Understanding Why: Statistical Techniques to Infer Causality are Underused in Computing Education Research

Seth Poulsen
University of Illinois
Urbana-Champaign
sethp3@illinois.edu

Binglin Chen
University of Illinois
Urbana-Champaign
chen386@illinois.edu

Craig Zilles
University of Illinois
Urbana-Champaign
zilles@illinois.edu

Matthew West
University of Illinois
Urbana-Champaign
mwest@illinois.edu

ABSTRACT

A common thread through education research is asking questions about how treatments applied to students affect their education, career, and other outcomes. For example: Will being taught in a certain way increase students' learning? Will taking a computer science course lead to higher job satisfaction in the future? Are remedial programs serving their intended purpose?

The most robust way to establish the causal effects of treatments is to perform randomized controlled trials. However, in the context of education, it would frequently be unethical or logistically impossible to simply assign students to take a certain class or participate in a certain program for the purpose of research.

As a result, we often take advantage of natural experiments or quasi-experiments. In such situations, the traditional method of analysis is to look at the correlation between the treatment and outcome variables. However, this doesn't tell us whether the outcome was caused by the treatment, as there are almost always substantial selection biases or confounding variables.

In the past few decades, advanced statistical methods have been developed to analyze the assignment of subjects to treatments as if it was random, allowing us to deduce the causal effect of the treatment. Such methods include difference-in-differences, instrumental variables, and regression discontinuity design.

In this paper we argue that these methods have been underused in computing education research. To encourage their increased use, we describe the methods and present selected examples of education studies where they have allowed researchers to bridge the gap from correlation to causation.

Proceedings of the 7th Educational Data Mining in Computer Science Education (CSEDM) Workshop, March 2023

Keywords

computing education research, research methods, regression discontinuity design, instrumental variables, difference-in-differences, econometrics

1. INTRODUCTION

Randomized control trials (RCTs) are considered the gold standard for experimental design. In an RCT, individuals in the sample population are randomly assigned to groups, either *treatment* groups or *control* groups that receive either no treatment or a *placebo* treatment that is expected to have minimal effect. Through randomized assignment, *confounding factors*, characteristics of members of the population (unrelated to the treatments) that potentially affect the property of interest, can be expected to be uniformly distributed across the treatments when sample populations are sufficiently large. This prevents treatment assignment from biasing measurements of the property of interest, so that we can assume a causal relationship between the treatments and the observed outcomes for the treated groups.

In computing education research (CER), however, like many human sciences, running RCTs can be too expensive, too difficult, or too limiting. When evaluated in the context of a reasonable baseline (e.g., current best known practices), the impact resulting from a particular treatment may only be reliably measurable after many tens or hundreds of hours of engagement by the learners and months (or even years) of their lives. Running *in-vitro* laboratory studies at this scale, where a researcher can control all of the variables including treatment assignment, is generally cost and effort prohibitive; feasible shorter studies may have effect sizes too small to measure or lack *ecological validity*, in that student behavior in a short laboratory study might not predict their behavior in the real world.

As a result, a significant amount of CER is performed *in-situ*, in the computing courses and activities in which stu-

dents engage as part of their normal lives. In these contexts, there are often barriers that prevent the researcher from controlling the assignment of subjects to treatments, making these *quasi-experimental* studies. These barriers can be practical, ethical, or legal. Most colleges and universities (as well as their Institutional Review Boards) have the expectation that students within a single section of a course are treated fairly and equitably; this prevents faculty from requiring a random half of the students to write essays in place of half of their programming assignments, for example. Even if it were permissible, some treatments (e.g., comparing traditional lecture vs. a peer instruction lecture using clickers) would be impossible when all students are attending the same lecture.

A commonly used strategy, therefore, is to apply different treatments to different offerings of a class, either different sections in the same semester or in different semesters. Different sections in the same semester are subject to either confounding variables impacting the assignment of treatment (e.g., students that register for classes early are more organized in general and are disproportionately represented in sections at preferred times) or confounding treatments (e.g., two lectures at the same time taught by different faculty), or both. Comparing treatments in different semesters is subject to either abrupt (e.g., Fall vs. Spring offerings of early courses having significantly different proportion of students with prior programming experience) or subtle (e.g., institutional admission priorities/policies changing over time) population variations.

The problem with quasi-experimental studies and *natural experiments*, observational studies where the researcher was not involved in constructing the experiment in any way, is that one needs to be very careful about inferring causality when a treatment is correlated to a property of interest. It could be that some other characteristic of the population causally affected group assignment and also causally affected the property of interest. Even when great effort has been taken to show that treatment groups are similar, it cannot be proven that some other, unmeasured variable isn't the confounding factor.

Because many phenomena in human sciences can only be studied through natural experiments and quasi-experiments, researchers have developed a number of statistical techniques to infer causality in these contexts. These techniques, however, have rarely been used in CER and are notably absent from the recent Cambridge Handbook of Computing Education Research [1]. We believe that they are likely being underused in CER because they are not yet well known by the community. As such, the contribution of this paper is to describe three of the most prominent techniques and explain their utility in the context of educational studies.

2. REGRESSION DISCONTINUITY DESIGN (RDD)

Regression discontinuity designs (RDD) [10] are useful in the case where there is a threshold that decides whether an intervention is applied. In a classic example, we might be interested in whether merit-based scholarships for college freshmen improve student degree outcomes. To understand the effect of the scholarships, we cannot simply compare

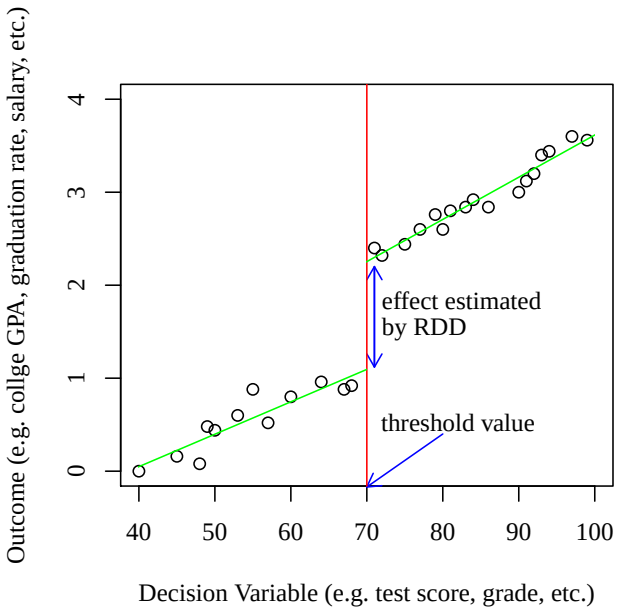

Figure 1: Example data for a Regression Discontinuity Design

the outcomes of those students who were awarded one to those who were not. This would be invalid, because students who do well in high-school are likely to *both* be awarded scholarships and to do well at college.

To use RDD for this situation, we instead focus our attention on the group of students who were *just below* the threshold for being awarded a scholarship and compare them to the group of students who were *just above* the threshold. It is often reasonable to assume that these groups are very similar in every way except the scholarship, so differences in group outcomes can be *causally attributed* to the intervention of giving the scholarship.

In practice, we can use a local linear regression near the threshold, where we fit one linear model for students just below the threshold and a second model above, as shown in Figure 1. Comparing the predictions of the two models at the threshold then gives a good estimate of the effect size. The same idea can also be used with more complex parametric models, such as fitting two quadratics on either side of the threshold and again comparing predictions at the threshold.

For an overview of RDD and a guide to its use in practice we refer to the excellent paper by Lee and Lemieux [10].

2.1 Case study: RDD for remedial programs

An excellent example of the use of regression discontinuity design is the work by Jacob and Lefgren [8], who used it to analyze the impact of remedial programs in K-12 education. Among other questions, they were interested in the effectiveness of "grade retention", where a student is required to repeat a grade (year) of school (that is, be "retained") if they have not demonstrated sufficient achievement. Prior to their work, many studies had found that students who had been retained in a grade scored lower than similar students who

were not retained [6] and were more likely to drop out of school [14].

However, these prior studies were correlational and, although they attempted to compare retained students to similar non-retained counterparts, it is likely that there were unobserved characteristics that both increased the likelihood that a student was retained and also made it more likely that they would under-perform or drop out.

To overcome this limitation of prior studies, Jacob and Lefgren [8] took advantage of a natural experiment that occurred in the Chicago Public School (CPS) system in 1996. Prior to 1996, students in the CPS system moved on to the next grade each year regardless of their achievement level. Starting in 1996, CPS retained students and forced them to repeat a grade if they did not meet certain preset levels of performance in standardized tests.

This new policy introduced a clear discontinuity, where students below the performance threshold were retained and students above the threshold moved on. However, it is reasonable to assume that students *just below* the threshold were essentially identical to students *just above*. More precisely, we assume that student characteristics are continuous across the threshold, so that differences in outcomes between the two sides of the threshold are *caused* by the grade retention policy.

By comparing these two groups of students in an RDD analysis, Jacob and Lefgren [8] showed that, in contrast to previous studies, grade retention had no negative impact on third-grade students and may have slightly helped them, while it had a mixed impact on sixth-grade students with no impact on math and a negative impact on reading. These results suggest that previous studies had substantially overestimated the negative impacts of grade retention, likely because of unobserved confounding variables.

2.2 Prospects for RDD in CER

Regression discontinuity designs are ideal in cases where a treatment is applied or not based on a somewhat arbitrary threshold in a continuous measurement, so that students just below the threshold (who receive the treatment) are essentially the same as those just above (who do not). Anytime this occurs, RDD allows us to make strong causal inferences about the effect of the treatment.

In computing education research, there are many questions where RDD could be used to help determine causal impact. For example: (1) Does streaming incoming students into a no-programming-background introductory course help them in subsequent courses? (2) Does giving a high-performing student a non-"A" grade increase the chance they will change majors? (3) Do one-on-one tutoring sessions for low-performing students help them pass the class? In all of these cases it is the *causal* effects that we are interested in.

3. DIFFERENCE-IN-DIFFERENCES (DID)

Another method to study the mean effect of an intervention from natural experiments or quasi-experiments is Difference-in-Differences (DiD). For example, we might want to know if a dedicated segmentation fault lecture improves students' ability to write segmentation-fault-free code by the end of the semester. We cannot derive any conclusion from pre-post tests over the semester on the group of students that received the intervention (the dedicated lecture) because the timespan between tests is too long and students have plenty of time to improve, even without the intervention. We would need another group of students who do not receive the intervention for comparison (conceptually, a control group). However, randomly assigning students to such groups within a semester is often infeasible, so this is not a random or matched control group.

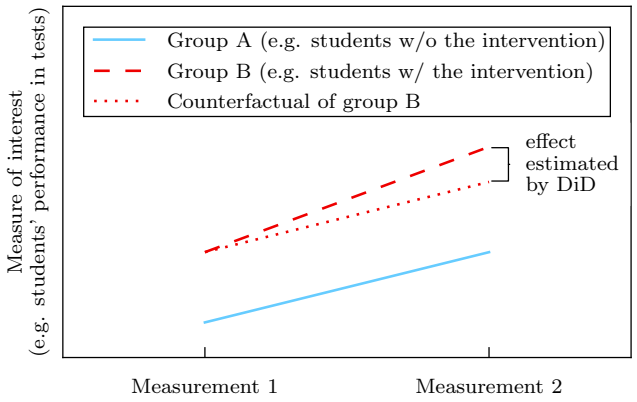

Figure 2: Illustration of an estimation carried out by Difference-in-Differences.

DiD allows us to derive a mean effect of the intervention in this case by collecting a pre-semester measurement 1 and post-semester measurement 2 of two groups, one without the intervention (Group A) and one with it (Group B). See Figure 2 for an illustration. For each group we calculate the *difference* between measurements 1 and 2. For Group A (the "control" group), this difference is the expected improvement without the intervention, while for Group B the difference includes both the normal expected improvement plus the effect of the intervention. To isolate the effect of the intervention, we thus subtract the Group A difference from the Group B difference, giving a *difference in differences* as the effect estimate.

The central assumption of the DiD method is called the *parallel trend assumption*. For DiD to be useful, it must be assumed that in the absence of the intervention, the groups will follow the same trend across measurements of the quantity of interest. Any divergence from this trend can then be attributed to the intervention. In the above example, the parallel trend assumption predicts that the two group of students will have the same improvements over the semester in the absence of the intervention. In order to make such a claim convincing, it would be desirable to have the same instructor teach the course in the same semester of two consecutive years and keep everything else in the course the same but the intervention.

3.1 Case study: DiD for cheating advantage

While DiD is often used in situations with multiple measurements across the time domain, it can also be used when measurements are made across other domains as long as the

parallel trend assumption is believed to hold. An example of such an application is Chen, West and Zilles's work on analyzing how different degrees of randomization affect the amount of benefit that cheaters can gain on asynchronous exams (i.e., when students are allowed to choose when they take an exam in a given time window) [5]. Here the multiple measurements were made on assessments that represented different degrees of randomization. Prior to the work, randomization was a recommended technique to safeguard asynchronous exams [12], but it was unclear how much randomization is necessary.

Chen, West and Zilles collected data from asynchronous exams comprised of randomly-parameterizing questions drawn from the homework as well as "hidden" problems only present on the exams. The questions fell into four categories: (1) hidden questions where every student received the same question, (2) homework questions where every student received the same question, (3) homework questions drawn from pools of two questions, and (4) homework questions drawn from pools of four questions [5]. They applied DiD to compare the performance of cheaters and non-cheaters on these four categories of questions and found that, when students' performance on questions in (1) was used as the baseline difference between cheaters and non-cheaters, cheaters had a significant advantage on questions in (2) but an insignificant advantage on questions in (3) and (4). These results suggest that randomization is indeed necessary for asynchronous exams and provide empirical evidence for the minimum amount of randomization required.

3.2 Prospects for DiD in CER

Difference-in-differences is a candidate method to consider in situations where randomized controlled trials are out of the question. It is best suited in cases where we are interested in the relationship between multiple measurements, e.g., whether a particular intervention improves students' performance for a period of time. The multiple measurements do not have to be across the time domain, but could be in different assessment settings instead. As long as the parallel trend assumption holds with reasonable confidence, DiD can be used to study the *causal* effects of interventions.

The use of DiD in computing education would give better estimates of the *causal* effect of interventions such as (1) the introduction of a block language for the first few weeks in CS1 before a switch to a typical programming language, (2) the teaching of spatial abilities before the regular course contents, and (3) the introduction of programming games as part of the course.

For a gentle overview of DiD and instructions on how to carry out the computation of the mean effect and its confidence interval, we recommend Agrist and Pishke's books [2, 3].

4. INSTRUMENTAL VARIABLES (IV)

Another way to estimate a causal effect from a dataset that was not the result of a a randomized controlled trial is by using a technique called *instrumental variables estimation*. Using this technique requires access to another variable, called the *instrumental variable*, in addition to the treatment and outcome variables. Ideally, the instrumental variable is chosen so that it cannot have an effect on the outcome variable *except* through its effect on the treatment variable. Assuming this is the case, the instrumental variable, once added to the regression, allows an estimation of the causal effect of the treatment variable on the outcome variable.

For example, suppose researchers wanted to understand the causal effect of alcohol consumption on the number of car crashes. Finding the correlation between the two will not reveal the causal effect, as there could be confounding factors, such as income, that effect both alcohol consumption and car crashes—and car crashes themselves may cause more alcohol consumption. On the other hand, the tax rate on alcohol may be used as an instrumental variable because it has no effect on car crashes *except* through its influence on alcohol consumption. Thus, we can estimate the causal effect of alcohol consumption on car crashes by examining the relationship between alcohol tax rate and car crashes.

Instrumental variables have been used to estimate effects of various treatments in education [13, 9, 4].

4.1 Case study: IV for more math classes

We will look more in depth at an example of how instrumental variables can be used to estimate the effect that taking an additional math class in high school has on college and career success.

In the 1980s, while Denmark was in the process of revamping its high school curriculum, there was a pilot program where students were given more flexibility in their course options, and were given the option to take advanced math combined with physics or chemistry, instead of being required to take advanced math with physics. Due to this additional curricular options, about one-third more students decided to take advanced math [9].

Joensen and Nielsen [9] saw this as an opportunity to use an instrumental variables approach to estimate the causal effect that the additional math class had on the students' college and career success. Their instrumental variable was if a student's school had joined the pilot program after the student had started attending. A case can be made that the instrumental variable meets the requirement of not affecting the outcome variable (except through the treatment variable) because the students couldn't have chosen the school for the opportunity to take more math—so their choice to attend that school is not likely to be correlated with other factors that influence their college outcomes, such as socioeconomic status and parental attitudes. Joensen and Nielsen also added additional controls to account for students' existing knowledge. The fact that schools self-selected to be a part of the pilot program introduces some school-level selection bias to the experiment, but they controlled for this using the average GPA of students at the school. They conclude that taking additional math and chemistry in high school caused students to complete a higher level of education and have higher earnings later on.

There have been multiple studies on the question of the effects of additional math in high school using various methods (including a Regression Discontinuity Design, a method discussed in Section 2), forming a consensus that additional

math in high school causes higher college enrollment and wages [13].

On the other hand, we have much less of an idea about the effect that taking computer science has on students' college and career success and satisfaction. A few studies have examined the relationship between taking computer science in high school and college performance, but as these studies only examine correlation between the variables and do not make any estimate of the causal effect of the computer science course [7, 16]. Furthermore, many people are arguing for all college students to take at least one computer science course [11, 15], but studies to support this practice don't yet exist.

In order to understand the effect that taking computer science classes has on students lives in college and beyond, we need multiple studies, using sound methods to establish a causal relationship between taking computer science and whatever the desired outcomes are—whether it be higher college completion rate, greater career satisfaction, or any other outcome.

4.2 Case study: IV for course data access

Instrumental variables have also proved very useful in education research in quasi-experimental settings.

Researchers at Stanford sought to find the effect on student performance of giving students access to additional data about courses in their catalog, including past GPA and student's reported time commitments for the courses [4]. In order to test the causal effect, they used a design known as *randomized encouragement*—randomly notifying a subset of a population about an available treatment, then using whether or not a subject was encouraged to accept the treatment as an instrumental variable to understand the causal effect of the treatment. In their case, this meant randomly notifying some students of the availability of additional course data. They were able to show that being given access to additional course information caused students GPA to decrease by a statistically significant amount.

4.3 Prospects for IV in CER

As stated, IV along with other methods are great candidates for helping us understand the causal relationship between taking computer science classes and college and career outcomes. New research in this area could act as a wake-up call for both private and public entities seeking to make funding decisions.

The randomized encouragement design pattern could open the door for studying the causal effect of many different treatments in computer science education—not only the effects of taking computer science courses, but also the effects of online learning tools, using TA office hours, taking second-chance exams [**?**], and more.

For further examples of instrumental variables estimation and more detail on how to perform one yourself see Agrist and Pishke's books [2, 3].

5. CONCLUSION

While analyzing the results of quasi-experiments or natural experiments by looking at the correlation between the treatment and outcome is a good initial step, it falls short of establishing causation. The methods that we have covered in this paper—regression discontinuity designs (RDD), difference-in-differences (DiD), and instrumental variables (IV)—can shed light on the causal effects of treatments in education research. They can work in many contexts, showing effects of large scale curricular decisions, instructional tooling, pedagogical decisions, and more. In many cases, more than one of these methods can and should be used to gain increased understanding of treatment effects. These methods have already been successful in increasing understanding of educational phenomena in many contexts, and it's time that the computing education research community uses these methods more fully to help us find the answers we seek.

6. REFERENCES

[1] *The Cambridge Handbook of Computing Education Research*. Cambridge Handbooks in Psychology. Cambridge University Press, 2019.

[2] J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[3] J. D. Angrist and J.-S. Pischke. *Mastering'metrics: The path from cause to effect*. Princeton University Press, 2014.

[4] S. Chaturapruek, T. S. Dee, R. Johari, R. F. Kizilcec, and M. L. Stevens. How a data-driven course planning tool affects college students' gpa: evidence from two field experiments. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, page 63. ACM, 2018.

[5] B. Chen, M. West, and C. Zilles. How much randomization is needed to deter collaborative cheating on asynchronous exams? In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, L@S '18, New York, NY, USA, 2018. Association for Computing Machinery.

[6] C. T. Holmes. Grade level retention effects: A meta-analysis of research studies. In L. A. Shepard and M. L. Smith, editors, *Flunking Grades: Research and Policies on Retention*. The Falmer Press, New York, 1989.

[7] K. E. Howard and D. D. Havard. Advanced placement (ap) computer science principles: Searching for equity in a two-tiered solution to underrepresentation. *Journal of Computer Science Integration*, 2(1):1–15, 2019.

[8] B. A. Jacob and L. Lefgren. Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, 86(1):226–244, 2004.

[9] J. S. Joensen and H. S. Nielsen. Is there a causal effect of high school math on labor market outcomes? *The Journal of Human Resources*, 44(1):171–198, 2009.

[10] D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 2(48):281 – 355, 2010.

[11] R. Libeskind-Hadas. Every college student should take

a computer science course, 2015. [Online; accessed 17-January-2020].

[12] M. R. Olt. Ethics and distance education: Strategies for minimizing academic dishonesty in online assessment. *Online journal of distance learning administration*, 5(3):1–7, 2002.

[13] S. Poulsen. The effect of additional math in high school on college success. *The Mathematics Educator*, 28(2), 2019.

[14] R. W. Rumberger. High school dropouts: A review of issues and evidence. *Review of Educational Research*, 57:101–121, 1987.

[15] R. Sedgewick. Should computer science be required?, 2019. [Online; accessed 17-January-2020].

[16] H. G. Taylor and L. C. Mounfield. Exploration of the relationship between prior computing experience and gender on success in college computer science. *Journal of Educational Computing Research*, 11(4):291–306, 1994.