11] | .944 |
| Intervention (WE vs SE) | -0.01 | [-0.12, 0.10] | .839 | 0.06 | [-0.05, 0.18] | .281 |
| Block × (AC vs WE) | -0.07 | [-0.20, 0.05] | .252 | -0.03 | [-0.16, 0.10] | .661 |
| Block × (AC vs SE) | -0.01 | [-0.14, 0.11] | .848 | 0.11 | [-0.02, 0.24] | .085 |
| Block × (WE vs SE) | 0.06 | [-0.07, 0.19] | .345 | 0.14 | [0.01, 0.27] | **.032** |

Note. AC = Analogical Comparison; WE = Worked Example; SE = Self Explanation; Boldface = $p < .05$.

## Bibliography


Adams, R. C., Challenger, A., Bratton, L., Boivin, J., Bott, L., Powell, G., Williams, A., Chambers, C. D., & Sumner, P. (2019). Claims of causality in health news: A randomised trial. *BMC Medicine*, *17*(1), 1–11.

Ainsworth, S., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science*, *27*(4), 669–681. https://doi.org/10.1207/s15516709cog2704_5

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*(1), 112.

Bakalar, N. (2016). Walkable Neighborhoods Cut Obesity and Diabetes Rates. *The New York Times*. https://well.blogs.nytimes.com/2016/05/24/walkable-neighborhoods-cut-obesity-and-diabetes-rates/

Bensley, D. A., Crowe, D. S., Bernhardt, P., Buckner, C., & Allman, A. L. (2010). Teaching and Assessing Critical Thinking Skills for Argument Analysis in Psychology. *Teaching of Psychology*, *37*(2), 91–96. https://doi.org/10.1080/00986281003626656

Blanchette, I., & Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. *Memory & Cognition*, *28*(1), 108–124.

Bleske-Rechek, A., Morrison, K. M., & Heidtke, L. D. (2015). Causal inference from descriptions of experimental and non-experimental research: Public understanding of correlation-versus-causation. *Journal of General Psychology*, *142*(1), 48–70. https://doi.org/10.1080/00221309.2014.977216

Bobek, E., & Tversky, B. (2016). Creating visual explanations improves learning. *Cognitive Research: Principles and Implications*, *1*(1), 1–14.

Bratton, L., Adams, R. C., Challenger, A., Boivin, J., Bott, L., Chambers, C. D., & Sumner, P. (2020). Causal overstatements reduced in press releases following academic study of health news. *Wellcome Open Research*, *5*.

Brownell, S. E., Wenderoth, M. P., Theobald, R., Okoroafor, N., Koval, M., Freeman, S., Walcher-Chevillet, C. L., & Crowe, A. J. (2014). How students think about experimental design: Novel conceptions revealed by in-class activities. *BioScience*, *64*(2), 125–137.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, *127*(4), 355.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120.

Cheng, P. C.-H., Lowe, R. K., & Scaife, M. (2001). Cognitive Science Approaches To Understanding Diagrammatic Representations. In A. F. Blackwell (Ed.), *Thinking with Diagrams* (pp. 79–94). Springer Netherlands. https://doi.org/10.1007/978-94-017-3524-7_5

Chi, M. T. (2013). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In *Advances in instructional psychology* (pp. 161–238). Routledge.

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

Cofield, S. S., Corona, R. V., & Allison, D. B. (2010). Use of causal language in observational studies of obesity and nutrition. *Obesity Facts*, *3*(6), 353–356. https://doi.org/10.1159/000322940

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*(4), 347.

Cummins, D. D. (1992). Role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 1103.

Evans, J. S. B., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382–389.

Fugelsang, J. A., & Thompson, V. A. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *54*(1), 15.

Fugelsang, J. A., & Thompson, V. A. (2001). Belief-Based and covariation-based cues affect causal discounting. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *55*(1), 70.

Garcia-Retamero, R., Müller, S. M., Catena, A., & Maldonado, A. (2009). The power of causal beliefs and conflicting evidence on causal judgments and decision making. *Learning and Motivation*, *40*(3), 284–297.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38.

Goedert, K. M., Ellefson, M. R., & Rehder, B. (2014). Differences in the weighting and choice of evidence for plausible versus implausible causes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 683.

Haber, N., Smith, E. R., Moscoe, E., Andrews, K., Audy, R., Bell, W., Brennan, A. T., Breskin, A., Kane, J. C., Karra, M., McClure, E. S., & Suarez, E. A. (2018). Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): A systematic review. *PLOS ONE*, *13*(5), e0196346. https://doi.org/10.1371/journal.pone.0196346

Halonen, J. S., Buskist, W., Dunn, D., Freeman, J., Hill, G., Enns, C., & others. (2013). APA guidelines for the undergraduate psychology major (version 2.0). *Washington, DC: APA.*

Han, M. A., Leung, G., Storman, D., Xiao, Y., Srivastava, A., Talukdar, J. R., El Dib, R., Morassut, R. E., Zeraatkar, D., Johnston, B. C., & others. (2022). Causal language use in systematic reviews of observational studies is often inconsistent with intent: A systematic survey. *Journal of Clinical Epidemiology*.

Haneef, R., Lazarus, C., Ravaud, P., Yavchitz, A., & Boutron, I. (2015). Interpretation of results of studies evaluating an intervention highlighted in google health news: A cross-sectional study of news. *PLOS ONE*, *10*(10), e0140889. https://doi.org/10.1371/journal.pone.0140889

Harrell, M. (2012). Assessing the efficacy of argument diagramming to teach critical thinking skills in introduction to philosophy. *Inquiry: Critical Thinking Across the Disciplines*, *27*(2), 31–39.

Hausmann, R. G., & VanLehn, K. (2007). Explaining self-explaining: A contrast between content and generation. *Frontiers in Artificial Intelligence and Applications*, *158*, 417.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology*, *93*(3), 579.

Kershaw, T. C., Lippman, J. P., & Fugate, J. (2018). Practice makes proficient: Teaching undergraduate students to understand published research. *Instructional Science*, *46*(6), 921–946.

Klahr, D., & Nigam, M. (2004). *The Equivalence of Learning Paths in Early Science Instruction: Effects of Direct Instruction and Discovery Learning*. https://journals.sagepub.com/doi/10.1111/j.0956-7976.2004.00737.x?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational Interventions to Advance Children's Scientific Thinking. *Science*, *333*(6045), 971–975. https://doi.org/10.1126/science.1204528

Koehler, J. J. (1996). The base rate fallacy reconsidered: Normative, descriptive and methodological challenges. *Behavioral & Brain Science*, *19*, 1–17.

Kreher, S. A., Pavlova, I. V., & Nelms, A. (2021). An active learning intervention based on evaluating alternative hypotheses increases scientific literacy of controlled experiments in introductory biology. *Journal of Microbiology & Biology Education*, *22*(3), e00172-21.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, *18*(4), 495–523.

Kuhn, D., & Dean Jr, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*(11), 866–870.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

LaCosse, J., Ainsworth, S. E., Shepherd, M. A., Ent, M., Klein, K. M., Holland-Carter, L. A., Moss, J. H., Licht, M., & Licht, B. (2017). An active-learning approach to fostering understanding of research methods in large classes. *Teaching of Psychology*, *44*(2), 117–123.

Lawson, T. J., & Brown, M. (2018). Using pseudoscience to improve introductory psychology students' information literacy. *Teaching of Psychology*, *45*(3), 220–225.

Leary, M. (2012). *Introduction to Behavioral Research Methods* (6th ed.). Pearson.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131. https://doi.org/10.1177/1529100612451018

List, A., Du, H., & Lyu, B. (2022). Examining undergraduates' text-based evidence identification, evaluation, and use. *Reading and Writing*, *35*(5), 1059–1089.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098.

McCrudden, M. T., Schraw, G., Lehman, S., & Poliquin, A. (2007). The effect of causal diagrams on text learning. *Contemporary Educational Psychology*, *32*(3), 367–388. https://doi.org/10.1016/j.cedpsych.2005.11.002

McLean, C. P., & Miller, N. A. (2010). Changes in critical thinking skills following a course on science and pseudoscience: A quasi-experimental study. *Teaching of Psychology*, *37*(2), 85–90.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*(1), 1–30.

Meijer, D. (2007). *Undergraduate psychology students' ability to transfer critical thinking skills* [M.A., California State University, Fresno]. https://www.proquest.com/docview/304706806/abstract/DBAD93969D1F475APQ/1

Michal, A. L., Seifert, C., & Shah, P. (2019). *Diagramming causal models improves correlation-causation discrimination*. 60th annual meeting of the Psychonomic Society.

Michal, A. L., Seifert, C., & Shah, P. (2021). *Effects of Prior Beliefs on Correlation-Causation Discrimination*. 62nd Annual Meeting of the Psychonomic Society.

Michal, A. L., Zhong, Y., & Shah, P. (2021). When and why do people act on flawed science? Effects of anecdotes and prior beliefs on evidence-based decision-making. *Cognitive Research: Principles and Implications*, *6*(1), 28. https://doi.org/10.1186/s41235-021-00293-2

Mill, J. S. (1872). *A System of Logic Ratiocinative and Inductive: 1* (Vol. 2). Longmans.

Mueller, J. F., & Coon, H. M. (2013). Undergraduates' ability to recognize correlational and causal language before and after explicit instruction. *Teaching of Psychology*, *40*(4), 288–293. https://doi.org/10.1177/0098628313501038

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 398–415. https://doi.org/10.1037/0278-7393.17.3.398

Owens, L. (2018). *Identifying Student Difficulties in Causal Reasoning for College-aged Students in Introductory Physics Laboratory Classes* [University of Cincinnati]. https://etd.ohiolink.edu/apexprod/rws_olink/r/1501/10?clear=10&p10_accession_num=ucin152231524077991

Parra, C. O., Bertizzolo, L., Schroter, S., Dechartres, A., & Goetghebeur, E. (2021). Consistency of causal claims in observational studies: A review of papers published in a general medical journal. *BMJ Open*, *11*(5), e043339.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.2307/2337329

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Penningroth, S. L., Despain, L. H., & Gray, M. J. (2007). A course designed to improve psychological critical thinking. *Teaching of Psychology*, *34*(3), 153–157. https://doi.org/10.1080/00986280701498509

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, *95*(4), 667.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Prentice Hall.

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

Renkl, A. (2014). *Learning from worked examples: How to prepare students for meaningful problem solving.*

Richey, J. E., & Nokes-Malach, T. J. (2013). How much is too much? Learning and motivation effects of adding instructional explanations to worked examples. *Learning and Instruction*, *25*, 104–124.

Richey, J. E., & Nokes-Malach, T. J. (2015). Comparing Four Instructional Techniques for Promoting Robust Knowledge. *Educational Psychology Review*, *27*(1), 181–218. https://doi.org/10.1007/s10648-014-9268-0

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. (2007). The incidence of "causal" statements in teaching-and-learning research journals. *American Educational Research Journal*, *44*(2), 400–413. https://doi.org/10.3102/0002831207302174

Seifert, C. M., Harrington, M., Michal, A. L., & Shah, P. (2022). Causal theory error in college students' understanding of science studies. *Cognitive Research: Principles and Implications*, *7*(1), 1–22.

Shah, P., Michal, A., Ibrahim, A., Rhodes, R., & Rodriguez, F. (2017). Chapter Seven—What Makes Everyday Scientific Reasoning So Challenging? In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 66, pp. 251–299). Academic Press. https://doi.org/10.1016/bs.plm.2016.11.006

Shanks, R. A., Robertson, C. L., Haygood, C. S., Herdliksa, A. M., Herdliska, H. R., & Lloyd, S. A. (2017). Measuring and advancing experimental design ability in an introductory course

without altering existing lab curriculum. *Journal of Microbiology & Biology Education*, *18*(1), 18–1.

Shi, J., Power, J., & Klymkowsky, M. (2011). *Revealing student thinking about experimental design and the roles of control experiments*.

Sibulkin, A. E., & Butler, J. S. (2019). Learning to Give Reverse Causality Explanations for Correlations: Still Hard After All These Tries. *Teaching of Psychology*, *46*(3), 223–229. https://doi.org/10.1177/0098628319853936

Sirum, K., & Humburg, J. (2011). The Experimental Design Ability Test (EDAT). *Bioscene: Journal of College Biology Teaching*, *37*(1), 8–16.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sweller, J., & Cooper, G. A. (1985). The Use of Worked Examples as a Substitute for Problem Solving in Learning Algebra. *Cognition and Instruction*, *2*(1), 59–89. https://doi.org/10.1207/s1532690xci0201_3

Thompson, V., & Evans, J. S. B. (2012). Belief bias in informal reasoning. *Thinking & Reasoning*, *18*(3), 278–310.

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies*, *4*(1), 1–17.

van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica*, *151*, 143–154. https://doi.org/10.1016/j.actpsy.2014.06.007

VanderStoep, S. W., & Shaughnessy, J. J. (1997). Taking a Course in Research Methods Improves Reasoning about Real-Life Events. *Teaching of Psychology*, *24*(2), 122–124. https://doi.org/10.1207/s15328023top2402_8

White, P. A. (1995). Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory & Cognition*, *23*(2), 243–254.