CAUSAL LANGUAGE AND STATISTICS INSTRUCTION: EVIDENCE FROM A RANDOMIZED EXPERIMENT

JENNIFER HILL
New York University
jennifer.hill@nyu.edu

GEORGE PERRETT
New York University
gp77@nyu.edu

STACEY A. HANCOCK
Montana State University
stacey.hancock@montana.edu

LE WIN
Infinitus Systems Incorporated
lyw9402@nyu.edu

YOAV BERGNER
New York University
yoav.bergner@nyu.edu

ABSTRACT

Most current statistics courses include some instruction relevant to causal inference. Whether this instruction is incorporated as material on randomized experiments or as an interpretation of associations measured by correlation or regression coefficients, the way in which this material is presented may have important implications for understanding causal inference fundamentals. Although the connection between study design and the ability to infer causality is often described well, the link between the language used to describe study results and causal attribution typically is not well defined. The current study investigates this relationship experimentally using a sample of students in a statistics course at a large western university in the United States. It also provides (non-experimental) evidence about the association between statistics instruction and the ability to understand appropriate causal attribution. The results from our experimental vignette study suggest that the wording of study findings impacts causal attribution by the reader, and, perhaps more surprisingly, that this variation in level of causal attribution across different wording conditions seems to pale in comparison to the variation across study contexts. More research, however, is needed to better understand how to tailor statistics instruction to make students sufficiently wary of unwarranted causal interpretation.

Keywords: *Statistics education research; Causal inference; Causal language; Introductory statistics; Statistics instruction*

1. INTRODUCTION

One of the most challenging aspects of teaching statistics is helping students understand how to interpret the results of statistical procedures and to effectively communicate this understanding to others (Ben-Zvi & Garfield, 2004). This involves a combination of mathematical knowledge, logical reasoning, and, typically, understanding of the context such as study details and some knowledge of the subject matter. It also requires precise use of language. Even simple data summaries such as measures of central tendency (mean, median, mode) have subtleties that necessitate careful wording if the audience is to comprehend the practical implications (Cooper & Shore, 2008; Ismail & Chan, 2015).

Statistics Education Research Journal, 23(1), Article 6. https://doi.org/10.52041/serj.v23i1.673

The challenge in accurately conveying statistical results is amplified when the potential exists for results to be interpreted causally. As a result, various communication approaches have evolved over the years to guard against unwarranted causal interpretations. Arguably, the most common strategy is to use the word "association" in situations where there is inadequate evidence for causal attribution. These wording choices, however, may be insufficient to prevent causal attribution by the reader. One reason for this is that there may be other linguistic signals in such a description that would still imply a causal connection. Another is that, even with careful wording, there is a strong proclivity in human reasoning towards causal attribution (e.g., Kahneman, 2011). Thus, it may be necessary to take extra precautions when conveying study results involving associations between variables.

In this study, we investigate whether the wording of study results impacts causal attribution through an experimental vignette study conducted in a large-enrollment introductory statistics course at a large Western university in the United States. Our findings reveal that the language used when relaying study results does indeed affect readers' confidence in the causal relationship between variables. Our findings also highlight startling variation in the extent to which results are interpreted causally based on the hypothetical study contexts explored, independently of the wording. These findings suggest potential changes in how we should teach students to interpret and communicate results and provide a starting point for future exploration in this area.

This paper begins with a review of relevant literature and then proceeds to discuss the sample and the experimental treatment manipulations and measures. Findings from the experiment are then presented and their implications discussed.

2. BACKGROUND

Why is it important for students to be able to distinguish between study results that provide evidence for a causal relationship versus those that are merely associational? The primary reason is that causal relationships can provide guidance for decision making in ways that associational evidence does not. For instance, if we know that there is an association between using an educational tool and subsequent test scores, it might be interesting to try to explain why that association is present. Are more prepared or confident students more likely to adopt the tool? Or is the tool adding value on its own? If instead we *knew* the relationship were causal, then we might be moved to adopt this tool in the classroom. If we adopted the tool for more general use based on merely a positive association and the causal effect was in fact zero or negative, then this would be a waste of time and money and other scarce resources (e.g., teacher goodwill) in the educational system.

There are several challenges to effectively communicating whether research findings warrant causal interpretation. We start by providing a more precise definition of causal effects and then discuss the challenges of communicating these properly. We close this section with a description of current teaching practices in introductory statistics classrooms around causal inference and a discussion of related work.

2.1. HOW ARE CAUSAL EFFECTS DEFINED?

While philosophers still debate the definition and meaning of causality, most of the world of statistics has converged on a counterfactual interpretation such as the one first proposed by Hume (1748), which has been further elaborated by other philosophers, most notably by Lewis (1973a; 1973b).[1] The counterfactual framework presents *causal effects as a comparison between what was observed to happen and what would have happened in a counterfactual world with a different treatment regime*. This type of counterfactual interpretation now pervades science and has strongly influenced the standards of evidence required in most fields as well as institutions focused on policy, practice, and regulation. For instance, this paradigm is implicit in the Federal Drug Administration requirement of

[1] Although Causal Directed Acyclic Graphs (DAGs), popularized for causal inference by Pearl (2009) are also widely used for causal inference, we consider those to have a counterfactual interpretation as well given that scholars (Richardson & Robins, 2013) have demonstrated a one-to-one mapping between the DAG framework and the potential outcomes framework (Rubin, 1978) that formalizes counterfactuals in statistics.

evidence from randomized experiments for approval of new drugs and medical devices (Fleming et al., 2017).

The field of statistics has formalized the idea of the counterfactual state in statistical notation referred to as potential outcomes (Holland, 1986; Rubin, 1978). The establishment of potential outcome notation for causal inference has allowed for the development of formal mathematical theory about causal inference that had previously been elusive. The counterfactual framework also provides a clearer understanding of the potential connection between our language and causal attribution. For our purposes, this framework may be useful in understanding where the disconnect occurs between the language and the concept that language is trying to convey.

We illustrate the role of counterfactuals with an example. Suppose an educator had developed a supplemental online reading tool for third grade children in the United States and wanted to understand if this reading tool was helpful for improving reading proficiency. As a trial, she might try it out on one student by first administering a reading pretest, allowing the student to use the tool for three hours per week, and then testing the student again after three months. Suppose that the reading score obtained after using the tool for three months was 10 points higher than that obtained in the pretest. Would that mean that the tool caused the increase in reading scores? It is difficult to know with certainty. Perhaps the child would have experienced this gain even in the absence of the tool purely by instruction received at school or other experiences in and out of school during those three months. To say with certainty that the tool caused the increase we would need to know what *would have happened* in the imaginary counterfactual world where the student never received access to the tool.

We can formalize counterfactual ideas with some notation. We use the random variable Z_i to denote treatment receipt for the *i*th student; $Z_i = 1$ indicates the treatment was received by the *i*th student (that student used the online tool) and $Z_i = 0$ indicates that it was not (student did not use the tool). Then, rather than just defining a random variable, for example Y_i, to denote the observed outcome (test scores three months after the decision was made to use or not use the online tool), potential outcomes encode the idea of what might happen in each of two potential worlds: the world with the treatment (where the online tool was used) and the world without the treatment (where the online tool was not used). For instance, in our example above, the potential outcome that occurs three months after treatment initiation would be denoted as $Y_i(Z_i = 1)$ (the outcome under the condition that online tool was used). We will use $Y_i(1)$ as a shorthand for this potential outcome. We can similarly define $Y_i(Z_i = 0) = Y_i(0)$ for the potential outcome corresponding to the condition that no treatment is received (online tool was not used). We can then define a causal effect formally as the difference between these two potential outcomes, $Y_i(1) - Y_i(0)$.

Unfortunately, we can never observe both potential outcomes for the same person. Thus, identifying individual-level causal effects is a nearly impossible task. Typically, researchers seek to estimate *average* causal effects instead. An average causal effect for a sample (of size *n*) can be defined simply as an average of the individual causal effects for all members of that sample,

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i(1) - Y_i(0)\big).$$

While estimation of this quantity still suffers from a missing data problem (half the potential outcomes are missing!), careful designs can be used to identify these average effects. For instance, randomized experiments provide a design solution to this missing data problem because they create independence between the potential outcomes and the treatment assignment (for more information please refer to Chapter 18 of Gelman et al., 2020). In practice that means that we can use the average outcomes for the treatment group—a random sample of the full sample—to get an unbiased estimate of,

$$\sum_{i=1}^{n} Y_i(1).$$

Similarly, we can use the average outcomes for the control group—a random sample of the full sample—to get an unbiased estimate of,

$$\sum_{i=1}^{n} Y_i(0).$$

The important take-away from this discussion is that the causal estimand represents a comparison across counterfactual states (receiving or not receiving the treatment) for the same person (for instance, the "*i*th" person) or the same group of people (for instance, the sample). What can be confusing to

researchers and their audiences is that the estimators used to estimate these causal effects (for instance, difference in means across treatment groups or a regression coefficient from a regression of the outcome on a binary treatment indicator) represent a comparison of means across two different groups of people, those who receive the treatment or do not receive the treatment. We can only use these comparisons across groups to estimate a causal effect under special circumstances, such as in the context of a randomized experiment. So, if we only know about the difference in average outcomes across two groups and we do not know the design that produced those results, it is inappropriate to attribute causality to such a comparison. Similarly, if we know only that two phenomena are associated, this is not sufficient to justify a causal interpretation.

2.2. LINGUISTIC BARRIERS

The first mechanism for misinterpretation considered is the language used to convey study findings. The linguistic features and implications of causal language have been discussed for many years within linguistics and psychology (e.g., Fausey et al., 2010; Haber et al., 2018; Solstad & Bott, 2017). Additionally, several statistics education scholars have argued that the language used to describe statistical methods and results is typically not sufficiently precise (Ancker, 2006; Cooper & Shore, 2008; Ismail & Chan, 2015), specifically when it comes to distinguishing between causal and non-causal phenomena (Thapa et al. 2020). Furthermore, Kaplan and colleagues (2009, 2010, 2012) have argued that when words commonly used in colloquial speech take on a specialized meaning in a discipline such as statistics, it creates a greater cognitive burden for the learner. They explored this "lexical ambiguity" by focusing on the words "association," "average," "confidence," "random," and "spread," and found a disconnect between the desired statistical interpretation and the perceived meaning among introductory statistics students. Richardson et al. (2013) performed a similar study with the word "significance," and Lavy and Mashiach-Eizenberg (2009) reported similar patterns in the Hebrew language.

When statistical methods are used in situations involving associations between variables, the risk of causal misattribution becomes even more salient. For instance, consider a situation where the online tool described above was evaluated using a study that compared students in two third-grade classrooms in the same school. The tool was incorporated into the learning environment in one of the classrooms but not in the other. At the end of the school year, researchers compared test scores across the two classrooms and saw that the students in the classroom that had used the tool fared far worse on a reading comprehension test compared to children in the classroom that did not use the tool. In fact, the scores of those who used the tool were 20 points lower on average. This finding might be summarized as a negative association between tool exposure and test score and be described in the following way: "The online tool was associated with a 20-point decrease in test scores."

What is wrong with this wording? After all, it includes the word "association." Shouldn't that be a strong enough signal that the relationship is not necessarily causal? One issue is that, although statisticians have a precise definition for the word "association," that does not formally imply causation, the technical meaning of the term (including lack of causal implication) is not always sufficiently well understood by students or other audiences. This lack of clarity is further complicated by the fact that the word "association" is used outside of statistics in a colloquial manner that has other meanings (Kaplan et al., 2009).

Moreover, this interpretation also includes the word "decrease," and thus seems to imply that the tool use decreased test scores for students (on average) by 20 points. A logical implication of such a causal interpretation would be that if the students who used the reading tool had instead *not* used the tool, they would have scored an average of 20 points higher than what we observed at the end of the year. Such an interpretation, however, is not merited. The only claim that is supported by the data is the conclusion that the two groups of children—those who used the reading tool and those who did not—scored differently at the end of the year on a reading assessment. *Why* this difference in means occurred is still elusive. Perhaps the students in the class that adopted the tool were performing at a lower level even before the tool was adopted. But what if we could measure test scores of the students before the tool was ever used and they were the same on average across classrooms? Would that evidence allow for causal attribution? The problem is that there are countless other reasons that the two groups could be different in terms of student or teacher characteristics that could account for the difference in test

scores at the end of the year. In general, pairing the statistical term "association" with words such as "change," "gain/loss," and "increase/decrease" may suggest to the reader a within-person comparison, either over time or across counterfactual states, despite the intention of suggesting a non-causal interpretation (Solstad & Bott, 2017).[2]

2.3. PSYCHOLOGICAL BARRIERS

Even if the language used to describe study results has been constructed to be more deliberately non-causal, the reader still may be inclined to interpret the evidence causally (Fry, 2018; Mueller & Coon, 2013; Sibulkin & Butler, 2019; Tunstall, 2018). This is because, psychologically, people are predisposed to these types of misinterpretations of the evidence due to the human tendency to make meaning of experiences through causal interpretations. Multiple studies support this claim. For instance, work by van den Broek (2010) and O'Brien and Myers (1987) suggested that people are more likely to understand and retain information described using causal connections. Related work has demonstrated that people explain and predict behavior similarly to how they understand and tell stories; this requires understanding the sequence of events and making inferences about the mechanism behind each event (Read, 1987). Relatedly, Gerstenberg et al. (2017) demonstrated how people unconsciously use counterfactual simulation to imagine how a situation could have played out differently.

Even more specific to our context, Kaplan (2009) showed how prior beliefs among undergraduates in an introductory statistics class influence the degree to which they accept study results as evidence. Tunstall (2018) observed similar patterns in an undergraduate quantitative literacy course, where "the majority of the students agreed with the author's misleading message" (p. 76) in an opinionated news article. In addition, Owens (2018) provided evidence about how (physics) students were likely to invalidate evidence in cases where that evidence conflicted with their previous belief. Depending on the magnitude of the impact of prior beliefs, it seems possible that even if non-causal study results are communicated clearly, the audience may still be inclined to interpret them causally if the causal conclusions are consistent with their prior beliefs.

2.4. CURRENT PRACTICE IN STATISTICS

Most students are not taught explicitly about foundational concepts in causal inference such as potential outcomes and counterfactuals. Nevertheless, most introductory (first or second semester) statistics students *are* generally taught about three related topics: randomized experiments, association/correlation, and linear regression. Let us consider how each topic is typically taught.

We surveyed the introductory statistics textbook *OpenIntro Statistics* (Diez et al., 2019), two of the *Statistics for Dummies* series textbooks (McCormick & Salcedo, 2015; Rumsey, 2016), as well as each textbook on the Example Textbook List for the Advanced Placement (AP) Statistics Course Audit (College Board, 2024a). The AP Statistics course is important because it is a college-level statistics course offered in secondary schools around the world; students may earn college credit or advanced placement at universities in over 100 countries with a high score on the AP Statistics Exam (College Board, 2024b). These books were selected as a representative sample of textbooks that may be used in a typical introductory statistics course—*OpenIntro Statistics* due to its similarity to the textbook used by the students in this study, the *Statistics for Dummies* series as a top hit among introductory statistics books on Amazon.com, and the textbooks on the AP Statistics example list due to the AP Statistics course's popularity and equivalency to an introductory college-level statistics course. Several of the textbooks on the AP Statistics example list are also used in university courses. For each textbook, we considered the textbook content that has a relationship to causal inference.

Randomized experiments. Randomized experiments are often described as the most reliable way to generate data that can be used to estimate causal effects. Generally, the capability to estimate causal

[2] Words such as gain/loss and increase/decrease are considered by linguists to be causative verbs in the English language. For related studies that capitalize on these connections, see, for example, Adams et al. (2017), Parra et al. (2021), and Thapa (2021). For an example of a research methods book in the field of psychology that clearly delineates between causal and non-causal language, see Morling (2017).

effects when subjects are randomized into treatment groups is motivated to students by a combination of intuition (analogies to lotteries and coin flipping) and sometimes allusion to more formal statistical properties (randomization creates independence between the treatment assignment and pre-treatment variables by balancing out confounding effects) (e.g., Tintle et al., 2020, p. 239). Unfortunately, unless they are taking a course in causal inference, students are rarely taught to understand the formal definition of a causal estimand (none of the books we surveyed did so) and thus presumably continue to think of randomized experiments as answering a question about comparisons across groups. This omission in instruction may make it easier for the line between causal and non-causal interpretations of any such across-group comparisons to blur.

Association and correlation. Most, if not all, introductory statistics textbooks caution that "association does not imply causation" (e.g., Moore et al., 2012, p. 130; Peck et al., 2016, p. 210). These discussions are typically accompanied by definitions and examples of confounding or lurking variables and often use news headlines such as "Do you spank? Studies indicate it could lower your kid's IQ" as examples of how not to report such results (Peck et al., 2016, p. 30). In general, this topic is taught cautiously. However, it is not unheard of for interpretations to be overly causal. For instance, in their section on correlation/association, the textbook *SPSS Statistics for Dummies* states, "positive relationships show that as you increase in one variable, you increase in the other variable" (McCormick & Salcedo, 2015, p. 250).

Regression. Contemporary instruction regarding regression analysis is often detrimental with respect to creating misunderstandings about causality. When interpreting regression coefficients, many commonly used textbooks are inconsistent in their guidance. For instance, although most descriptions of regression that we examined point out that "correlation is not causation" and provide associational interpretations, many then go on to make a variety of troubling claims. For instance, commonly prescribed interpretations of the slope of a least-squares regression line also tend to include language that *is* commonly interpreted causally, including "change," "gain," "loss," "increase," and "decrease" (Thapa et al., 2020). For example, Tintle et al. (2021) include a "key idea" that "the slope … is interpreted as the predicted *change* in the average response variable for a one-unit *change* in the explanatory variable" (p. 587, emphasis added). In another popular textbook, the reader is told that the slope "represents the predicted change in the response variable y given a one unit increase in the explanatory variable x" (Lock et al., 2017, p. 128). Similarly, Utts and Heckard (2015) wrote, "The slope tells us how much of an increase (or decrease) there is for the predicted or average value of the y variable when the x variable increases by one unit" (p. 76). As discussed in Section 2.2, interpreting the slope as a "change" or "increase" may well lead the reader to assume a causal relationship. What would be preferable is language that makes explicit that regression can only elucidate the difference between adjacent subgroup means—and even then, only then when the appropriate assumptions hold. A more honest interpretation of a regression coefficient advocated for in texts such as Gelman and Hill (2007), would take the following form: "The estimated coefficient, b, on variable, x, represents the difference between the means of subgroups of our sample that are 1 unit apart from each other on the variable z."

Causal interpretations of regression abound, in particular when textbooks describe results from applied examples. For instance, Lock et al. (2017) provide the following interpretation: "The slope of 0.182 indicates that the tip is predicted to go up by about \$0.182 for a one dollar increase in the bill" (p. 128). As another example, an interpretation of the slope of the regression line between a college student's family income and the gift aid received reads, "For each additional \$1,000 of family income, we would expect a student to receive a net difference of \$1,000 × (–0.0431) = –\$43.10 in aid on average, that is \$43.10 *less*" (Diez et al., 2019, p. 321, emphasis in the original). Use of the word "less" might be considered non-causal, but when combined with a focus on just one student this interpretation feels like a within-person comparison of different states. A few sentences later, the authors wrote:

> We must be cautious in this interpretation: while there is a real association, we cannot interpret a causal connection between the variables because these data are observational. That is, increasing a student's family income may not cause the student's aid to drop. (p. 321)

At this point, it is possible the damage has already been done.

These are not isolated examples. Similar language abounds throughout popular statistics texts. Often these texts also include a warning about unwarranted causal interpretations of regression results, and some have additional chapters that focus on randomized experiments or causal inference more broadly. Moreover, often the examples given are ones where a functional or physical relationship might plausibly exist (consider examples of the tip for a bill or aid received versus family income), and some texts only explicitly use these interpretations in examples based on randomized experiments. For instance, Starnes and Tabor (2014) provided a causal interpretation of a regression coefficient but for data that arose from a randomized experiment. The connection between the interpretation and the study design, however, is never made explicit. Presumably, the problem is not that these authors did not understand that it is inappropriate to interpret regression results causally, they possibly did not understand that the specific language choices used to describe these results might have such a powerful impact. Indeed, before undertaking this research, we had no idea either of the impact of such language or the baseline proclivity of students to infer causality inappropriately.

2.5. TEACHING ABOUT CAUSATION: CHALLENGING BUT IMPORTANT

Given the barriers described above, it seems likely that the ability to carefully reason about causation may be difficult for students to develop. Indeed, in addition to the linguistic and psychological barriers described previously, researchers have documented the struggles students demonstrate when attempting to distinguish causation from correlation (Fry, 2018; Mueller & Coon, 2013; Sibulkin & Butler, 2019; Tunstall, 2018). When we additionally consider the strong tendency for human beings to attribute causation where there may be none, we can make a strong argument for increased instruction in causal inference. Velleman (2008) stated, "We should teach students to resist jumping to conclusions, extrapolating, and proposing explanations for associations that assume causation" (Section 11). Lübke et al. (2020) contended that causal inference should be explicitly taught in statistics courses in order to "overcome the mantra 'Correlation does not imply Causation'" (Abstract) and offer examples of how to do so using linear regression with simulated data. This recommendation is echoed in Cummiskey et al. (2020), who discussed how to integrate causal inference into the introductory statistics course through the use of causal diagrams. Causal inference is also featured prominently in Horton's editorial note (2023) and recent collection of papers that highlight approaches to teaching multivariate thinking (2022). In addition to the arguments for emphasizing causal inference in the classroom, researchers have also contributed classroom activities for teaching causal inference (e.g., Bennett, 2014; Cummiskey et al., 2020; Delport, 2023; Gelman et al., 1998; Lübke et al., 2020; Lu et al., 2023; Tunstall, 2016; Witmer, 2021).

Although the literature reflects the argument for incorporating a deeper understanding of causal inference into the introductory course and offers examples of classroom activities for doing so, there is a lack of research into *how* students comprehend causation in study descriptions. Without this knowledge it is difficult to know how best to advise students to report non-causal findings. The current study addresses this gap by examining how students' tendencies for causal attribution vary across language cues and study contexts.

3. RESEARCH QUESTIONS AND DESIGN

We collected data from university students in the United States to help us better understand the impact of word choices on causal interpretation. This data collection took the form of a vignette experiment administered over two time points. This section describes the following aspects of our study in more detail: the sample recruited for our study, the context within which the data were collected, the design and content of the vignette experiment, and the measures used to collect data. Our goal with this design is to address the following research questions.

1. Does the choice of wording when describing study results impact the level of causal attribution for participating students?
2. Is the choice of topic of study results associated with level of causal attribution for participating students?

3. Is variation in causal effects by experimental condition (word choice) moderated by vignette topic for participants?

3.1. SAMPLE

A survey was administered in two waves to undergraduate students enrolled in an introductory statistics course at a large university in the United States in the Spring 2022 semester. The second wave was included to help understand whether we might see different effects of question wording after students had been exposed to material presented in class about topics related to causal inference. The course was split into nine sections of approximately 90 students each. Because the course used a flipped format—students watch videos, read the textbook, and take a short quiz on the material outside of class, then work through activities in groups during class—each section was led by one lead instructor, one statistics graduate student, and two undergraduate teaching assistants. The class material addressed issues of causality regularly and was largely the same as in previous years (see below for more details about course content).

At the start of the semester, 767 students were registered for the course, of which 77 (10%) withdrew prior to the end of the semester. The majority of the student population were first or second year students (76%); 45% self-identified as female and 55% as male; 18% reported they were first generation college students; and the student population predominitely self-described as white (84%). There were a total of 57 majors represented in the student population, with Business (29%), Nursing (7%), Ecology (6%), and Computer Science (6%) as the top four.

In the first survey wave administered during the second class of the semester (January), 721 students took the survey and 661 of these (91.7%) consented to participate in the study. All respondents at Wave 1 who agreed to participate completed the entire survey instrument. The second survey wave was administered at the end of week 14 (May) in the 15-week semester. In Wave 2, 504 students took the survey and 504 (100%) consented to participate in the study. Since 31 (6.2%) respondents at Wave 2 did not complete the entire survey, they were dropped from the analysis, leaving a total of 473 respondents. The demographics of respondents at Wave 2 differed slightly from Wave 1, as shown in Table 1.

Table 1. Sample descriptive statistics for Waves 1 and 2

Characteristic		Wave 1	Wave 2
Sample size (*n*)		721	473
Average age (years)		20.1 (*SD* = 3.1)	21.0 (*SD* = 7.6)
English as first language		647 (97.9%)	456 (96.4%)
Prior statistics course		200 (30.3%)	163 (34.4%)
Gender identity	Female	298 (45.1%)	221 (46.7%)
	Male	346 (52.3%)	239 (50.5%)
	Nonbinary	13 (2%)	7 (1.5%)
	Other	2 (.3%)	4 (.9%)

3.2. STUDY CONTEXT

The introductory statistics course from which we recruited our study participants followed a relatively modern curriculum, using an open source textbook that was modeled on the *Introduction to Modern Statistics* by Çetinkaya-Rundel and Hardin (2021) (Hancock et al., 2021). Sampling and study design were taught in the first two weeks, followed by descriptive statistics and data visualization in the next two weeks. The remainder of the semester focused on both simulation-based and theory-based hypothesis tests and confidence intervals for a single proportion, difference in proportions, difference in means, paired mean difference, and simple linear regression slope or correlation.

Scope of inference (which includes topics such as whether causal attribution is warranted and can generalize to a larger population) was heavily emphasized on nearly all assignments, classroom activities, and exams. Students watched videos and read material about whether results can be

generalized to a larger population and whether a study design allows for causal conclusions, during the second week of the semester, when the topics were introduced across three consecutive class periods. The first survey wave was purposefully conducted prior to students' exposure to this material.

Random sampling, sampling bias, and their relation to study generalizability were introduced in the first in-class activity on scope of inference. This activity started with examples of four different studies. For each study, students were asked to identify the target population, the sample, the variable(s) being measured, and whether selection bias, response bias, or non-response bias could potentially be a problem. The activity then led students through selecting a sample of words "by eye" from a famous American Indian speech, calculating the sample mean length of the words, and comparing the distribution of sample means to the mean word length of the entire speech. Students could then discover that sampling "by eye" introduces selection bias, since larger words tend to be overrepresented in the students' samples. During an activity in the next class period, students revisited the same speech, but this time selected samples of words using a random number generator.

During the third class period of the second week, students completed a lab that reexamined scope of inference principles in a different context. Observational studies, confounding variables, randomized experiments, and the purpose of random assignment were the focus of this activity. As with the first activity, students read example studies and were asked to identify the explanatory and response variables and the study design. Next, students used the Rossman and Chance "Randomizing Subjects" applet (2021) to simulate randomly assigning subjects to placebo and treatment groups, then explored the distribution of the difference in proportion of males in each group and the difference in mean heights between the groups, noticing that random assignment tends to "balance out" potential confounding variables. Scope of inference continued to be discussed with every study introduced in the class and was included as an aspect of most student assessments.

3.3. VIGNETTE EXPERIMENT

To assess the impact of the wording of research findings on the extent to which students view the implications as causal, we designed an experimental vignette study, or vignette experiment, embedded within each survey (Waves 1 and 2). This experiment randomized students to be exposed to one of several options for the wording of a description of hypothetical study findings. In any given administration of the survey, we presented each student with vignettes on each of four different topics about hypothetical data that had been collected on a given sample of individuals (for instance, adults or grade school students) where a relationship between two variables (for instance, vaping and anxiety or an afterschool program and test scores) was observed. The four topics used in the second wave differed from those in the first wave, for a total of eight topics across the study. (Additional information about the topics is included below.) We followed each vignette with questions aimed at revealing the extent to which the reader understood the study results to be implying a causal relationship. Our focus on the presentation of research findings follows in the tradition of scholars who have performed similar experiments using wording from actual studies and reporting on studies (see, for instance, Adams et al., 2019; Haber et al., 2018; Haber et al., 2022). We focused instead on hypothetical studies so that we could control the wording of the experimental conditions more precisely.

We summarize examples of the wording choices in Table 2 and describe our thinking regarding our classifications of the examples in this section (complete language is available in Appendix A). Within each wave, students were randomly assigned to one of six distinct types of descriptions, or vignettes, of the hypothetical study findings. Students each experienced the same experimental condition (type of wording) for all four topics at a given wave. The language used in these descriptions, which ranged from strongly causal to purely descriptive, constituted our treatment conditions.

Table 2. Vignettes (treatment conditions) presented to students for the vaping/anxiety topic

	Condition Name	Level of Implied Causality
Researchers have found that vaping caused an increase in anxiety levels among college students.	Caused	Explicitly causal
Researchers have found that vaping led to an increase in anxiety levels among college students.	Led to	Explicitly causal
Researchers found that among college students vaping was associated with an increased level of anxiety.	Increased/decreased	Implied causality
Researchers found that vaping was associated with higher levels of anxiety among college students.	Higher/lower	Implied causality
Researchers found that college students who vaped regularly had higher levels of anxiety compared to those who didn't vape. It's possible that students who decided to vape had higher levels of anxiety when they made the decision to start vaping.	Possible that	Explicitly non-causal
Researchers found that college students who vaped regularly had higher levels of anxiety compared to those who didn't vape. Skeptics argue that this difference in outcomes could be explained by the fact that students with higher levels of anxiety are more likely to decide to vape.	Skeptics argue	Explicitly non-causal

3.4. MEASURES

Within a section devoted to a given topic after reading the description of findings randomly assigned to them, the study participant was asked a series of follow-up questions. Responses to these questions were used as covariates and outcomes in our analyses as described next.

Primary outcome. The question corresponding to our primary outcome of interest assessed the degree to which the student considered the relationship described in the study findings to be causal, which we will henceforth refer to as the "level of causal attribution." For example, in the vaping/anxiety topic, the survey question was, "Based on these findings, how confident are you that vaping made the students more anxious?" The respondents selected their responses on a scale from 0 ("not confident") to 100 ("most confident") using a slider.

Given the multitude of choices regarding how to understand whether our study participants interpreted the findings presented in the vignettes causally, it is worth discussing both our specific word choices for as well as the scale of our measure. Regarding our word choice, why not ask the students more directly about causal attribution by using the word "cause" in the question? The choice to avoid use of the word "cause" was driven by concern about the issue of "demand characteristics." This is a term used to describe signals to participants about the true underlying aim of a study. Psychological research has documented that when these study features are present, participants may respond to social pressures to be "good participants" and provide responses that confirm the study's hypothesis (see, for instance, Nichols & Maner, 2008). In our case, if we made it explicit that we were studying causality, then that might unintentionally alter the responses in ways that did not reflect the respondents' native understanding of the language they were exposed to in the vignettes. We avoided use of the word "cause" in our outcome measure for this reason.

This choice, however, left us with a decision to make about how to phrase the question for our outcome measure in a way that clearly implied causality. We chose the word "made" (e.g., "Vaping *made* the students more anxious") because linguistically, verbs of the form "to make" are considered to have very strong causal connotations (see, for instance, Nadathur & Lauer, 2020). Moreover, there is empirical evidence that readers understand this connection (Adams et al., 2017). Additionally, in informal cognitive testing with students in our lab, we assessed that the phrasing of the questions was consistently interpreted as reflecting a causal connection.

Finally, we address our choice of scale for our outcome measure. There has been repeated debate over the past two or more decades about the tradeoffs in terms of usability and reliability of using Likert-based scales versus continuous slider or related Visual Analogue Scale (VAS) measures, particularly as technology increasingly allows for easier interface with and scoring of slider-type approaches to measurement. The literature suggests that this debate has not been unequivocally resolved, but recent positive evidence suggesting similarity in performance and a desire for a straightforward analytic model encouraged us to use a continuous measure (Roster et al., 2015; Simms et al., 2019).

Moreover, as an informal means of assessing whether this measure was appropriate for our goals, we asked the students in the causal inference lab associated with one of the authors to evaluate the questions based on understandability. We also asked them whether they preferred a scale from 0 to 100 or a seven-level Likert scale as a means of expressing their confidence regarding whether the relationship is causal. They uniformly preferred the continuous measure. These assessments led us to believe that the scale from 0 to 100 was preferable for assessing causal attribution.

Secondary outcomes. To support the information solicited by our primary outcome measure, we included several secondary outcome measures to understand students' prior beliefs or knowledge about each of the topics described by the vignettes. Our concern was that our participants' level of confidence in the causality of a relationship might be strongly influenced by these prior beliefs or knowledge. For example, we suspected that the topic of vaping might be tied to strong opinions. To guard against these prior characteristics dominating our results we included two additional aspects to our study design. Our first strategy was to present vignettes for several different types of topics that would allow for variation in students' prior opinions and knowledge. These topics addressed a variety of relationships and populations. In the first wave, these topics included: 1) vaping and anxiety among teens, 2) an afterschool tutoring program and reading scores among grade school students, 3) yoga and falls resulting in broken bones among senior citizens, and 4) participation in study abroad programs and graduation rates among college students. We used different topics for the four vignettes presented in the second wave so that students' previous experience with these topics in the first wave could not influence their second wave results: 1) nutritional supplements and muscle mass among senior citizens, 2) meditation and anxiety among adults, 3) reading science fiction in middle school and majoring in a STEM (science, technology, engineering, and math) field among college students, and 4) game-based learning and math skills among elementary school students.

As our second strategy, we explicitly asked about students' presumed prior knowledge or affinity with the topic. In particular, we asked each participant follow-up questions of the form, "How much do you know about vaping and its health implications?" and "How much do you care about vaping and its health implications?" For each of these questions, the respondent could choose one option on a five-point Likert scale that ranged from "not at all" to "a great deal." Since these questions were asked after the students received the experimentally manipulated prompt, they may have been influenced by the wording, in which case they should not be used as covariates or moderators in our analyses. We found no strong relationship between the experimental manipulation and responses to the "know" or "care" questions.

Socio-demographics and prior experience. We collected information on students' age (in years) and gender identity (male, female, non-binary/third gender, prefer not to say, or other). We waited to collect additional information about study participants until the end of the survey to avoid any potential that responding to these questions might affect responses to the experimental manipulation. These questions asked about previous coursework in statistics and whether English was their primary language. Summaries of these variables by wave are presented in Table 1.

To avoid potential bias due to lack of attention to the content of the survey, we included an explicit attention check. At the beginning of the survey (first question after the consent process), we asked survey participants to report their age in years. At the end of the survey (after the experimental manipulation), we asked them to report what year they were born. If this information did not match, we did not use the survey results. At both waves, no respondents reported conflicting information about

age, and all respondents reported being over 18 years old, suggesting that students were attending to the content of the questions.

4. METHODS AND RESULTS

The primary goal of our analyses was to understand whether the wording of research findings affects levels of causal attribution in our sample. To put these results in context, we first report descriptive summaries of our findings. We then present results from models that explore variation in responses based on both question wording and vignette topic. Finally, we present findings for our primary goal and explore whether these effects are moderated by vignette topic or prior statistical background. We discuss the methods used at the same time we present results from those analyses to allow for easier access to the model specifics when interpreting the findings. Throughout, all analyses were fit separately for each of the two waves because, due to the anonymity of the survey responses, we had no way of linking students' responses between survey waves, and pooling all responses into a single analysis would have introduced non-identifiable dependencies.

4.1. DESCRIPTIVE DIFFERENCES ACROSS VIGNETTE TOPICS

We began our analysis with unadjusted comparisons of the effect of experimental factors (wording used for descriptions of results) and vignette topic on confidence in causal attributions. Figure 1 displays boxplots that highlight features of the distribution of the level of causal attribution for each topic. The top panel shows the measures from the four vignette topics presented in the first wave (beginning of semester). The bottom panel shows the measures from the four vignette topics presented in the second wave (end of semester).

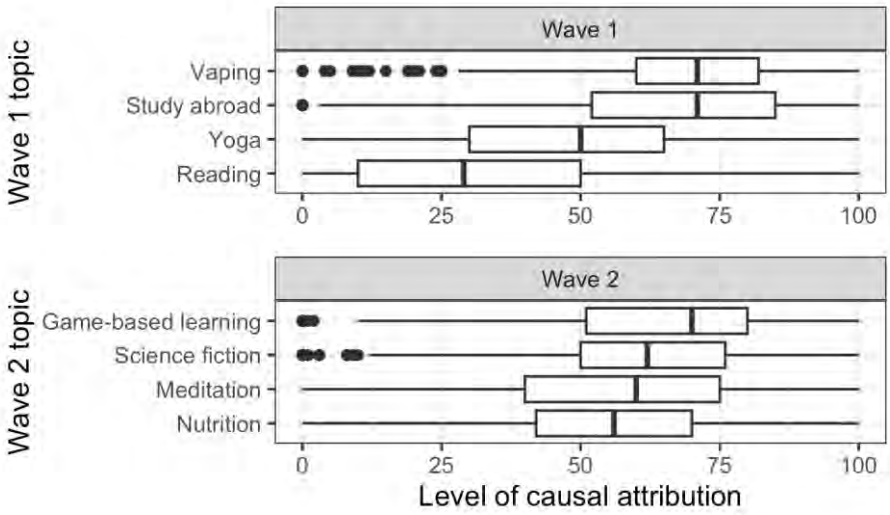

Figure 1. Boxplots of causal attribution by vignette topic for both waves, with vignette topics ordered by mean level of causal attribution (highest to lowest from top to bottom) for each wave

These boxplots reveal substantial variation in causal attribution irrespective of our experimental manipulation. In the first wave, participants were less likely to have high confidence in a causal relationship for the yoga and reading topics compared to the vaping and study abroad topics. In the second wave, there was less variation in the level of causal attribution across the vignette topics. Because different topics were used in the second survey administration, we have no way of knowing if this difference is a function of the subject matter of the topics or of the fact that the students had been exposed to more statistical reasoning about causal inference and experiments at that point in the semester. Overall, across waves, we found the strength of the level of causal attribution to be rather surprising given that two-thirds of these respondents received wording meant to indicate only associational evidence.

4.2. DIFFERENCES ACROSS EXPERIMENTAL CONDITIONS

Figure 2 presents boxplots of the levels of confidence in a causal relationship by experimental condition. Here, several important patterns begin to emerge. In Wave 1, the level of causal attribution is not noticeably higher when results are explicitly presented as being causal compared to when the relationship is described as an association. Conditions that include caveats in the description ("possible that," "skeptics argue"), however, lead to a marked decrease in the level of causal attribution. In Wave 2, explicitly causal presentations appear to increase confidence in a causal relationship. However, no strong differences in average level of causal attribution emerge between those settings in which the relationships between variables are not described using the word "cause."

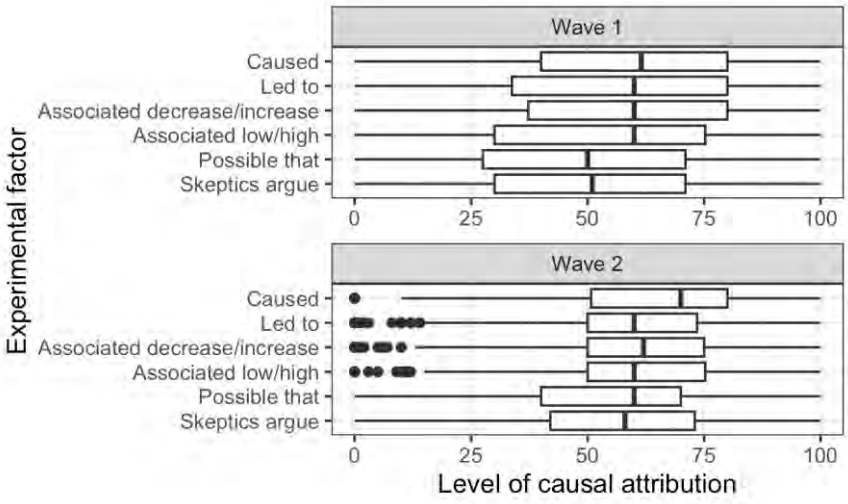

Figure 2. Boxplots of causal attribution by experimental condition for both waves, with the conditions in each wave ordered by presumed relative causality of wording (most to least from top to bottom)

4.3. EFFECT OF EXPERIMENTAL CONDITION AND CONTEXT

After these descriptive analyses, a separate multilevel linear regression analysis for each wave was used to formally test differences in causal attribution across vignette topics and experimental factors of the wording used to describe results for our study participants. We fit mixed effects models to estimate treatment effects. To account for dependencies within a given student's responses within a single survey wave, a variable for student ID[3] was included as a random effect to preserve degrees of freedom and yield more efficient estimators (Gelman & Hill 2007). Vignette topics were also included as a separate random effect.

The choice to model these terms as random effects rather than fixed effects (given that there are only six categories) was motivated by several considerations. The first is that the subset of topics included in our study can be loosely considered to be drawn from a population of all possible research topics. The second consideration is that fitting vignette topic as a random effect allows us to easily make direct comparisons between all topics within a given time period rather than comparing topics to a single designated reference class. This is desirable given that there is neither an a priori ordering of these topics with regard to their expected level of causal attribution nor a clear "reference" category. Finally, the fact that we are making a total of 12 topic comparisons could lead to multiple comparisons issues that would leave us over-confident in our results. Fitting a random effects model has been shown to alleviate concerns of multiple comparisons in this situation without compromising efficiency (Gelman & Hill, 2007; Gelman et al., 2012).

[3] Although we can identify individual survey responses within a single wave, we do not have a unique identifier to link students across waves.

Experimental factors (wording used for descriptions of results) were included as fixed effects. In this initial model, covariates were included to increase the efficiency of our estimates. We did not, however, allow the experimental effects to vary across vignette topics. The covariates included were age, English as a first language, gender, the extent to which students care about the topic of the vignette, the extent to which students know about the topic of the vignette, and whether the student had received prior statistics training. In this initial model we do not allow the experimental effects to vary across vignette topics.

Our mixed effects models were fit in *R* using the `rstanarm` package with the default weakly informative priors (Goodrich et al., 2022). In essence, that prior specification reflects the fact that our information about the difference between groups (for the topic random effects) or individuals (for the individual level random effects) does not give us any reason to believe a priori that there are specific differences between them. So, without data, our best guess would be that there are no differences. Our model specification simultaneously captures the fact that we have a great deal of uncertainty about that best guess. In practice we have enough data that this priori should be "swamped" by the information it contains (that is, the estimates should be driven primarily by the data, not the priori). Thus, our Bayesian analysis should converge to the maximum likelihood estimates from a frequentist analysis while maintaining the added flexibility and advantages described above (easier comparisons with appropriate calibration of uncertainty to account for multiple comparisons). Models were fit by drawing from the posterior distribution using 10 Markov chain Monte Carlo (MCMC) chains using 1000 iterations with 1000 burn in iterations with no thinning. The goal is to get independent draws from a posterior distribution (the Bayesian equivalent of a sampling distribution) that can be used to evaluate how likely various hypotheses are relative to observed data. "*R* hat" values were used to check model convergence. All *R* hat values were very close to 1.0, indicating convergence of MCMC chains (Gelman & Rubin, 1992).

Contrasts between the least causal experimental setting and all other settings are displayed in Figure 3. These contrasts adjust for all covariates. Points indicate the difference in the mean level of causal attribution between the experimental condition specified on the *y*-axis and the most explicitly non-causal experimental factor ("skeptics argue"). These conditions are ordered by our prior expectation of how likely they were to be interpreted causally. Lines represent 95% uncertainty intervals, and the shade denotes whether the results were from the first or second wave. 95% uncertainty intervals in a Bayesian analysis can be thought of as representing the probability that the parameter in question lies in the given interval. Note that the average level of causal attribution for the baseline category ("skeptics argue") is 54.3 for Wave 1 and 58.7 for Wave 2 on the 100-point scale.

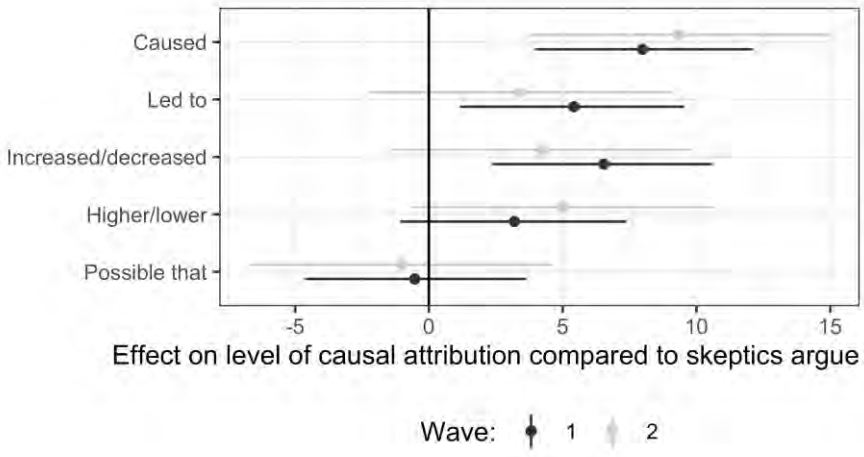

Figure 3. Causal effects of question wording (points) and 95% intervals (line segments) compared to the "skeptics argue" condition, with Wave 1 results in black, Wave 2 results in gray, and experimental conditions ordered by presumed relative causality of wording (most to least from top to bottom)

The results displayed in Figure 3 indicate that, in our setting, research findings that use language including the word "cause" are those that are rated most causal by the study participants, with estimated effects (relative to the "skeptics argue" condition) of about seven points on the 100-point scale. Neither of the 95% intervals for the difference between this condition and the "skeptics argue" condition overlap with zero. Noticeable effects (ranging from about 3 to 6.5) occur for the next two conditions as well, with intervals for Wave 1 that do not include zero. These conditions do not use the word "cause," but do, however, include words with causal implications ("increased/decreased" and "led to"). Use of words without explicit causal meaning "higher/lower" does not substantially alter effect estimates (estimates of about three or four points), and both associated uncertainty intervals cover zero. Results from the other condition that offers a caveat to the initial statement ("possible that") are quite similar to those from our comparison condition ("skeptics argue"). If our level of causal attribution outcome measure is inducing measurement error, it is possible that our effect sizes have been attenuated.

Given that each participant at each survey wave responded to questions about the same four topics, we have the opportunity to better understand whether the level of causal attribution was impacted by the subject matter of the vignette. Although vignette topics were not randomly assigned to participants (each participant was exposed to all of them), we have the advantage that all four topics were seen by each person, allowing for within person comparisons at the same point (modulo a few minutes). To interpret these effects causally, we do have to assume that a vignette on any given topic did not cause any change in the individual that would influence their response to the next vignette (on a different topic). For instance, we have to assume that reading hypothetical results about a study on vaping and anxiety would not influence a student's response to a vignette focused on study abroad. As a reminder, participants each experienced the same experimental condition (type of wording) for all four vignette topics at a given wave.

The same multilevel regression was used to understand the effect of vignette topic on confidence in a causal relationship. Figure 4 displays results as mean differences across pairs of topics with the corresponding uncertainty intervals. Because we have no a priori beliefs about which vignette topic would elicit the highest level of causal attribution on average, we present a summary of all six pairwise comparisons within each wave (distinguished again by black and gray shading). Comparisons are ordered from largest effect estimates to smallest (top to bottom).

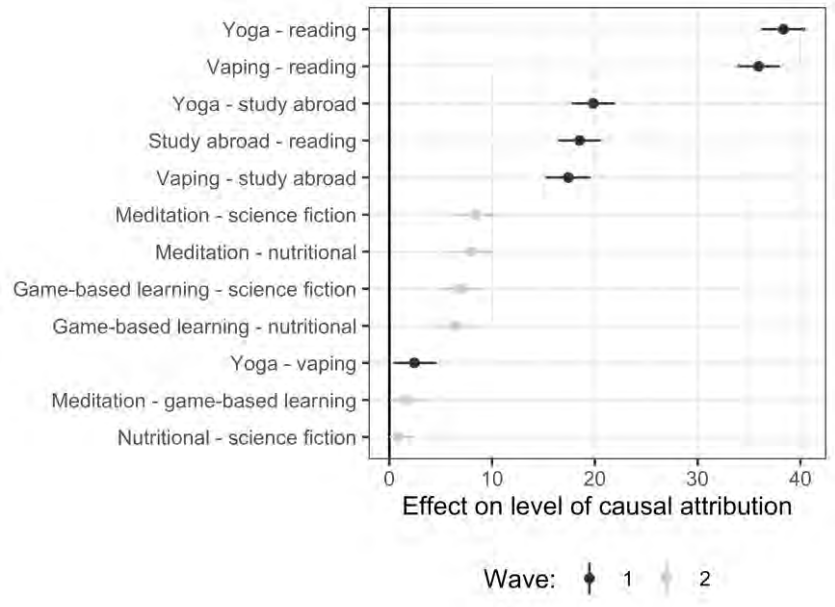

Figure 4. Mean differences in levels of causal attribution (points) and corresponding 95% uncertainty intervals (line segments) across pairs of vignette topics, with Wave 1 results in black, Wave 2 results in gray, and ordered by the size of the difference (highest to lowest from top to bottom)

We are reluctant to read too much into the specific topics chosen and their potential effect on a participant's degree of confidence in a causal connection. Rather, we display these comparisons as a

contrast to the results from our experimentally manipulated wording conditions. The largest treatment effect we saw from our wording conditions for our study participants was about seven points. In contrast, the "effect" of vignette topic varies from close to zero to nearly 40 points. Half of these "topic effects" were as large or larger than the largest wording effect. It is also worth pointing out that, although the uncertainty intervals are much narrower for these comparisons relative to the experimental comparisons, that is not surprising given that we had six times as many subjects available to estimate these effects (the full subject pool at each wave) relative to those available for the experimental condition comparisons. These results provide ballast for explanations favoring the evidence for psychological explanations of these effects.

4.4. MODERATORS

Our results thus far help us understand overall impacts for our sample. It is, however, likely that the effects of our experimental conditions vary based on the specifics of the individuals exposed to them. We explore two such moderation mechanisms in this section.

Vignette topics as moderators. Given that we know that levels of causal attribution vary across vignette topics and that moderately sized average treatment effects for our sample were induced by some of our experimental conditions (relative to the least causal option), it seemed worth exploring whether our experimental treatment effects varied across vignette topics. To investigate, we expanded the model described in Section 4.3 to include a varying slope component to explore the variation in treatment effect estimates across vignette topics.

Figure 5 displays our results in a similar manner to Figure 3, but now allowing each treatment effect to vary across vignette topic. Descriptively, these results demonstrate variation in treatment effect estimates across vignette topics; however, the pattern of differences is not consistent across experimental settings (wording choices). Moreover, all of these uncertainty intervals overlap with each other, so it is difficult to draw strong conclusions about differences in treatment effects across vignette topics.

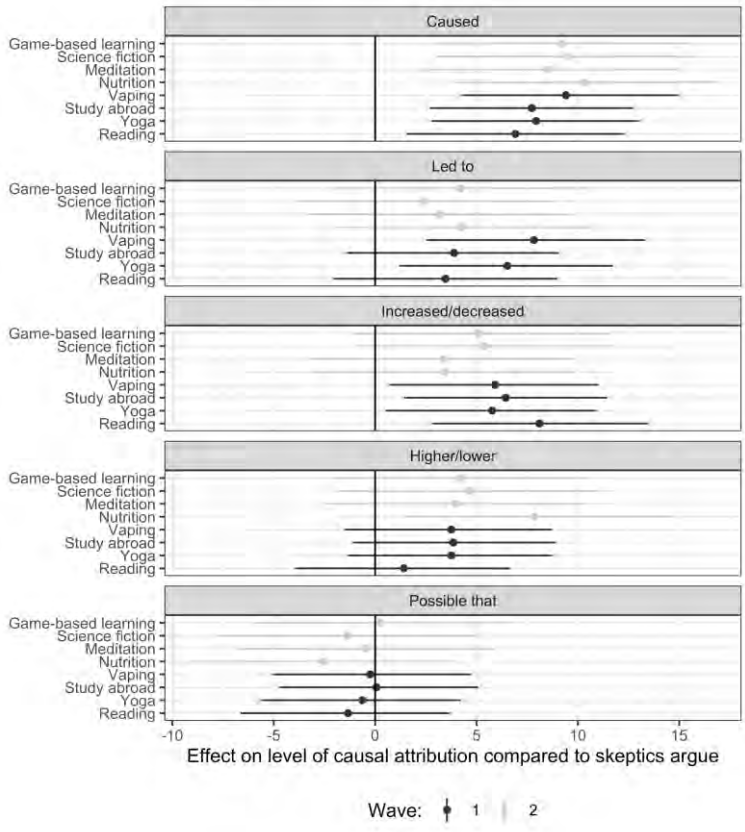

Figure 5. Variation in wording effects estimated separately for each vignette topic, with Wave 1 results in black and Wave 2 results in gray and where "skeptics argue" is the counterfactual condition

Prior statistical training as moderators. It is possible that the size of the effect of wording choice in our sample varies based on whether students had previously been exposed to training in statistics. At a minimum, these students should be more likely to understand the connotations of statistics terms such as "association" and possibly would be more cautious in general about attributing causality, particularly at the start of the semester.

Our results are suggestive of these patterns. Figure 6 displays boxplots of the causal attribution variable separate for each experimental condition and wave, and additionally breaks this out by whether students have received statistical training prior to their current course. Descriptively, we see the biggest differences between these two groups in Wave 1 within the "associated higher/lower" and the "skeptics argue" conditions. In Wave 2, while the medians of each distribution are similar across experimental conditions, the distributions for causal attribution are more prominently skewed to the left for four of the conditions, suggesting that at least some of those with previous training were apt to be more cautious in their interpretations.

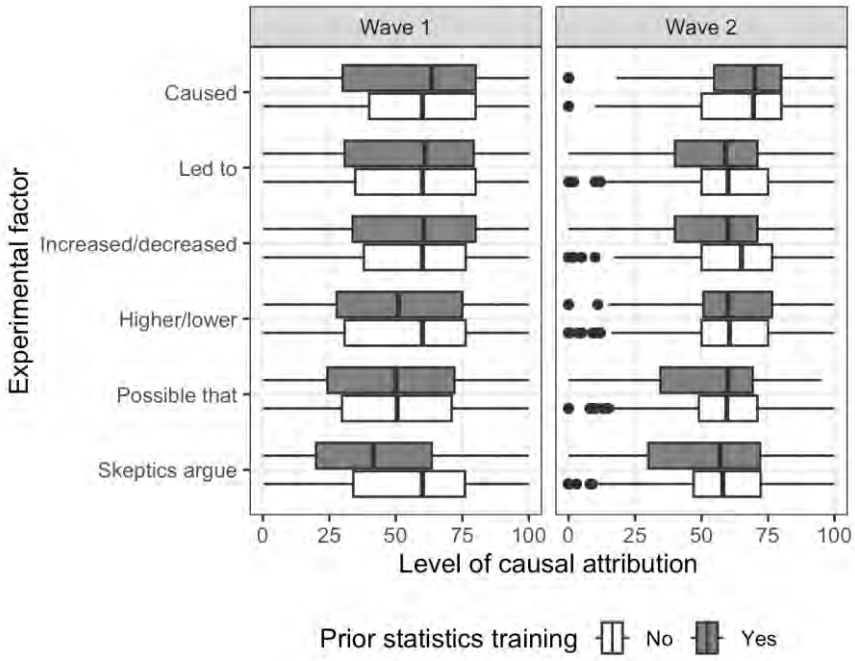

Figure 6. Boxplots of level of causal attribution by experimental factor, wave, and prior statistics training, with darker boxes representing responses from students with prior statistics training and lighter boxes representing responses from students without any prior statistics training

We also extended our model from above to allow for direct tests of the effects of question wording in these subgroups. The results are displayed in Figure 7 as treatment effect estimates in each wave for each of the experimental conditions relative to "skeptics argue" condition, as with Figure 5. In this plot, however, these effects are further broken out by subgroups defined by prior statistical coursework. Here we see causal effects with 95% uncertainty intervals that exclude zero for all (wave, coursework) subgroups for the condition where the word "caused" was used in the description of the findings (relative to "skeptics argue" reference condition). That is, for each causal effect, at least 97.5% of the corresponding posterior distribution is greater than zero. However, only the students with prior statistical training who participated in Wave 1 also display evidence that the other wording conditions led to effects on causal attribution (compared to the "skeptics argue" condition) with uncertainty intervals that exclude zero. The posterior distributions for each of these effects have a smaller mean in the second wave and each of the 95% uncertainty intervals includes zero.

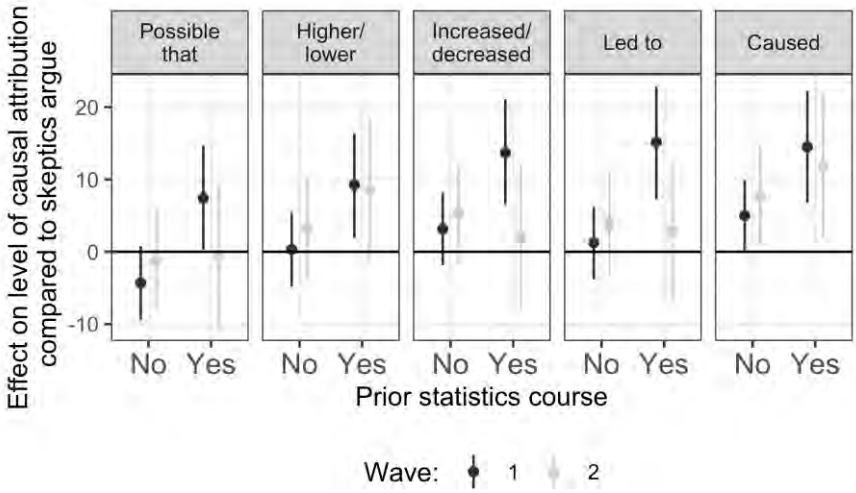

Figure 7. Variation in wording effects (points) by prior statistical training relative to the "skeptics argue" condition, with Wave 1 results in black and Wave 2 results in gray

5. LIMITATIONS

While this study has many advantages, most importantly the ability to randomly assign wording conditions, it is not without limitations. One key limitation is the lack of unique identifiers for students. Without this information we cannot run analyses on the full sample because the observations are not independent of each other. Moreover, the fact that we used different vignette topics at each wave means that we cannot distinguish between survey wave and vignette topic salience. We chose this approach because we were concerned that students might remember the settings from one wave to the next, which could bias the Wave 2 results. In future work, we will obtain student identifiers and reverse the ordering of the vignette topics between waves to address both issues. We would also hope to randomly assign the order of the vignette topics within a given wave in case the information obtained for one has an influence on the response given for another.

Although our sample is reasonably large, it consists entirely of undergraduates from a single university who were taking a course in introductory statistics. We cannot guarantee that our results will generalize to students who would not take this introductory course or students from another university. We also cannot generalize to individuals who are the same age but are not students, or individuals from other age groups. Finally, we can only draw conclusions about people residing in the United States. Wording choices and proclivity towards causal attribution may well vary substantially across cultures, nationalities, and languages, among other individual characteristics.

An additional limitation arises based on the fact that the measure used for our outcome variable has not been psychometrically validated. Using a scale of "confidence in causal attribution" from 0 to 100 has advantages in terms of modeling choices and the related ease of interpretability of model parameters. It may, however, not be the optimal choice for understanding the extent to which a survey participant views a statement as conveying a causal relationship. In fact, the arbitrary nature by which participants might choose their responses on this scale may introduce measurement error. If so, our results may be attenuated, and we may have less power to distinguish between groups. More research is needed to understand the best way to measure this construct.

Another limitation of this study is that it focuses solely on language and does not explore whether students' understanding of how a study was carried out might inform their level of causal attribution. For instance, if a well-trained student knew that a study randomized treatments they might feel (justifiably) comfortable with a causal interpretation even in the absence of causal language. Conversely, the student might be wary of the use of causal language if it was known that the observational study design did not warrant any causal conclusion. Our current study focuses solely on the language used and thus cannot speak to how a student's perception might be altered if they additionally were informed about the study design. This could be an interesting avenue to pursue in future research.

Finally, we recognize that the specific wording and topic choices incorporated in our study represent just a small subset among an infinite number of options. Although our choices regarding wording were tied to language that we have seen used in practice, there are many other options available that could be explored. Moreover, it is difficult to learn much about the specifics of what types of contexts are more prone to causal misattribution using this current study.

6. DISCUSSION AND NEXT STEPS

Our results suggest that the wording used to describe study findings does appear to impact the degree to which students in our sample understand relationships to be causal. Unfortunately, switching wording from explicitly causal ("A caused B") to language that statisticians consider to be associational ("A was associated with a change/difference in B") does not seem to eliminate the proclivity to interpret the relationship between key variables causally. Our study, however, does provide support for the fact that additional caveats that provide alternative explanations can, in some settings, at least *reduce* the propensity towards causal attribution.

Perhaps more surprising, if we interpret the variation in level of causal attribution across vignette topics causally, it appears that students' tendency to interpret findings causally may have more to do with the subject matter context of the findings than the wording. This finding suggests that students may rely more on their own prior beliefs about the relationships between variables than the empirical evidence, even when they are explicitly told to only use the evidence presented when assessing their confidence in a causal relationship. This importance of context for student understanding is also reflected in the literature. On the one hand, a meaningful study context has the potential to promote statistical reasoning and engagement (Langrall et al., 2006; Yilmaz et al., 2023; Zapata-Cardona, 2023). Indeed, the third recommendation in the *Guidelines for Assessment and Instruction in Statistics Education* College Report reads, "Integrate real data with context and purpose" (GAISE College Report ASA Revision Committee, 2016, p. 3). On the other hand, students may use study context in a way that is not necessarily productive or helpful to the task at hand (Langrall et al., 2006), and their beliefs or opinions about a certain context are sometimes hard to separate from their statistical reasoning (Wroughton et al., 2013). Context appears to show this same range from helpful to unhelpful when determining causal attribution in scientific studies. This study provides some insight into how changes in language and context relate to students' understanding of causation. Our results suggest that there may be a proclivity among introductory statistics students towards causal attribution when interpreting research findings that is difficult to change. One implication of this is that statistics instruction should likely be more careful and explicit when teaching students how to interpret non-causal findings and perhaps should also include more content on the distinctions between causal and non-causal estimands. For instance, it may be useful to present study results in a variety of ways, demonstrating the range of language students might encounter when studies are reported in the media and discussing how subtle changes in language (e.g., "vaping *increases* anxiety" versus "those who vape have *higher* levels of anxiety") can signal different levels of causal attribution. Additional research is needed to understand what strategies might be more successful in reducing the proclivity of students (and other individuals who digest research findings) to attribute causality to findings that are explicitly only descriptive in nature.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Institute of Education Sciences, R305D200019. The authors would also like to acknowledge help in data collection efforts from Jade Schmidt and Melinda Yager and helpful feedback from members of the thinkCausal lab at New York University.

REFERENCES

Adams, R. C., Sumner, P., Vivian-Griffiths, S., Barrington, A., Williams, A., Boivin, J., Chambers, C. D., & Bott, L. (2017). How readers understand causal and correlational expressions used in news headlines. *Journal of Experimental Psychology: Applied*, *23*(1), 1–14. https://doi.org/10.1037/xap0000100

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 randomized trial. *BMC Medicine*, *17*(1), Article 91. https://doi.org/10.1186/s12916-019-1324-7

Ancker, J. S. (2006). The language of conditional probability. *Journal of Statistics Education, 14*(2), Article 5. https://doi.org/10.1080/10691898.2006.11910584

Bennett, K. A. (2014). Using a discussion about scientific controversy to teach central concepts in experimental design. *Teaching Statistics, 37*(3), 71–77. https://doi.org/10.1111/test.12071

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning and thinking: Goals, definitions and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 3–15). Kluwer Academic Publishers. https://doi.org/10.1007/1-4020-2278-6_1

Çetinkaya-Rundel, M., & Hardin, J. (2021). *Introductory to modern statistics*. OpenIntro. https://openintro-ims.netlify.app/

College Board. (2024a). *AP Central: AP Statistics course audit*. https://apcentral.collegeboard.org/courses/ap-statistics/course-audit

College Board. (2024b). *AP Central: The course AP Statistics*. https://apcentral.collegeboard.org/courses/ap-statistics

Cooper, L. L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education*, *16*(2), Article 1. https://doi.org/10.1080/10691898.2008.11889559

Cummiskey, K., Adams, B., Pleuss, J., Turner, D., Clark, N., & Watts, K. (2020). Causal inference in introductory statistics courses. *Journal of Statistics Education, 28*(1), 2–8. https://doi.org/10.1080/10691898.2020.1713936

Delport, D. H. (2023). The development of statistical literacy among students: Analyzing messages in media articles with Gal's worry questions. *Teaching Statistics, 45*(2), 61–68. https://doi.org/10.1111/test.12308

Diez, D., Çetinkaya-Rundel, M., & Barr, C. D. (2019). *OpenIntro statistics* (4th ed.). OpenIntro. https://www.openintro.org/book/os/

Fausey, C. M., Long, B. L., Aya, I., & Boroditsky, L. (2010). Constructing agency: The role of language. *Frontiers in Psychology, 1*, Article 162. https://doi.org/10.3389/fpsyg.2010.00162

Fleming, T. R., Demets, D. L., & McShane, L. M. (2017) Discussion: The role, position, and function of the FDA: The past, present, and future. *Biostatistics*, *18*(3), 417–421. https://doi.org/10.1093/biostatistics/kxx023

Fry, E. (2018). *Introductory statistics students' conceptual understanding of study design and conclusions* (Publication No. 10689030). [Doctoral dissertation, University of Minnesota]. ProQuest Dissertations Publishing.

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education college report 2016*. American Statistical Association. http://www.amstat.org/education/gaise

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel models.* Cambridge University Press.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. https://doi.org/10.1017/9781139161879

Gelman, A., Hill, J. & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., Nolan, D., Men, A., Warmerdam, S., & Bautista, M. (1998). Student projects on statistical literacy and the media. *The American Statistician, 52*(2), 160–166. https://doi.org/10.1080/00031305.1998.10480556

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science, 28*(12), 1731–1744. https://doi.org/10.1177/0956797617713053

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2022). *rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.3.* https://mc-stan.org/rstanarm/

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), Article e0196346. https://doi.org/10.1371/journal.pone.0196346

Haber, N. A., Wieten, S. E., Rohrer, J. M., Onyebuchi, A. A., Tennant, P. W. G., Stuart, E. A., Murray, E. J., Pilleron, S., Lam, S. T., Riederer, E., Howcutt, S. J., Simmons, A. E., Leyrat, C., Schoenegger, P., Booman, A., Dufour, M.-S., K., O'Donoghue, A. L., Baglini, R., Do, S., … Fox, M. P. (2022). Causal and associational language in observational health research: A systematic evaluation. *American Journal of Epidemiology, 191*(12), 2084–2097. https://doi.org/10.1093/aje/kwac137

Hancock, S., Carnegie, N., Meyer, E., Schmidt, J., & Yager, M. (2021). *Montana State introductory statistics with R*. Montana State University. https://mtstateintrostats.github.io/IntroStatTextbook/. [Adapted from Çetinkaya-Rundel, M. & Hardin, J. (2021). *Introduction to modern statistics*.] OpenIntro. https://openintro-ims.netlify.app/

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960. https://doi.org/10.2307/2289064

Horton, N. J. (2022, September 8). *Collection of papers on teaching Simpson's paradox, confounding, and causal inference*. Taylor & Francis Online. https://www.tandfonline.com/journals/ujse20/collections/teaching-simpsons-paradox

Horton, N. J. (2023). Teaching causal inference: Moving beyond "correlation does not imply causation." *Journal of Statistics and Data Science Education, 31*(1), 1–2. https://doi.org/10.1080/26939169.2023.2178778

Hume, D. (1748). *An enquiry concerning human understanding*. A. Millar.

Ismail, Z., & Chan, S. W. (2015). Malaysian students' misconceptions about measures of central tendency: An error analysis. *AIP Conference Proceedings, 1643*, 93–100. https://doi.org/10.1063/1.4907430

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kaplan, J. J. (2009). Effect of belief bias on the development of undergraduate students' reasoning about inference. *Journal of Statistics Education*, *17*(1), Article 3. https://doi.org/10.1080/10691898.2009.11889501

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2009). Lexical ambiguity in statistics: What do students know about the words association, average, confidence, random and spread? *Journal of Statistics Education*, *17*(3), Article 6. https://doi.org/10.1080/10691898.2009.11889535

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2010). Lexical ambiguity in statistics: How students use and define the words: association, average, confidence, random and spread. *Journal of Statistics Education*, *18*(2), Article 6. https://doi.org/10.1080/10691898.2010.11889491

Kaplan, J. J., Rogness, N. T. & Fisher, D. G. (2012). Lexical ambiguity: Making a case against spread. *Teaching Statistics, 34*(2), 56–60. https://doi.org/10.1111/j.1467-9639.2011.00477.x

Langrall, C., Nisbet, S., & Mooney, E. (2006). The interplay between students' statistical knowledge and context knowledge in analyzing data. In A. Rossman & B. Chance (Ed.), *Working cooperatively in statistics education. Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Bahia, Brazil.* International Statistical Institute.

Lavy, I. & Mashiach-Eizenberg, M. (2009). The interplay between spoken language and informal definitions of statistical concepts. *Journal of Statistics Education, 17*(1), Article 4. https://doi.org/10.1080/10691898.2009.11889502

Lewis, D. (1973a). *Counterfactuals.* Blackwell.

Lewis, D. (1973b). Causation. *The Journal of Philosophy*, *70*(17), 556–567.
Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2017). *Statistics: Unlocking the power of data* (2nd ed.). John Wiley & Sons.
Lübke, K., Gehrke, M., Horst, J., & Szepannek, G. (2020), Why we should teach causal inference: Examples in linear regression with simulated data. *Journal of Statistics Education, 28*(2), 133–139. https://doi.org/10.1080/10691898.2020.1752859
Lu, Y., Zheng, Q., & Quinn, D. (2023). Introducing causal inference using Bayesian networks and do-calculus. *Journal of Statistics and Data Science Education, 31*(1), 3–17. https://doi.org/10.1080/26939169.2022.2128118
McCormick, K., & Salcedo, J. (2015). *SPSS statistics for dummies* (3rd ed.). John Wiley & Sons.
Moore, D. S., McCabe, G. P., & Craig, B. A. (2012). *Introduction to the practice of statistics* (7th ed.). W. H. Freeman and Company.
Morling, B. (2017). *Research methods in psychology* (3rd ed.). W. W. Norton.
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
Nadathur, P. & Lauer, S. (2020). Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: A Journal of General Linguistics, 5(1),* Article 49. https://doi.org/10.5334/gjgl.497
Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *Journal of General Psychology, 135*(2), 151–65. https://doi.org/10.3200/genp.135.2.151-166
O'Brien, E. J., & Myers, J. L. (1987). The role of causal connections in the retrieval of text. *Memory & Cognition, 15*(5), 419–427. https://doi.org/10.3758/BF03197731
Owens, L. (2018). *Identifying student difficulties in causal reasoning for college-aged students in introductory physics laboratory classes* (Publication No. 10901947) [Doctoral dissertation, University of Cincinnati]. ProQuest Dissertations Publishing.
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). Article e043339. https://doi.org/10.1136/bmjopen-2020-043339
Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
Peck, R., Olsen, C., & Devore, J. L. (2016). *Introduction to statistics and data analysis* (5th ed.). Cengage Learning.
Read, S. J. (1987). Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology, 52*(2), 288–302. https://doi.org/10.1037/0022-3514.52.2.288
Richardson, A. M., Dunn, P. K., & Hutchins, R. (2013). Identification and definition of lexically ambiguous words in statistics by tutors and students. *International Journal of Mathematical Education in Science and Technology, 44*(7), 1007–1019. https://doi.org/10.1080/0020739X.2013.830781
Richardson, T., & Robins, J. (2013). *Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality* [Working Paper]. Center for Statistics and the Social Sciences at the University of Washington. http://www.csss.washington.edu/Papers/wp128.pdf
Roster, C., Lucianetti, L. & Albaum, G. (2015). Exploring slider vs. categorical response formats in web-based surveys. *Journal of Research Practice*, *11*(1), Article D1. http://jrp.icaap.org/index.php/jrp/article/view/509/413
Rossman, A., & Chance, B. (2021). *Rossman/Chance applet collection 2021*. http://www.rossmanchance.com/applets/index2021.html
Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *1*(6), 34–58. https://doi.org/10.1214/aos/1176344064
Rumsey, D. (2016). *Statistics for dummies* (2nd ed.). Wiley Publishing.
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), 233–229. https://doi.org/10.1177/0098628319853936

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31*(4), 557–566. https://doi.org/10.1037/pas0000648

Solstad, T., & Bott, O. (2017). Causality and causal reasoning in natural language. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 619–644). Oxford University Press.

Starnes, D. S., & Tabor, J. (2014). *The practice of statistics* (5th ed.). W. H. Freeman.

Thapa, D. K., Visentin, D. C., Hunt, G. E., Watson, R., & Cleary, M. (2020). Being honest with causal language in writing for publication. *Journal of American Nursing, 76(6)*, 1285–1288. https://doi.org/10.1111/jan.14311

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T., & VanderStoep, J. (2020). *Introduction to statistical investigations* (2nd ed.). Wiley.

Tunstall, S. L. (2016). Fostering comprehension of risk and causation through media case studies. *Teaching Statistics: An International Journal for Teachers*, *38*(2), 65–66. https://doi.org/10.1111/test.12099

Tunstall, S. L. (2018). Investigating college students' reasoning with messages of risk and causation. *Journal of Statistics Education*, *26*(2), 76–86. https://doi.org/10.1080/10691898.2018.1456989

Utts, J. M., & Heckard, R. F. (2015). *Mind on statistics* (5th ed.). Cengage Learning.

van den Broek, P. (2010). Using texts in science education: Cognitive processes and knowledge representation. *Science*, *328*, 453–456. https://doi.org/10.1126/science.1182594

Velleman, P. F. (2008). Truth, damn truth, and statistics. *Journal of Statistics Education, 16*(2), Article 7. https://doi.org/10.1080/10691898.2008.11889565

Witmer, J. (2021). Simpson's paradox, visual displays, and causal diagrams. *The American Mathematical Monthly, 128*(7), 598–610. https://doi.org/10.1080/00029890.2021.1932237

Wroughton, J. R., McGowan, H. M., Weiss, L. V., & Cope, T. M. (2013). Exploring the role of context in students' understanding of sampling. *Statistics Education Research Journal, 12*(2), 32–58. https://doi.org/10.52041/serj.v12i2.303

Yilmaz, Z., Ergül, K., & Asik, G. (2023). Role of context in statistics: Interpreting social and historical events. *Statistics Education Research Journal, 22*(1), Article 6. https://doi.org/10.52041/serj.v22i1.72

Zapata-Cardona, L. (2023). The role of contexts in supporting early statistical reasoning in data modeling. *Statistics Education Research Journal, 22*(2), Article 5. https://doi.org/10.52041/serj.v22i2.448

JENNIFER HILL
Applied Statistics, Social Science, and the Humanities
New York University
246 Greene St., 3rd Floor
New York, NY 10003

APPENDIX A

Appendix A shows all combinations of topics and experimental factors. This appendix is organized by experimental condition.

Wave 1

Experimental condition 1: "Skeptics"

Researchers found that elementary school children who participated in an afterschool reading program had lower reading scores, on average, at the end of the school year compared to elementary school children who did not participate in the afterschool reading program. Skeptics argue that the difference in outcomes could be explained by the fact that children who are worse at reading are more likely to participate in such programs.

Researchers found that senior citizens who participated in yoga once a week had lower rates of falls resulting in broken bones, on average, compared to senior citizens who did not participate in yoga. Skeptics argue that the difference in outcomes could be explained by the fact that senior citizens who are in better physical condition are more likely to participate in yoga.

Researchers found that college students who participated in a study abroad program had higher rates of on-time graduation compared to those who didn't. Skeptics argue that the difference in outcomes could be explained by the fact that students who are on track to graduate on time have more freedom to participate in study abroad programs.

Researchers found that college students who vaped regularly had higher levels of anxiety compared to those who didn't vape. Skeptics argue that this difference in outcomes could be explained by the fact that students with higher levels of anxiety are more likely to decide to vape.

Experimental condition 2: "Possible that"

Researchers found that elementary school children who participated in an afterschool reading program had lower reading scores, on average, at the end of the school year compared to elementary school children who did not participate in the afterschool reading program. It's possible that students who participated in the program had lower test scores than the non-participants before beginning the program.

Researchers found that senior citizens who participated in yoga at least once a week had lower rates of falls resulting in broken bones, on average, compared to senior citizens who did not participate in yoga once a week. It's possible that senior citizens who participated in yoga were in better physical condition before participating in yoga.

Researchers found that college students who participated in a study abroad program had higher rates of on-time graduation compared to those who didn't. It is possible that the students who participated in study abroad programs had a higher graduation rate than the non-participants before the beginning of the program.

Researchers found that college students who vaped regularly had higher levels of anxiety compared to those who didn't vape. It's possible that students who decided to vape had higher levels of anxiety when they made the decision to start vaping.

Experimental condition 3: "Higher/lower"

Researchers found that participating in an afterschool reading program was associated with lower reading scores among elementary school children at the end of the school year.

Researchers found that, among senior citizens, participating in yoga at least once a week was associated with lower levels of falls resulting in a broken bone.

Researchers found that participating in a study abroad program while in college was associated with higher rates of on-time graduation.

Researchers found that vaping was associated with higher levels of anxiety among college students.

Experimental condition 4: "Increased/decreased"

Researchers found that participating in an afterschool reading program was associated with a decrease in reading scores at the end of the school year among elementary school children.

Researchers found that, among senior citizens, participating in yoga at least once a week was associated with a decrease in probability of falling and breaking a bone.

Researchers found that participating in a study abroad program while in college was associated with an increased probability of on-time graduation.

Researchers found that among college students vaping was associated with an increased level of anxiety.

Experimental condition 5: "Led to"

Researchers found that participating in an afterschool reading program led to a decrease in reading scores among elementary school children who participated in the program.

Researchers found that, among senior citizens, participating in yoga at least once a week led to a decreased probability of falling and breaking a bone.

Researchers found that participating in a study abroad program while in college led to an increased probability of graduating on time.

Researchers have found that vaping led to an increase in anxiety levels among college students.

Experimental condition 6: "Caused"

Researchers found that participating in an afterschool reading program caused a decrease in reading scores among elementary school children who participated in the program.

Researchers found that, among senior citizens, participating in yoga at least once a week caused a decrease in falls with broken bones.

Researchers found that participating in a study abroad program while in college caused an increase in on-time graduation rates.

Researchers have found that vaping caused an increase in anxiety levels among college students.

Wave 2

Experimental condition 1: "Skeptics"

Researchers found that senior citizens who take nutritional supplements every day had greater muscle mass, on average, compared to those who did not take nutritional supplements. Skeptics argue that this

difference in outcomes could be explained by the fact that senior citizens who exercise more regularly are more likely to decide to take nutritional supplements.

Researchers found that adults who meditate at least three times a week had lower levels of anxiety, on average, compared to those who did not meditate. Skeptics argue that this difference in outcomes could be explained by the fact that adults with lower levels of anxiety are more likely to choose to meditate.

Researchers found that those who read science fiction regularly in middle school had higher rates of majoring in a STEM (science, technology, engineering, and math) field in college, on average, compared to those who didn't read science fiction. Skeptics argue that this difference in outcomes could be explained by the fact that those who were already interested in science and related fields in middle school would also be more likely to then choose to read science fiction.

Researchers found that elementary school children who take game-based learning for math had higher levels of math confidence, on average, compared to those who do not participate. Skeptics argue that this difference in outcomes could be explained by the fact that elementary school children who already liked math and were confident in their math ability might be more likely to be assigned by teachers to a game-based learning program for math.

Experimental condition 2: "Possible that"

Researchers found that senior citizens who take nutritional supplements every day had greater muscle mass, on average, compared to those who did not take nutritional supplements. It's possible that senior citizens who decided to take nutritional supplements were also those who more regularly exercised.

Researchers found that adults who meditate at least three times a week had lower levels of anxiety, on average, compared to those who did not meditate. It's possible that adults who decided to meditate already had lower levels of anxiety when they chose to start meditating.

Researchers found that those who read science fiction regularly in middle school had higher rates of majoring in a STEM (science, technology, engineering, and math) field in college, on average, compared to those who didn't read science fiction. It's possible that those who choose to read science fiction regularly in middle school were also disproportionately those most interested in science classes since kindergarten.

Researchers found that elementary school children who participate in a game-based learning program for math had higher levels of math confidence, on average, compared to those who did not participate in the program. It's possible that those elementary school children assigned to the game-based learning program for math were, on average, more interested in math and confident in their ability from the outset.

Experimental condition 3: "Higher/lower"

Researchers found that taking nutritional supplements every day was associated with a higher muscle mass among senior citizens.

Researchers found that meditating at least three times a week was associated with a lower level of anxiety among adults.

Researchers found that reading science fiction regularly in middle school was associated with higher rates of majoring in a STEM (science, technology, engineering, and math) field in college.

Researchers found that participating in a game-based learning program for math was associated with higher levels of math confidence among elementary school children.

Experimental condition 4: "Increased/decreased"

Researchers found that taking nutritional supplements every day was associated with increased muscle mass among senior citizens.

Researchers found that meditating at least three times a week was associated with decreased levels of anxiety among adults.

Researchers found that reading science fiction regularly in middle school was associated with increased rates of majoring in a STEM (science, technology, engineering, and math) field in college.

Researchers found that participating in a game-based learning program for math was associated with increased levels of math confidence among elementary school children.

Experimental condition 5: "Led to"

Researchers found that taking nutritional supplements every day led to an increase in muscle mass among senior citizens.

Researchers found that meditating at least three times a week led to a decrease in anxiety among adults.

Researchers found that reading science fiction regularly in middle school led to an increased chance of majoring in a STEM (science, technology, engineering, and math) field in college.

Researchers found that participating in a game-based learning program for math led to increased math confidence among elementary school children.

Experimental condition 6: "Caused"

Researchers found that taking nutritional supplements every day caused an increase in muscle mass among senior citizens.

Researchers found that meditating at least three times a week caused a decrease in anxiety among adults.

Researchers found that reading science fiction regularly in middle school caused an increased chance of majoring in a STEM (science, technology, engineering, and math) field in college.

Researchers found that participating in a game-based learning program for math caused an increase in math confidence among elementary school children.