Scientific Research Publishing

Psychology, 2024, 15, 123-144
https://www.scirp.org/journal/psych
ISSN Online: 2152-7199
ISSN Print: 2152-7180

Rethinking the Pedagogy of Evaluating Causal Claims in the Psychology Curriculum

Richard M. Wielkiewicz

College of Saint Benedict, Saint John's University, St. Joseph, MN, USA
Email: rmwielk@gmail.com

How to cite this paper: Wielkiewicz, R. M. (2024). Rethinking the Pedagogy of Evaluating Causal Claims in the Psychology Curriculum. *Psychology, 15,* 123-144.
https://doi.org/10.4236/psych.2024.151009

Received: December 8, 2023
Accepted: January 23, 2024
Published: January 26, 2024

Abstract

Unfounded causal claims from the internet, the fact that randomized control trials (RCTs) cannot address many critical issues, and reports that scientific studies fail replication attempts suggest reconsidering how students learn to evaluate causal claims. Traditionally, students learn RCTs are at the top of the research methods hierarchy, and that they cannot infer causation from associations (e.g., correlations). Both traditions are debatable. Students need to learn how to evaluate causal claims they encounter in daily life as well as claims supported by scientific evidence. Students will become better critical thinkers from learning a definition of evidence that applies inside and outside the psychology laboratory and how to use anecdotes, associations, and RCTs in defense of causal claims. They must learn to question all evidence and seek patterns of supporting evidence for causal claims.

Keywords

Causal Claims, Evidence, Critical Thinking, Research Methods, Statistics, Pedagogy, Correlation and Causation

1. Introduction

Social media emit a flood of causal claims supported by weak or nonexistent evidence. For example, some Americans argue that COVID-19 vaccinations might make one magnetic (Schwarcz, 2021) or contain an electronic tracking chip (Schoolov, 2021). No causal pathway exists for magnetism or a tracking chip to be transferred via a liquid vaccine and no evidence supports those claims. Yet, these beliefs persist among some people who refuse a medical intervention that would benefit them, while demanding medications that will not help (Taccone et al., 2022). The traditional curricular emphasis on well-designed studies with random assignment leaves students unprepared to evaluate such claims.

A randomized control trial (RCT) is an experiment that has random assignment of subjects to conditions which demonstrates the researcher has manipulated the independent variable. Traditionally, students learn RCTs are at the top of the research methods hierarchy and that RCTs demonstrate a causal relationship between the independent variable and the dependent variable. However, many scientific studies do not work when they are repeated or replicated, which means interpretation of scientific studies, even those with random assignment, must be approached with skepticism. Whether based on observations, associations, or RCTs, a valid causal claim needs support from a pattern of evidence, not a single study. APA Guidelines for the Undergraduate Psychology Major (American Psychological Association, 2013: p. 21) strongly emphasize scientific method and pursuit of scientific knowledge. The research methods foundation can be strengthened by adjusting pedagogy to account for failures to replicate studies and expanding beyond the pursuit of scientific knowledge to the pursuit of knowledge involving all types of causal claims including baseless conspiracy theories. Critical thinking is the most frequently addressed career skill in introductory psychology (Richmond et al., 2021) but students should learn to evaluate the entire range of causal claims from conspiracy theories to RCTs.

Most students learn research methods using the Campbell model (e.g., Campbell & Stanley, 1966) which emphasizes the value of random assignment, categorizes threats to validity of a causal claim, and examines a study's internal versus external validity. Some assumptions of this model can be challenged, especially the ideas that correlation cannot support a causal claim and placing RCTs at the top of a hierarchy of scientific methods. RCTs are essential evidence in many contexts such as evaluating the efficacy of treatments, but in many other contexts RCTs are impossible or unethical and associations and observations must be used to support causal claims. Further, many scientific studies do not produce the same results when they are repeated or replicated. These failures to replicate indicate students should learn to skeptically evaluate all scientific studies. Evidence supporting a causal claim may come from anecdotes, observations, associations, *and/or* RCTs, but all forms of evidence need to be interpreted with caution.

The paper begins with a definition of evidence applicable to a wide variety of causal hypotheses followed by an overview of research methods and statistical concepts that assist in evaluating the strength of a causal claim. Next, the paper addresses the question of what is meant by a causal claim with examples of critical thinking about causal claims that do not come from laboratory studies. The paper then reviews why replication failures occur and their implications for making causal claims from scientific evidence. The last section summarizes guidelines for critically evaluating causal claims.

2. Rethinking Critical Thinking Pedagogy for Psychology Students

First, I propose a definition of evidence that applies to issues students face in

daily life *and* scientific studies. The next section reviews core concepts in research methods that assist in evaluating the validity of causal claims. The remaining sections illustrate how to evaluate causal claims while placing less emphasis on RCTs.

2.1. Definition of Evidence

I propose psychology courses begin by defining *evidence* as a *verifiable or repeatable observation that is falsifiable*. Thus, evidence can be anecdotes, fingerprints, surveys, eye-witness testimony, DNA, videos, RCTs, associations, frequencies, or any documented record of events. Evidential requirements vary from situation to situation. Judging guilt or innocence on a jury requires different evidence than establishing efficacy of a psychological treatment. Thus, evidence provides a broader basis from which to argue causal claims and prepares students to look at the evidence whether the claim emerges from social media, a jury trial, or a RCT.

Consider the causal claim that Person X committed a crime. Finding fingerprints of Person X at the crime scene is one piece of evidence. The observation consists of comparing the crime scene prints with known prints from Person X. If the prints match, Person X is likely to have been at the crime scene. The observation is verifiable because a judge, jury members, or experts can assess whether the prints from Person X match those at the crime scene. Other evidence may consist of DNA found at the crime scene, eyewitness testimony, identification of the suspect in a photo lineup, etc. The more observations that associate Person X with the crime, the more likely Person X committed the crime. However, skepticism applies to all evidence. Experts and nonexperts might disagree whether fingerprints taken at two separate locations and times actually match. A *pattern* of credible evidence should support the claim that Person X committed (caused) the crime.

Although observations are the main evidence in a criminal jury trial, some circumstances demand a RCT. An anecdotal observation may lead to the discovery of a treatment, such as the accidental discovery of the antibiotic, penicillin. However, the antibiotic effect must hold up in RCTs for the treatment to become mainstream. For medical or psychological treatments, RCTs are the "gold standard" for evaluating treatment efficacy. However, one must take a skeptical view of a single RCT and seek a pattern of supporting evidence.

Psychologists dismiss nonsystematic observations and anecdotes in favor of more sophisticated evidence such as RCTs. However, observations can support causal arguments. For example, a video of a person committing a crime provides convincing evidence that the person in the video is guilty. In contrast, eye-witness testimony is unreliable evidence (Loftus, 2019). An observation on video would have more weight than an eye-witness report with no corroborating evidence. Observations varying in their degree of sophistication and strength are the evidence for all causal claims.

2.2. Pedagogical Ideas

Students need to understand the definition of evidence, a *verifiable or repeatable observation that is falsifiable*. A video record of an observation has the element of verifiability, whereas an eye-witness report is more difficult to verify. An eye-witness account of primate tool-use by a trained observer might be compelling evidence that cognition was a causal factor in the use of the tool. Falsifiability comes from the fact that others have the opportunity to observe under similar conditions and either verify or disconfirm the original finding. However, some events are so unusual, such as a claimed encounter with extraterrestrials, that they are essentially unfalsifiable. While one reported sighting of extraterrestrials may be unfalsifiable, other physical evidence such as extraterrestrial technology, videos, and radar records, could be used to support the claim that extraterrestrials caused the findings. Multiple independent sightings lend additional support to the original report. Students need to subject all evidence to a critical thinking process to evaluate the strength of support it provides for the causal claim. The goal is to seek a pattern of evidence that supports the causal claim. Either weak evidence or no evidence invalidates the claim.

Quizzes can verify that students understand the definition of evidence. Students can discuss examples of evidence in small groups and evaluate the strength of support for the underlying causal claim. Either the course content or students could provide the examples. Following is a list of questions to which each small group can provide a response.

- What is the evidence?
- What is the underlying causal claim?
- Is the evidence verifiable or repeatable?
- Is it possible to falsify the evidence or claim?
- On a scale from Extremely weak to Extremely Strong, what is the strength of this evidence?
- Why did you reach this conclusion?

3. Basic Research Methods and Statistics Concepts

The Task Force on Statistical Inference (Wilkinson et al., 1999) solidified the place of Null Hypothesis Significance Testing (NHST) in research. To read research articles or become researchers, students must understand NHST logic. The strength of evidence represented by a scientific study is linked to components of NHST such as probability, alpha level, effect size, and power. A doubtful emphasis of the Task Force was on the importance of random assignment which "allows for the strongest possible causal inferences free of extraneous assumptions" (p. 595). Random assignment does not always accomplish this, and one RCT alone cannot validate a causal claim. See Wielkiewicz (2022) for a detailed discussion of the concepts discussed below.

3.1. Measured and Manipulated Variables

The introductory course should define measured and manipulated variables

(Meltzoff & Cooper, 2018; Morling, 2021). A manipulated variable is a necessary component of RCTs. Random assignment to groups indicates a manipulated variable. If a variable is manipulated and the result is statistically significant, this is evidence that the manipulated variable causes the changes in the dependent variable. Dependent variables in a RCT, and subject variables such as gender, age, or diagnosis are measured variables. A study assessing the relationship between or among measured variables is only capable of testing the association or correlation among the variables. Although no single study can establish a causal relationship, other things being equal, a RCT is stronger evidence for a causal claim than a study that assesses associations or correlations. A successful replication of a RCT greatly strengthens the underlying causal claim. However, a single study demonstrating an association requires a broader pattern of evidence than a simple replication to validate the causal claim. In order to understand the nature of a scientific study manipulated and measured variables need to be labeled correctly. The meaning of the terms independent and dependent variable varies with the context. Identifying measured and manipulated variables is a key to comprehending the design of a study and the strength of support it lends to a causal claim.

3.2. Cognitive Biases

Cognitive biases may influence almost any aspect of the design or interpretation of research or an anecdotal observation. In other words, cognitive biases interfere with scientific objectivity. Stapleton (2019) reviews cognitive biases including confirmation bias, seeing patterns in place of randomness, confusing an association with causation, ignoring the importance of sample size and representativeness, and not accounting for the base rate. Whether data are from a large-sample survey, a literature review, a RCT, or personal experience, cognitive biases may influence collection or interpretation of observations.

For example, confirmation bias is the tendency to select information that supports a pre-existing belief and ignore information that contradicts the belief. Stapleton (2019) asserts that confirmation bias may be the most difficult cognitive bias to overcome in both everyday life and in research activities. Those who do not accept that climate change is real may ignore data that supports the predictions made by climate scientists while focusing on information that supports their own belief such as cold weather days, snowstorms, people ice fishing, and other isolated cold weather events. Further, such individuals gather their biased evidence and consider the issue decided. They do not look for reports prepared by climate scientists, and they see those who demand action to avert irreversible global heating as being misled or falling for a "hoax." Confirmation bias can be countered by slowing down decision processes and looking for contradictory evidence.

Lilienfeld et al. (2009) suggested that bias could be counteracted by instructions to seek counterexamples, slowing down decision processes, and specific education regarding cognitive biases. In an empirical study, Čavojová et al.

(2020) found that scientific reasoning skills, measured by an author-developed scale, were correlated negatively with susceptibility to bias, dogmatism, and holding intellectually suspect beliefs. Higher scientific reasoning skills were associated with less susceptibility to bias, less dogmatism, and lower amounts of suspect beliefs. The study had high statistical power and medium effect sizes, suggesting favorable conditions for future replication. Thus, the type of training in scientific method provided by the psychology curriculum is associated with avoiding cognitive bias. Echoing the standard conclusion to many empirical studies, "further research is needed." Avoiding the impact of bias and noting potential sources of bias are vital components of evaluating causal claims in any context.

Another source of bias is human evolutionary history. Most human evolution occurred in an environment in which humans were both prey and predator (Mealey, 2000). Boudry et al. (2015) argue that some biases in human cognition result from mechanisms that benefited humans in primitive times. For example, if an individual observed an association between rustling leaves and seeing a predator, the most adaptive behavior was to treat this relationship as causal and get away from rustling leaves. These inherited mechanisms err on the side of caution because the cost of a false positive (expending energy to avoid the situation) is minor compared to the cost of a false negative (death or injury). The evolutionary advantage of making causal inferences from associations increases the likelihood of accepting the validity of causal claims supported by associations. Students should learn about this evolutionary bias, seek patterns of supporting evidence for all causal claims, and avoid inferring causality from a single association.

For example, if a student obtains relief every time they take a medication for a headache, they are likely to infer a causal relationship between taking the medication and headache relief (Blanco, 2017; Matute et al., 2015; Matute et al., 2011). Imagine the student omits taking medication and finds that the headache continues. This evidence strengthens the causal claim but threats to validity remain. So, the student attempts another intervention. The next time they experience a headache, the student stops reading and gets a glass of water but does not take medication. Disappearance of the headache breaks the causal link to medication and other hypotheses become viable. This type of causal reasoning is common so learning to skeptically evaluate such causal claims will strengthen students' critical thinking skills. A useful in-class exercise is having students discuss the evidence needed to conclude a medication, food supplement, or psychological intervention was improving their quality of life.

Kahneman (2011) writes about System 1 (fast and intuitive) and System 2 (slow, deliberate) cognitive processes. See Martín and Valiña (2023) for an updated review of this approach to understanding cognition. According to Kahneman, conspiracy theories and other irrational beliefs take advantage of System 1 by pairing words that evoke negative emotion (pedophile, rapist, corruption, hoax, etc.) with policies opposed by some groups, such as teaching Critical Race

Theory or addressing the climate crisis. The emotional conditioning blocks System 2 from being engaged. There is an "illusion of validity" (Kahneman, 2011: p. 209ff), a belief that persists, even in the face of contradictory evidence.

Detailed instruction in evaluating causal claims will be useless unless students let go of the fast and intuitive processes that lead to false conclusions. Social media tend to elicit fast and intuitive System 1 processes (Moravec et al., 2018). To counter these cognitive biases, people need to slow down, evaluate the available evidence, and apply critical thinking to all causal claims. Further, when one event follows another, students should resist the evolved tendency to link them in a causal chain without additional supporting evidence.

3.3. Type I Errors, Type II Errors, Power, and Effect Size

Statistical decisions made about research outcomes are based upon probability, so errors are inevitable. Alpha is the probability that the null hypothesis is incorrect assuming the null is accurate. A low probability (i.e., $p < 0.05$) of the null being correct, leads to rejection of the null, and support for the research hypothesis. Incorrect rejection of the null hypothesis means a Type I error (false positive) has occurred. Type I errors are difficult to identify because they occur under the same conditions as a statistically significant result. If an unidentified Type I error has occurred, attempts to replicate the finding will most likely fail, because the original finding represents a statistical aberration. Thus, a single RCT provides weak support for the underlying causal hypothesis unless the study has been replicated or factors such as a large effect size indicate a Type I error is unlikely. A Type II error occurs when the results are not statistically significant but there is actually an undetected effect in the populations, i.e., a false negative. Researchers suspect a Type II error when results are not statistically significant, but the effect size is medium or large suggesting the study did not have enough statistical power. Statistical power is the probability of not making a Type II error. The larger the sample size, the more power, and the greater the probability of not making a Type II error, meaning that the results are likely to be statistically significant, if there is an effect in the populations.

Effect size (*ES*) measures the strength or importance of a result. Introductory classes usually define *ES* as a difference between populations in units of the *SD* for the population of individuals. The correlation coefficient is also a measure of *ES*, along with measures of variance accounted for. Studies with large effect sizes are likely to be robust and replicable, whereas those with very small effect sizes provide weak support for the underlying causal claim. Type I and Type II errors, *ES*, alpha, and statistical power can explain why studies may be difficult or impossible to replicate. They are all part of the toolbox that students need to evaluate causal claims.

3.4. Pedagogical Ideas

Ensuring psychology students have adequate knowledge of statistics and research methods is a challenging instructional problem because only psychology

majors typically take statistics and research methods. Psychology minors, nursing students, pre-med students, and others do not. Thus, what level of competence in statistics and methodology should non-majors have? One approach is to cover a standard unit on statistics and methodology in introductory psychology and reinforce the principles in other psychology courses. Topics covered in such a unit might include the definition of evidence, a discussion of using evidence to support causal claims, measured versus manipulated variables, basic research designs, Null Hypothesis Significance Testing, the role of probability and statistical significance in research, Type I and Type II errors, and replication failures. Instructors should integrate these topics under the concept of using a pattern of evidence to support causal claims.

4. What Is a Causal Claim?

Reiss (2009) analyzed the problem of developing a unified model of causality for the social sciences. After examining counterfactual, regularity, probabilistic, mechanistic, and interventionist accounts of causality, Reiss concluded that none of these works in all situations. Instead, Reiss proposed that these accounts of causal models do not define causality so much as set a context or set of "test conditions" or evidence for the existence of causal relationships, an approach which Reiss called "evidential pluralism" (p. 27).

> The idea behind evidential pluralism is that evidence of a variety of kinds—say, probabilistic, mechanistic, regularity—can bear on a causal hypothesis and strengthen it. Especially when evidence from two or more different sources speaks in favor of the hypothesis, our confidence in the hypothesis should be boosted… Since any given method is fallible—as shown by the counterexamples to the various accounts—the epistemically responsible strategy is to bring as much evidence as possible to bear on the [causal] hypothesis at stake, and confirmation from a number of independent methods is one and perhaps the only way to be reasonably confident about the truth of the [causal] hypothesis (Reiss, 2009: p. 27).

For example, one purpose of causal analysis is to establish that manipulation of X causes changes in Y. We trust in the efficacy of clinical treatments evaluated in double-blind, placebo-controlled trials with random assignment (i.e., RCTs). RCTs, however, may establish the efficacy of a medication but fail to establish boundaries of how to prescribe it (Worrall, 2010). For example, while a RCT is essential to establish efficacy of a medication, random assignment will not contribute to determining the appropriate dose by age. Age is a measured characteristic which a researcher cannot randomly assign and cannot manipulate. Generally, claims that manipulation of X causes a change in Y require RCTs. But other research methods have important roles in establishing the details of treatment and supporting the causal claim.

Worrall (2010) argues that placing types of evidence in a hierarchy with RCTs

at the top is an illusion. RCTs have too many constraints in terms of the subjects selected, how long the study lasts, dosages, etc. to generalize to the real world of treatment. Thus, even a single-subject pretest-posttest design can provide insight into conditions that influence treatment efficacy. Drake et al. (2004) also concluded that RCTs are useful to establish treatment efficacy but other designs including qualitative and observational studies can contribute to clinical decisions. In sum, students should learn to value *all* types of evidence in making causal arguments and seek a pattern of evidence in support of a causal claim.

4.1. Breaking down the Classic Experiment and Making Causal Claims from Associations

"You can't infer causation from correlation."

"Correlation is not causation."

"A causal claim cannot be successfully defended with a single correlation study."

These phrases represent a principle that psychology students learn but they oversimplify a complex problem in critical thinking (Pearl, 2018). Many questions about causality (e.g., the climate crisis, criminal guilt or innocence, conspiracy theories, smoking and human health, etc.) cannot be argued using RCTs (Kenny, 2019). The details of designing scientific experiments using random assignment, threats to validity, and precise control over the independent variable are essential topics. However, the complexity of the world outside the laboratory requires a broader understanding of constructing causal arguments. RCTs cannot be used to study some issues (i.e., cigarette smoking and human health), which leads to making causal arguments using associations and other research techniques.

Hatfield et al. (2006) argue that the phrase "correlation does not imply causation" is misleading because the statistical method of correlation could involve either manipulated or measured variables. The rule "correlation does not imply causation" should state that one study involving an association between measured variables does not imply causation. If the evidence consists of associations, a pattern of evidence consisting of a variety of studies and observations is needed to support a causal claim.

4.2. Lecture Example: Breaking down Complexities of Causal Claims

Figure 1 summarizes important aspects of the correlation-causation quandary and would be an excellent lecture slide. The character entered a statistics class believing correlation implies causation. After the class, the character rejects this belief. Then, the character rejects the causal claim attributing the change to taking the class by agreeing that "maybe" the class was helpful. This is a single-case pretest-posttest design, a design that Morling called a "really bad experiment" (Morling, 2018: pp. 308-309; Morling, 2021), because it has so many threats to the validity of a causal claim. The character could have learned the principle

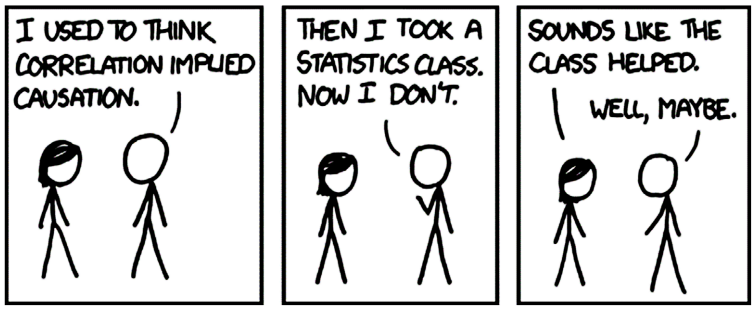

Figure 1. Correlation and Causation. https://xkcd.com/552/ CC BY-NC 2.5.

from casual reading or another class, undermining the claim that taking the class caused the change in the character's understanding of correlation and causation. However, this character could construct a strong causal claim based upon their experience. If they remember learning the principle from the class and not another source, a causal argument has support based upon anecdotal testimony. The course syllabus and/or exam questions could provide more support. A survey of students in the class could add more evidence. Thus, a pattern of evidence could support the claim that classroom instruction caused the change in knowledge of correlation and causation.

Continuing the thought experiment, imagine a researcher randomly assigned the character to the *experimental* group of a study investigating the correlation-causation issue. From the character's point of view, nothing has changed. Random assignment indicates an experimenter was able to manipulate presence versus absence of correlation/causation instruction, supporting the causal claim that instruction improves students' understanding of correlation and causation while in theory eliminating alternate explanations. However, support of the causal claim relies on obtaining a statistically significant difference between group *means*. For any individual, other explanations (in theory, equalized across groups by random assignment) are possible. A person in either group could acquire the knowledge outside the study, whereas an individual in the experimental group could fail to acquire the knowledge because of inattention or absence.

This fictional example illustrates how a researcher can manipulate "instruction in correlation and causation," and that instruction can cause differences in student learning. However, questions remain because the effect is a difference in group means. Motivation, attention, intelligence, attendance, and other variables are likely to be associated with learning about correlation and causation. Further, support of the causal claim assumes that random assignment created equivalent groups. This is not known for certain unless the researcher measures potential confounding variables and subjects them to more complex causal analyses (Pearl, 2018). Undergraduates are unlikely to employ complex mathematical causal models while thinking critically about everyday issues. Instead, they should learn to evaluate the evidence and reach a conclusion based upon whether the pattern of evidence supports a causal claim. For example, a pattern of evidence might consist of several studies that show an association between the

cause and effect using different measures. Further support for the causal claim might come from anecdotal reports in which the cause (i.e., a psychological treatment) preceded the predicted effect (improvement in social functioning). Another way to address these uncertainties is to replicate the original finding and empirically investigate the role of other variables. With each replication, it is less likely that extraneous confounding variables account for the basic finding.

Pedagogical Ideas

Instructors could use **Figure 1** as a prompt in a large- or small-group discussion, or an individual quiz. Students could respond to these questions:

- Summarize the content of the cartoon.
- Explain the meaning of the character's response of "WELL, MAYBE" to the question of whether the class helped.
- List examples of evidence that could support the claim that the character learned about correlation and causation from the class.
- Explain the implications of assuming the character was in the experimental group in a study of correlation and causation instruction.

4.3. Climate Change Example

For over a century, experiments have shown CO_2 causes heat to be retained in a system instead of being reflected back into space (Marx et al., 2017) and evidence human activity is causing a climate crisis is vast (Intergovernmental Panel on Climate Change, 2022). Thus, decreasing CO_2 emissions should cause a decline in global heating. This causal claim is made without any possibility of conducting a RCT. There is no "control" earth. Measures confirming that CO_2 concentrations are rising as a result of human activity support the causal claim. Further, increasing global temperatures, melting of glaciers and sea ice, rising sea levels, more destructive storms, drought, and other evidence is associated with CO_2 increases. A single association between the predicted effects of global heating and rising CO_2 is inadequate to validate the hypothesis that human activity is causing global heating. The pattern and extent of evidence validate the causal claim. Further, climate scientists have accurately predicted increasing global temperatures, more intense storms, sea level rise, and other events providing essential validation of global climate change models. Accurate predictions about future events from a causal hypothesis provide affirmative evidence.

Various claims denying the climate crisis have proliferated (Biddlestone et al., 2022). Some such claims deny climate change is happening, perhaps reflecting a cognitive bias in how or what evidence is interpreted. Disinformation and political polarization create a context for the climate crisis in which it is difficult to focus on the evidence (Lewandowsky, 2021). Asking advocates to present their evidence is one way to challenge such claims. Often such claims are made without any evidence or by citing one inadequate observation to counter the massive amount of evidence supporting the claim that human activity is causing the climate crisis. For example, a senator brought a snowball into the U.S. Senate as proof that global heating was not occurring (Barrett, 2015). Although meeting

the definition of evidence, this single observation is not adequate to counter the massive amount of evidence that shows global heating is occurring. Further, a single cold day or snowfall is within the expected variability of temperature and cannot be interpreted as evidence of a downward trend. To cope with issues involving measured independent variables, students must weigh the totality of evidence, including associations, in judging the validity of a causal claim, while also identifying weak, unverifiable, or unfalsifiable evidence.

RCTs also cannot resolve the question of what causes a psychological disorder. It is impossible to randomly assign human subjects to groups with and without a disorder, because disorders are a measured characteristic. Further, if a researcher hypothesized that some factor such as a history of child abuse, was a causal factor in clinical depression, it would not be ethical to randomly assign children to abusive and non-abusive environments or to subject them to conditions that might lead to a psychological disorder. Instead, such factors need to be measured through self-reports. RCTs have limited use in evaluating causal hypotheses about causes of psychological disorders, and associations are an essential component of arguing causal claims when RCTs are impossible.

Pedagogical Ideas

The climate crisis is relevant in many psychology classes, especially environmental or conservation psychology. In statistics classes, databases of temperatures or atmospheric CO_2 concentration can be used as statistical examples. In principles of learning, behavioral principles can explain energy usage and how behavioral incentives like tax credits can modify energy use or foster purchase of more efficient appliances and vehicles (Wielkiewicz, 2016). In social and community psychology, the social impact of climate change (migration, flooding of island nations, etc.) can be discussed. These examples provide opportunities to discuss the causal claim that CO_2 emissions are causing a climate crisis.

Using psychological disorders as an example, students could also discuss other research questions that cannot be addressed with RCTs.

4.4. Three Directions of Causality

The three directions of causality provide simple rules for evaluating a causal claim based upon an association. Aron et al. (2011) state: "If two variables have a clear linear correlation, we normally assume that there is something causing them to go together. However, you can't know the **direction of causality** (what is causing what) just from the fact that two variables are correlated" (p. 82). Given a correlation between X and Y, three hypotheses about causation are viable, the hypothesis that X causes Y, the hypothesis that Y causes X, and the hypothesis that a third variable could be causing both X and Y. By considering each possibility, one has a method for evaluating causal claims based upon associations. For example, a positive correlation exists between atmospheric CO_2 concentrations and global heating. Experiments demonstrate the direction of causality is from increased atmospheric CO_2 to increasing temperatures (Marx et al., 2017) and numerous studies have eliminated variables that might cause both

increasing CO_2 and increasing temperatures (Intergovernmental Panel on Climate Change, 2022). The causal claim that global heating results from human-caused emissions of CO_2 has overwhelming support from a pattern of evidence.

4.5. An Example of a Causal Claim without a RCT

An example of causal reasoning relevant in many psychology classes is the thesis of Edward L. Thorndike (1911) who began a new era in the study of learning cited in many modern texts. The context of Thorndike's experiments (his word choice) was a popular movement reporting stories about intelligent animal behavior, such as cats that could open doors or latches. The belief was that such examples supported Darwin's (1964) theory of evolution because the beginnings of human reasoning could be seen in animals. Thorndike pointed out that anecdotes are often recorded for entertainment and not scientific discovery and anecdotes about animal intelligence reported animals at their best and ignored the fact that hundreds of animals sit helplessly meowing and yowling but no one turned the event into a circulating anecdote. Although anecdotes are based upon observation, most are one-time events that cannot be verified or subjected to falsification. By themselves, they are a weak form of evidence.

Thorndike placed cats and other animals into situations like those that anecdotes reported they could intelligently solve. Anecdotes about animal intelligence predicted the puzzles would be solved quickly and suddenly. When this prediction was falsified, the animal intelligence hypothesis received a contradiction that has stood the test of time and led to the law of effect, which states that animal learning results from a gradual process, not insight. Thorndike described his puzzle boxes in detail and recorded the time it took each animal to trigger the escape mechanism and exit the puzzle box for each trial—all operations that could be verified or repeated. In fact, these operations have been repeated with electromechanical devices by innumerable researchers (Chance, 1999). Thorndike's work illustrates that carefully designed control groups and/or random assignment are not necessary for arguing causal claims.

Citing Thorndike's puzzle box studies in a course context provides an opportunity to help students learn about causal claims. Students can be asked what type of research methodology Thorndike's studies represent. They can also be asked to determine what causal claim is being tested and the strength of support provided by these studies. Thorndike's studies illustrate that animal learning is slow and gradual, a function of experience. The puzzle boxes were not solved suddenly as though the animal was analyzing the problem and solving it via reasoning. These studies can be classified as within-subjects demonstrations with each animal serving as its own control. The studies showed animals learned slowly with no evidence of insight and no RCT was involved. This example can be used to illustrate weaknesses of anecdotal observations, why anecdotal observations should be subjected to additional testing, and the process of arguing a causal claim in the absence of a RCT.

4.6. Pattern and Parsimony, and Conspiracy Theories

Pattern and parsimony (Morling, 2018) or the method of signatures (Abelson, 1995) is one way to interpret a large volume of evidence pointing to a causal relationship. Given a pattern of evidence, the most parsimonious explanation may be that a causal explanation underlies the evidence. The classic example is the causal hypothesis that cigarette smoking causes lung cancer, which is impossible to ethically study in humans with RCTs. A pattern of associations between cigarette smoking and lung cancer in humans using various indices supports the causal claim. Whether based on associations or RCTs, a valid causal claim needs support from a pattern of evidence. Conspiracy theories violate the principle of parsimony because the causal pathways they argue lack plausibility, testability, and evidence. Further, failed predictions often characterize conspiracy theories, which drastically undercut their plausibility. When advocates of a causal claim cannot cite credible evidence or refuse to consider that evidence is relevant, the claim is most likely false.

4.7. Pedagogical Ideas

Students who spend substantial time engaged in social media will encounter obtuse conspiracy theories and other claims that endanger themselves and society and they should be prepared to argue effectively about such claims. A broad approach to the kinds of evidence that can support a causal claim will prepare students to skeptically evaluate causal claims they encounter via the internet, social media, and daily living. Further, it will contribute to meeting "Goal 3. Ethical and Social Responsibility in a Diverse World" of the APA Guidelines for the Undergraduate Psychology Major (American Psychological Association, 2013: p. 26).

5. Failures to Replicate

Researchers assume that scientific studies, particularly RCTs, are convincing evidence in support of causal claims. However, it has been challenging to replicate results of published studies in psychology (Ioannidis, 2005; Marek et al., 2022; Open Science Collaboration, 2015), biomedical research (Errington et al., 2021), and other areas (National Academies of Sciences, Engineering, and Medicine, 2019). The Open Science Collaboration successfully replicated only 35 experimental and correlational studies out of 97 attempts (Open Science Collaboration, 2015). Errington et al. (2021) in the Reproducibility Project: Cancer Biology, attempted to replicate 158 high impact experimental effects. Only 42 of 97 positive effects successfully replicated despite higher sample sizes in the replications. Replication failures are a normal component of scientific progress. Students should understand that replication is a key to establishing a pattern of evidence supporting a causal claim.

Several issues contribute to replication failures: the complexities of replicating a study from a published description, publication bias, lack of statistical power in

the replication study, abuse of NHST, questionable research practices, and flaws in random assignment.

5.1. Complexities of Successful Replication

Independent replication of a published study is challenging because it is difficult to describe every relevant detail when journals have limited pages. For example, if the laboratory facilities of an institution are well-lighted, modern, and close to classrooms, the mood of subjects may be more positive than subjects in a replication conducted in a poorly lit basement. Thus, a replication may fail because the replication environment differs from the original study. Further, limited journal space may require that key details be omitted, though they may play a role in the outcome. Differences in mechanical equipment, measurement techniques, room temperature, paint colors, and other details may cause failures to replicate original findings.

5.2. Publication Bias

Publication bias occurs because most studies selected for publication have statistically significant results. A published study may have a larger effect size (*ES*) than typical of published and unpublished studies exploring the phenomenon. Editors publish the one extreme study, a Type I error, while other studies of the same problem remain unpublished. Another name for this is the file-drawer problem which is the possibility that a statistically significant study is a Type I error because unpublished attempts to demonstrate the same phenomenon are filed away, unpublished. Meta-analyses include a statistical test estimating the probability that the results are based upon a few statistically significant studies while the majority of effects are unpublished (File Drawer Problem, 2007). The more studies or effects included in the meta-analysis, the less likely the file-drawer problem is an issue.

5.3. Statistical Power

Maxwell et al. (2015) state that exact replications may fail because the replication study is underpowered. They believe a replication needs power of 0.90 to 0.95 (versus 0.80 power in most original studies). They suggest the confidence limit of the reported ES that is closest to zero should guide sample size choices, leading to larger sample sizes. If the goal of a replication is to show that the *ES* is zero, the replication study needs sample sizes ranging from 1714 to 10,000+ per group. A failure to replicate is likely to be underpowered and unlikely to present solid evidence that the *ES* is zero due to low power. If the power of a study is close to 50%, which occurs when the probability of the result is just below the cutoff value of 0.05, a successful replication will require much more statistical power. Typically, this would require a substantial increase in sample size.

5.4. Abuse of NHST and Questionable Research Practices

Flora (2020) says the "primary culprit" for replication failures is overreliance on

and abuse of Null Hypothesis Significance Testing (NHST) with too much emphasis placed on statistical significance while ignoring effect sizes. Flora contrasts NHST which determines whether there is any effect size at all to using the *ES* and its confidence interval (*CI*) as the leading element and focus of interpretation. Questionable Research Practices (QRPs) represent an abuse of NHST (John et al., 2012). QRPs include testing for significance and then deciding whether to add more subjects, running a study several times and reporting only the significant result, using significance testing to make decisions about outliers, and rounding *p*-values so they are significant. QRPs distort the validity of the 0.05 criterion for statistical significance and lower the probability of successful replication.

5.5. Problems with Random Assignment

Random assignment supposedly equalizes groups on an infinite number of potential confounding variables, so the manipulated variable is the only difference between the groups. However, random assignment can fail, even with large sample sizes (Goldberg, 2019; Schmidt & Oh, 2016; Worrall, 2010). Small sample sizes (~24 or fewer) fail to adequately represent the populations which decreases the likelihood of group equivalence (Schmidt & Oh, 2016). If sample sizes are too small, variability between samples contributes to difficulty in replication. Thus, in some cases a statistically significant result in a RCT might be accounted for by a confounding variable instead of the manipulated variable (which has no effect). A future replication is then likely to fail. In sum, random assignment does not always work perfectly, and use of random assignment does not stand alone as a cornerstone of causal arguments. In contrast, although a single RCT might have an unidentified confounding variable, each subsequent successful replication lowers the odds that a confounding variable explains the results.

Pedagogical Exercise

Although the present context emphasizes flaws of random assignment, it remains an essential component of studies supporting causal claims. A study by Sawilowsky (2004) used a Monte Carlo design to reportedly demonstrate successful random assignment with a sample size of only two per group ($n = 2$). Sawilowsky created a dataset for which each individual in the population had 7500 scores representing potential confounding variables. Then a sample of $n = 4$ was drawn and randomly assigned to two groups of $n = 2$ each. Independent groups *t*-tests performed on all 7500 variables determined whether the two groups differed on any of the 7500 potential confounding variables. With only 33 variables out of 7500 (0.44%) showing statistically significant differences, Sawilowsky concluded that random assignment is effective even with this minimum sample size. In advanced courses, students could read this article and subject it to a critical analysis. Many questions can be asked about this finding.

- Is this finding surprising or unique?
- What are the implications of the extremely low sample size?

- What alpha level did the study use?
  - Does this have any bearing on interpreting this result?
- Does the article discus effect sizes and how might knowledge of effect sizes influence interpretation of this finding?
- Evaluate the potential for successful replication of this study.
- What additional information would make a stronger case for the success of random assignment in this study?
- What degree of support does the article provide for the causal claim that random assignment was successful in this study?

Instead of the traditional alpha of 0.05, the study employed an alpha of 0.01. This may have led to an underestimate of the number of significant differences. Second, using easily available online independent *t*-test and effect size calculators, the effect size required for statistical significance under these conditions is about 9.8, when alpha is set to 0.01. Under the procedures reported by Sawilowsky, huge effect sizes that certainly represent the possibility of confounding variables would be ignored and counted as successful random assignments. Are these Type II errors? The claim made by the author would have more credibility if the author discussed effect sizes and had compared the number of significant differences with alpha set at 0.05 and 0.01. This exercise provides an opportunity for a broad and detailed discussion of random assignment.

5.6. Meta-Analysis: Trust No Single Study

Meta-analysis is a statistical technique in which effect sizes from many studies are combined to show the average *ES* for a question in the research literature. Meta-analysis also provides an opportunity to determine what characteristics of studies are associated with the magnitude of the *ES*. Schmidt and Oh (2016) argue a robust meta-analysis literature shows that successful replications of studies are abundant. However, publication bias and questionable research practices (Fiedler & Schwarz, 2015; John et al., 2012) may be leading to an excess of statistically significant findings that do not replicate and create complications for meta-analyses. A flawed study that is not replicable will distort a meta-analysis. Tryon (2016) states that: "*No single study should ever be trusted*. No single report should ever be considered sufficient to establish any scientific fact" (p. 236). There is too much variability in samples and other details to expect an exact match in a replicated study. Instead, meta-analysis is needed to determine the *ES* and a *CI* around that value. However, meta-analyses also produce inconsistent results (Sharpe & Poets, 2020). Researchers can perform meta-analysis meticulously and carefully, but the analyses remain subject to biases in selection of studies, computation of effect sizes, analysis of variability, interpretation, and other factors.

5.7. Implications for Teaching: Lindsay's "Troubling Trio"

Lindsay (2015), editor of *Psychological Science*, advised psychologists to avoid the "troubling trio": "(a) low statistical power, (b) a surprising result, and (c) a *p*

value only slightly less than 0.05" (pp. 1827-1828). The presence of all three elements indicates the result will be difficult to replicate. Discussing the replicability of studies can begin by placing studies in their context. Do they represent new and/or surprising results or do they represent well-established findings? Then, evaluate the study for low power and a probability close to 0.05. Keeping students aware of the potential for replication issues will sharpen their critical thinking skills and help them understand the uncertainty inherent in scientific progress.

Pedagogical Ideas

Only the most advanced students destined for graduate programs would be expected to retain full knowledge of the reasons for replication failures. On the other hand, all psychology students, majors and non-majors, should be able to apply Lindsay's (2015) "troubling trio" consisting of a surprising finding, low power, and borderline statistical significance and associate the trio with a study that should be viewed cautiously because it may prove difficult to replicate. Although it may be impractical to review every empirical study cited in a course, cited studies should regularly analyzed using Lindsay's (2015) three criteria. Instructors should take every opportunity to model critical thinking for their students (Wagner, 2022).

6. Integrating a Pedagogy of Causality into the Psychology Curriculum

Improving students' critical thinking skills should be a shared responsibility among instructors of all psychology courses. Table 1 summarizes principles of evaluating causal claims. Instructors may present the table to students as either a handout or slide. In sum, two main ideas students should learn about causal claims are:

- Trust no single piece of evidence to establish the truth of a causal claim.
- A *pattern of supporting evidence*, preferably using different methods, is essential to establish the truth of a causal claim.

Table 1. Guidelines for evaluating causal claims.

Evidence Type	Possible Methods of Evaluation
Is there *any* evidence?	• If not, ask for evidence. Reject the claim if no evidence is produced. • If yes, determine the type of evidence and evaluate its credibility.
Anecdotes/observations	• Has the observation(s) been recorded so it can be verified? • Can the observation be verified through replication? • Is the observation falsifiable? ○ If the observation cannot be repeated, is there corroborating evidence? • What biases could have influenced the observer's objectivity?
Associations	• Evaluate whether the cause must logically precede the claimed effect. • Have studies or observations with different measures and populations found the same results?

Continued

	• What third variables might explain the association? • Have third variables been controlled with multiple regression, quasi-experimental designs, or other methods? • For evidence with statistical results, ask whether a unique finding is reported, whether power is low, and whether the obtained probability is close to 0.05. Presence of all three criteria indicate a low probability of replication. • Is there evidence of bias in design or interpretation?
Experiments with Random Assignment	• Identify the manipulated and measured variables so the direction of causality is clear. • Was the obtained *p*-value close to 0.05, indicating low power and difficulty in replication? • Does the study show a surprising result, or does it fit with pattern of prior similar results? • Has the finding been replicated? • Each replication adds to the strength of evidence the study represents. • Is there evidence of bias in design or interpretation?
Evaluate the strength of the causal argument and evidence.	• Is it possible to identify *a pattern of evidence*, i.e., several studies/observations with different methods that support the claim? • Does a meta-analysis show support for the causal claim? • Does the evidence include a replicated study with random assignment, given more weight in the argument? • Studies of association require a greater number and variety of studies to validate the claim. • Does the evidence include failed predictions which undermine the causal claim? • Does the evidence include successful predictions which strengthen the causal claim? • Are there alternate interpretations of the findings that need to be tested? • Examine the findings for signs of bias in the design or interpretation.

Acknowledgements

I would like to thank Pamela L. Bacon, Michael G. Livingston, John N. Moritsugu, and Alexis Swanz for helpful comments on earlier drafts of this manuscript.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Abelson, R. P. (1995). *Statistics as Principled Argument*. Lawrence Erlbaum Associates.

American Psychological Association (2013). *APA Guidelines for the Undergraduate Psychology Major: Version 2.0*. Author. https://tinyurl.com/2c5ybmcs

Aron, A., Coups, E. J., & Aron, E. N. (2011). *Statistics for the Behavioral and Social Sciences: A Brief Course* (5th ed.). Pearson.

Barrett, T. (2015, February 27). *Inhofe Brings Snowball on Senate Floor as Evidence Globe Is Not Warming*. CNN. https://www.cnn.com/2015/02/26/politics/james-inhofe-snowball-climate-change/inde

x.html

Biddlestone, M., Flavio Azevedo, F., & van der Linden, S. (2022). Climate of Conspiracy: A Meta-Analysis of the Consequences of Belief in Conspiracy Theories about Climate Change. *Current Opinion in Psychology, 46,* Article ID: 101390. https://doi.org/10.1016/j.copsyc.2022.101390

Blanco, F. (2017). Positive and Negative Implications of the Causal Illusion. *Consciousness and Cognition, 50,* 56-68. https://doi.org/10.1016/j.concog.2016.08.012

Boudry, M., Vlerick, M., & McKay, R. (2015). Can Evolution Get Us off the Hook? Evaluating the Ecological Defense of Human Rationality. *Consciousness and Cognition, 33,* 524-535. https://doi.org/10.1016/j.concog.2014.08.025

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally.

Čavojová, V., Šrol, J., & Jurkovič, M. (2020). Why Should We Try to Think like Scientists? Scientific Reasoning and Susceptibility to Epistemically Suspect Beliefs and Cognitive Biases. *Applied Cognitive Psychology, 34,* 85-95. https://doi.org/10.1002/acp.3595

Chance, P. (1999). Thorndike's Puzzle Boxes and the Origins of the Experimental Analysis of Behavior. *Journal of the Experimental Analysis of Behavior, 72,* 433-440. https://doi.org/10.1901/jeab.1999.72-433

Darwin, C. (1964). *On the Origin of Species by Charles Darwin: A Facsimile of the First Edition with an Introduction by Ernst Mayr*. Harvard University Press.

Drake, R. E., Latimer, E. A., Leff, H. S., McHugo, G. J., & Burns, B. J. (2004). What Is Evidence? *Child and Adolescent Psychiatric Clinics of North America, 13,* 717-728. https://doi.org/10.1016/j.chc.2004.05.005

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021, December 7). Investigating the Replicability of Preclinical Cancer Biology. *eLife, 10,* e71601. https://doi.org/10.7554/eLife.71601.sa2

Fiedler, K., & Schwarz, N. (2015). Questionable Research Practices Revisited. *Social Psychological and Personality Science, 7,* 45-52. https://doi.org/10.1177/1948550615612150

File Drawer Problem (2007). In N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (Vol. 1, pp. 353-354). SAGE.

Flora, D. B. (2020). Thinking about Effect Sizes: From the Replication Crisis to a Cumulative Psychological Science. *Canadian Psychology, 61,* 318-330. https://doi.org/10.1037/cap0000218

Goldberg, M. H. (2019). How Often Does Random Assignment Fail? Estimates and Recommendations. *Journal of Environmental Psychology, 66,* Article ID: 101351. https://doi.org/10.1016/j.jenvp.2019.101351

Hatfield, J., Faunce, G. J., & Soames Job, R. F. (2006). Avoiding Confusion Surrounding the Phrase "Correlation Does Not Imply Causation". *Teaching of Psychology, 33,* 49-51.

Intergovernmental Panel on Climate Change (2022). *Summary for Policymakers*. https://tinyurl.com/fneejxkm

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine, 2,* e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling. *Psychological Science, 23,* 524-532. https://doi.org/10.1177/0956797611430953

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kenny, D. A. (2019). Enhancing Validity in Psychological Research. *American Psychologist, 74,* 1018-1028. https://doi.org/10.1037/amp0000531

Lewandowsky, S. (2021). Climate Change Disinformation and How to Combat It. *Annual Review of Public Health, 42,* 1-21. https://doi.org/10.1146/annurev-publhealth-090419-102409

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving Debiasing Away. *Perspectives on Psychological Science, 4,* 390-398. https://doi.org/10.1111/j.1745-6924.2009.01144.x

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science, 26,* 1827-1832. https://doi.org/10.1177/0956797615616374

Loftus, E. M. (2019). Eyewitness Testimony. *Applied Cognitive Psychology, 33,* 498-503. https://doi.org/10.1002/acp.3542

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A J., Cordova, M., Doyle, O., & Dosenbach, N. U. F. (2022). Reproducible Brain-Wide Association Studies Require Thousands of Individuals. *Nature, 603,* 654-660. https://doi.org/10.1038/s41586-022-04492-9

Martín, M., & Valiña, M. D. (2023). Heuristics, Biases and the Psychology of Reasoning: State of the Art. *Psychology, 14,* 264-294. https://doi.org/10.4236/psych.2023.142016

Marx, W., Haunschild, R., Thor, A., & Bornmann, L. (2017). Which Early Works Are Cited Most Frequently in Climate Change Research Literature? A Bibliometric Approach Based on Reference Publication Year Spectroscopy. *Scientometrics, 110,* 335-353. https://doi.org/10.1007/s11192-016-2177-x

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of Causality: How They Bias Our Everyday Thinking and How They Could Be Reduced. *Frontiers in Psychology, 6,* Article No. 888. https://doi.org/10.3389/fpsyg.2015.00888

Matute, H., Yarritu, I., & Vadillo, M. A. (2011). Illusions of Causality at the Heart of Pseudoscience. *British Journal of Psychology, 102,* 392-405. https://doi.org/10.1348/000712610X532210

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is Psychology Suffering from a Replication Crisis? What Does "Failure to Replicate" Really Mean? *American Psychologist, 70,* 487-498. https://doi.org/10.1037/a0039400

Mealey, L. (2000). *Sex Differences: Developmental and Evolutionary Strategies*. Academic Press.

Meltzoff, J., & Cooper, H. (2018). *Critical Thinking about Research: Psychology and Related Fields* (2nd ed.). American Psychological Association. https://doi.org/10.1037/0000052-000

Moravec, P. L., Kim, A., & Dennis, A. L. (2018). Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media. *Information Systems Research, 31,* 987-1006. https://doi.org/10.2139/ssrn.3269902

Morling, B. (2018). *Research Methods in Psychology* (3rd ed.). Norton.

Morling, B. (2021). *Research Methods in Psychology* (4th ed.). Norton.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. The National Academies Press.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science, 349,* aac4716. https://doi.org/10.1126/science.aac4716

Pearl, J. (2018). Challenging the Hegemony of Randomized Controlled Trials: A Commentary on Deaton and Cartwright. *Social Science & Medicine, 210,* 60-62. https://doi.org/10.1016/j.socscimed.2018.04.024

Reiss, J. (2009). Causation in the Social Sciences: Evidence, Inference, and Purpose. *Philosophy of the Social Sciences, 39,* 20-40. https://doi.org/10.1177/0048393108328150

Richmond, A. S, Boysen, G. A., Hudson, D. L., Gurung, R. A. R., Naufel, K. Z., Neufeld, G., Landrum, R. E., Dunn, D. S., & Beers, M. (2021). The Introductory Psychology Census: A National Study. *Scholarship of Teaching and Learning in Psychology, 7,* 163-180. https://doi.org/10.1037/stl0000277

Sawilowsky, S. S. (2004). Teaching Random Assignment: Do You Believe It Works? *Journal of Modern Applied Statistical Methods, 3,* 221-226. http://digitalcommons.wayne.edu/coe_tbf/16 https://doi.org/10.22237/jmasm/1083370980

Schmidt, F. L., & Oh, I.-S. (2016). The Crisis of Confidence in Research Findings in Psychology: Is Lack of Replication the Real Problem? Or Is It Something Else? *Archives of Scientific Psychology, 4,* 32-37. https://doi.org/10.1037/arc0000029

Schoolov, K. (2021, October 1). *Why It's Not Possible for the Covid Vaccines to Contain a Magnetic Tracking Chip That Connects to 5G*. CNBC. https://www.cnbc.com/2021/10/01/why-the-covid-vaccines-dont-contain-a-magnetic-5g-tracking-chip.html

Schwarcz, J. (2021). *Can Vaccines Make Our Body Magnetic*? McGill Office for Science and Society. https://tinyurl.com/2evdeavy

Sharpe, D., & Poets, S. (2020). Meta-Analysis as a Response to the Replication Crisis. *Canadian Psychology, 61,* 377-387. https://doi.org/10.1037/cap0000215

Stapleton, P. (2019). Avoiding Cognitive Biases: Promoting Good Decision Making in Research Methods Courses. *Teaching in Higher Education, 24,* 578-586. https://doi.org/10.1080/13562517.2018.1557137

Taccone, F. S., Hites, M., & Dauby, M. (2022). From Hydroxychloroquine to Ivermectin: How Unproven "Cures" Can Go Viral. *Clinical Microbiology and Infection, 28,* 472-474. https://doi.org/10.1016/j.cmi.2022.01.008

Thorndike, E. L. (1911). *Animal Intelligence*. Hafner.

Tryon, W. W. (2016). Replication Is about Effect Size: Comment on Maxwell, Lau, and Howard (2015). *American Psychologist, 71,* 236-237. https://doi.org/10.1037/a0040191

Wagner, P. A. (2022). Tools for Teaching and Role-Modeling Critical Thinking. *Psychology, 13,* 1335-1341. https://doi.org/10.4236/psych.2022.138086

Wielkiewicz, R. M. (2016). *Sustainability and Psychology* (2nd ed.). Main Event Press. https://www.amazon.com/dp/B012LJ9ACQ/

Wielkiewicz, R. M. (2022). *A Quick Review of Statistical Thinking* (4th ed.). Main Event Press. https://www.amazon.com/Quick-Review-Statistical-Thinking-ebook/dp/B09YBYXD4L/

Wilkinson, L., & the Task Force on Statistical Inference (TFSI) (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist, 54,* 594-604. https://doi.org/10.1037/0003-066X.54.8.594

Worrall, J. (2010). Evidence: Philosophy of Science Meets Medicine. *Journal of Evaluation in Clinical Practice, 16,* 356-362. https://doi.org/10.1111/j.1365-2753.2010.01400.x