RESOURCES AND TENSIONS IN STUDENT THINKING ABOUT STATISTICAL DESIGN

KELLY FINDLEY
University of Illinois Urbana-Champaign
kfindley@illinois.edu

BREIN MOSELY
Harvard University
breinmosely@g.harvard.edu

AARON LUDKOWSKI
University of Chicago
aludko2@uchicago.edu

ABSTRACT

Reform efforts in statistics education emphasize the need for students to develop statistical thinking. Critical to this goal is a solid understanding of design in the process of collecting data, evaluating evidence, and drawing conclusions. We collected survey responses from over 700 college students at the start of an introductory statistics course to determine how they evaluated the validity of different designs. Despite preferring different designs, students offered a variety of productive arguments supporting their choices. For example, some students viewed intervention as a weakness that disrupted the ability to generalize results, whereas others viewed intervention as critical for identifying causality. Our results highlight that instruction should frame design as the balancing of different priorities: namely causality, generalizability, and power.

Keywords: *Statistics education research; Causality; Sampling; Generalizability; Power*

1. INTRODUCTION

For statistics educators, design is a pivotal topic for engaging students in discussions about the evaluation of data-based claims. Authors of the *Curriculum Guidelines for Undergraduate Programs in Statistical Science* (American Statistical Association [ASA], 2014) advocated that statistics majors develop a deep understanding in the "design of studies … and issues of bias, causality, confounding, and coincidence" (p. 11). Furthermore, the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) emphasized the importance of teaching basic principles of design at all levels (GAISE College Report ASA Revision Committee, 2016).

Although there is agreement in curricular reform documents that design is important, it is not necessarily clear what form or depth this topic should take. The Advanced Placement (AP) Statistics course for secondary students (14–18-year-olds) offers a comprehensive introductory curriculum that many United States universities count toward college credit if students pass the AP Statistics examination. The AP curriculum primarily addresses statistical design through randomization in sampling and in experimental group sorting, while acknowledging the dangers of causality and generalizability outside of these contexts (College Board, 2020). Reviews of many widely used college statistics curricula reveal similar approaches to design (Fry, 2017; Schield, 2018). Although many textbooks discuss observational studies and non-random sampling methods, few address how arguments for causality and generalizability may be constructed when the standard for randomization is not met—what Schield (2018) termed a curricular "bait-and-switch." Randomization may provide a productive starting point for evaluating good design, but design in the real world is often messier and deeply intertwined with context (Easterling, 2004; Watson et al., 2020).

Statistics Education Research Journal, 22(3), Article 8. https://doi.org/10.52041/serj.v22i3.662

In this study, we offer a contribution to curricular and instructional efforts in statistics by considering how students enrolled in college introductory statistics courses weigh and compare imperfect design choices before instruction on design. This research was motivated by our observation that students in introductory college courses often express nuanced understandings of design that go beyond considerations of randomization. As a preliminary step toward reconceiving our curriculum, we wanted to document the conceptual resources and tensions students had about common design elements (Smith et al., 1994). We also consider how a unit on design at an introductory level might best organize student thinking toward statistical design evaluation principles that go beyond randomization as a sole criterion for judging validity.

2. CONCEPTUAL FRAMEWORK

In this study, we have chosen to examine students' reasoning about design through a *resource* framing. Resources may be conceptualized as "knowledge-in-pieces," representing fine-grained intuitions drawn from experiences and activated in multiple contexts in which the learner identifies potential connections (diSessa, 1988; Smith et al., 1994). As an example, diSessa presented the notion of "closer means stronger" as a resource students use to make sense of the physical world (e.g., sound or heat is heard or felt stronger as one moves closer to the source). This fundamental pattern, however, can also be applied in inappropriate contexts (e.g., the temperature is hotter in summer because our hemisphere faces the sun more directly, not because Earth is closer to the sun). Resources, when blended appropriately, serve as building blocks for constructing complex conceptions and supporting flexible application of ideas in novel situations (Hammer, 1996; Smith et al., 1994).

In contrast to resource views of learning, a misconception framing identifies the "fallacies, misunderstandings, misuses, or misinterpretations of concepts, provided that they result in a documented systematic pattern of error" (Sotos et al., 2007, p. 99). Although misconceptions could be a helpful entry point to understanding student thinking, we find this perspective to be in tension with constructivist views of learning. This tension arises from labeling the concept itself as problematic, rather than seeing student thinking as the combination of many fine-grained resources that may simply be misapplied or incomplete. Constructivist views of learning instead focus on identifying the heuristics that guide student conceptions, rather than on the belief that incorrect ideas need to be simply replaced with correct ones (Garfield et al., 2015; Smith et al., 1994).

In the context of statistics education, Findley and Lyford (2019) examined the resources students used when reasoning about the behavior of a sampling distribution for the sample mean. Findley and Lyford identified the resource, *Stabilizing*, as an observation that larger samples can be thought of as creating more stabilization—much like the intuition for the law of large numbers. Another resource students expressed was *Growing Possibilities* to indicate that with samples of large sizes, there would be a larger set of possible values and perhaps even a larger range of values in each sample. This resource could be valuable if paired with the notion of *Growing More Accurate*, which suggests that larger samples should produce more accurate sample means.

Fauconnier and Turner's (2002) notion of conceptual blending further complements the theoretical basis for this research. When reasoning about a new question, we may need to pull together multiple schemata into a single, unified structure, leading to a new schema from which to analyze and make sense of incoming information. Likewise, pulling together certain conceptual resources while lacking others may lead to incorrect conclusions. For example, Findley and Lyford (2019) found that one student, David, activated the Stabilizing resource in combination with Growing Possibilities to believe that sampling distributions would become "more level, more gradual, and less steep" (Findley & Lyford, p. 15). A misconceptions framing on this issue may assume David needs to erase his conceptions and start over. A resource framing, however, recognizes two productive resources that guide this response. If David could pair these ideas with the notion of Growing More Accurate, he may be able to build a conceptually strong model for the behavior of a sampling distribution by now recognizing that the sample means are stabilizing in their convergence toward the population mean.

We approach this work with the belief that students are capable sense-makers who have valuable ideas to bring to discussions about design. With this guiding approach, we aim to look past labeling students' misconceptions about good statistical design and instead identify what is productive about their observations. By taking this lens, we hope to bridge the gap between the formal concepts that students see in our courses and the informal intuitions and heuristics they had formed already.

3. BACKGROUND

3.1. STATISTICAL THINKING AS SCIENTIFIC THINKING

Instruction on design and the evaluation of claims exists at the intersection of statistics and science education; therefore, there is much to learn from science pedagogy on this topic (Watson et al., 2020). Science educators have cautioned against teaching scientific inquiry as a series of procedures that must be followed—most notably by critiquing the scientific method in science curricula (Berland & Reiser, 2009; Windschitl et al., 2008). Windschitl et al. outlined several concerns: for one, not all inquiries can be tested through experimentation—data can come from varying sources and should be judged appropriately. For another, the scientific method over-contains and linearizes what should be an expansive and revision-filled pursuit to building knowledge. Although testing ideas through intervention is a valuable tool of science, it should also be accompanied by the practices of offering explanations, revising ideas, arguing from data, and valuing non-experimental data and observation appropriately (Next Generation Science Standards Lead States, 2013; Smith et al., 2000).

The failings of a procedural approach to inquiry and investigation are related to the tension in statistical thinking outlined by Tintle and colleagues (2015) and shown in Figure 1. The authors explain that a lack of scientific understanding for building knowledge may leave students to objectify and dichotomize the process of evaluating statistical claims.

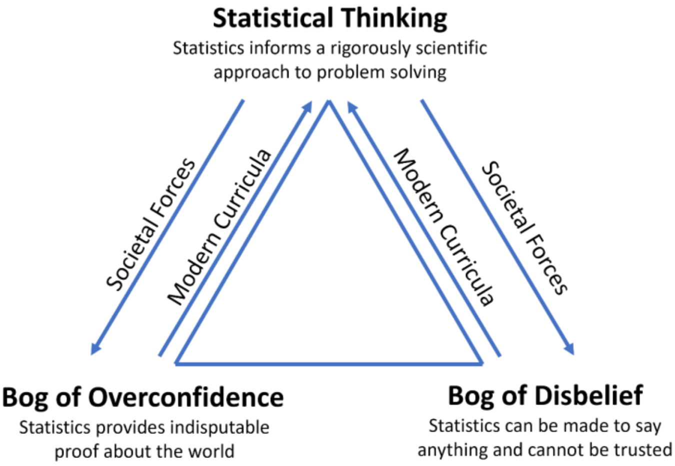
Statistical Thinking
Statistics informs a rigorously scientific approach to problem solving
Societal Forces
Modern Curricula
Modern Curricula
Societal Forces
Bog of Overconfidence
Statistics provides indisputable proof about the world
Bog of Disbelief
Statistics can be made to say anything and cannot be trusted

Figure 1. Tensions in promoting statistical thinking (Adapted from Tintle et al., 2015, p. 363).

Tintle and colleagues (2015) identified traditional curricula and societal forces as factors that push students toward one of two reactions to statistical claims. On one hand, students may fall into a bog of disbelief—distrusting statistical claims when the simplified criteria for good studies (e.g., random sampling, experimental design) are not met. On the other hand, students may fall into a “bog of overconfidence” when assuming these criteria are met or when failing to recognize other considerations when evaluating a statistical claim.

Consistent with this view of statistical thinking as scientific thinking is the importance of linking elements of design to reasoning in context (Bennett, 2015; Easterling; 2004). Choosing such things as an experimental unit, a target population, intervention procedures, and an appropriate control condition or even whether one should use an experimental design at all are inherently contextual choices. For example, Easterling (2004) cited an example question of comparing Fertilizer A with Fertilizer B in relation to the growth of eleven tomato plants in a row. A superficial examination of this question may simply end with randomizing the fertilizers and noting average growth. But context brings a number of considerations forward: how might we reasonably guard against the bleeding of fertilizer across plants or ensure no bias in watering, weeding, and insect treatment to these plants during the experiment? Are eleven plants enough to detect any differences that might exist? Should we trust pure random assignment or consider a potentially significant blocking factor such as position? These questions bring attention to the fact that the data we are analyzing often has a storied past of decisions, and the conclusions we draw depend on more than a *p*-value.

3.2. THE RELEVANCE OF TEACHING DESIGN

Curricular inclusion of design. GAISE emphasized design as a fundamental component of an introductory course that sets up statistical literacy, multivariable thinking, and an appreciation for randomization in inference (GAISE College Report ASA Revision Committee, 2016). Design also represents an important cross-cutting topic for students in other disciplines who take a statistics course. For example, undergraduate students in the biological sciences need to understand validity in experimental design and its connection to claims of causality and generalizability (Dasgupta et al., 2014; Hiebert, 2007). Likewise, students in the social sciences need adequate coverage of causal inference in the context of observational studies (Imai et al., 2011; Schield, 2018). Even students studying statistics should have rich understandings of confounding variables and validity starting at the introductory level (Cummiskey et al., 2020; Tintle et al., 2013). But in practice, many introductory curricula introduce confounders while offering randomization as the only valid antidote (Fry, 2017). This, however, may create a disconnect for students if many of the observational contexts used to showcase confounders cannot be addressed with experimental randomization, creating a curricular “bait-and-switch” (Schield, 2018, p. 2).

Although inclusion of causal inference and experimental design varies from curriculum to curriculum, there is clearer consensus on the inclusion of sampling in introductory course topics (GAISE College Report ASA Revision Committee, 2016). At a basic level, curricula may contrast the simple random sample with a convenience sample, whereas other textbooks may go into more detail about different sampling schemes (Fry, 2017). Sampling is an important principle for guiding students to consider the generalizability of sample results to a larger population. It should, however, be noted that the validity of a generalization may go beyond the sampling units to also include aspects of the study itself. For example, in the context of biological experimental design, Dasgupta et al. (2014) considered setting aspects (e.g., temperature, habitat) as important considerations for determining the generalizability of sample results to a larger population.

Often linked to discussions of sampling are considerations of sample size. Although the composition of a sample relates to the generalizability of study results, the size of a sample relates to the power of a study to detect an effect or association. There seems to be little formal guidance from statistics education curricula and reform documents on the importance of teaching power at an introductory level, but several studies acknowledge sample size as an important component to tease out from sample composition (Dasgupta et al., 2014; Derry et al., 2000).

Student difficulties with design. Several researchers have examined students’ evaluations of statistical claims based on design and data sources. Many of these studies specifically focus on students’ beliefs around randomization in group sorting and sampling situations, or how students relate group sorting to claims of causality and sampling to claims of generalizability. For example, Derry et al. (2000) documented that honors students in an education program struggled to identify the value of random assignment as a means of making a controlled comparison, even after an instructional unit that emphasized this idea. Students may exhibit lack of trust around randomization as a fair way to create balanced groups (Deane et al., 2014; Wagler & Wagler, 2013). The secondary students (14–18-year-olds) in Wagler and Wagler’s (2013) study also believed that equal group sizes created a valid comparison, or that the use of randomization at any stage supported causal claims.

Research has also suggested that study context may affect students’ evaluations. For example, Kaplan (2009) found that more controversial findings may elicit a student belief bias that can potentially override their evaluation of the design itself. Specifically, in the context of a study testing the existence of extra-sensory perception (ESP), introductory statistics students who believed ESP was real were more likely to judge a claim favorable to ESP as compared to students who did not believe ESP was real. This trend was further replicated in the context of sampling with Wroughton et al. (2013), where a significant number of students in introductory statistics courses based their evaluations on personal beliefs and opinions, rather than on the data or the sampling methods.

3.3. EVALUATING STATISTICAL CLAIMS IN A MODERN LANDSCAPE

In March 2020, when researchers and clinicians were scrambling for solutions to the COVID-19 pandemic, a study published in the *International Journal of Antimicrobial Agents* reported a statistically significant link between the drug hydroxychloroquine and the eradication of the SARS-CoV-2 virus (Gautret et al., 2020). It is believed that this study was the impetus for United States President Trump's announcement that the federal government would begin stockpiling the drug (Crowley et al., 2020). But over the next several months, many studies began finding no effect from hydroxychloroquine and, in some cases, negative effects from the drug that would eventually end the Federal Drug Administration's emergency authorization of the drug's use with COVID-19 patients (Kupferschmidt, 2020). But in the United States, as Kupferschmidt reported, political battle lines and arguments over misinformation had already been drawn. So how did this happen, and how might an understanding of statistical design help the everyday consumer of news navigate this charged terrain?

To investigate the issue, Gautret and colleagues (2020) compared 20 patients who took a daily regimen of hydroxychloroquine to a control group of 16 patients who received standard care. The results showed a stunning difference in viral clearance by Day 6 between the two groups: 70% of the treatment group tested negative whereas only 12.5% of the control group tested negative (*p*-value = .001). To an everyday consumer, this may sound like a robust result, but this study (as most scientific studies do) had limitations. Unfortunately, not all consumers have the statisitcal acumen to be able to understand and evaluate the limitations presented in a scientific study such as this one.

First, there is a threat to causality due to potential group differences. The treatment group was comprised of patients at one particular hospital who consented to the treatment. The control group included the patients who declined the treatment at that same hospital, in addition to a few patients at a different hospital who were also included for comparison. This introduces two confounding threats: a) there may be differences in the standard of care at each hospital, or possibly in testing procedures, that differentiate results and b) patient willingness may be a proxy for some underlying difference between these groups. The study also mentioned that the treatment group initially started with 26 patients, but six patients dropped out—four of these patients still tested positive when dropping out, and several of them dropped out due to admission to the intensive care unit or increased nausea. There is also a question of generalizability to a larger population based on this sample. With all patients in the treatment group receiving care from the same hospital, it is hard to determine if cases that happen to be at this hospital would represent cases at other hospitals.

What might statistical thinking look like in evaluating the data on hydroxychloroquine in relation to Tintle and colleagues' (2015) framework? Statistical thinking avoids overconfidence by recognizing that each group may be different in ways other than the use of hydroxychloroquine or because this use of convenience sampling may not generalize well to a broader population. Statistical thinking also avoids the cynicism of disbelief by recognizing that imperfect studies may still be informative. For example, we might consider how the authors presented a case for group equivalency by arguing for demographic equivalency. But the context of this study—participant willingness to try experimental medication—may help the reader realize that a demographic comparison may not capture this key, systematic difference. The study's limitations should not automatically invalidate its claims, but contextual reasoning helps the reader recognize why similar studies trying to replicate these results may offer a more complete picture and, perhaps, change scientific concensus.

Design should not be viewed as a set of truths that are mapped objectively onto tasks—rather, good design is grounded in principles to be weaved into data investigations as a means to set up strong, but often imperfect, claims. Students should leave our courses with a mindset that helps them navigate evidence such as the example above and evaluate what strengths or limitations might accompany a study's claims, as well as how to compare the affordances and limitations of different designs.

4. RESEARCH AIMS

Our work identifies the types of arguments students constructed when evaluating and comparing imperfect designs. We chose to investigate this through the lens of *resources* with the intention to identify the productive observations students make when grappling with different design features. With resources in mind, we introduce our paper's first two research questions:

1. What resources do introductory statistics students use when evaluating a randomized controlled experiment in comparison with a pre-post design?
2. What resources do introductory statistics students use when evaluating a randomized controlled experiment in comparison with a larger observational study?

We also considered how students may have balanced competing priorities through the notion of conceptual blending. We wanted to identify how student arguments point to the awareness of choice in design based on how different priorities may be valued: namely causality, generalizability, and power. Furthermore, conceptual blending in design might consider how the context of the task students were working with might influence their priorities on these matters. This leads to a third research question:

3. What tensions do students encounter when evaluating these designs? How might these tensions relate to the contextual features of each prompt?

5. METHODS

5.1. DATA COLLECTION

At the beginning of the Spring 2021 semester, we contacted approximately 1,200 students in an introductory statistics course (which we will refer to as Course A) and approximately 300 students in an introduction to statistics for life sciences course (Course B). Course A has no statistics prerequisite and generally attracts students who need a general education quantitative credit. Class data revealed this to be a first statistics course for approximately 87% of enrolled students. Course B also has no statistics prerequisite, but this course includes more students who reported completing AP statistics or another introductory course. In Course B, only 59% of enrolled students identified this as their first statistics course. We did not specifically gauge students' prior knowledge of design, but it is likely that a portion of these students (especially in the statistics for life science class) may have seen experimental design content before. However, students enrolled in introductory college level courses are likely to have varying levels of prior knowledge, and our results reflect that knowledge diversity.

Both courses were taught at a large public university in the Midwestern United States. Students in both classes were offered extra credit for completing a short survey on making conclusions from a statistical study. Out of these 1,500 students contacted, a total of 755 students filled out the survey and gave consent, providing us with a response and consent rate of 50.3%. Although students who did not complete the survey or give research consent may be different in some ways from those who did, we believe that we have a fairly diverse sample of responses. This sample included students from almost every major on campus, but representation largely favored programs from Applied Studies, Biology, Chemistry, Environmental Science, Health Science, Social Science, and the Humanities.

Students were incentivized to complete the survey for a small amount of extra credit. We did not want to overburden students with too many comparisons and risk superficial or incomplete responses. Therefore, we chose to only pose two prompts related to design. For each prompt, students read a scenario that asked them to compare two different design options for answering a singular research question. They were then asked to choose one design as more effective or choose both designs as equally effective. Each prompt was strategically created to offer two design choices with different strengths and weaknesses. After each multiple-choice question, students were asked to briefly explain their answer to the previous question, or to state "don't know" if they guessed.

The purpose of this study was to understand *how* students were making their evaluations and what they were noticing to make their decisions. As a result, we crafted design comparisons that offered multiple features with different advantages, rather than trying to gauge students on a difference from a single feature in isolation. The result was to offer two designs that had limitations but that might still yield insights into students' decisions about design. The first prompt on the survey had students compare a randomized controlled experiment with a pre-post design without a control group (Figure 2). This design choice presented a strategic difference for students to compare. The pre-post design has the advantage of tracking all 200 participants' responses to the medication, rather than only 100 in the multi-group design. It also includes pre- and post-measurements for each participant, allowing the researcher to track individual change. The pre-post design has definite advantages in power and may also have certain contextual advantages if one wanted to understand for whom the medication is most effective. The multi-group design better addresses whether the medication is effective on average. By comparing results for the group receiving medication against a placebo group, the researchers are better positioned to make a causal argument for the medication's effectiveness, rather than leaving open the possibility for some other change over time or a placebo effect.

> *A researcher is studying the use of a new medication (a tablet taken by mouth once a day) that is designed to lower blood pressure. This researcher gathers 200 participants with above average blood pressure and randomly assigns 100 of them to try the new medication for 2 weeks and the other 100 take a placebo pill (a pill that has no actual effects) for 2 weeks. The researcher hopes that this comparison will reveal if the new medication is effective at lowering high blood pressure.*
>
> *Now compare that to a pre-post study, where all 200 participants have their blood pressure taken first, then all take the experimental medication for 2 weeks, then have their blood pressure taken again. In this design, the researcher directly compares each person's pre and post measurement to find the average within-person change.*
>
> *Which design do you think is more effective for assessing the effectiveness of this experimental drug?*
>
> *A. The multi-group design is more effective than the pre-post design*
> *B. Both designs are equally effective*
> *C. The pre-post design is more effective than the multi-group design*

Figure 2. Prompt 1.

The second prompt also included a randomized controlled experiment, but in comparison with an observational study in the form of a cross-sectional survey (Figure 3). The context also changed from blood pressure to sleep, which presented a scenario that could make sense in both observational and experimental contexts. The survey design included a much larger sample size, which would theoretically give the survey more power to detect a difference. It may also be argued that the survey better generalizes to everyday life because being in a sleep intervention study may introduce psychological changes in the participants that do not generalize to their everyday lives.

> *Study 1: A research team gathers survey responses from a representative sample of 1,000 U.S. college students and asks them "How many hours of sleep do you get on an average weeknight?" The research team also asks students whether or not they keep a regular bedtime on weeknights. They find that the average amount of sleep for those who keep a regular bedtime is 6.9 hours per night, while the others reported 6.2 hours of sleep per night on average.*
>
> *Study 2: A different research team gathers a representative sample of 200 college students who do not ordinarily keep a regular bedtime on weeknights to be in the study. They randomly choose 100 of these people to set a regular bedtime for 2 weeks, and the other 100 are asked to live normally. At the end of the 2 weeks, they ask each student how much sleep they got on average that week. Those assigned to keep a regular bedtime reported 6.9 hours of sleep per night on average, while the others reported 6.2 hours of sleep per night on average.*
>
> *Which study do you think provides better evidence for this claim: "Keeping a regular bedtime on weeknights may directly help college students get more sleep on average."*
>
> *A. Study 1 provides stronger evidence for this claim*
> *B. Both studies provide equivalent evidence for this claim*
> *C. Study 2 provides stronger evidence for this claim*

Figure 3. Prompt 2.

The experiment theoretically provides a stronger causal argument. By randomly assigning students to a treatment or control group, the experiment can argue for an equivalent group comparison, whereas the survey respondents who choose to keep a regular bedtime may be systematically different than those who do not. We also chose to report the results of each study in the prompt because we were curious if seeing the results would affect students' evaluation of the design. We can informally see the effect of having the results in the prompt by observing how many students made an argument strictly from the results being the same, rather than from design features.

After collecting the survey data, the second author completed follow-up interviews with six students. The students were selected based on the depth of their survey response and to allow for some variety of arguments based on our initial examination of the data. During the interview, respondents were shown the prompt, followed by their response, and asked to share more about how they were thinking about the comparison. These follow-up interviews offered some additional validity evidence for our interpretations of students' responses, in addition to some limitations. We discuss these in detail in Section 7.2.

5.2. METHODS OF ANALYSIS

Before thinking in terms of resources, we began our analyses with a round of descriptive coding and a round of pattern coding that helped us better understand the data in general (Saldaña, 2021). Descriptive and pattern coding were completed for Prompt 1 responses, followed by Prompt 2 responses. Finally, we reviewed the data again and coded in full for resources. This process is visually summarized in Figure 4.

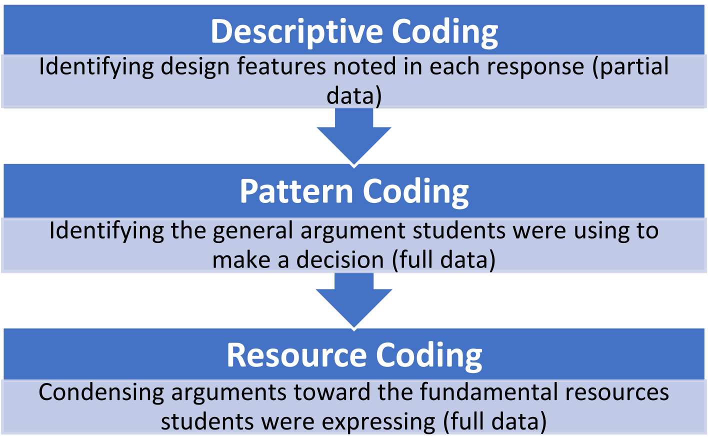
Descriptive Coding
Identifying design features noted in each response (partial data)
Pattern Coding
Identifying the general argument students were using to make a decision (full data)
Resource Coding
Condensing arguments toward the fundamental resources students were expressing (full data)

Figure 4. Outline of response analysis phases

Our descriptive codes focused on features that students noticed (Saldaña, 2021). Examples included "control group," "placebo," "baseline," and "survey." The first and second authors completed two rounds of coding for each prompt. For each round of coding, we randomly selected 60 student responses and wrote our descriptive codes independently. We then met to discuss our codes and formalize themes we each noticed. We repeated this process again with 60 more responses. The descriptive coding was not an aim in itself, but rather a step in helping us to understand what students noticed and to generate ideas for pattern codes. After coding 120 responses from each prompt, our discussions of descriptive codes naturally began evolving into discussions of students' arguments; thus, we ended the descriptive coding stage and began the work of pattern coding the arguments students were making.

We began drafting pattern codes based on some of the descriptive codes we had discussed from the first reviewed responses. For example, "baseline" and "directly compare" helped us see a common theme for students to value the pre-post design for its ability to directly compare participants' blood pressures against their own baselines. This idea would eventually be labeled under the pattern code *Baseline Comparison*. Other examples of pattern codes included *Group Imbalance*, *Individual Variation*, and *Equivalent Evidence*. Although many responses were represented with just one pattern code, some included several codes if multiple arguments or concepts were present in responses. Blank responses and responses that only said, "Don't know," or some variation of that, were coded as *No Argument*. Others that were either too difficult to interpret or provided too superficial an explanation (e.g., "because they compared one group to another") were coded as *Unclear*.

After creating an initial codebook with descriptions and examples, the first and second author independently pattern coded a random sample of 30–60 responses. We then met after each batch to discuss difficult-to-code responses, refine existing codes, create new codes as needed, and come to agreement on responses we coded differently. Using 60 uncoded responses for each prompt, we reached over 80% agreement in our codes and came to agreement on discrepant codes. We then divided the remaining responses up to finish the pattern coding. This process resulted in 18 pattern codes for Prompt 1 and 21 pattern codes for Prompt 2.

Although the pattern codes offered insights into the different kinds of arguments students made, we decided to complete a final iteration of coding to focus on the fundamental resources students drew from to form their arguments. In this stage, we focused on identifying productive observations, rather than coding for argument styles and themes. Some of the pattern codes were already in the form of a resource, but many pattern codes were combined or reconceived to better represent a general heuristic rather than a stylistic difference of argument. For example, the pattern codes *Individual Movement* and *Individual Variation* were reconceived into the resource code *Individual Effects*. We also developed an *Other* category to include general arguments that did not show sufficient evidence of a clear underlying resource guiding the student's thinking. For example, the pattern code *Vague Experiment* reflected responses that preferred the experimental design in Prompt 2 because it used an experiment but did not include sufficient explanation for why being an experiment would provide better evidence.

The first author led the task of identifying and describing resource codes, with consultation from the third author. After reviewing the data, we identified three unique resource codes for the first prompt, six unique resource codes for the second prompt, and two resource codes that appeared in both prompts. The first and third author followed a similar process for resource coding as for pattern coding. We once again reached above 80% consistency in our decisions by using 60 responses for each prompt. We then coded the remaining responses separately and discussed any difficult-to-code responses together.

For a complete listing of resources, including descriptions and student examples, please see Tables 1 and 2. All student quotes are tagged with a response ID that can be traced in the data file that accompanies this report, which includes all student responses along with the abbreviations displayed in Tables 1 and 2 for the appropriate resource code. Furthermore, readers may also notice examples of responses tagged as Other in the data file, which represent responses that we did not find to have a clear or articulate enough argument to tag with a resource code.

Table 1. Resources identified from Prompt 1

Resources	Descriptions	Student Examples
Fair Comparison (FC)	Using individuals' pre-treatment measurements allows us to make a direct, fair comparison. Comparing two independent groups' post measurements may not be a fair, equivalent comparison.	It's better to see the before and after result for each individual because each individual is different. The more diverse the test subjects are, the more accurate the results would be. [A-172]
Individual Effects (IE)	Taking pre- and post-measurements for each person allows us to measure individual-level effects. This may either come from an argument that the pre-post can better answer how much change the treatment causes, or more generally to better track individual reactions and experiences.	You can see how the drug directly effected each person and by how much. [A-290]
Sample Size (SS)	Taking before and after measurements from each person provides twice as much data. The multi-group design will have only half of the participants take the medication, whereas the pre-post allows us to see how everyone responds to the medication.	The pre-post design has more data to compare to each other. [B-11]

Table 1 (Cont.). Resources identified from Prompt 1

Resources	Descriptions	Student Examples
Confounding Variables (CV)	A good comparison will isolate the drug as the only difference between each set of measurements. The pre-post design may allow for other differences between the two sets of measurements (e.g., sources of stress, changes in diet), whereas the multi-group design makes a controlled comparison.	If everyone takes the medication, then there is no way to tell if their blood pressure is changing because of the medication or other factors. By having a group that doesn't take the medicine, there is a group that you can compare to. [A-498]
Placebo Effect (PE)	The multi-group design better guards against a placebo effect, making a better argument that the medication is truly effective, rather than detecting any psychological effect.	The response can be attributed to the treatment itself and not the idea of the treatment. [A-399]

Table 2. Resources identified from Prompt 2

Resources	Descriptions	Student Examples
Natural Habits (NH)	The data better represents reality if we study people in their natural habits, rather than after an intervention where we have disrupted their natural habits.	Study one provides stronger evidence because there is no manipulation with subjects, they are living normally rather than the other study where people had their bedtimes changed. [A-329]
Representative (RP)	Our sample should include all types of people in our population of interest. Only looking at those with irregular bedtimes does not represent the whole population.	Study 2 only uses a sample of college students that don't regularly keep a sleep schedule while study 1 uses a representative sample of college students in the US. [B-200]
Sample Size (SS)	Having a larger sample size or more information in general provides a clearer picture from which to base our conclusion.	There is a much larger sampling size in Study 1. [B-157]
Intervention (IN)	It is better to enact an intervention or change in routine in order to see a direct effect; We can control the groups and see what happens.	Study 2 provides stronger evidence for this claim because participants changed their routine to keep a regular bedtime and as a result they slept longer at night. [A-64]
Confounding Variables (CV)	A good comparison will isolate the sleep schedule as the only difference between each group. There could be confounding factors when only observing what people normally do. Keeping a regular bedtime may just be correlated with more sleep, rather than the causal reason.	In Study 1, it is difficult to make this claim because the survey doesn't ask about other circumstances and outside factors. Study 2 has two randomized groups with a standard set of directions to directly compare the independent and dependent variables. [B-148]
Homogeneous Groups (HG)	It's better to compare two groups with similar starting characteristics.	Because the participants of two groups are under a more similar condition. The effectiveness of outcomes is stronger. [A-373]
Measurement Accuracy (MA)	Survey data might have more bias or estimation, as opposed to making measurements, or recording data with full awareness at the time.	The students in the second study track their exact times over several days as opposed to the first group who likely doesn't record the exact amount they sleep. [A-460]
Equal Group Sizes (EG)	It's better (or perhaps, necessary) to have equal group sizes. Having an unequal number in each comparison group may create a biased comparison.	Study 2 has an equal amount of participants in each group while Study 1 doesn't state how many people are in each group. [A-468]

6. FINDINGS

6.1. SURVEY RESULTS FOR PROMPT 1

Resource breakdown. In terms of overall design preferences, a plurality of students (43.7%) chose the pre-post design as being more effective in determining the effectiveness of the experimental drug. There were 36.3% of the students who preferred the multi-group design, and 20.0% who chose both as being equally effective.

After analyzing the open-ended responses, we identified 32.9% of responses as representing at least one of five resources listed in the Resources column of Table 3. Nineteen student responses were coded with more than one resource; for this reason, the percentages from all resource categories in each column of Table 3 add up to slightly more than 32.9%. In addition, we also coded 29.1% of student responses as providing *No Argument*, and the remaining 38.5% of responses as *Other*. The first three resources listed in Table 3—*Fair Comparison, Individual Effects,* and *Sample Size*—were associated with arguments in favor of the pre-post design. The last two resources—*Confounding Variables* and *Placebo Effect*—were associated with arguments for the multi-group design.

Table 3. Results from Prompt 1 by course and overall

Resources	Course A	Course B	Overall
Fair Comparison	7.6%	14.8%	9.7%
Individual Effects	10.8%	8.3%	10.1%
Sample Size	4.3%	0.9%	3.3%
Confounding Variables	5.4%	9.3%	6.4%
Placebo Effect	3.5%	9.7%	5.2%
Other	39.1%	37.0%	38.5%
No Argument	31.5%	23.1%	29.1%

When comparing responses by course, there are a few notable differences. We identified resources in approximately 40% of Course B responses, as compared to approximately 30% in Course A. Students in Course B were also more likely to discuss a placebo effect threat or confounding threats when comparing the designs, whereas Course A students were slightly more likely to argue in favor of individual effects. These differences might be explained by the higher proportion of Course B students who had reported taking a statistics course previously. Course B students were also much more likely to be enrolled in a science program such as Biology, so prior exposure to experimental design may be an influence on their reasoning as well.

Resource discussion. According to *Fair Comparison*, it is fairer to compare people to themselves than to compare them to a separate group of people. Some students expressing this resource may have a distrust for (or lack of understanding about) random assignment in creating equivalent groups. Student B-191 explained, "I think the pre/post-test is the most effective as it directly measures the results of the medication from the initial and not the results of the medication from the control, which should not be related at all." But even with random assignment, there may still be differences between the two groups created as a result of random chance. "The measurement is more specific to each person in the experiment and can produce a more accurate result." [B-52] Although students using this resource may not realize other systematic differences that could be present in the pre-post design, they do recognize and value equivalency as a key principle.

Individual Effects values the pre-post design for more clearly measuring effectiveness of the medication, or for tracking how effectiveness might be different for different people. For example, "The pre-post design is more effective because the researchers can then garner the total change that the medication has made on each patient instead of just the end results." [B-201] Another idea tied into this resource is that determining effectiveness may involve more than just a numeric change in blood pressure. The pre-post design might be better at tagging individual reactions in general, as Student A-43 identified, "Individual reactions to medication is important."

Sample Size was one of the two resources that appeared in students' responses to both prompts. In the context of Prompt 1, the appeal to the Sample Size resource can be seen in observations that the pre-post study produces twice as many measurements as the multi-group design. Student A-120 articulated, "Finding the average difference in blood pressure in all members is best, since there will be more data to compare."

Confounding Variables and *Placebo Effect* were originally coded as the same resource but were eventually split into two separate codes. A placebo effect is technically a specific case of confounding, but the contextual features of the first prompt seemed to be key in eliciting that particular observation. Students who used Placebo Effect would sometimes discuss how other factors can confound the comparison as well (in which case, both resources were coded), but more often, students talked about either the placebo effect or other confounding factors. In both cases, students using these resources value the importance of isolating the medication as the only systematic difference between each set of measurements.

Other responses. Even though 38.5% of the responses were marked *Other* during the resource coding phase, most of these responses still had an identifiable pattern of argument. One such pattern argument was *Aggregate Differences*, which is represented well in Student A-157's response: "I think the multi-group design is more effective because you are comparing the groups as a whole to each other instead of comparing each person in the groups because everyone has different effects to things." We tagged 2.1% of all responses with this pattern code, but we did not recognize this as a resource because the argument did not actually highlight anything productive about the multi-group design. This argument is a misplaced concern that the pre-post design cannot be used to make an aggregate comparison. We discuss this argument more in a later section concerning its conceptual tension with Individual Effects.

There were also many responses that highlighted the value of taking before-and-after measurements in the pre-post design but without a clear explanation for why. Student A-371, who chose the pre-post design, wrote, "We could better see the changes in blood pressures as they are taken before and after medicine taken." Many of these students might be thinking in terms of Fair Comparison or Individual Effects, but there was not enough evidence to make these connections clear. We had several pattern codes overlapping with this theme, but we estimate that an additional 8.8% of all responses based their argument on some form of this observation.

Other responses included experimental language to justify the multi-group design as more effective. Many of these responses were tagged Other if they lacked clear reasoning for why that mattered. For example, "The multi-group design is more effective because the treatment and control groups can be directly compared to determine the effectiveness of the experimental drug." [A-356] Although this response is generally true, it does not explain why pre- and post-measurements cannot also be directly compared to determine effectiveness. There were also a large number of students that chose the multi-group design because it had a control group or because it had a placebo; many of these, however, did not clearly explain why those factors were strengths. These vague experimental arguments align closely with findings from Derry et al. (2000); students may be familiar with this vocabulary but fail to articulate the value of experimental design. Vague experimental arguments comprised 12.7% of all responses.

The remaining Other responses varied. Some expressed misunderstandings of the designs; some were difficult to interpret and code; and some lacked evaluative language in general. For example, 4.2% of all responses were tagged with the pattern code of *Equivalent Evidence*, which is demonstrated in Student A-398's response: "I believe that both are effective because they are just different moderations of the effectiveness of the drug." Many responses tagged with this code identified both designs as valid because they made a comparison, or because they both used data that was relevant to the question.

6.2. SURVEY RESULTS FOR PROMPT 2

Resource breakdown. With this comparison, a clear majority of students (53.4%) preferred the smaller sample experiment. This was followed by 27.2% of students who preferred the larger sample survey and 19.5% who chose both designs as providing equivalent evidence.

After analyzing the open-ended responses, we coded 3.2% as providing *No Argument* and 39.1% as *Other*. The remaining 57.9% of responses represented at least one of eight resources listed in the Resources column of Table 4. As with the first prompt, the percentages in each column will add up to more than 57.9% to reflect that some students elicited more than one resource in their response. We tagged 11% responses for this prompt as referencing at least two resources. The first three resources listed in Table 4—*Natural Habits*, *Representative*, and *Sample Size*—were associated with arguments in favor of the observational study. The last five resources—*Intervention*, *Confounding Variables*, *Homogeneous Groups*, *Measurement Accuracy*, and *Equal Group Sizes*—were associated with arguments for the experiment.

Table 4. Results from Prompt 2 by course and overall

Resources	Course A	Course B	Overall
Natural Habits	10.2%	7.4%	9.4%
Representative	2.8%	3.2%	2.9%
Sample Size	17.1%	13.9%	16.2%
Intervention	19.7%	27.3%	22.0%
Confounding Variables	3.3%	9.7%	5.2%
Homogeneous Groups	4.6%	4.2%	4.5%
Measurement Accuracy	8.3%	8.3%	8.3%
Equal Group Sizes	0.9%	1.9%	1.2%
Other	41.4%	33.3%	39.1%
No Argument	3.2%	3.2%	3.2%

There are, again, minor differences between Course A students and Course B students. These differences mirror those discussed with Prompt 1, where Course B students were more likely to use a resource than Course A students. We also see higher percentages of Course B students using an Intervention or Confounding Variables resource, and slightly less likely to use a Natural Habits or Sample Size resource. As with Prompt 1, Course B students seem slightly more likely to favor the randomized controlled experiment.

Resource discussion. *Natural Habits* stems from a heuristic that sleep and bedtime routines might be difficult things to disrupt in an experimental format. Student A-184 explained, "Study 1 allows students to maintain their habits instead of changing their schedules and allowing their body to adjust as performed in Study 2." This concern does not necessarily threaten the experiment's ability to make a fair comparison, but it does dampen any benefits of a regular bedtime if studied over a short time period. It may also be a concern that participants who know they are in a study may behave differently and threaten the generalizability of results (i.e., Hawthorne Effect). We believe this is a productive observation for students to make because it targets the discussion of when a controlled intervention will appropriately generalize to everyday life.

The *Representative* resource focuses on the sample makeup of each study. By only looking at students with irregular bedtimes in the experiment, it seems that the observational study is better at representing the population of interest by including students with different bedtime routines. Several students who used this argument described Study 1's sample as more "randomized" or suggested that Study 2's sample was biased from the beginning. Although there may be some blind spots with this resource in isolation, it does suggest a basic awareness of generalizability in linking the sample's composition to the population at large.

Sample Size was a common and explicit resource in the second prompt due to the obvious difference in sample size. Out of the 109 responses that we tagged with this resource, we identified 75 as using this resource uniquely in their arguments, whereas 34 paired this observation with at least one other argument. We also believe that the explicit sample size difference might be a reason why we identified fewer No Argument responses for Prompt 2.

In contrast to *Natural Habits*, many students viewed *Intervention* as a strength in favor of the experiment. When coding for Intervention, we were looking for explicit language around the idea of having participants make a change or complete an action. For example, "Study 2 shows the direct effects of changing to a regular bedtime, while Study 1 might be more of a correlation." [B-59] We did not tag responses with this resource if students simply identified Study 2 as an experiment without the intervention component. For example, we did not tag the following response: "Study 2 actually performed an experiment, and this allowed for their results to represent the claim." [B-211] We tagged responses such as these with a pattern code that are addressed later. Although the Intervention resource does not provide as sophisticated an argument as Confounding Variables, it does represent a starting point for evaluating causation. It values the assignment of a treatment in order to observe and identify the mechanism of change.

Confounding Variables also applies to the second prompt. The observational study leaves open the possibility that students with a regular bedtime might be different in other ways; thus, we cannot be sure that a regular bedtime affects sleep, "Because it is possible that the 6.9-hour group in Study 1 is just more health conscious and it has nothing to do with if they keep to a schedule." [B-215] In contrast, the experiment isolates the regular bedtime as the only systematic difference between each group: "Study 2 allows us to conclude that the decision to keep a regular bedtime is causing the effect, and nothing else." [A-163]

Homogeneous Groups recognizes that in the experimental design, we are comparing two groups of people with similar habits. This makes for a cleaner comparison as compared to two groups who may have other starting differences. Many of the responses tagged with this resource were short, direct observations of this situation. For example: "the sample has the same habits," [A-264] and "It takes students who started out the same as opposed to just two different groups." [B-15] In many cases, students might have the more sophisticated understanding of Confounding Variables—but consideration of the effects of confounding was not made explicit in their responses.

With *Measurement Accuracy,* some students expressed general concerns over survey data as prone to bias, lying, or inaccurate information. Student A-463 stated: "Study 1 is just a survey, where the information given can be false." Many paired this idea with the belief that an experiment has measurement advantages in getting more accurate data. As Student A-139 explained, "Study 2 is using direct measured evidence in a two-week experiment, while study one is going off what the students say." Many students identified how the experiment was collecting data day-by-day, whereas the survey depended more on memory recall and lack of specificity.

Equal Group Sizes was typically expressed as a concern over the uncertainty in group sizes for the observational study: "Study 2 has an equal amount [sic] of participants in each group while study 1 doesn't state how many people are in each group." [A-468] Although the case that equal group sizes makes for the experimental design is somewhat dubious, we do believe it has a productive element. A larger sample size would provide power in a statistical comparison, but an extreme imbalance could leave the observational study with less power to accurately represent the sleep averages of a particular group.

Other responses. One distinction between the first and second prompts was the addition of a statistical result. In the second prompt, Study 1 and Study 2 each found the same statistical result, with college students who hold a regular bedtime averaging 6.9 hours of sleep, and those who do not hold a regular bedtime reporting 6.2 hours on average. This feature resulted in the pattern argument *Same Results*—the justification that both studies provide equal evidence for the claim because they both reported the same outcome. In total, we tagged 13.7% of all responses with this pattern code. Related was the pattern code *Equivalent Evidence*—the justification that both studies are equally valid, but without referencing the results being the same (2.0%). We did not regard either of these pattern arguments as resources because we did not find them to be productive entries in evaluating each design. The focus on the statistical results alone seemed to side-step engagement with the designs.

There were also many responses in support of Study 2 that were based on its experimental features or organization in general. Consider the following three examples: "Study 2 because they have a control group and an experimental group" [A-4]; "I think Study 2 is more organized and focused on the main idea" [A-382]; and "Study 2 is an experiment. Study 1 reads like an observational study." [B-126] We tagged 11.9% of all responses as having a vague experimental argument without enough clarity or justification to be tagged with a resource. Similar to these were references to randomization. These were largely in favor of Study 2 but occasionally were attributed to Study 1, as demonstrated in these two responses: “Study 2 because it is more random” [A-240], and “Study 1 is more randomized. Study 2 could have some bias.” [A-425] This pattern code appeared alone in 2.4% of responses.

6.3. TENSIONS IN PROMPT 1

Power versus causality. Several of the identified resources, and argument patterns in general, reflected an explicit prioritizing of generalizability, causality, or power. Although most responses focused on no more than one of these evaluative dimensions, there were several responses that revealed students grappled with different strengths (or weaknesses) of each design explicitly. While pattern coding, we noted responses that reflected a *Direct Tension* argument—the recognition that both studies have different strengths or different weaknesses. For most responses that reflected a tension, students would choose both designs as equally effective.

With the first prompt, some students recognized that the pre-post design had stronger power and the multi-group design had a stronger claim to causality. For example, consider Student A-66’s response valuing both Sample Size and Placebo Effect: “Both designs are roughly the same in efficacy as pre-post study will generate a lot of data while the one with a placebo will have less data but will check against a placebo group.” In addition to Sample Size, we found that Fair Comparison could also be interpreted as an argument for power, as it is with Student B-52’s response: “The measurement is more specific to each person in the [pre-post design] and can produce a more accurate result.” This perspective suggests that both designs can make a comparison, but the pre-post design can make a cleaner comparison that reduces person-to-person variation.

Responses tagged with Fair Comparison could also take a slightly different lens. Many students preferred the pre-post design because they believed that a comparison with two independent groups could not provide an appropriate comparison. For example: “you have to have a starting point, otherwise you don’t know if the medication actually helped, or you just had people with lower blood pressure in the experimental group.” [A-241] With this perspective, Fair Comparison becomes a causal argument for the pre-post design.

Generalizability. Some students also compared the studies from a generalizability perspective. A common tension emerged around whether the overall group differences reported by the multi-group design were more generalizable or whether tracking individual responses provided better information for that purpose. This tension is well expressed by Student A-516:

> *I think that the multi-group option works best to assess the effectiveness of the experiment drug because you want to see how effective it can be to the general public. By using the pre-post design, you’ll be able to identify how effective it can be to individual participants and base the results on their stats.*

The first part of this response reflects the Aggregate Differences argument, whereas the second highlights advantages of the pre-post design through an Individual Effects argument.

As mentioned earlier, we do not see Aggregate Differences as a resource in itself because it focuses on a deficit understanding of the pre-post data. Also, as discussed earlier, the Aggregate Differences argument may also be a result of a wording inconsistency that led students to think that a measure of aggregate differences may simply not be considered when using the pre-post design. Regardless, we found it interesting to consider this tension through one of making generalizable claims. Aggregate Differences would argue that generalizable claims can only be drawn from aggregate measures rather than individual data. The Individual Effects resource values how having individualized data helps us make more specific generalizations to sub-populations but is not necessarily in contrast to valuing aggregate measures.

6.4. TENSIONS IN PROMPT 2

Generalizability versus causality. The second prompt produced two clear tensions in valuing generalizability versus causality. First, we see Natural Habits making an argument in direct contrast to Intervention, where students were deciding whether intervention was a good or bad thing to try in this study context. Natural Habits values generalizability of habits and environment. By observing student behavior during an intervention, the results we see may not generalize to how such a change would truly affect college student behavior. Intervention instead appeals to the importance of causality to determine if the consistent bedtime would change total sleep times. Student A-490 demonstrated this tension:

> *Study 2 is having the students maintain a regular sleep schedule that they are supposed to follow, so consequently, the results show that those with a set sleep schedule will be more likely to get more seep [sic]. However, in Study 1, the results of average amount of sleep among both of the groups are more accurate because the students are being asked to report their experiences, rather than change them for an experiment.*

The nature of sleep and routine in this prompt seemed key for many students making these arguments. Students also commonly noted the length of intervention time (two weeks) as being too short to set and see results from a new habit.

There is also an inherent tension between Representative and Homogeneous Groups. In the case of Representative, there is a concern that Study 2 only represents students with irregular bedtimes rather than all students. Homogeneous Groups view that commonality as a strength for making an equivalent comparison between the treatment and control group. Representative is a critical component to the observational study, but the experiment does not lack generalizability if we view the population as students with irregular bedtimes. We view Representative as a proper heuristic in general, but not particularly appropriate in this context to argue that the experiment had improper representation.

Power versus causality. With the explicit difference in sample size, many students recognized this as an advantage in some way for observational study. But students also recognized causal strengths to the experimental design. Seven students expressed this tension, and all seven concluded that both studies provided equal evidence. For example, "Study one has a much larger sample size, so the results are reliable, but study two also provided solid evidence because the students were told to change their behavior and when they did the results supported the claim." [B-116] It is difficult to judge how others might have valued the sample size because many students simply did not acknowledge that in their responses. But these seven responses do suggest that some students may not know how to evaluate how sample size stacks up against other study features when weighing evidence.

7. DISCUSSION

7.1. IMPLICATIONS FOR TEACHING

Implications for curriculum. The data we examined provides a glimpse at the general arguments and conceptual resources students used to make quick, evaluative comparisons between different designs. Some students took a more radical, either–or evaluation of each design by suggesting that only one of the two studies provided data that could address the guiding question. For example, some who favored the pre-post design believed that measurements from two independent groups could not be validly compared. Student A-165 explained, "The post-measurements have no significance unless they are compared to the pre-measurements." Although pre-measurements would likely strengthen the power of the multi-group design, its absence does not make the comparison invalid. This more radical evaluation resembles the bog of disbelief identified by Tintle et al. (2015) where studies that have an apparent flaw or limitation may be disregarded completely.

We also saw examples of students who believed both studies to be valid and equivalent for surface-level reasons. For example, many students equally valued both the large observational study and experiment in Prompt 2 because they yielded the same statistical results. Others rated each design as equal because, as Student A-17 argued, "both designs will put out data so you can gather information for both." These more widely accepting responses reflect Tintle et al.'s (2015) bog of overconfidence in ascribing validity to almost every study that collects data in relation to the guiding question.

Some responses were closer to striking a balance, either through an explicit tension, or through other language suggesting both studies offered at least some insight. The primary limitation for some was a lack of clarity about *how* to evaluate each design. Few were proactive or clear about describing what each design might afford and fail to afford, with many simply choosing both as equally effective whenever they encountered pros and cons for each.

In all cases described above, students may benefit from an organizing framework to evaluate design. We propose framing these various strengths and weaknesses through the dimensions of causality, generalizability, and power. Traditional design topics, such as randomization in sampling and group assignment, might still be productive, but do not engage students in other valid resources they may elicit (e.g., Sample Size, Fair Comparison, Natural Habits). Framing these discussions in terms of causality, generalizability, and power would create a space for students to name and understand the role of many of their observations and orient students to how different design features may affect different evaluative dimensions.

We see two important benefits to this approach to teaching design. First, it would help students realize that statistical *claims* are not objective. Instead, drawing claims from data is situated in a cycle of inquiry (Windschitl et al., 2008). Data-based claims offer insights that can be evaluated with different lenses. Some studies have better generalizability; some show causality more clearly; and power affects how small an effect or association we can confidently identify. Second, it would demonstrate that the *process* of designing and carrying out a statistical analysis is also not fully objective. An experiment may be less generalizable in certain contexts, and baseline data may introduce confounding effects in some situations. Design decisions are embedded in context (Bennett, 2015).

Implications for instruction. From our experience working with introductory students, we were not surprised to see that students had different design preferences. We were, however, struck by the variety of arguments students made for each design, including the sensible contextual arguments in favor of the pre-post design and observational study. Students who preferred the pre-post design over the randomized controlled experiment did not necessarily lack an understanding of confounding; instead, some valued the ability to track the variation in medication response across different individuals. Similarly, some students drew concern about whether a short-term intervention could show generalizable results for something such as bedtime routines and sleep. In these and other cases, students offered reasonable critiques by engaging with the context of each question and inquiring about what we might gain or lose with each design. Contextual reasoning is a hallmark of expert thinking that is a goal in and of itself for statistics instruction (Zieffler et al., 2008). Thus, the prior ideas students expressed were valuable and should be considered in instruction on design.

In our teaching, evaluation of design has often been focused on causality above all else. We have often framed randomized controlled experiments as a gold standard for design and observational studies as a design that should be undertaken with caution. We now worry that this type of framing, absent of context, may shut down the authentic inquiry that should take place around crafting a design. In many ways, it encourages students toward overconfidence when evaluating randomized controlled experiments (Tintle et al., 2015), much like the initial confidence ascribed to the early randomized controlled experiment for hydroxychloroquine (Gautret et al., 2020). We do not believe the solution is to teach all designs as equally imperfect because that would only push students toward general disbelief and disregard toward statistical claims. Rather, students need a healthy understanding of the limitations that accompany different designs *in context*.

The most important thing we have learned from this investigation is that instruction in design must be deeply embedded in contextual examples (Bennett, 2015; Easterling, 2004; Watson et al., 2020; Zieffler et al., 2008). We view this link as critical to supporting statistical thinking as an opportunity for inquiry, rather than as an exercise in objectivity. Randomized controlled experiments can offer strong claims of causality, but some contexts may make this design difficult to implement or problematic to generalize from. We now plan to present contextual examples that inspire students to propose different design features we have learned, rather than teach named designs first. Lastly, students should recognize that research does not progress linearly with each new study. Instead, students should develop the perspective that research proceeds through inquiry, and different designs may complement one another in addressing a question.

7.2. LIMITATIONS AND FUTURE RESEARCH

The data we collected for this study is quite limited in scope. First, our data were limited to only two contextual examples that students were asked to consider. It is likely the case that students would draw on more resources if presented with a different context (e.g., Hiebert, 2007). Similarly, we chose two particular design comparisons that were intended to be simple and accessible for students. Had features been different, such as making the observational study the same sample size as the experiment or having the multi-group design also take pre-measurements, it would have been interesting to see how students' choices and arguments might have changed.

We also acknowledge limitations to the wordings for each prompt that may have affected student responses. For Prompt 1, we realized only after data collection that the nature of the statistical comparison (e.g., a comparison of post-intervention means) for the randomized controlled experiment was not explicit. Whereas for the pre-post study, the statistical comparison is more explicit as measuring the average change in blood pressure before and after medication. The six students we interviewed post-hoc expressed no ambiguity over confusion regarding the timing of measurements in the multi-group; however, we acknowledge that this may have confused some survey-takers and potentially increased the number of unclear argument patterns. Had it not been for this ambiguity, more students might have expressed a resource, or at least made a clearer argument. We do not, however, expect that the overall *set* of resources we identified would likely have changed had the wording for the randomized controlled experiment been clearer.

We presented all students with the same ordering of prompts and design orderings within prompts. It is possible that students may have answered differently or noticed differently had the order been changed. Like the wording issue above, we believe an ordering effect might slightly affect the quantitative results reported but is unlikely to have affected the actual set of resources identified.

As discussed in the Methods section, we chose to only compare two prompts with various feature differences. We made this decision to better understand what features students noticed and valued most, rather than comparing designs with a single targeted difference. We also designed a shorter survey to avoid burnout and encourage more engagement and elaboration from students in their responses. With the various resources identified in this study, it may be advantageous to construct a more targeted set of comparisons that leverage the tensions identified in our data.

In addition, the nature of a survey likely dissuaded students from presenting comprehensive arguments representing multiple resources and tensions explicitly. By limiting the survey to only two comparisons, we tried to reduce the likelihood of burnout and superficial responses. Still, we found several of the interviewed students expressed richer considerations than were truly captured in their written responses. Although the survey was useful for identifying the set of resources that may be elicited by students with these tasks, interviews might yield more depth to explore resource blending and to identify all features and factors influencing their choices. A more carefully prepared survey might also do the same if students were asked to rate the importance of various design features in making their evaluation.

As statistics educators, we quickly noticed how most evaluative responses could be traced back to the issues of causality, generalizability, and power. We acknowledge that this may not be a comprehensive categorization for all design considerations. For example, different designs may highlight questions of instrumentation, or the suitability of methods to address a question. Future research may target these broader design considerations with prompts intended to highlight such features.

After completing this study, we are curious to what extent students recognize evaluations of power (e.g., sample size) as being quite different from evaluations of causality and generalizability. Many students made an argument solely from the Sample Size resource, in addition to others who had difficulty judging how to weigh Sample Size against Intervention. Future research may be useful to explore whether students believe larger sample sizes make a study's claims more generalizable or improve the argument for causality.

8. CONCLUSION

Our study examines how introductory statistics students at the beginning of the course make evaluative decisions about basic design choices. When presented with a large observational study versus a smaller randomized controlled experiment, students heavily favored the experiment. The justifications students provided for this choice were also typically clearer, with over 96% making at least some response. Choices were more mixed when comparing the randomized controlled experiment to a pre-post design with no control group; there was also a much higher proportion of students who offered no argument or an unclear argument in justifying their choices.

From a curricular perspective, we are interested in how coverage of design can reflect the ideas that students are noticing in a cohesive way. We identified a variety of resources that students drew from when making their evaluations. Most students, however, focused on only one resource or argument to justify their response, and those who grappled with multiple arguments struggled to see how these separate features often represented different evaluative criteria (e.g., causality, generalizability, power). For this reason, we believe students might benefit from seeing an organizing framework that explicitly discusses these dimensions of evaluation. Such a framework might guide students to see how the features they notice affect the strength of a claim and how these features may often be on different evaluative dimensions.

In addition to channeling student ideas through the dimensions of causality, generalizability, and power, we also see how this data demonstrates the critical role of context in inspiring student inquiry. Many students fall victim to more objective views of design in viewing studies as either valid or invalid with little middle ground. Instead, students should recognize how different designs may offer different advantages. Furthermore, these advantages depend on the questions we ask and the kinds of data we can collect. For this reason, high-quality instruction on design should deeply embed students in context to construct and reason through design choices. Ultimately, we want students to see statistics as a scientific process of building and presenting arguments, reasoning contextually, and applying nuance to the claims that can be made from data. Design is a great opportunity to build that notion into students' statistical conceptions.

ACKNOWLEDGMENTS

We thank the students who completed the survey. We also thank Karle Flanagan for helping with the coordination of the data collection process and for assisting with the interrater reliability coding.

REFERENCES

American Statistical Association Undergraduate Guidelines Workgroup. (2014). *Curriculum guidelines for undergraduate programs in statistical science*. American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/edu-guidelines2014-11-15.pdf

Bennett, K. A. (2015). 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

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55. https://doi.org/10.1002/sce.20286

College Board. (2020). *AP Statistics course and exam description*. https://apcentral.collegeboard.org/media/pdf/ap-statistics-course-and-exam-description.pdf

Crowley, M., Thomas, K., & Haberman, M. (2020, April 5). Ignoring expert opinion, Trump again promotes hydroxychloroquine. *New York Times*. https://www.nytimes.com/2020/04/05/us/politics/trump-hydroxychloroquine-coronavirus.html

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

Dasgupta, A. P., Anderson, T. R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE—Life Sciences Education*, *13*(2), 265–284. https://doi.org/10.1187/cbe.13-09-0192

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2014). Development of the biological experimental design concept inventory (BEDCI). *CBE—Life Sciences Education*, *13*(3), 540–551. https://doi.org/10.1187/cbe.13-11-0218

Derry, S. J., Levin, J. R., Osana, H. P., Jones, M. S., & Peterson, M. (2000). Fostering students' statistical and scientific thinking: Lessons learned from an innovative college course. *American Educational Research Journal*, *37*(3), 747–773. https://doi.org/10.3102/00028312037003747

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. B. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Erlbaum.

Easterling, R. (2004). Teaching experimental design. *The American Statistician*, *58*(3), 244–252. https://doi.org/10.1198/000313004X1477

Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

Findley, K., & Lyford, A. (2019). Investigating students' reasoning about sampling distributions through a resource perspective. *Statistics Education Research Journal*, *18*(1), 26–45. https://doi.org/10.52041/serj.v18i1.148

Fry, E. B. (2017). *Introductory statistics students' conceptual understanding of study design and conclusions*, [Doctoral dissertation, University of Minnesota]. https://iase-web.org/documents/dissertations/17.ElizabethBrondosFry.Dissertation.pdf

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

Garfield, J., Le, L., Zieffler, A., & Ben-Zvi, D. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics*, *88*, 327–342. https://doi.org/10.1007/s10649-014-9541-7

Gautret, P., Lagier, J.-C., Parola, P., Hoang, V. T., Meddeb, L., Mailhe, M., Doudier, B., Courjon, J., Giordanengo, V., Vieira, V. E., Dupont, H. T., Honoré, S., Colson, P., Chabrière, E., La Scola, B., Rolain, J.-M., Brouqui, P., & Raoult, D. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial. *International Journal of Antimicrobial Agents*, *56*(1), 1–6. https://doi.org/10.1016/j.ijantimicag.2020.105949

Hammer, D. (1996). Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions. *The Journal of the Learning Sciences*, *5*(2), 97–127. https://doi.org/10.1207/s15327809jls0502_1

Hiebert, S. M. (2007). Teaching simple experimental design to undergraduates: Do your students understand the basics? *Advances in Physiology Education*, *31*(1), 82–92. https://doi.org/10.1152/advan.00033.2006

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765–789. https://doi.org/10.1017/S0003055411000414

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

Kupferschmidt, K. (2020). Big studies dim hopes for hydroxychloroquine. *Science*, *368*(6496), 1166–1167. https://doi.org/10.1126/science.368.6496.1166

Next Generation Science Standards Lead States. (2013). *Next generation science standards: For states, by state*s. The National Academies Press.

Saldaña, J. (2021). *The coding manual for qualitative researchers*. SAGE Publications.

Schield, M. (2018). Confounding and cornfield: Back to the future. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018)*. International Statistical Institute/International Association for Statistical Education. https://icots.info/10/proceedings/pdfs/ICOTS10_1C2.pdf?1531364185

Smith, C. L., Maclin, D., Houghton, C., & Hennessey, M. G. (2000). Sixth-grade students' epistemologies of science: The impact of school science experiences on epistemological development. *Cognition and Instruction*, *18*(3), 349–422. https://doi.org/10.1207/S1532690XCI1803_3

Smith, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(2), 115–163. https://doi.org/10.1207/s15327809jls0302_1

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98–113. https://doi.org/10.1016/j.edurev.2007.04.001

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2013). Challenging the state of the art in post-introductory statistics. *Proceedings of the 59th World Statistics Congress of the International Statistical Institute* (pp. 295–300). https://core.ac.uk/download/pdf/214190061.pdf

Tintle, N., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2015). Combating anti-statistical thinking using simulation-based methods throughout the undergraduate curriculum. *The American Statistician*, *69*(4), 362–370. https://doi.org/10.1080/00031305.2015.1081619

Wagler, A., & Wagler, R. (2013). Randomizing roaches: Exploring the 'bugs' of randomization in experimental design. *Teaching Statistics*, *36*(1), 13–20. https://doi.org/10.1111/test.12029

Watson, J., Fitzallen, N., & Chick, H. (2020). What is the role of statistics in integrating STEM education? In J. Anderson & Y. Li (Eds.), *Integrated Approaches to STEM Education. Advances in STEM Education*. Springer, Cham. https://doi.org/10.1007/978-3-030-52229-2_6

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, *92*(5), 941–967. https://doi.org/10.1002/sce.20259

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

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40–58. https://doi.org/10.52041/serj.v7i2.469

KELLY FINDLEY
605 E. Springfield Ave., Room 151
Champaign, IL 61820