6-6-2023

Identifying Advantages to Teaching Linear Regression in a Modeling and Simulation Introductory Statistics Curriculum

Kit Harris Clement
Portland State University

Identifying Advantages to Teaching Linear Regression in a Modeling and Simulation

Introductory Statistics Curriculum

by

Kit Harris Clement

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Mathematics Education

Dissertation Committee:
Eva Thanheiser, Chair
Jennifer Noll, Co-Chair
Stephanie Casey
Robert Fountain
Antonie Jetter

Portland State University
2023

Abstract

Statistical association is a key facet of statistical literacy: claims based on relationships between variables or ideas rooted in data are found everywhere in media and discourse. A key development in introductory statistics curricula is the use of simulation-based inference, which has shown positive outcomes for students, especially in regards to statistical literacy and conceptual understanding. In this dissertation project, I investigate students from the Change Agents for the Teaching and Learning of STatistics (CATALST) curriculum in activities I designed for learning statistical association and linear regression. First, I analyzed the informal line fitting strategies of CATALST students. Findings suggest that students still face many challenges in informal line fitting, but their use of the offsetting distances criterion may be a future point of focus for teaching and activity development. Next, I compared student outcomes in a traditional course and a CATALST course on their ability to recognize the need for inference and hypothesis testing. Results revealed that CATALST students were more prepared to learn inference in their course and made greater gains by the end of the linear regression unit. Finally, I examine CATALST students' inferential reasoning in light of frameworks that identify challenges in learning simulation-based inference. Based on the success CATALST students demonstrated, I propose technology innovations to the simulation software so that the classroom can better focus on learning statistics rather than technology. Overall, this dissertation provides insights into activities that expand the existing CATALST curriculum to include linear regression and shares the benefits of leveraging this simulation-based curriculum while highlighting challenges these students experienced and directions for future work to address these challenges.

Acknowledgements

This dissertation project would not have been possible without the support I have received from my advisors at Portland State. I cannot express how grateful I am for my advisor, Jennifer Noll. Despite your numerous responsibilities and having no other ties remaining to Portland State, you have set aside so much time for me and my growth as a researcher and educator. My teaching philosophy has profoundly changed in my time working with you and teaching your course. Even with the many speedbumps along the way, I would not have chosen any other path to where I am today. I am also extremely grateful to Eva Thanheiser, who has probably endured more statistics than she would have liked for her whole career. Thank you for carrying the torch for Jen in being my advisor at Portland State and always being available to support me in this process.

I would also like to thank my fellow colleagues and graduate students for their support in my journey. Thank you to Jason Dolor and Dana Kirin for your support in the teaching and researching of the CATALST course; you both have pushed me to think about my teaching and statistical thinking in many ways, and have shaped me throughout this process. I am also grateful for Jack Miller and Brenda Gunderson for encouraging me to embark on this journey while I was a master's student. Thank you to the graduate students in the Math Education program at Portland State who have supported me in various ways, and have challenged me to grow and evolve as an academic.

Finally, I would like to thank my friends and family for their support in this process. I could not have made it through this without the support of my wife, Lauren. She experienced all the highs and lows of this process with me, provided reassurance and

comfort in all of the most stressful times, and only asked me when I would be finishing my PhD a few times over these eight years. I could not be here today either without the support of my parents, who were truly my first teachers. There are so many people who have supported my personal growth and interest in mathematics and statistics that I could not possibly list all of these people here. If this includes you, then I thank you from the bottom of my heart; you have shaped the work I have done and I hope to continue as an educator.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

"Proximity to freeways increases autism risk, study finds." This headline, published in the Los Angeles Times, appears to be claiming a cause-and-effect relationship. However, the study's researchers were quoted directly in that article, stating that "this study isn't saying exposure to air pollution or exposure to traffic causes autism," revealing that the study merely found a correlation which may be spurious (Roan, 2010). While we can hope that readers will think critically about what is being presented to them in the entire article, many in the age of social media do not even engage with news beyond a headline. On Twitter, 59% of links that are shared are never clicked, and most of the remining links have fewer than 1 in 1000 followers click that link. (Gabielkov et al., 2016). But even for those that do engage with the full news content, readers would need to recognize the difference between correlation and causation to be able to challenge the claim made in the headline.

This small anecdote of one news headline highlights the importance of a generally literate public. But literacy goes beyond the importance of just reading and writing – it is also vitally important to give citizens the tools to evaluate information and think critically, especially in the age of "fake news" where sources of information are often misleading and may contradict each other. Quantitative reasoning is required of readers in order to challenge these types of claims rather than accepting them at face value. This kind of reasoning is a social empowerment that literacy alone cannot provide: "mathematics should be taught so as to … enable learners to function as numerate critical citizens, able to use their knowledge in social and political realms of activity, for the

betterment of both themselves and for democratic society as a whole" (Ernest, 2015, p. 191). Many see statistics as being truly at the heart of this social empowerment that Ernest describes, with notable author H.G. Wells claiming that:

> Endless social and political problems are only accessible and only thinkable to those who have had a sound training in mathematical analysis … for complete initiation as an efficient citizen of one of the new great complex world-wide states that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write" (Wells, 1904, p. 192).

Such a claim seems especially prophetic considering the relative infancy of statistics as a field then. And as technology has evolved, views on the importance of being able to reason with data and statistics have become more centered around technology: "As information becomes ever more quantitative and as society relies on computers and the data they produce, an innumerate citizen today is as vulnerable as the illiterate peasant of [the 15th century]" (Steen, 1997, p. xv). We now live in a world of "big data," with seemingly endless amounts of statistics and information to process and analyze, many of which are conflicting and challenging to relate and internalize in totality without the proper skills.

This idea of reading, writing, and critiquing claims based in data is known as *statistical literacy*. Gal (2002) defines statistical literacy as being made up of two interrelated components: "people's ability to *interpret and critically evaluate* statistical information, data-related arguments, or stochastic phenomena" (p. 2) and "their ability to *discuss or communicate* their reactions to such statistical information, such as their understanding of the meaning of the information, their opinions about the implications of this information, or their concerns regarding the acceptability of given conclusions" (p.

3). This definition draws a parallel to general literacy, with each of the components describing one's ability to "read" statistics and "write" or communicate statistics to others. Writers of the aforementioned Los Angeles times article did not appropriately communicate their reactions to the study on autism and freeways, and thus leaves those without the ability to critically evaluate these claims susceptible to mis-information. Thus, it is not only important for the consumers of media to be statistically literate, but the producers of media should also be careful not to spread such misinformation.

Statistical literacy is central to curricular standards for statistics, with the preK-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) stating that their ultimate goal is promoting statistical literacy (Franklin et al., 2007). While the term of statistical literacy is no longer present in the college level GAISE (Carver et al., 2016), researchers argue that statistical literacy is still at the forefront of these guidelines (Schield, 2017). I argue that the most central concept to statistical literacy is statistical association, as it is key to understanding how real-world events and processes are linked together. "Knowing whether events are related, and how strongly they are related, enables individuals to explain the past, control the present, and predict the future" (Crocker, 1981, p. 272). McKenzie and Mikkelson (2007) state that covariational reasoning is one of the most important activities that humans perform. Understanding relationships and making connections between different phenomenon based on data is necessary to understand the world and how different aspects of it are connected. It is a necessary skill to have to be able to critically analyze the arguments from the Los Angeles Times article, which analyzed an association between incidence rates of autism and factors associated with proximity to freeways.

To meet the aims of addressing statistical literacy in regards to the topic of covariation and statistical association, this study investigated classroom activities designed to support students learning of statistical association through linear regression. Understanding linear regression is central to addressing the need for students to be able to process and analyze data-based claims. These activities take place in the context of the Change Agents for Teaching and Learning Statistics (CATALST) curriculum, which currently does not include content on association between two numerical variables, typically analyzed by linear regression techniques. I believe that it is important to expand this curriculum to include such a fundamental topic because it is a topic recommended for introductory statistics courses at the collegiate level (Carver et al., 2016), there are clear advantages to the use of modeling and simulation in the CATALST curriculum, and there is a clear societal benefit to understanding statistical association generally through empowering students statistical literacy.

The recognition of statistical association is just one part of covariational reasoning, which is a broad topic that lies at the intersection of fields like psychology, mathematics, science, and statistics. The key element of covariational reasoning in statistics is the use of data in multiple variables, which can be used to support or question claims of association, especially those made in the media. Unfortunately, students are not apt to question these types of claims, even when such claims are not supported with evidence like data or graphs (Watson & Moritz, 1997). Even when data are present, prior beliefs about an association often take priority in making conclusions – this is known as an *illusory correlation*, and has been shown to be a major element in the formation of stereotypes (Hamilton & Gifford, 1976). Not only is there benefit in promoting statistical

association and covariational reasoning in promoting a well-informed society, but it can also provide a benefit to society by combating the kinds of reasoning that lead to unjust stereotyping.

This study focuses on CATALST-based activities for linear regression with the goal of building on students' existing conceptions of association identified in the literature, while also upholding the modeling and simulation philosophy of the CATALST curriculum. The CATALST curriculum is rooted in the modeling of probability-based situations and using simulation-based statistical methods to draw conclusions (Garfield et al., 2012). Research has shown that simulation-based methods provide many advantages to students' learning of statistical topics, especially inferential reasoning (Chance et al., 2016, 2018; Hildreth et al., 2018; Tintle et al., 2014). The CATALST curriculum is notable among simulation-based curricula as because it allows students to create and model their own simulation processes, unlocking the supporting rationale for the statistical methods being used. This study examines potential advantages of using this simulation and modeling approach with the topic of linear regression while also drawing comparisons to more traditional curricula.

Overview of Chapters

This dissertation follows a three-paper model. In my first paper, I investigate CATALST students' strategies for informally fitting a line of best fit to scatterplots in various data scenarios. The second paper compares students from both the CATALST curriculum and a traditional curriculum on how they determine whether a data scenario yields a significant linear relationship, with a focus on if the students recognize the need

for using a hypothesis test. Finally, the last paper examines CATALST students' experiences modeling and carrying out test for the least squares line and their conceptual understanding of their probability models, the null hypothesis, and the *p*-value. This paper also presents potential technology innovations to better support students learning.

Students' Knowledge about Lines of Best Fit in a Modeling and Simulation Introductory Statistics Curriculum. Students hold various conceptions about statistical association that can interfere with learning the line of best fit. Three of these conceptions that this study focuses on are the univariate conception, the localist conception, and prior beliefs (Batenero et al., 1996; Estepa et al., 1999; Moritz, 2004). When students attempt to fit lines to data, these conceptions can lead to students potentially biasing toward lines that are upward sloping when not appropriate, fitting a line based only on a few cases in the data, or fitting a line based on their own prior knowledge of the data context. Previous work has focused on middle school students and pre-service teachers informal line fitting strategies, and has found these existing conceptions have influenced their strategies (Casey, 2015; Casey & Wasserman, 2015). This study aims to add to this literature by focusing on a novel population of college students using the CATALST curriculum. There is reason to believe that CATALST students may have success in fitting lines to scatterplots informally, not only because of the advantages of simulation-based curricula (Chance et al., 2018, 2022; Hildreth et al., 2018; Tintle et al., 2012, 2014), but also because of the potential advantages of CATALST's focus on modeling (Noll et al., 2018), which may aid in understanding the line of best fit as a model itself. To assess this hypothesis about the CATALST curriculum, I investigated the following research

question: What are CATALST students' intuitive strategies for placing lines of best fit before and after formally learning about least squares criterion?

Analysis of the data leveraged the coding scheme from the aforementioned studies on informal line fitting (Casey, 2015; Casey & Wasserman, 2015), with additional codes added to reflect the strategies that emerged in the data collected. In many cases, students' strategies still often reflected many of these known conceptions identified in the literature, which indicates more work needs to be done to improve instruction on informal line fitting. One novel strategy that emerged among CATALST students in the interview was the use of offsetting distances, where students attempted to group all the data into sets where their residuals appeared to sum to 0 visually. This strategy, which is a necessary but not sufficient condition of the least squares criterion, may be an approachable way to teaching students informal line fitting that aligns with the concepts of least squares. One additional result that emerged from students' informal line fitting was regarding the impact of outliers. Students seemed to account well for outliers that appeared in the corner of graphs, but did not account for outliers that had large residuals but appeared in the middle of the graph.

Comparing Student Outcomes on Testing for a Statistical Association for Traditional and Simulation-Based Curricula. Cobb's (2007) call for reforming the introductory statistics course to emphasize conceptual pillars of inference and leverage modern technology has brought about the rise of simulation-based inference courses. Numerous studies comparing student outcomes across traditional and simulation-based curricula have yielded many benefits, especially on tasks regarding the purpose and

interpretation of inferential techniques (Chance et al., 2018, 2022; Hildreth et al., 2018; Tintle et al., 2012, 2014). This study adds to this wealth of comparison literature by focusing on models and modeling in two ways: first, by studying students who used the modeling-focused CATALST curriculum, and by examining their understanding of hypothesis testing as it pertains to linear regression models. In the introductory statistics course, the unit on linear regression typically is taught with many various diagnostic and descriptive measures, such as correlation, r-squared, residual standard error, and the slope/intercept of the least squares line, all of which can be used to evaluate the relevance of a linear relationship. Given that in the traditional classroom that the procedures for hypothesis testing or any of these descriptive measures are often all reliant on using technology to produce statistical output, I hypothesis that students may have trouble making distinctions in their purposes and interpretations. In the CATALST curriculum, the methods for conducting a hypothesis test are quite distinct from descriptive statistics, as students are responsible for constructing a probability model for carrying out a simulation as well as building up a sampling distribution to find the p-value. This may give them a stronger conceptual understanding of inferential techniques in linear regression. To test this hypothesis, my study aims to answer the following research question: do students from a traditional curriculum and the CATALST curriculum recognize the need to use a hypothesis test for evaluating the statistical significance of a linear relationship? How do students' approaches compare across these two curricula?

Students participated in pre/post-surveys during their introductory statistics course that asked them to describe how they would carry out a hypothesis test for specific data context on linear regression. Selected students were invited to participate in interviews

where they were asked to carry out the hypothesis test and draw conclusions based on their results. Analysis of the survey responses began with an open coding procedure to determine interesting features of students' strategies for determining a significant linear relationship. These codes were refined into a coding scheme that categorized students' responses as reflecting a hypothesis test, descriptive statistics, or non-statistical method, with further codes in each category to capture the detail of their response. Data from one interviewed student from each curriculum were also analyzed to provide a more detailed look at two students with similar survey responses. Results from the study revealed that CATALST students not only made greater gains from the pre-survey to post-survey, but CATALST students were often more prepared to describe a hypothesis test before formally learning this content. Interview data also revealed that CATALST students also were more apt in determining the difference in purpose between the correlation value and a hypothesis test, and often exhibited greater conceptual understanding of hypothesis testing. These results have implications for the teaching of linear regression and distinguishing the purpose and interpretation of measures like correlation from inferential techniques. It also raises questions about how CATALST may compare to other simulation-based curricula that do not emphasize modeling.

Evidence for Further Development of TinkerPlots to Support Inferential Reasoning with Linear Regression. The CATALST curriculum and TinkerPlots software provide students with a fertile modeling environment for expressing their statistical reasoning. This environment is powerful in providing a true transparent experience of simulating from probability models that allows students to have full ownership of the process. However, the original design of the CATALST curriculum

does not cover all topics covered in a typical introductory statistics course, such as linear regression. I designed CATALST-inspired activities for linear regression that leverage TinkerPlots, and detail the clumsy workarounds required to use TinkerPlots in this way. Suggestions for future improvements to TinkerPlots to avoid this workaround are provided. These suggested improvements are based on research-based recommendations for the choice of simulation-based software (Rossman & Chance, 2014) as well as empirical results of students understanding of inference who learned using these TinkerPlots activities and the workaround. These empirical results assessed students' understanding of hypothesis testing in linear regression through the analysis of classroom assessments. This analysis aimed to answer the following research question: How does using TinkerPlots for conducting a hypothesis test for the least squares line aid students' inferential reasoning and address common challenges faced when using simulation?

Analysis of the data leveraged Case and Jacobbe's (2018) framework on the common difficulties students experience when interpreting simulation-based inference techniques, as well as work on connecting study design to the interpretations of hypothesis testing, especially with experiment-to-causation inference (Pfannkuch et al., 2015). These two works provided three areas of focus in analyzing students' assessment work: how they connect the null hypothesis to their sampler model in TinkerPlots, how they determine their choice of replacement and connect this to the study design, and how they interpret their p-value and the results of the hypothesis test. Results showed that an overwhelming majority of the CATALST students provided responses that exhibited an understanding of the null hypothesis and their p-values. There were some students that did not recognize the appropriate study design in their choice of replacement, but these

students still often provided reasoning consistent with their choice. These results provide

evidence for the success of emphasizing modeling in simulation-based curricula using

TinkerPlots, and add support for the development of further technology innovations with

TinkerPlots to improve student experiences with linear regression and other typical

introductory statistics topics.

Discussion

These papers add to statistics education literature by expanding the existing use of

the CATALST curriculum to linear regression. They detail the benefits of this expansion,

especially with regards to inference and hypothesis testing, while highlighting potential

improvements that can be made in students' strategies for informally fitting lines of best

fit. The following three chapters present these three studies and are followed up by a

concluding chapter that discusses and synthesizes the relevance of this work.

References

Batenero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, *27*, 151–169.

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Roswell, G., Velleman, P., Witmer, J., & Wood, B. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association.

Case, C., & Jacobbe, T. (2018). A framework to characterize student difficulties in learning inference from a simulation-based approach. *Statistics Education Research Journal*, *17*(2), 9–29.

Casey, S. A. (2015). Examining student conceptions of covariation: A focus on the line of best fit. *Journal of Statistics Education*, *23*(1).

Casey, S. A., & Wasserman, N. H. (2015). Teachers' knowledge about informal line of best fit. *Statistics Education Research Journal*, *14*(1).

Chance, B., Mendoza, S., & Tintle, N. (2018). Student gains in conceptual understanding in introductory statistics with and without a curriculum focused on simulation-based inference. *Proceedings of the 10th International Conference on Teaching Statistics*. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3B2.pdf

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centered on simulation-based inference. *Statistics Education Research Journal*, *21*(3), Article 3. https://doi.org/10.52041/serj.v21i3.6

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education*, *24*(3), 114–126. https://doi.org/10.1080/10691898.2016.1223529

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin*, *90*(2), 272–292. http://dx.doi.org/10.1037/0033-2909.90.2.272

Ernest, P. (2015). The social outcomes of learning mathematics: Standard, unintended or visionary? *International Journal of Education in Mathematics Science and Technology*, *3*(3), 187–192.

Estepa, A., Batanero, C., & Sanchez, F. T. (1999). Students' Intuitive Strategies in Judging Association When Comparing Two Samples. *Hiroshima Journal of Mathematics Education*, *7*, 17–30.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report*. American Statistical Association.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016, June 14). *Social Clicks: What and Who Gets Read on Twitter?* ACM SIGMETRICS / IFIP Performance 2016. https://hal.inria.fr/hal-01281190

Gal, I. (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review*, *70*(1), 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898. https://doi.org/10.1007/s11858-012-0447-5

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*(4), 392–407. https://doi.org/10.1016/S0022-1031(76)80006-6

Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing student success and understanding in introductory statistics under consensus and simulation-based curricula. *Statistics Education Research Journal*, *17*(1). https://iase-web.org/documents/SERJ/SERJ17(1)_Hildreth.pdf?1526347238

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33–61. https://doi.org/10.1016/j.cogpsych.2006.04.004

Moritz, J. (2004). Reasoning about covariation. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227–255). Springer.

Noll, J., Clement, K., Dolor, J., Kirin, D., & Petersen, M. (2018). Students' use of narrative when constructing statistical models in TinkerPlots. *ZDM*, *50*(7), 1267–1280. https://doi.org/10.1007/s11858-018-0981-x

Pfannkuch, M., Budgett, S., & Arnold, P. (2015). Experiment-to-causation inference: Understanding causality in a probabilistic setting. *Reasoning about Uncertainty: Learning and Teaching Informal Inferential Reasoning*, 95–128.

Roan, S. (2010, December 16). Proximity to freeways increases autism risk, study finds. *Los Angeles Times*. https://www.latimes.com/archives/la-xpm-2010-dec-16-la-he-autism-20101217-story.html

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(4), 211–221. https://doi.org/10.1002/wics.1302

Schield, M. (2017). GAISE 2016 promotes statistical literacy. *Statistics Education Research Journal*, *16*(1), 50–54.

Steen, L. A. (1997). *Why Numbers Count: Quantitative Literacy for Tomorrow's America*. College Entrance Examination Board.

Tintle, N., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2014). *Quantitative evidence for the use of simulation and randomization in the introductory statistics course*. 9th International Conference on Teaching Statistics, Flagstaff, Arizona. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21.

Watson, J. M., & Moritz, J. B. (1997). Student analysis of variables in a media context. *Papers on Statistical Education Presented at ICME-8*, 129–147. https://www.stat.auckland.ac.nz/~iase/publications/12/Watson%20&%20Moritz.pdf

Wells, H. G. (1904). *Mankind in the Making*. C. Scribner.

Chapter 2: Students' Knowledge about Lines of Best Fit in a Modeling and Simulation Introductory Statistics Curriculum

Abstract: Students' hold a wide variety of conceptions regarding statistical association. These conceptions pose challenges when summarizing the relationship displayed in a scatterplot through informally placing a line of best fit. This study examined college students' strategies for fitting a line to a scatterplot informally through surveys and task-based interviews. The students represented a novel population of students from a simulation-based curricula who engaged with activities specifically designed to consider informal line fitting before learning the least squares criterion. Results from this study revealed that many students leveraged a unique strategy of using offsetting distances when informally fitting lines, and that students' placement of their line of best fit revealed a differing perspective on outliers that appear in a corner of the scatterplot as opposed to the middle of the scatterplot. Students in this learning environment also exhibited reasoning reflecting the previously known conceptions of association, which has implications for future work on how to best teach students lines of best fit.

Introduction

Research has shown that simulation-based methods provide many advantages to students' learning of statistical topics (Chance et al., 2016, 2018; Hildreth et al., 2018; Tintle et al., 2014), especially in regards to inferential techniques and drawing conclusions from data. The CATALST curriculum (Zieffler, 2012) was inspired by Schoenfeld's (1998) call for mathematics curriculum to focus on teaching students how to "cook" rather than just follow recipes, and by TinkerPlots (Konold & Miller, 2018), the modeling and simulation tool that acts as the ideal statistics "kitchen." TinkerPlots

provides students opportunities beyond just the ability to simulate sampling distributions; it allows students to create the devices and models used to simulate data. The CATALST curriculum and TinkerPlots provide students with opportunities to create and model their own simulation processes. Students' narratives built around data contexts and the statistical models they build can unlock the supporting rationale for the statistical methods being used (Noll et al., 2018; Noll & Kirin, 2017).

However, the CATALST curriculum was originally designed to only focus on key concepts of inferential statistics, and does not cover every topic traditionally taught in an introductory statistics course. While this limited selection of topics was by design to narrow the focus on inference and promote statistical literacy (Justice et al., 2020), statistical association is potentially one of the most fundamental concepts of statistical literacy. Understanding relationships and making connections between different phenomenon based on data is necessary to understand the world and how different aspects of it are connected. "Knowing whether events are related, and how strongly they are related, enables individuals to explain the past, control the present, and predict the future" (Crocker, 1981, p. 272). McKenzie and Mikkelson (2007) state that covariational reasoning is one of the most important activities that humans perform. Science is rooted in understanding these relations, with Halley's (1686) observations about barometric pressure and altitude being among the first statistical associations observed that led to the study of meteorology. As society continues to face the effects of climate change, understanding relationships between global temperatures and other variables that contribute to Earth's warming is vital not just to scientists and policy makers, but also the general population who will face this warming and resulting climate impacts. An
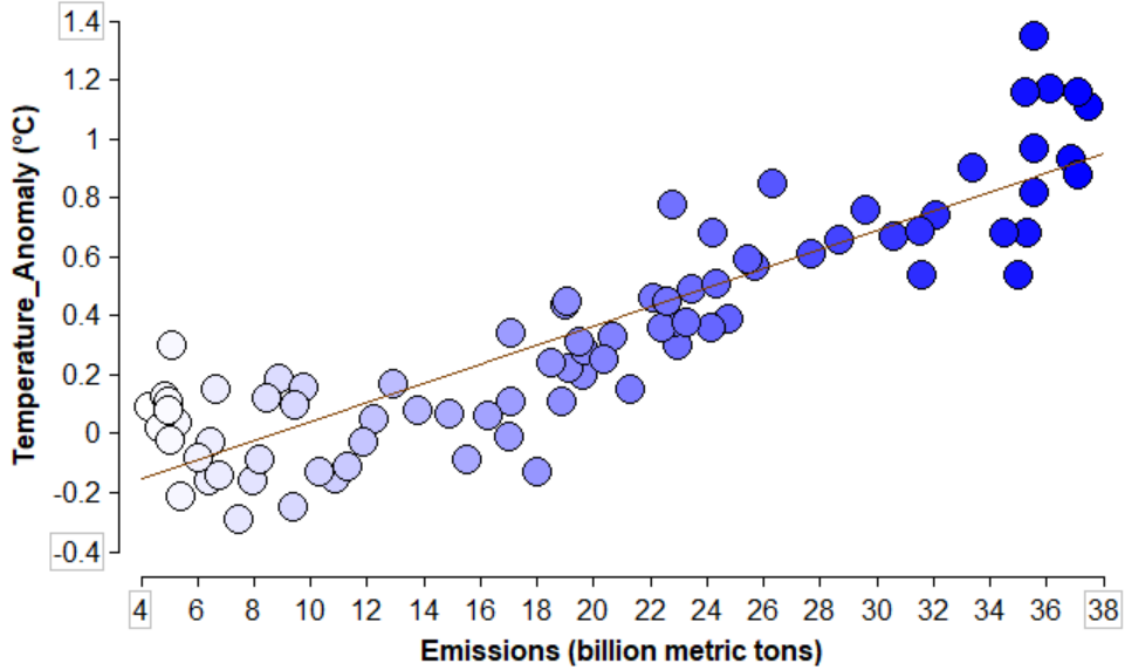
Figure 1. Scatterplot of CO_2 emissions and temperature anomaly from 1901-2000 average.

example of one such relationship can be seen in the scatterplot shown in Figure 1 relating

global CO_2 emissions to the global temperature anomaly. Understanding scatterplots is a

vital tool for both science and the general public, as they are ubiquitous. It is estimated

that at least 40 percent of data visualizations across all scientific publications relate two

or more variables, such as scatterplots (Tufte, 2001).

One tool used to summarize linear relationships is the line of best fit, seen

superimposed on the scatterplot in Figure 1. The line of best fit is the most formal topic

for statistical association in the collegiate introductory statistics class outlined by the

Guidelines for Assessment and Instruction in Statistics Education (GAISE) and is a major

focus for research on students' statistical literacy and thinking (Carver et al., 2016;

Garfield & Ben-Zvi, 2004). Exploratory bivariate data analysis is presented as an

introductory unit in AP Statistics and many other collegiate level textbooks (Agresti et al., 2017; *AP Statistics*, 2006; Utts & Heckard, 2014). Given the unique nature of CATALST as a statistics curriculum that gives students the ability to cook, not only is statistical association a natural and necessary addition to the CATALST curriculum, but there is reason to believe that this curriculum is a potentially ideal environment for students to learn this topic. Research has already identified clear advantages for the CATALST curriculum for learning statistical inference, and TinkerPlots makes for a powerful research tool for revealing students' statistical thinking. Features of this software could also provide a fertile environment for students to learn about lines of best fit and explore the criteria used to determine them through trial and error. Statistical literacy should be an art of cooking, not just reading recipes, as there is no one recipe for making judgments about statements that use any form of statistical association in the news, social media, or other various sources. Students who have this ability to "cook" can critique potentially misleading claims made based on data, and not just accept data visualizations or data-based conclusions at face value.

The goal of this study is to explore students' knowledge about lines of best fit before and after learning the material in a college level introductory statistics course. These students learned with a version of the CATALST curriculum that was expanded to introduce concepts of statistical association and linear regression. Students place lines of best fit informally on scatterplots using TinkerPlots software (Konold & Miller, 2018) before examining more formal methods that use a criterion to optimally place the line of best fit. While collegiate-level standards do not typically emphasize the use of informal line of best fit, they are recommended by Common Core State Standards in grade eight as

a way to gain foundational understanding to later learn the least squares regression line, and are also recommended at a similar level in the preK-12 GAISE standards. While fitting lines of best fit informally is not discussed in the collegiate GAISE standards, there are clear gaps in student knowledge at the preK-12 level (Biehler et al., 2018; Shaughnessy, 2007). These gaps can make learning formal concepts like least squares regression problematic when these students learn these topics at the collegiate level. GAISE guidelines explicitly spell out that it is not appropriate to teach more formal concepts to students without experiences at more foundational levels, highlighting the importance of studying informal line of best fit in collegiate statistics (Franklin et al., 2007, p. 13). Additionally, marginalized groups of students typically have less access to STEM fields like statistics through their K-12 education (Basile & Murray, 2015), which further highlights the need to address these foundational concepts to create an equitable learning experience at the collegiate level. This is especially relevant considering that this study was conducted at Portland State University, an urban institution with a high percentage of students from traditionally underrepresented groups (e.g. first-generation, ESL, black/African American, Hispanic, women). These groups of students tend to fail or drop the course in far higher numbers, with 41% of black/African American PSU students dropping the course from 2009-2013 compared to 21% of their white counterparts. Universities need to better prepare these underrepresented students to become statistically literate for a data-driven society, and curricular advancements along with research investigating their impact is necessary to improve learning for these groups of students. Focusing on informal lines of best fit can give students an environment to use their common or out-of-school knowledge, which is a productive environment for

transitioning to more formal concepts in many mathematical settings (Gueudet et al., 2016). With the implementation of statistics concepts at the K-12 level by Common Core being a relatively recent development, it will take significant time for the benefits of these curriculum changes to bear fruit. Professional development for primary and secondary teachers does not happen overnight, and the young students who benefit from learning statistics at a young age are several years away from collegiate courses.

Only a single study on informal lines of best fit with college students presently exists in the literature, which gave an informal presentation on one classroom of introductory level students' initial strategies for line fitting and suggested strategies for motivating the least squares lines with students (Sorto et al., 2011). These initial strategies were the basis for many studies with K-12 students on informal line fitting (Casey, 2015; Casey & Nagle, 2016; Casey & Wasserman, 2015), but little has been done on collegiate students since. This study aims to add to that pool of literature while also introducing a novel population in CATALST students. Looking at students' conceptions after the end of the course can aid in assessing how well the CATALST-inspired activities on regression address common conceptions students hold about statistical association and the line of best fit.

Literature Review and Research Question

To help frame the goals of this research study, I present three conceptions identified in research literature that students hold about statistical association relevant to the lines of best fit, which are outlined in the first subsection. The next subsection

provides the motivation for studying lines of best fit in the CATALST curriculum and why this may lead to successful learning outcomes for students.

Students' Conceptions of Statistical Association

Moritz (2004) describes three different types of covariation: logical, numerical, and statistical. Logical and numerical covariation involve more deterministic, mathematical forms of covariation, where logical covariation defines how the truth status of events varies the truth of other events (e.g. "not A = B"), and numerical covariation is how one quantity or variable defines a specific variation in another variable (e.g. $y = x^2$). Students are frequently exposed to these ideas of covariation through expressions and functions in their K-12 mathematics education, but it is not clear if students view this as covariation. Student images of a given function tend to be focused on a visual graph using Cartesian coordinates, which is problematic when analyzing compositions of functions such as $f(g(x))$, which require a deeper understanding of covariational reasoning (Thompson, 1994).

It is thus not surprising that students struggle with ideas of statistical association, which requires students to think not only about how two quantities or variables may change, but to think about this in a stochastic manner, where there is not necessarily a perfect relationship. Batenero et al. (1996) studied pre-university students' conceptions about statistical association with categorical data in 2x2 contingency tables and identified some conceptions that students have when thinking about covariational reasoning that interfere with analyzing this kind of data appropriately. These conceptions include the unidirectional and localist conceptions, which I detail in the subsections that follow. The

final subsection explores research that shows the importance of considering students' prior beliefs about the data.

Unidirectional Conception of Association. Historically, it was not immediately obvious to statisticians that statistical association could be summarized in two directions. When Francis Galton first defined the idea of a numeric value for correlation (now known as the Pearson correlation coefficient), he made the following statement defining the idea of correlation to the Royal Society on December 5th, 1889: "Two variable organs are said to be correlated when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction" (Pearson, 1920, p. 39). Galton had not considered the possibility of a negative correlation when defining the measure, and it turns out that students often do not consider this possibility either.

Batenero et al. (1996) noticed the unidirectional conception of association when students interpreted data regarding a drug's effect on reducing digestive troubles. The table showed a relationship between the drug and reducing digestive troubles, but students recognized these two variables as independent or showing no association rather than the inverse relationship that was presented. Cognitive psychologists studied association with similar tasks on adults, and found that adults reason poorly about covariation when the presence of one variable tends to correspond with the absence of another (Beyth-Marom, 1982). Moritz (2004) gave a task to students from third to ninth grade that involved producing a graph of students' spelling test scores based on the statement "People who studied for more time got lower scores." This statement would describe a negative association, but intuition about studying and test scores would

suggest a positive association. Many of the students produced graphs that showed a positive association in spite of the data that were described in the task – it's unclear if this was due to the influence of prior beliefs (a topic to be discussed later) or this unidirectional conception of association, but it is very possible that both played an influence.

When focusing on this conception as it relates to the line of best fit, students often have difficulty in placing lines of best fit for data that have a flat or negative slope. Casey (2015) noted this when studying 8th graders' ability to place an informal line of best fit on scatterplots. While students tended to have somewhat less accuracy in placing lines on negatively sloped data, this struggle was incredibly prominent with data that exhibited no association. For the task on placing lines with no association, numerous students drew informal lines through the data that were clearly upward sloping. Even many pre-service and in-service teachers who attempted this same task presented similar upward sloping lines to fit the data, with one teacher who refused to place a line on the data at all (Casey & Wasserman, 2015). Students conducting middle school science experiments struggled in situations where their data showed that the dependent variable did not covary with the independent variable. Multiple investigations with situations of association and no association are recommended to ensure that students are comfortable drawing conclusions in both situations, rather than just when there is an association (Kanari & Millar, 2004).

Thus, it seems that there are inherent difficulties in recognizing a negative association or a lack of association in data. One possible conjecture for the source of this

existing conception is how students' mathematics education shaped students' knowledge of slopes and rates of change. Teuscher and Reys (2010) noted that when secondary students are introduced to the ideas of slope, that concepts of *steepness* and *slope* can be often seen as the same concept, but slope carries information about a sign or direction where steepness does not. Teuscher and Rays provided a common example determining the steepness of a roof, in which it is possible to ignore the sign of the slope.

> "We must help our students understand that the slope of a line is calculated according to a particular orientation and that the sign of the slope indicates whether the line goes up or down… one way to extend this example and help students focus on the sign in addition to the steepness is to ask them to find the slope of the other side of the roof and compare it with the original" (Teuscher and Reys, p. 523).

The concepts that students build when working with ideas of non-negative rates like speed and steepness may be a source of this unidirectional conception when students transition to analyzing statistical association. But it might simply be more natural to think of positive associations before negative ones – even a statistician like Galton initially did not consider these cases.

Localist Conception of Association. Students often try to summarize a relationship between two variables by focusing on just a few data cases or just one of the variables presented. Batenero et al. (1996) attributed the idea of a localist conception to students whose strategies for analyzing contingency tables focused solely on a single cell of the table or only one relative frequency or percentage to draw their conclusions about association. For example, one student in this study claimed that there was no dependence on smoking and bronchial disease because there were a higher percentage of non-smokers in the study. Moritz (2004) also noted this tendency to focus on just one variable in tasks

designed to have students produce graphs that describe a particular relationship. In a context about studying time and test scores, students would often create graphs that focused on only one of the two variables, thus insufficiently conveying any idea of association.

This localist conception is not unique to covariational reasoning though, and is a common approach that students take when analyzing even univariate data. Bakker et al. (2004) noted this orientation as case-oriented, whereas an expert in statistics would analyze and interpret data with an aggregate perspective. This view is informative in summarizing the strategies students use for fitting lines of best fit to data presented in scatterplots, as it informs whether they see the line as relative to just a few points in the scatterplot or representative of the entire data set. Many students often attend to just a few points in a scatterplot when producing an informal line of best fit; these are typically a set of points that are nearly collinear or the two most extreme points (Casey, 2015). Outlier points are also a source of case-oriented thinking when placing a line of best fit. When fitting lines to data with outliers present, adult participants seemed to place lines that overstated the effect of outliers, even when explicitly asked to identify and disregard them when placing the line (Ciccione et al., 2022). Ideally, students should be fitting lines of best fit by placing them as close as possible to all data points simultaneously, as this represents an aggregate view of data, while also reflects the logic of fitting lines based on the ordinary least squares method. Given the complexity of managing the relationship between two variables, localist approaches seem natural, as they avoid the complexities of analyzing two variables simultaneously. Educators should consider how to best use

these initial localist conceptions to scaffold toward an understanding of covariational

reasoning.

Prior Beliefs. One last difficulty with regards to statistical association is

emphasizing the use of data for drawing conclusions about associations and avoiding

prior beliefs about the association to impact those conclusions. One study examined

abilities of the "intuitive psychologist" with respect to covariational reasoning. Jennings

and colleagues interviewed a group of Stanford undergraduates with no collegiate

statistics course, determining their ability to judge the level of association in two

scenarios. They presented both data on two categorical variables with no context, as well

as contexts based on existing studies with no data provided. These students massively

understated the actual level of association when data was provided, but quite frequently

overestimated the level of association when just provided contextual information. Their

main conclusion:

> "When no objective, immediately available, bivariate data can be examined, but
> prior theories or preconceptions can be brought to bear, the intuitive psychologist
> is apt to expect and predict covariations of considerable magnitude – often of far
> greater magnitude than are likely to have been presented by past experience or to
> be borne out by future experience" (Jennings et al., 1982, p. 224).

These results agree with previously discussed findings from Moritz (2004) in which

students created graphs that showed positive associations, reflecting their own beliefs

rather than the negative association that was conveyed in the task. Batenero et al. (1996)

found similar results when students claimed associations that matched intuition despite

the data reflecting no association. Prior beliefs about a causal relationship are also a

source of confusion in interpreting statistical association, as Estepa et al. (1999) found

that some students analyzing scatterplots would only identify association if there was a

known causal link between the variables based on their previous experience. Estepa & Sánchez Cobo (2001) also found that some students were likely to interpret strong correlation coefficients with causal statements. This causal conception of association is a common fallacy of statistical associations necessitating a causal link, summarized by the statistics instructor's mantra of "correlation does not imply causation." Working with concepts like critical inference or mapping relationships between variables with causal diagrams may help students to manage their prior beliefs with data-based assumptions about these confounding relationships. This gives students the ability to know when a causal link can be drawn based on data given that potential confounding sources are controlled (Cummiskey et al., 2020).

On the one hand, it seems that prior beliefs should be avoided in these bivariate contexts. Wild and Pfannkuch warn that "whenever students have contextual knowledge about a situation… they will come up with a range of possible causal explanations with little or no prompting" (Wild & Pfannkuch, 1999, p. 238). Psychological studies on placing lines of best fit on scatterplots of contextless data seem to reflect that adults' ability to perform "mental regression" is quite strong, although the lines placed better reflect minimizing the orthogonal distance from points to the line rather than the vertical distance as done with least squares (Ciccione & Dehaene, 2021). Despite this relative success in placing lines of best fit informally on contextless data, this should not be an argument to remove statistics questions from their contexts. Contextual reasoning gives meaning to the statistics problems at hand, and real statistical problems are always entrenched in the contextual world. Gil and Ben-Zvi (2011) found that students were more engaged in problems with data contexts chosen to match student interests and

expertise. Culturally relevant teaching practices agree with this recommendation but stress that this implementation is not trivial; relevant contexts should be integrated into larger, open investigations that allow students to process and apply their knowledge in order to promote engagement and independent learning (Hammond, 2014). Place-based education is especially engaging for students conducting science experiments on-site, and is especially effective with students from underrepresented groups (PEEC, 2010; Leonard et al., 2016). Day-to-day life presents challenges that require covariational reasoning, and these situations are not absent of context. Given that students have difficulty questioning claims of association made in the media (Watson & Moritz, 1997), it is important to integrate these contextual aspects in a meaningful way for students and emphasize the use of data rather than beliefs in making conclusions. Biases in data analysis and data themselves are ever present in a data-rich society, and this presents numerous ethical and social justice issues. An example of such an issue can be seen with software like PredPol, a predictive policing method that uses data to predict where crimes will happen. But since historical crime data is collected by police forces that have targeted and heavily patrolled areas where marginalized groups often live, the source of this data is inherently based in these biases (D'ignazio & Klein, 2020). This highlights the importance for students to work with data sets in relevant contexts and think about potential sources and biases in the data collection. To address the difficulty of managing both the contextual and statistical in the classroom, Moritz (2004) suggests emphasizing to students having them temporarily set aside their beliefs about the data, and then once the covariation in the data is understood, integrate the contextual aspects and their own experiences to be able to properly question any conclusions made, or how the data were collected. On the other

hand, having students explicitly make conjectures about the data based on their beliefs and experiences and revisit them after conducting a statistical analysis may also be promising, and reflects the design principles of the CATALST curriculum (Cobb & McClain, 2004; Garfield et al., 2012).

To summarize these three conceptions, Table 1 provides a description and examples of each. These three conceptions were used as a foundation for task development and analysis in this study.

Table 1. Summary and examples of existing conceptions of statistical association.

Conception	Description	Example
Localist	Characterizing a statistical association through focusing on only a few cases or only one variable.	Informally determining the line of best fit on a scatterplot by connecting just a few nearly collinear points.
Univariate	A biased view of statistical associations toward those that are positively associated, leading to a mischaracterization of unassociated or negatively associated variables.	Informally fitting an upward sloping line of best fit on data that have little to no association.
Prior Beliefs	Determining a statistical association by the contextual details rather than the data presented.	Informally fitting an upward sloping line on data based on one's belief about the two variables, despite the actual data showing no correlation.

Background and Motivation

To address common student conceptions about covariation such as univariate, localist and prior beliefs, I created statistical association activities to align with the learning theories of the original CATALST materials. The key characteristic of the CATALST curriculum that sets it apart from other simulation-based curricula is its use of modeling probability-based situations. This is achieved through the TinkerPlots software (Konold & Miller, 2018), which gives students the ability to create "sampler" devices

using commonly understood chance devices like spinners or containers of balls akin to a lottery machine. TinkerPlots allows students to have a great deal of control and customization of the simulation process, and animates these devices in order to connect with their physical counterparts. This degree of exploration and visualization allows students to gain a deeper understanding of how simulations generate results that can be used to draw statistical conclusions. The activities that I developed leverage TinkerPlots so that students could explore informal line fitting through placing diagonal lines freely on a scatterplot. This gave students a way to propose criteria to evaluate their lines of best fit. Students could then use this criteria to calculate a global measure of distance from the line, leading to criteria like least absolute deviations. CATALST students are already familiar with the concept of using distance or differences as a measure of interest through working on guided reinvention activities that focus on concepts like the mean absolute deviation to measure variability, or taking a difference of two means or percentages to draw comparisons. Employing a similar strategy to lead students to using least absolute deviation allows students to explore ideas of fit visually and informally before moving into more formal measures for determining the best fit for a line, like least squares.

While literature does not prescribe an ideal simulation-based curriculum, there are some advantages provided to the CATALST curriculum over other simulation-based and traditional curricula, notably with the success rate of students in the CATALST course and with understanding the purpose of using simulation and randomization methods (Hildreth et al., 2018). Students' understanding of the simulation itself can likely be explained by the novel use of TinkerPlots samplers. Giving students autonomy over model construction for simulating data reveals that students hold a variety of conceptions

on the purpose of simulating data as well as how data should be simulated (Noll & Kirin, 2017). Students may not fully grasp how data are generated when simulating data in a black box environment, and thus do not easily see how the simulation can be used to carry out statistical inferences. Even in-service and pre-service teachers with statistics experience referred to the simulation process done as "magic" and hand-waved the details of this process in a sequence of MEAs designed to conduct inference with simulation, leading the researchers to recommend that an emphasis be placed upon students creating their own models to simulate data (Lee et al., 2016). Additionally, the creation of student models often reflects narrative perspectives students hold with respect to the problem's contextual details, reinforcing their understanding of the data that is being simulated (Noll et al., 2018).

Most other simulation-based curricula that exist use an "applet"-based approach, where the design of the simulation itself is pre-constructed. One such simulation-based curriculum is based off the *Introduction to Statistical Investigations* textbook (Tintle et al., 2015), which has been analyzed in the literature heavily in comparison studies with traditional statistics curricula. While the research typically reveals many advantages to this simulation-based curricula over traditional statistics curricula, it is notable that studies that examine student performance by topic area reveal that simulation-based curricula are not significantly better than traditional curricula for bivariate data, with the pre-post gains often being larger for the traditional curricula (Tintle et al., 2011, 2012, 2014, 2018).

While the lone comparison study on CATALST did not include comparisons by topic, there is reason to believe that CATALST students may have better success with topics of bivariate data and lines of best fit. Simulation itself isn't directly relevant to informally placing lines of best fit; however, there are aspects of modeling and simulation that do apply. The line of best fit allows students to simplify noisy data into a summarized linear relationship, just as probability models often take complex random outcomes and simplify them to their most necessary aspects. CATALST students have experience with the process of building probability-based models based on certain assumptions they identify, often associated with a null hypothesis. When placing a line of best fit, students also need to identify an assumption by defining criteria that determine how well a line fits. Like the assumptions placed upon probability-based models, students need to identify criteria they see as important in fitting lines to evaluate the connection between the summarized model and the data as a whole.

In summary, I hypothesize that CATALST students may be best equipped for learning strategies for informal line fitting. First, their experience with statistical modeling and managing the assumptions made within their TinkerPlots samplers could potentially transfer for managing criteria to evaluate a line of best fit. Second, these CATALST students have already experienced guided reinvention activities that focus on using global differences as a measure, which may prepare them for the activities that are constructed for reinventing least absolute deviations. For these reasons, studying this population of students is an interesting and novel avenue for research.

In light of the difficulties students face in learning topics of statistical association, this study aims to leverage the modeling-focused features of the CATALST curriculum and line-fitting capabilities of TinkerPlots software for teaching the line of best fit in the collegiate introductory statistics classroom. The research question for this study is: What are CATALST students' intuitive strategies for placing lines of best fit before and after formally learning about least squares criterion?

Methodology

Data Collection Instruments

To answer the research questions posed, individual surveys and task-based interviews were conducted with students. First, I provide the rationale for why using instruments targeted at individual students is appropriate, and then detail the tasks that were used on each instrument.

Rationale. The CATALST curriculum leverages carefully scaffolded activities that have students work in groups to uncover statistical concepts in TinkerPlots. Within the classroom, I take a social constructivist view to learning. This view assumes that students come with many pre-conceived notions about statistical associations and the line of best fit from their own experiences. It also prepares students to be able to discuss statistical ideas with others, and critique statistical claims that are ubiquitous in today's society. However, a student's individual knowledge is just as important for evaluating these kinds of statistical claims once they complete the course. Additionally, for better or worse, most higher education institutions still assess students via grades at the individual level. To improve the achievement gaps for typically underrepresented groups previously

identified in the introductory statistics course, individual students' knowledge must be assessed and addressed.

For the purposes of assessing students' learning in this study, I am thus interested in assessing individual knowledge, representing a cognitive approach to learning. These two learning perspectives can be viewed as compatible, as social constructivism involves students shuffling between interpsychological and intrapsychological levels, where students bring their individual experiences to a social setting and center learning within a group of students. When students learn in groups, the experiences they bring to the course and the experiences they share with their classmates during the course affect their individual experiences with the activities and the data contexts. Thus, the individual instruments used in this study can still capture the results of students shared experiences in the course. The pre-survey aims to establish a baseline by capturing these out-of-course experiences of individual students and how they impact their reading and understanding of data. The post-survey and interview reflect what knowledge they constructed working with other students and their various perspectives, but again at an individual level. Thus, individual surveys and individual task-based interviews are an appropriate choice to capture students' learning considering these perspectives.

Survey. Individual surveys were administered to students both before and after learning linear regression content in the course. These pre and post surveys contained three questions related to selecting the most appropriate line of best for the data. These questions each had six choices of lines of best fit to pick from, with one of the lines being the least squares line. Students were also asked to justify their choice. These task-based

instruments on the survey were designed to elicit student conceptions about lines of best

fit as well as the conceptions that they constructed throughout the course. For more

details on the survey tasks and how the design of the tasks connects to the conceptions

identified in the literature review, see Appendix 1. A summarized list of the survey tasks

and their details is shown in Table 2. The second task listed that focused on elementary

students' shoe size and height, originally used in two previous studies (Casey, 2015;

Casey & Wasserman, 2015), is placed in the survey to potentially target students' prior

beliefs. Students often think there is an association between these two variables as there

would be among adults, but the data presented here for elementary students show no such

association.

Table 2. Summary of survey tasks on informal line-fitting.

Task	Direction	Correlation	Targeted Conception(s)
Adult age and heights	None	$r = -0.07$	Localist, Unidirectional
Child shoe size and height	None	$r = -0.03$	Localist, Unidirectional, Prior Beliefs
Athlete height and long jump distance	Positive	$r = 0.82$	Localist

Interview. The interview tasks had similar goals of targeting these existing

conceptions of statistical association. These task-based interviews featured line-fitting

tasks similar to those in the survey, but with students able to fully control the placement

of the line in TinkerPlots rather than picking from one of six choices. A summary of the

four line fitting tasks from the interviews can be seen in Table 3, with the full tasks and

protocol shown in Appendix 2. While the survey aimed to capture students' conceptions

of the line of best fit before and after learning the content in the course, the purpose of the

Table 3. Summary of interview tasks on informal line-fitting.

Task	Direction	Correlation	Targeted Conception	Outliers?
Attendance and Grades	Positive	$r = 0.66$	Localist (collinearity)	No
Ocean Temperature and Salinity	Negative	$r = -0.84$	Localist (outliers)	Yes (5)
UberEATS delivery distance and tip	None	$r = 0.05$	Univariate	No
Accidental deaths by truck and bed	Negative	$r = -0.64$	Prior Beliefs	Yes (1)

interview was to provide a richer perspective of these conceptions. The interview format

better allows for students to follow-up and provide more detail for their rationale in

placing lines of best fit. This also acts as a way to triangulate conceptions observed in the

survey data, which were often based on responses that are brief in nature.

With the open-nature of these interview tasks allowing students to freely place

their lines of best fit, the design of the task to target certain conceptions lies in the choice

of data, rather than the prescribed choices of lines in the surveys. The first task thus had

several places that would allow students to place a line that followed some collinear

points that did not closely align with the least squares line. The second task targeted

localist conceptions by having students need to fit a line to data that was mostly linear,

with a cluster of five outliers[1] that somewhat broke off from a very clear linear trend.

This aimed to determine if students would appropriately incorporate the outliers into the

---

[1] There is not a strict definition I am using to identify outliers, and instead use the term informally based on visual separation. Students may recognize other data points in these tasks as outliers rather than natural variations in the data, and may not label points as outliers that were intended to be by the task design.

placement of their line, if they would ignore the outliers completely, or if their line would be placed too heavily toward the outliers. The third UberEATS task aimed to determine how students recognized unassociated data, and if they would place a flat line or place a line representing some association not displayed in the data. Finally, the accidents task had students work with data that exhibited a purely spurious correlation, which may lead students to think the data should be unassociated based on their own beliefs.

After placing their lines, students were asked a series of follow-up questions to understand their thinking. First, students were asked "Why did you place your line in that location, and why do you think that best fits the data?" to understand generally their criteria for placing the lines. To get more specific answers for this, students would be asked "Did you use any specific criteria for placing your line?" if no criteria were provided initially. To get students to recall how they might have learned about placing lines of best fit from their CATALST course, students were asked "Do you think this reflects how the line of best fit was determined in your class? Did your class use different criteria for determining the line of best fit?" For the final task on accidents with the spurious correlation, students were asked "Does this plot indicate that more deaths from falling out of bed in a given year causes there to be fewer deaths by truck crashing into stationary objects?" in order to get them to think critically about the context and the meaning of the line, as well as challenge their understanding of the difference between correlation and causation.

Participants

This study focuses on one CATALST classroom of 23 students in the second 10-week course of an undergraduate introductory statistics sequence. Of those 23 students, 21 consented to participate in the study. This course is targeted at non-statistics majors, most of which come from a social science background. Some students in the course may have had some prior statistics knowledge from high school or other courses in their own departments, but for the most, this course is their primary exposure to statistics in college.

To encourage participation in the surveys, these were assigned as a homework assignment to students to introduce students to ideas of line fitting and capture their conjectures, and again to revisit these scenarios again after formally learning the content. Students who did not consent to research completed the assignments for the purpose of the course, but their responses were removed before analysis. Of the 21 students who agreed to participate in the research, there were 18 who participated in both pre and post surveys and are included in the analysis of this study. A subset of those 18 students were then selected to participate in interviews. The selection of interviewed students was done purposefully based on their survey responses to obtain a pool of students with a wide variety of conceptions on the line-fitting survey tasks. Thus, the interview sample is somewhat biased toward conceptions that were uncommon, and is intended to show the full range of possible student conceptions rather than be a representative sample. Eight students were invited to participate in interviews approximately 1-2 months after the course's completion, five of whom participated. These five students are referenced in this study by the pseudonyms Dabney, Dene, Garnett, Morgan, and Riley.

Analysis

Analysis of the survey responses began with the development of a coding structure for student justifications. The justifications that students gave in the survey lined up with codes used to characterize student responses during task-based interviews in Casey (2015), which served as a basis for the coding structure. However, since this previous study used task-based interviews rather than surveys, student survey answers were often brief and only justified the choice based on the apparent overall direction of the relationship. These brief responses could be investigated further in an interview setting with follow-up questions to determine their criteria or rationale for that specific choice. In this setting, to characterize these kinds of vague responses, codes for recognizing that the data held a specific association were created. The final coding structure used for these tasks is shown in Table 4. Codes were not mutually exclusive for a given response, so students could be assigned multiple codes or none at all. All student responses were coded by the author and a second coder, and all disagreements in coding were discussed until an agreement could be reached. Analysis of the student interviews began with the creation of transcripts. These transcripts were then read through for interesting discussions, with major points of interest being the criteria students used to characterize their line, how they responded to perceived outliers, how they characterize the line of best fit itself, and how their beliefs impacted the placement of the line. This process was iterative in nature, with moving back and forth between the transcripts themselves and the summaries/themes that emerged from the transcripts. Based on these observations, one additional reasoning code for offsetting distances was added for the interviews, which was verified by the second coder. Additionally, since the interviews

Table 4. Summary and examples of each code for characterizing explanations from survey tasks.

Code	Description	Examples
Prior beliefs	Characterizes the relationship based upon contextual details and previous experience, clearly not based on the given data.	"It shows that as the person gets older, they get shorter and it also covers more data on the line." "It allows you to see a trend… allowing for an easier understanding that shoe size and heigh show correlation."
Recognizes ____ (+/-/0) correlation	Describes the data as having a positive, negative, or no relationship. This may be implicit by choice of line, if the data "follow" the line or something similar.	"The line is not increasing just like the data. And the line goes in the direction the data is going." "I chose this to be the line of best fit because it goes in the same direction and incline as the dots."
Equal above and below	Describes a desirable quality of their chosen line to have equal number of points above and below it. May also reference ideas of median or characterize the line as representing the median.	"This seemed to go through the middle of the data." "I chose this line because it seems as though the data is relatively split between the higher and lower sides, so it makes sense the line is in the middle."
Closest to points	Describes how the line is the closest to the points or a closest fit, or focuses on vertical distances in a way that implies this. Could also characterize the line as representing the mean or being in between the points.	"The data means are going to hover right around that 170.0 height across the ages, giving relatively close to a zero slope." "The line pretty evenly divides the data… with similar average vertical distances between the dots and the lines."
Collinear/Localist	Describes a desirable quality of the line to go through or go near to a selection of a few data points.	"This choice seemed to connect through many of the points in the graph as well." "It includes the majority of the student's heights and distance they jumped."

provided richer responses than the surveys, the codes for recognizing the direction of the relationship were not applied. The final coding scheme applied to these tasks can be seen in Table 5. Codes were applied at the task level if that reasoning was used in finding their line of best fit for that given task. When multiple codes were selected for a task, the coders also selected one code to be the primary code that best reflects the main reasoning a student used in placing their line. As with the surveys, both the author and second coder determined their codes independently and resolved all disagreements. To aid in tying together the method of analysis and the literature, Table 6 summarizes the codes applied

Table 5. Summary of each code for characterizing explanations from interview tasks.

Code	Description	Examples
Prior beliefs	Characterizes the relationship based upon contextual details and previous experience, clearly not based on the given data.	"They have nothing to do with one another… Just looking at the scenario, I don't see why there would be any correlation whatsoever." "My mental model is that the further you drive the more tip you would get, but at the same time though, I don't know that people honestly think a lot about it."
Equal above and below	Describes a desirable quality of their chosen line to have equal number of points above and below it. May also reference ideas of median or characterize the line as representing the median.	"There's like three below here, three above, they're roughly equal number above and below here." "They were balanced on each side and almost running through the center of the dots."
Closest to points	Describes how the line is the closest to the points or a closest fit, may also characterize the line as representing the mean or being in between the points.	"[I'm] determining the distance between each line at the line and each point. The smallest distance possible, approximately, but I don't have the calculations." "You're trying to find the center of the data… trying to represent the average of the data points."
Collinear /Localist	Describes a desirable quality of the line to go through or go near to a selection of a few data points.	"When I look at information or data points always feel like they have to, at least for me, like, it's easier to understand if the line is going through data." "[If] this was the starting point… then I think I'd be trying to touch more of these dots with the line."
Offsetting Distances	Describes their line as being placed so that there are many pairs or groups of points whose residuals balance each other out.	"You want the average distance to be balanced between the two sides, the average distance to the line" "These two [distances] right here are roughly the same as these two, these two [distances] over here don't have any counterparts. But neither does this one [data point], which is a bit further away from the line."

to surveys and interviews and how they connect back to the conceptions of association identified in the literature.

Students' final lines of best fit for each task were also compiled and analyzed against the least squares line in each scenario to reveal any interesting differences. While the logic of least squares was not readily intuitive, appropriate informal methods of fitting a line should come close to this line. Examining differences between students' informal lines and the least squares line highlighted some interesting differences in how students dealt with outliers, which led to a greater focus on students' comments about outliers in future readings.

Table 6. Summary and examples of existing conceptions of statistical association.

Conception	Description	Example	Associated Code
Localist	Characterizing a statistical association through focusing on only a few cases or only one variable.	Informally determining the line of best fit on a scatterplot by connecting just a few nearly collinear points.	Collinear/Localist
Univariate	A biased view of statistical associations toward those that are positively associated, leading to a mischaracterization of unassociated or negatively associated variables.	Informally fitting an upward sloping line of best fit on data that have little to no association.	Recognizes (+/-/0) association
Prior Beliefs	Determining a statistical association by the contextual details rather than the data presented.	Informally fitting an upward sloping line on data based on one's belief about the two variables, despite the actual data showing no correlation.	Prior beliefs

Results

This section details the results from the surveys and interviews conducted in this study. After reviewing the overall results from the surveys and interviews, transcripts from the interviews are presented and have been organized into subsections by common themes.

Surveys

Overall, students generally had success in picking the choice that represented the least squares line in the survey tasks. Table 7 shows the summary tallies of students whose chosen line of best fit was the least squares line for the data. This relative success was prevalent in both the pre-survey and the post-survey; on two of the tasks, approximately 80% of the 18 students selected the correct line on both surveys. The only task that proved to be troublesome for students was the first task on adults' age and heights, whose least squares line was essentially flat. However, the task on elementary students' shoe size and height was also a flat-lined relationship, yet far more students identified the least squares line as the line of best fit when they did not on the task for adults' age and heights. There were some gains made from pre to post survey on the age

Table 7. Tallies and percentages of students who identified the least squares line.

Task	Pre-Survey	Post-Survey	Difference
Adult age and heights	10/18 (55.6%)	13/18 (72.2%)	3/18 (16.7%)
Child shoe size and height	15/18 (83.3%)	14/18 (77.8%)	-1/18 (-5.6%)
Athlete height and long jump distance	15/18 (83.3%)	15/18 (83.3%)	0/18 (0%)

Table 8. Tallies for the number of students that used a particular reasoning on a given survey task.

Code	Adults' age and heights			Elementary students' shoe size and height			Athletes height and long jump distance		
	Pre	Post	Diff	Pre	Post	Diff	Pre	Post	Diff
Prior Beliefs	1	1	0	1	1	0	0	0	0
No Correlation	8	8	0	8	11	3	0	0	0
Positive Correlation	1	0	-1	2	2	0	9	12	3
Negative Correlation	5	3	-2	0	1	1	0	0	0
Equal Above and Below	4	7	3	4	8	4	10	9	-1
Closest to Points	5	2	-3	3	4	1	2	5	3
Collinear	0	2	2	0	1	1	2	1	-1
Uncoded	2	2	0	3	0	-3	0	-3	-1

and height task, but other tasks saw little change, with one task having one fewer student picking the least squares line.

Despite this relative success in choosing the least squares line, the students' reasoning on the survey was typically not backed by detailed reasoning. A summary of the reasoning codes applied to each of the tasks is shown in Table 8. Typically, students used reasoning summarized by the equal above and below or closest to points codes on agiven task 50% of the time or less, and for some tasks, this was far less often. This type of reasoning may be infrequent due to the nature of open-ended survey questions, as students may have not shared their thought process fully in the prompt if they felt like a simple explanation like "The data seem to follow this line best" was sufficient given the other choices available. Of course, it's also possible that students actually had trouble articulating their reasoning in the prompt, or something else entirely. Reviewing the

interview data provides more depth of students' understanding of informally placing a line of best fit.

Students were generally more apt to use reasoning reflecting the equal above and below or closest to line codes on the task with an upward sloping least squares line as opposed to the two tasks with flat least squares lines. Students rarely leveraged prior beliefs in their survey responses, but when they did, the reasoning only appeared on tasks with flat least squares lines as well. This was expected with the shoe size and height context based on the potential expectation for these variables to be associated, but students did not commonly show this reasoning. When looking across the surveys from pre to post, the reasoning that students gave did not change in any meaningful way. The elementary students' shoe size and height task was the only task where students used the two ideal reasoning codes more often, especially the equal above and below code increasing by 4 students from pre to post. The other two tasks were generally mixed, with one type of reasoning decreasing and the other increasing. It's not readily clear why these shifts occurred. Shifting from only talking about the correlation or direction of the line to using reasoning like equal and above and below may reflect efficacy of the classroom intervention, but opposing shifts are curious. This may reflect the timing of the post survey being near finals week in the course, and students may have been less motivated to provide more detail that might have reflected a better picture of their reasoning. But it is also possible that these shifts are attributable to the classroom intervention too.

To investigate this at the student level, tallies for the number of students that used a particular reasoning at least once on either the pre-survey or post-survey are given in

Table 9. Tallies of students who used a particular reasoning on any of the survey items on either the pre-survey or post-survey.

Code	Pre-Survey	Post-Survey	Difference	Both Pre and Post
Prior Beliefs	2	2	0	0
Equal Above and Below	11	11	0	8
Closest to Points	5	5	0	3
Collinear	2	3	1	1

Table 9. This table also shows how many students gave this reasoning code in both their pre-survey and post-survey, in order to track if the same students were consistently using that type of reasoning. These counts reveal that overall, the number of students giving a certain type of reasoning did not change dramatically from the pre-survey to the post-survey. However, for the more troubling conceptions like leveraging prior beliefs or connecting collinear points to determine the line of best fit, all but one of the students that used this type of reasoning did it on just either the pre or post survey. Thus, some students ceased using this kind of reasoning after learning the relevant content in their course, but others began using this kind of reasoning in their post-survey only. It is not readily apparent why these opposing shifts in reasoning occurred with these students.

Interviews

As previously mentioned, students were chosen for interviews to obtain a group with a wide variety of conceptions on the line-fitting survey tasks. The five students selected and the codes for their survey responses are provided in Table 10. Among these students, all codes arose on at least one task for either the pre or post survey, which indicates a decent variability in responses. Codes like prior beliefs and collinear that represent troublesome conceptions for line fitting came only from Garnett, who also

Table 10. Codes assigned to survey responses for students selected for interviews.

Student	Adults' age and heights		Elementary students' shoe size and height		Athletes height and long jump distance	
	Pre	Post	Pre	Post	Pre	Post
Dabney	Neg. corr.	No corr., Equal above and below	No corr.	No corr., Equal above and below	Pos. corr., Equal above and below	Pos. corr., equal above and below
Dene	Neg. corr.	Neg. corr., Equal above and below	No corr.	No corr.	Pos. corr., Equal above and below	Pos. corr., equal above and below
Garnett	Uncoded	Uncoded	Uncoded	Prior beliefs, collinear	Uncoded	Pos. corr, closest to points
Morgan	Neg. corr., Prior beliefs, Equal above and below	Closest to points	No corr.	No corr., Closest to points	Pos. corr., Equal above and below	Closest to points
Riley	No corr., closest to points	Equal above and below	No corr.	Equal above and below, closest to points	Equal above and below	Closest to points

provided many responses that did not reflect reasoning captured by the coding structure, due to a lack of clarity in the response. Morgan and Riley's surveys seem to show a shift toward more expert reasoning like closest to points or equal above and below, especially with Morgan, whose reasoning was coded as closest to points and cited the least squares criterion in their reasoning throughout the post-survey. Dabney and Dene shifted to using the equal above and below reasoning more frequently on their post surveys, but did not use reasoning reflecting the closest to points code. Overall, these five students seem to meet the aims of the selection process to provide a wide variety of conceptions for informal line fitting.

On the interview, students were able to freely place their informal lines of best fit using TinkerPlots. The lines they placed for each of the four tasks are shown in Figure 2. The least squares line is also placed on each plot in red. Overall, students seemed to place
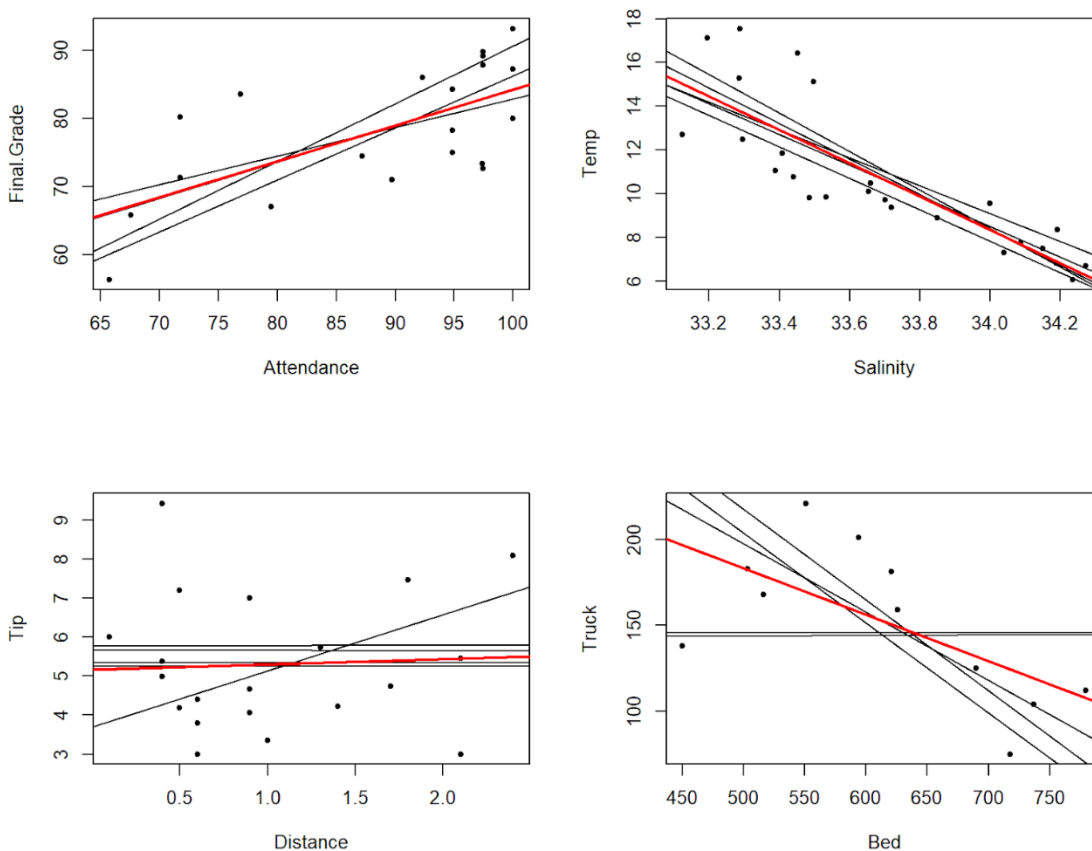
Figure 2. Students' informal lines of best fit and the least squares line for each task.
lines informally that generally matched the least squares lines, with the exception of one

upward sloping line on the UberEats tip and distance task, and two flat lines on the truck

and bed accidental deaths task. However, for students that did have a downward sloping

line on the accidental deaths task, they did seem to have a much steeper slope than the

least squares line.

The reasoning students gave for their informal lines varied from student to

student, and even task to task for many. Codes applied to students on each interview task

are shown in Table 11. In this table, each type of reasoning that appears in students'

explanations are marked with an "X". Additionally, the primary reasoning that best

characterizes how they determined their lines of best fit is highlighted in black in the

Table 11. Students assigned codes for placing their informal lines of best fit on each task.

Student	Task	Equal Ab/Bel	Closest to Points	Offsetting Distances	Prior Beliefs	Collinear/ Localist
Dabney	Grades	**X**	X			
	Ocean		**X**	X		
	UberEATS			**X**		
	Accidents			**X**		
Dene	Grades	**X**	X			
	Ocean		**X**			X
	UberEATS		**X**			
	Accidents			**X**		
Garnett	Grades					**X**
	Ocean					**X**
	UberEATS				X	**X**
	Accidents				X	**X**
Morgan	Grades	X	**X**			
	Ocean		**X**			
	UberEATS		**X**			
	Accidents		**X**			
Riley	Grades	X	X	**X**		
	Ocean	X	X	**X**		
	UberEATS	X	**X**		X	
	Accidents	X			**X**	

table. For each task, the student's code that best describes their primary or overall approach is highlighted in black. Two of the students were very consistent in their reasoning across all their tasks, where the other three students had varied justifications depending on the context presented and often by how their thinking evolved throughout the interview.

The following subsections will examine the transcripts of students working through these tasks. The first section will focus on students that used the more ideal criteria like closest to points or offsetting distances. The next subsection will look at how students dealt with the tasks that included outliers. The final subsection will examine

students whose reasoning aligns with previously-known, troubling conceptions of statistical association.

Students with Closest to Points or Offsetting Distances Reasoning. Morgan's reasoning was quite fixed throughout the interview. They heavily used closest to points reasoning, and were the only student to reference the least squares criterion in the interview. Their approach to each task was to emulate how least squares might place the line, and defined the least squares criterion appropriately: "There's a calculation where the smaller it is, the better it fits. It's the space -- If you see my pointer, from [the line] to [a data point], and then they square it and then they add them all together." Other students did leverage reasoning that tried to get the line as close to all the points as possible, but did not explicitly describe least squares in this way.

One of the codes that only emerged in the interviews was the use of offsetting distances. This is a good line of reasoning for students to leverage in their intuitive placement of the line of best fit. The condition of having the sum of residuals for your line equal to zero is at least necessary for least squares, but it is not a sufficient condition. In Dabney's interview, they first employed a strategy of having an equal number of points above and below the line for the grades task, a strategy that works well for symmetric data like the grades task has. However, this approach was forced to change for the ocean task, as there were many outliers, highlighted by the black dots in Figure 3.

Dabney: What makes this challenging is like, right, you know, if I were to move this line here, like all this data fits like beautifully on this line. And then we've got these [five outliers in black] right here. In my mind, that's going to skew our line a little bit, it's going to pull our line up.
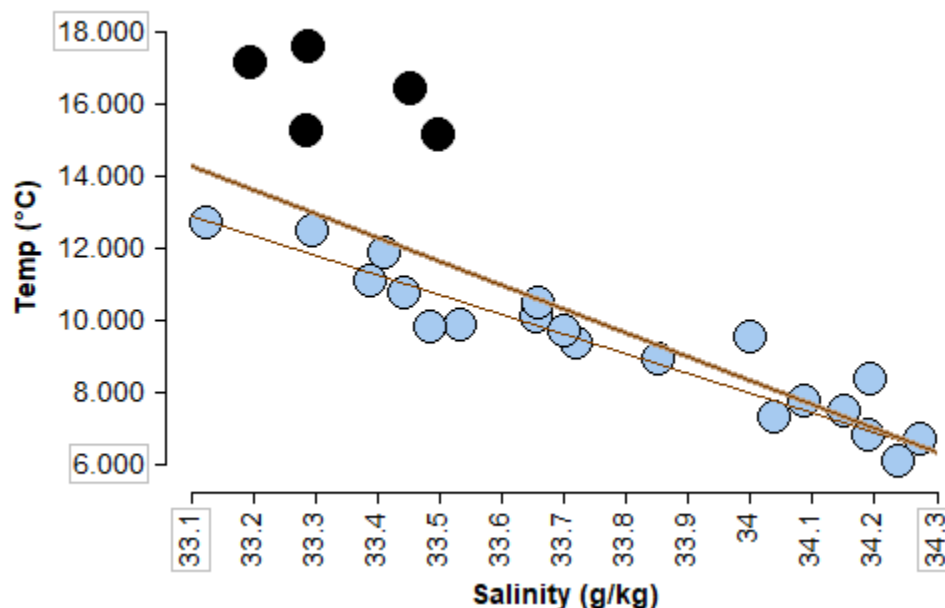
Figure 3. Dabney's proposed informal lines of best fit, with five outliers in data set highlighted in black.

…

> This is making me actually question that idea of trying to make them symmetrical, like, through their data points, because I don't think that there's really a way to do that with this data set.

…

> That makes me lean more towards using almost like the average of those data points, the average, dragging it as close to that middle.

Dabney comes to the realization that for the ocean task that using a "symmetrical" approach is not ideal, as the outliers in the upper left of the plot of the ocean task make it challenging to characterize the relationship appropriately. Thus, Dabney pivots and takes a more "average" focused approach to fit their line:

Dabney: I'm almost using like the pull of these [outliers marked in black in Figure 3] to modify the line.

…

There's very clearly a lot of data along this [thin line in Fig. 3] that I've drawn here. Yep. And then I've got these outliers here [in red]. So in order to also accommodate those, the line has to come up [to the thick line in Figure 3].

…

And really, what I'm doing is I'm trying to visually discern the weights of the value points of those data points. And so I guess by weight, in my mind, the further away from the group is more weight a data point would have… You want the average distance to be balanced between the two sides, the average distance to the line.

Dabney now describes each point as having an amount of "pull" to the line itself. This seems to be how Dabney reckoned with managing the line as representing an average of some kind. Building on this idea of the data points pulling the line, they mention having the "average distance balanced between the two sides," which was the emergence of this offsetting distance reasoning. On the tasks that followed, Dabney almost exclusively used this reasoning to justify the placement of their line of best fit, as can be seen on the UberEATS task. Figure 4 shows the UberEats task with several of the points that Dabney referenced during the interview excerpt.

Dabney: I'm not trying to cut the data points in half. I'm kind of using that same logic for like, these ones [in the rectangle in Fig. 4] definitely have more pull, because they're much higher up.

…

Maybe another way to describe it is by using [a weight] analogy. I am trying to balance it so you know like this ["100" dot in Fig. 4] pulls 100 pounds and [the "40" and "30" dots] pull, I don't know, 40 and 30 [pounds], and so on and so forth. That's the balance I'm trying to find, so that it's being pulled the same amount on both sides.

…
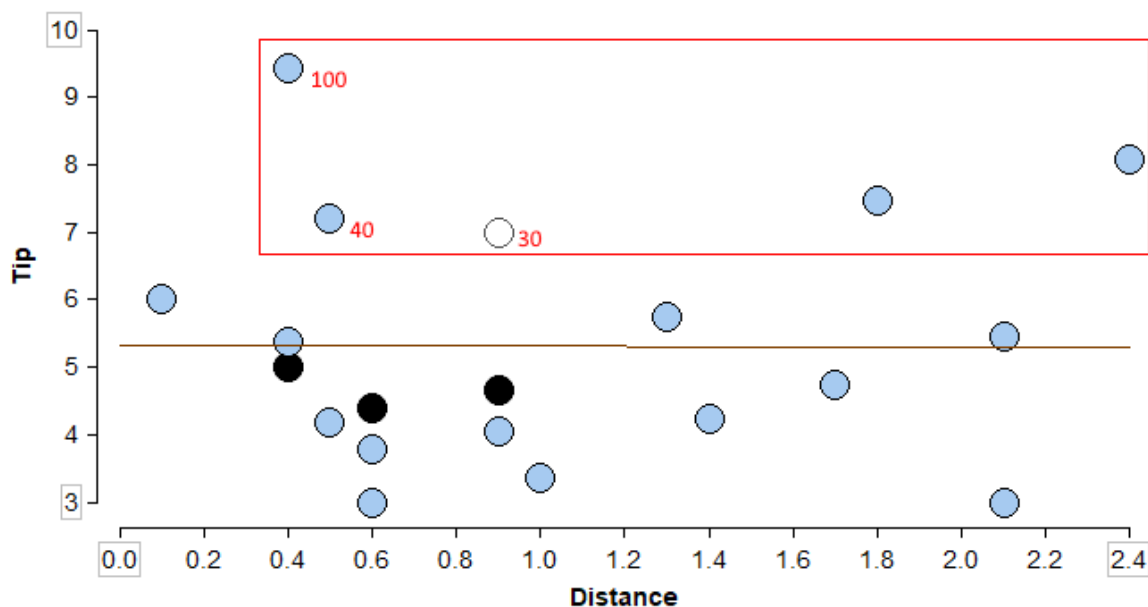
Figure 4. Plot of the UberEats task with certain dots that Dabney describes highlighted.

> Maybe I'm just reiterating the same concept, but this [white] dot here is gonna be the same weight wise as these three [black] dots here.

Unlike the previous interview segment, Dabney's reasoning here focused solely on the distances between points and if the distances on each side were balanced, rather than the idea of an average and getting as close as possible to all the points. Dabney uses informal weight measures to call this concept out, and then shows an example where dots on opposite sides of the line have equivalent weights, matching up one dot above the line with three dots below the line. Dene showed a similar trajectory of reasoning throughout their interview as well, shifting from equal above and below to closest to points and finally to offsetting distances as they moved through the tasks.

Students' Attentiveness to Outliers. Another interesting feature of the ocean and accidents tasks were their use of outliers. On the ocean task, students were very quick to notice the presence of the cluster of 5 points that were in the upper left corner of the plot.
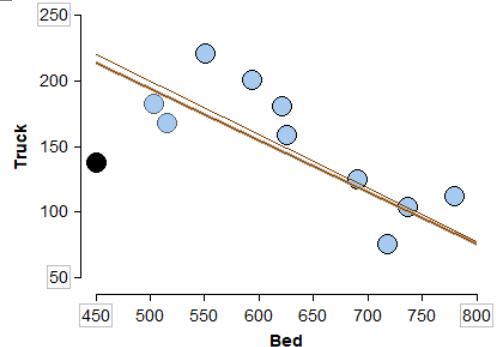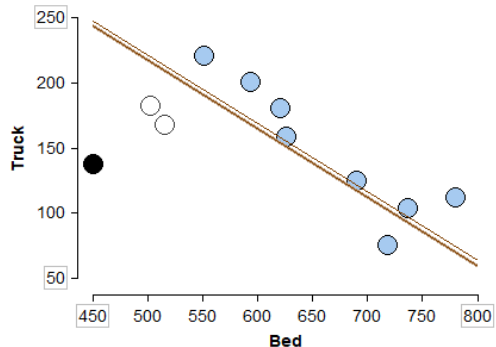
Table 12. Students' responses to outliers in the ocean task and their plotted lines.

Student	Excerpt	Plot
Dene	I don't know the correct wording for it. But [the black dots] appear [to be] pretty far out outliers. So I think because there's a lot of outliers up here, the line would be up a little bit more towards those [black dots].	
Morgan	So I think that, like, if there's a couple of points that are not really generally within the data, and they're rare, then we shouldn't follow them when determining the line of best fit. … I should probably consider them or maybe just at least two, because you know, not considering [these five black dots] is a lot. Maybe they're a little bit valid in the data, but maybe to consider them I would go like this. (adjusts line from thin line to thick line) A little bit like that.	
Riley	Even before I adjust the line, I'm noticing that you can almost draw a diagonal line [for the light blue points]. And then this is weird cluster [of black points] off to the top left that feels like it's going to influence that line. So the first thing I do is basically try to fit the line to the data, ignoring the data on the top … Now I have to take into account [the black points]. (adjusts line from thin line to thick line) So I am totally guessing at this point that it would raise the line by some number.	

In addition to Dabney recognizing this, three other students explicitly mentioned them, often as soon as they began working on the task. These three students lines and reasoning for their placement can be seen in Table 12. Morgan and Riley made significant adjustments to their line upon noticing the impact of the black outlier points. Dene did not have a similar adjustment in their interview, but did initially place their line with a very steep slope, so much that the line itself seems to touch one of the outlier points.

In contrast to this task, students were much less attentive to the outlier in the accidents task. Only two of the students made mention of this outlier, and in both cases seemed to make minimal adjustments to their line when reasoning through it. These two students' reasoning and plots of their lines can be seen in Table 13. While both students

Table 13. Students' responses to outliers in the accidents task and their plotted lines.

Student	Excerpt	Plot
Dabney	So I'm going to pull it down a little bit. This guy is a pretty big outlier… But I'm in my mind wondering how much would this [black] outlier pull this down? … I'm adjusting for this big outlier… (adjusts from thin line to thick line) and I think I'm getting pretty close to it.	
Dene	I was just trying to take into account that this [black dot] would really drag the data down, but I don't think it would drag it down so dramatically, because even these two [white dots] are pretty close, like, pretty in sync with all of the other data… (adjusts from thin line to thick line) I'll probably put it there, because I think this would have more of an effect on it than what I originally had it.	

made some adjustments to their downward sloping lines after recognizing the outlier's

impact on the line, the actual adjustments they made were minimal. Additionally, it can

be seen from Figure 2 that all three students who chose downward sloping lines were still

too steep relative to the least squares line, thus not fully accounting for the effect of that

outlier value. The third student who chose a downward sloping line, Morgan, did not

specifically call this out as an outlier at all. The other two students placed completely flat

lines instead, thus not seeing this point as an outlier from their perspective.

Students with Reasoning Reflecting Previously-Known Conceptions. There

were a few cases of students whose reasoning reflected some of the conceptions of

statistical association identified in previous literature. An example of this can be seen in

the Ocean task with the cluster of outliers. The only student that did not explicitly call out

these outliers was Garnett. As shown in Table 11, Garnett's reasoning was primarily

localist in nature, and this is in part why this was not called out specifically. Initially,

Garnett had placed a line that went directly through the cluster of points in the corner as

shown by the thin line in Figure 5. They gave this justification:

> Garnett: I feel like I focus more on the clusters towards the end of the data.
> Because I feel like that probably leads to a better explanation to
> where it kind of starts going down.
>
> … (adjusts line from the thin line in Figure 5 to the thick line)
>
> It kinda looks weird like this to me. It doesn't seem like it'll take
> all the information… not all the dots are going to be that close to it.

Garnett's initial intuition was to draw the line from corner to corner on the graph, but

made some changes after some back and forth with the interviewer. Still, Garnett is

uncomfortable with the placement of this line. The interview evolved into a discussion
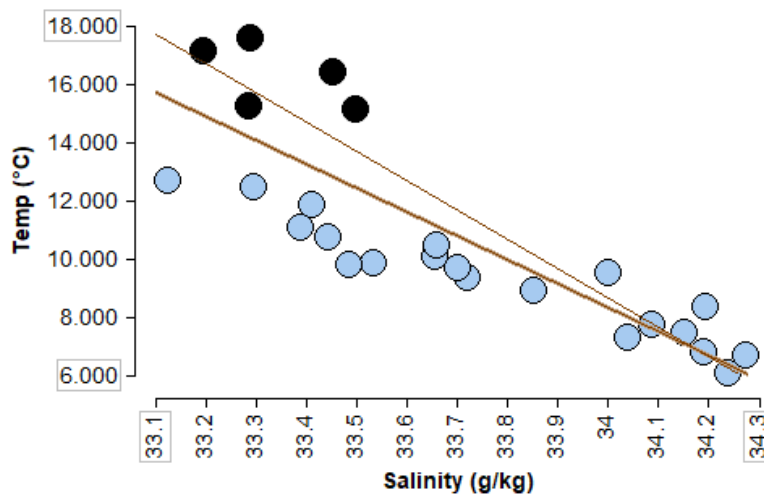
Figure 5. Garnett's initial and final line placed for the ocean task.

about Garnett's desire for having a "starting point" to place the line, and how this data

did not allow for an easy selection of a starting point.

Garnett: I think about where the placement starts. I mean, I can just assume that it's kind of like, decreasing towards the right. But if someone was also looking at this, thought it was increasing upwards. I don't know how that would work. But I think that, you know, this line kind of leads you to not put a starting point to it.

Interviewer: Can you clarify what you mean by like, putting a starting point to it?

Garnett: If I think that the starting point is [in the cluster of black points], and it's going down, I'm probably going to put [the line] up [in the middle of that cluster] towards going down. But if I'm trying to not label a starting point, but like trying to see if there's a relation with the dots not touching the line, then I think most of the time, we'll put it where it's kind of even [between the cluster of black points and the rest of the data].

…

Interviewer: Okay, so having the line go through a bunch of points is definitely an important criteria for you. But it sounds like also here, you're also weighing trying to go through the middle of the points.

Garnett: Yeah. Well, I feel like when I look at information or data points, I always feel like … it's easier to understand if the line is going through data. Just because like visually, [the five black points]

stand out to me. So if I had, let's say, in this case, if I had this line [going through the cluster], and it's kind of touching [a dot in the outlier cluster] now. But if it's kind of touching [a dot in the lower right corner], then like, it makes me interested in comparing [the two previously mentioned points] and trying to find out the relation between both data points, and what that has to do with the problem.

Garnett is having a struggle with placing their line while balancing between this idea of a "starting point" and going through a majority of the points on the graph. While Garnett placed the line for this task in an appropriate way, which balanced the outliers with the majority of the data, their initial reasoning and preferences for placing lines was heavily based on finding two critical points to connect with the line of best fit. When the resulting line did not go through a large number of points, Garnett decided to use a line that split the difference between these two overall localist ideas. Their sensemaking about placing lines of best fit seems to be centered around trying to compare dots at opposite ends of the line that the line goes through, and placing the line in the middle does not seem to give Garnett a way to interpret their line in a meaningful way.

Garnett ran into a similar situation with the accidents task in being torn between two different starting points. They were debating between drawing a mostly downward sloping line, ignoring the outlier marked in black from the plots of Table 10, or placing a completely flat line that used that outlier point as a starting point. Garnett's initial reasoning led them to place a downward sloping line, as they seemed to recognize the downward trend in this plot. However, Garnett then changes to a flat line, as the line did a good job of "separating the data" for them. This task's context would suggest that the two variables are completely unrelated, but Garnett did not make it explicit that this was their prior belief, only mentioning that the scenario was "weird" without discussing the link

between these variables. This suggests that their reasoning was more grounded in placing

the line based on a starting point and how they believe it best fits the data.

Riley was the other student to place a completely flat line on the accidents task.

However, unlike Garnett's localist reasoning, Riley's line of reasoning was firmly rooted

in prior beliefs.

Riley:	I think that there won't be any correlation between these two things? They have nothing to do with one another. … I mean, just looking at the scenario, I don't see why there would be any correlation whatsoever. … I would just placed the line as a horizontal line. If I didn't know what the data was, I would probably bifurcated this way, based upon what I've talked about before.
Interviewer:	If I had just shown you this plot, with no contextual information at all, would you have placed this line any differently?
Riley:	Either flat like this, because that sort of bifurcates the data or, again, using my ignoring the extremes at first and basically trying to divide the data in two, I would do something [downward sloping].
	…
Interviewer:	Do you feel like what you know about the situation, should that impact how you placed the line?
Riley:	Yeah, I think so. Absolutely. I mean, otherwise, I'm just like fumbling in the dark. I mean, if you if you look at like sugar and diabetes, we know in the world that has a relationship, right? … But if you look at like, beds and trucks, they that one has nothing to do with the other one apart from they both cause people to have trouble. … So it feels like you have to factor in a hypothesis when you look at this there. And the hypothesis is there is some relationship between these two things. And if you just can't believe your hypothesis in the first place, because it just seems like a random thing to say then it doesn't feel like you should be able to play that line with any sort of surety whatsoever.

Riley was adamant on not trying to use the data alone to guide their judgment on placing

what they believed the line of best fit to be, sticking to a flat line. It was not reasonable to

Riley to place any relationship between two variables whose connection seemed nonsensical. When asked how Riley would place the line if there was no context presented in this scenario, Riley did place a downward sloping line, but did not think it appropriate to hypothesize such a relationship in this context and stuck with the flat line.

The last conception of association identified by previous literature was the univariate conception of association, where students often struggle to properly identify associations that are non-positive, especially with unassociated data. The UberEATS task was the lone task that provided students an opportunity to fit a line to unassociated data. Four of the five students were successful in placing a flat line, with Garnett placing the lone upward sloping line. Their explicit reasoning was in line with the reasoning they gave on both the Ocean and Accidents tasks with connecting a line to a starting point. They also briefly referencing a belief about the context that a longer distance should result in a higher tip. It is impossible to decisively know if Garnett had an internal bias to look for an association in this unassociated data based on their reasoning, but it is reasonable to believe this could be the case given their beliefs about the situation and what is known based on previous literature.

Discussion

Students' conceptions of the line of best fit revealed by this analysis yield some promising features of what they gained from this CATALST course, but also highlight the many challenges in learning statistical association topics. This discussion will highlight three main themes: the existing conceptions from previous literature that still

persist, the use of offsetting distances as a line of reasoning, and students' approaches to tasks that feature outliers.

Existing Conceptions

While students were mostly successful in choosing an appropriate informal line of best fit in both the surveys and interviews, the reasoning that students gave for their choices did not always reflect reasoning that is consistent with how the least squares line is placed. In the surveys, students did not commonly use ideal criteria like equal above and below or closest to points. This could potentially be a result of students giving vague responses without the opportunity to follow-up, as the most common reasoning students gave was just observing the overall direction of the line as upward, downward, or flat. However, there was little change in the prevalence of equal above and below or closest to points reasoning from the pre-survey to the post-survey, which may have been expected after learning this content in the CATALST course. It may seem that students have strong beliefs based on their existing conceptions when reasoning through these tasks – do students know the criteria for lines of best fit and just experience difficulty applying it informally? While students were able to informally fit lines in TinkerPlots and measure how well it fits the data as they adjusted it, it may not be necessarily obvious or intuitive to connect these activities to the least squares criterion. Squaring distances between the dots and the line is not simple to do visually, and so students may rely on some sort of heuristic for doing this. This suggests that future work in this area should focus on how to improve students' informal line-fitting strategies, given that learning least squares alone does not seem to provide students with a rich understanding of how to visually fit a line

and justify its placement. It may be useful to emphasize other criteria for fitting lines to support students' statistical literacy; I suggest the offsetting distances criteria, which will be discussed in the next section.

Students in the surveys and interviews still exhibited known conceptions of association identified in previous research literature, primarily related to the univariate conception, localist conception, and prior beliefs. In the surveys, two of the tasks were overall relatively flat lines of best fit, yet many students' justifications for these tasks were based on recognizing some positive or negative correlation in the data. Thus, it seems that even after working with activities on informally fitting lines to scatterplots, students will still seek associations in data even if they are not present. It is also worth noting that while the use of prior beliefs was generally rare, it was more common in student reasoning in tasks with unassociated data. On the surveys, one student used prior beliefs on each of the two uncorrelated tasks in both the pre-survey and post-survey, but reasoning with prior beliefs was completely absent on the track athletes task which exhibited positive correlation. On the interviews, Garnett's informal line given on the UberEATS task was upward sloping, where they referenced their prior beliefs about the distance and tip being related to each other in their justification. It seems that when students are faced with uncorrelated data, they may leverage other sources of reasoning like prior beliefs to validate what they are seeing in the data, even if that may validate an inaccurate view of the correlation, which reflects the literature on the univariate conception and prior beliefs (Casey, 2015; Casey & Wasserman, 2015; Moritz, 2004). While beliefs about data are valuable in analyzing data, as seen with Riley objecting to the notion of the spurious correlation in the accidents task being anything meaningful,

these prior beliefs can also lead to biases in data analysis. Prior beliefs that are based in social prejudice can impact data analysis, leading to companies, policy makers, or others that hold power having a negative impact on marginalized groups. Statistical literacy should incorporate prior beliefs in such a way so that these biases can be recognized and challenged appropriately.

Offsetting Distances

Students use of offsetting distances reasoning is a new finding in this study. In previous studies, a similar form of reasoning was applied to pairs of offsetting points, exclusively when pairs of points had similar residuals but in opposite directions. In Casey & Wasserman (2015), only two teachers out of 19 used this kind of reasoning. Offsetting distances reasoning was used by 3 out of the 5 interviewed students in the present study, and was used by students in a way that expanded beyond just pairs of data points. The students that used offsetting distances typically would group one data point with a large residual with multiple points with small residuals in the opposite direction. This was exhibited with Dabney's reference to one point "pulling 100 pounds" and another pair of points pulling "30 or 40 pounds" each in the opposite direction. Thus, it seems that this conception extends to the idea that in least squares, the sum of residuals should be zero. It is possible that students gained this conception through the interactive tools they used in TinkerPlots, which gave students the ability to adjust and tweak their lines of best fit, visually seeing the residuals on the screen. While this is not a sufficient condition for least squares, it is at least a necessary condition, and thus is a reasonable and very visual strategy that students can employ that is consistent with least squares.

It is also worth noting how students offsetting distances reasoning evolved throughout the interview, with this criterion emerging out of necessity in the presence of data with outliers. Both Dabney and Dene's primary reasoning for the first task on grades data was based in placing an equal number of points above and below the line. As this task featured data that was relatively symmetric and without outliers, doing this strategy would produce a line that is in relatively close agreement with least squares; however, as the second task on ocean data introduced outliers, this necessitated a change in reasoning from both students. Both students in this task leveraged reasoning consistent with the line representing an average or being as close to all points as possible. But Dabney also used offsetting distances reasoning in this task, and leveraged this throughout the rest of the interview as their primary type of reasoning. Dene's trajectory to using offsetting distances reasoning was a bit slower in comparison to Dabney, but emerged by the final task on accidents. This developmental process that happened in two of the interviews may suggest that students leverage this kind of reasoning by the necessity principle. Both the ocean and accidents tasks feature outliers in the data, which are tricky to balance and account for when determining the line of best fit. The timing of offsetting distances emerging in each interview may suggest that students needed to leverage some other criteria in order to justify their informal line of best fit for data with these outliers present.

Considering the unintuitive nature of least squares, and how students seemed to pick up offsetting distances reasoning intuitively through the interviews, this method may be intriguing to use in the classroom when students fit lines of best fit informally. While there are many examples of inappropriate lines that have their sum of residuals equal to zero, use of this criterion along with recognizing the general direction of the trend would

likely produce a relatively strong approximation of the least squares line. This should not replace the use of more formal criterion like least squares, but it may be a more approachable method for students when learning about placing the line informally. The use of technology like TinkerPlots can then be used to interactively adjust lines while criteria like least absolute deviation or least squares updates as the line is adjusted, providing a transition to more formal methods of fitting lines to data.

Corner and Middle Outliers

Another feature that emerged in these interviews is the difference in how students handled outliers depending on where they appeared in the plot. The ocean and accidents tasks both presented students data that were negatively correlated and had outliers on the left side of the plot. For the ocean task, there were 5 outliers above the line. and in the accidents task there was one outlier below the line. This gives two different visual appearances to these values: in the ocean task, the five outliers appear in the far corner of the graph, where the single outlier in the accidents task is in the middle of the graph relative to the y-axis. This difference seemed to impact how students placed their lines of best fit. All five students placed accurate lines on the ocean task relative to the least squares line. Many students here leveraged a strategy of trying to place the line disregarding outliers first, and then adjust the line toward the outliers in a way that balances the differences between the outliers and the rest of the data. However, for the three students that did place a downward sloping line on the accidents task, all placed a line that was much steeper than the least squares line, seemingly not accounting for this outlier value appropriately. The previous transcripts revealed that two of these three

students explicitly mentioned this value as an outlier. Thus, it does seem that students are recognizing outliers in the middle of the graph, despite not accounting for its effect on the line of best fit enough. This reveals a point of emphasis in teaching lines of best fit and the impact outliers have on them, as it seems that students may not fully recognize the impacts of the middle outlier. It is important to note though that since the ocean task had 5 outliers where the accidents task only had one, the number of outliers may be playing a factor too in how students accounted for this in their informal lines. Future research that specifically focuses on students' perceptions of corner and middle outliers that removes confounding factors like this may be further informative on this topic and how it should inform teaching about outliers in this setting.

Conclusions and Future Work

On the whole, lines of best fit and understanding how to informally place one is a challenging concept for students. Even with students learning in a simulation-based curricula with interactive activities that allow for students to experience informal line fitting through the TinkerPlots software, students exhibited reasoning that reflected many already known conceptions of association. Some may believe that informally fitting lines of best fit is not a necessary skill for students that have already learned to fit lines with least squares using technology. However, I would argue that in order to be statistically literate, one should be able to read and interpret scatterplots. Being able to recognize the correlation in a scatterplot is one step to statistical literacy, but being able to visually fit a line or other statistical model to data is crucial as well to being able to process noisy data into a signal and summarize the relationship. It also leads into a deeper understanding of the line of best fit as a conditional mean function of the y-variable, and how residuals act

as the variance from this model. To that effect, there is much work to do in order for students to reason with informal lines of best fit effectively. Future work should focus on more interventions that can successfully teach students these strategies. The use of the offsetting distances criteria seems like a natural start for students given its connection to least squares and how intuitively students used it in this study. TinkerPlots already supports an interactively updating sum of residuals, which is displayed similarly to the sum of absolute residuals that students use as the criteria to place their line informally.

The results on outliers here also provides an interesting avenue for future research. Real world data is often messy with noise and outliers, and how students recognize and account for outliers in their interpretations of data is very relevant to statistical literacy. This study posits that students may not appropriately account for outliers on scatterplots that appear in the middle of the plot when assessing the statistical relationship between the two variables, even if they recognize the data points as outliers. Studies that remove more of the confounding variables that appeared in the tasks from the present study like the number of outliers could help to confirm this finding and further explore students' understanding of outliers in scatterplots.

References

Agresti, A., Franklin, C. A., & Klingenberg, B. (2017). *Statistics: The art and science of learning from data*. Pearson.

AP Statistics: The Course | AP Central – The College Board. (2006, July 10). AP Central. https://apcentral.collegeboard.org/courses/ap-statistics/course?course=ap-statistics

Bakker, A., Biehler, R., & Konold, C. (2004). Should young students learn about box plots. *Curricular Development in Statistics Education: International Association for Statistical Education*, 163–173.

Basile, V., & Murray, K. (2015). Uncovering the need for diversity among K–12 STEM educators. *Teacher Education and Practice*, *28*(2/3), 255–268.

Batenero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, *27*, 151–169.

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory & Cognition*, *10*(6), 511–519.

Biehler, R., Frischemeier, D., Reading, C., & Shaughnessy, J. M. (2018). Reasoning about data. *International Handbook of Research in Statistics Education*, 139–192.

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Roswell, G., Velleman, P., Witmer, J., & Wood, B. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association.

Casey, S. A. (2015). Examining student conceptions of covariation: A focus on the line of best fit. *Journal of Statistics Education*, *23*(1).

Casey, S. A., & Nagle, C. (2016). Students' use of slope conceptualizations when reasoning about the line of best fit. *Educational Studies in Mathematics*, *92*(2), 163–177. https://doi.org/10.1007/s10649-015-9679-y

Casey, S. A., & Wasserman, N. H. (2015). Teachers' knowledge about informal line of best fit. *Statistics Education Research Journal*, *14*(1).

Chance, B., Mendoza, S., & Tintle, N. (2018). Student gains in conceptual understanding in introductory statistics with and without a curriculum focused on simulation-based inference. *Proceedings of the 10th International Conference on Teaching Statistics*. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3B2.pdf

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education*, *24*(3), 114–126. https://doi.org/10.1080/10691898.2016.1223529

Ciccione, L., Dehaene, G., & Dehaene, S. (2022). *Outlier detection and rejection in scatterplots: Do outliers influence intuitive statistical judgments?*

Ciccione, L., & Dehaene, S. (2021). Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots. *Cognitive Psychology*, *128*, 101406.

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, 375–395.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin*, *90*(2), 272–292. http://dx.doi.org/10.1037/0033-2909.90.2.272

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

D'ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT press.

Estepa, A., Batanero, C., & Sanchez, F. T. (1999). Students' Intuitive Strategies in Judging Association When Comparing Two Samples. *Hiroshima Journal of Mathematics Education*, *7*, 17–30.

Estepa, A., & Sánchez Cobo, F. T. (2001). Empirical research on the understanding of association and implications for the training of researchers. *In C. Batanero (Ed.), Training Researchers In The Use of Statistics*, 37–51.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report*. American Statistical Association.

Garfield, J., & Ben-Zvi, D. (2004). Research on Statistical Literacy, Reasoning, and Thinking: Issues, Challenges, and Implications. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 397–409). Springer.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898. https://doi.org/10.1007/s11858-012-0447-5

Gil, E., & Ben-Zvi, D. (2011). Explanations and Context in the Emergence of Students' Informal Inferential Reasoning. *Mathematical Thinking and Learning*, *13*(1–2), 87–108. https://doi.org/10.1080/10986065.2011.538295

Gueudet, G., Bosch, M., DiSessa, A. A., Kwon, O. N., & Verschaffel, L. (2016). *Transitions in mathematics education*. Springer Nature.

Halley, E. (1686). An historical account of the trade winds, and monsoons, observable in the seas between and near the Tropicks, with an attempt to assign the physical cause of the said winds. *Philosophical Transactions of the Royal Society of London*, *16*(183), 153–168. https://doi.org/10.1098/rstl.1686.0026

Hammond, Z. L. (2014). *Culturally Responsive Teaching and The Brain: Promoting Authentic Engagement and Rigor Among Culturally and Linguistically Diverse Students* (1st edition). Thousand Oaks, CA: Corwin Press.

Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing student success and understanding in introductory statistics under consensus and simulation-based curricula. *Statistics Education Research Journal*, *17*(1). https://iase-web.org/documents/SERJ/SERJ17(1)_Hildreth.pdf?1526347238

Jennings, D., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based vs. theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211–230). Cambridge University Press.

Justice, N., Le, L., Sabbag, A., Fry, E., Ziegler, L., & Garfield, J. (2020). The CATALST Curriculum: A Story of Change. *Journal of Statistics Education*, *28*(2), 175–186. https://doi.org/10.1080/10691898.2020.1787115

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769.

Konold, C., & Miller, C. (2018). *TinkerPlots* (2.3.4). LearnTroop. http://www.tinkerplots.com

Lee, H. S., Doerr, H. M., Tran, D., & Lovett, J. N. (2016). The Role of Probability in Developing Learners' Models of Simulation Approaches to Inference. *Statistics Education Research Journal*, *15*(2), 216–238.

Leonard, J., Chamberlin, S. A., Johnson, J. B., & Verma, G. (2016). Social Justice, Place, and Equitable Science Education: Broadening Urban Students' Opportunities to Learn. *The Urban Review*, *48*(3), 355–379. https://doi.org/10.1007/s11256-016-0358-9

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33–61. https://doi.org/10.1016/j.cogpsych.2006.04.004

Moritz, J. (2004). Reasoning about covariation. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227–255). Springer.

Noll, J., Clement, K., Dolor, J., Kirin, D., & Petersen, M. (2018). Students' use of narrative when constructing statistical models in TinkerPlots. *ZDM*, *50*(7), 1267–1280. https://doi.org/10.1007/s11858-018-0981-x

Noll, J., & Kirin, D. (2017). TinkerPlots model construction approachs for comparing two groups: Student perspectives. *Statistics Education Research Journal*, *16*(2). http://iase-web.org/documents/SERJ/SERJ16%282%29_Noll.pdf

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25–45. JSTOR. https://doi.org/10.2307/2331722

Place-Based Education Evaluation Collaborative (PEEC). (2010). The benefits of place-based education: A report from the Place-based Education Evaluation Collaborative (2nd ed.).

Schoenfeld, A. H. (1998). Making mathematics and making pasta: From cookbook procedures to really cooking. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 299–319). Lawrence Erlbaum.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Information Age Editing, Inc. and NCTM.

Sorto, M. A., White, A., & Lesser, L. M. (2011). Understanding student attempts to find a line of fit. *Teaching Statistics*, *33*(2), 49–52. https://doi.org/10.1111/j.1467-9639.2010.00458.x

Teuscher, D., & Reys, R. (2010). Slope, rate of change, and steepness: Do students understand these concepts? *Mathematics Teacher*, *103*(7).

Thompson, P. W. (1994). Images of rate and operational understanding of the fundamental theorem of calculus. *Educational Studies in Mathematics*, *26*(2–3), 229–274.

Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T., & VanderStoep, J. (2015). *Introduction to statistical investigations*. Wiley Global Education.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T., & VanderStoep, J. (2018). Assessing the association between precourse metrics of student preparation and student performance in introductory statistics: Results from early data on simulation-based inference vs. nonsimulation-based inference. *Journal of Statistics Education*, *26*(2), 103–109. https://doi.org/10.1080/10691898.2018.1473061

Tintle, N., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2014). *Quantitative evidence for the use of simulation and randomization in the introductory statistics course*. 9th International Conference on Teaching Statistics, Flagstaff, Arizona. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21.

Tintle, N., VanderStoep, J., Holmes, V.-L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1). https://doi.org/10.1080/10691898.2011.11889599

Utts, J. M., & Heckard, R. F. (2014). *Mind on Statistics* (5th edition). Brooks Cole.

Watson, J. M., & Moritz, J. B. (1997). Student analysis of variables in a media context. *Papers on Statistical Education Presented at ICME-8*, 129–147. https://www.stat.auckland.ac.nz/~iase/publications/12/Watson%20&%20Moritz.pdf

Wild, C. J., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Zieffler, A. (2012). *Statistical thinking: A simulation approach to modeling uncertainty*. Catalyst Press.

Appendix A: Survey Questions

The following section details the tasks presented to the students. Comments in the tables are provided to give context to the design of the tasks and were not presented to students.

Task 1: Adult age and height

The image below shows a plot of 25 adults and their ages and heights (in centimeters).
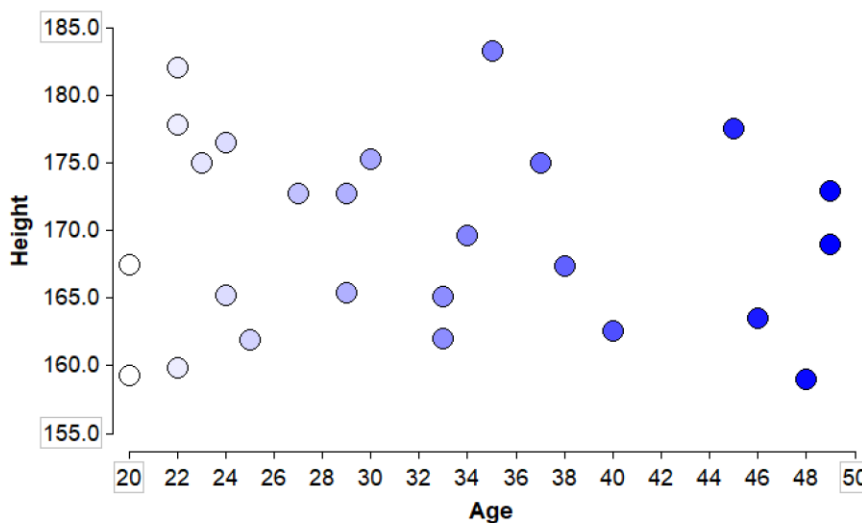
Figure 6. Age and height scatterplot as shown in survey.

The choices below all show the same plot of 25 adults' ages and heights with a line drawn over the data. Which of the six lines do you think is the line that best fits the given data?

Table 14. Answer choices for age and height scatterplot in survey.

Line Choice	Comments
	Reflects the least squares line for this data. One of two relatively flat line choices.
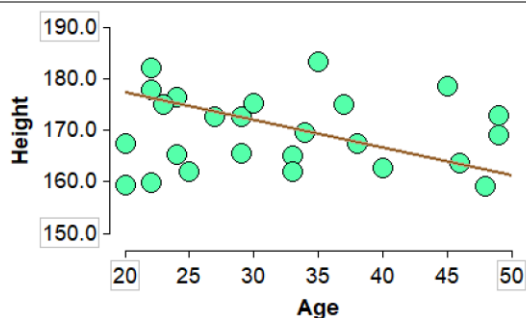
Choice goes through several collinear points, targets a localist conception of association. One of two negatively sloped line choices.
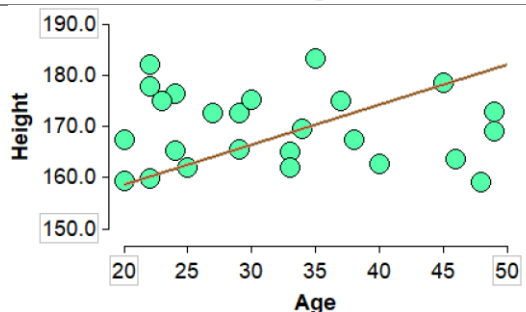
Choice goes through several near-collinear points, targets a localist conception of association. One of two positively sloped line choices.
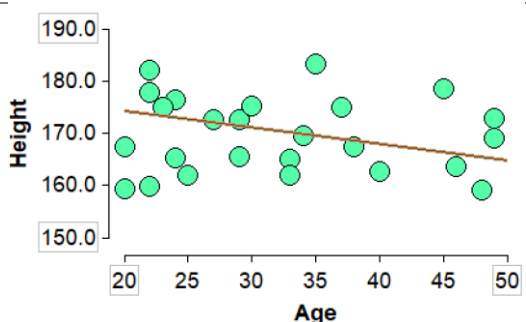
One of two negatively sloped line choices, but does not have collinear points. Not very steep though, so this choice may detect if students believe there is a slight negative association to the data.
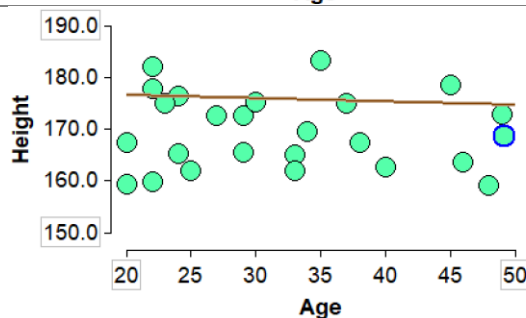
Choice goes through several near-collinear points, targets a localist conception of association. One of two relatively flat line choices.
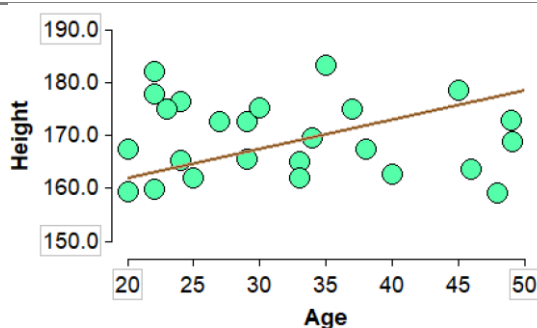
One of two positively sloped line choices, but does not have collinear points. Not very steep though, so this choice may detect if students believe there is a slight positive association to the data.

Task 2: Child shoe size and height

The image below shows a plot of 8 elementary students shoe size and their height in inches. Each circular point represents an observation for one elementary student.
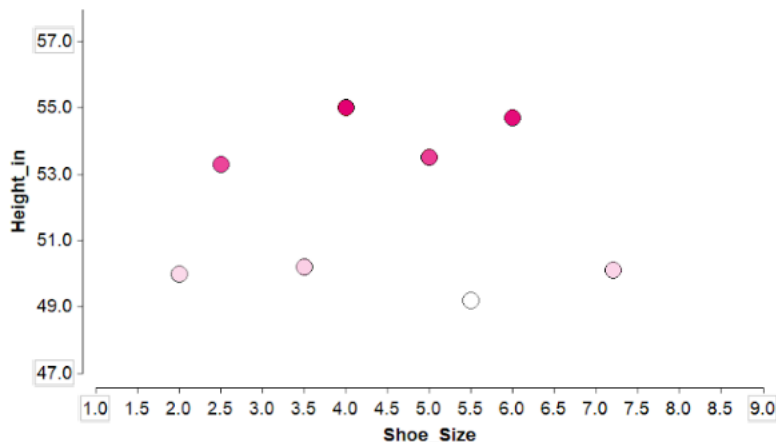
Figure 7. Shoe size and height scatterplot as shown in survey.

The choices below all show the same plot of elementary students' shoe size and height with a line drawn over the data. Which of the six lines do you think is the line that best fits the given data?

Table 15. Answer choices for shoe size and height scatterplot in survey.

Line Choice	Comments
	Choice goes through three collinear points, targets a localist conception of association. One of two negatively sloped line choices.
	Reflects the least squares line for this data. One of two relatively flat line choices.

One of two positively sloped line choices, targets a univariate conception of association without the collinear points present. Students who believe that shoe size should be related to height may also choose this based on their prior beliefs.
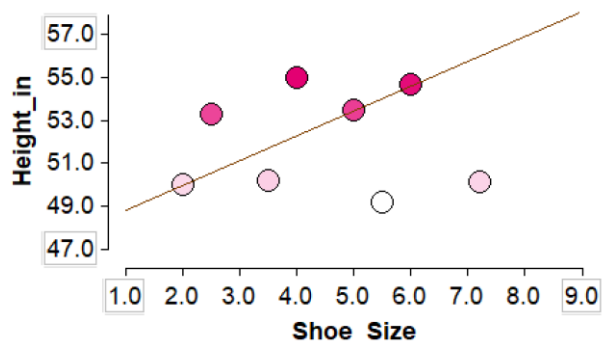
Choice goes through three collinear points, targets a localist conception of association. One of two positively sloped line choices, targets a univariate conception of association. Students who believe that shoe size should be related to height may also choose this based on their prior beliefs.
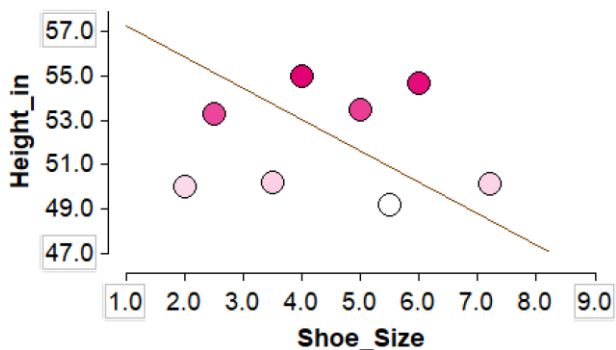
One of two negatively sloped line choices, but does not have collinear points.
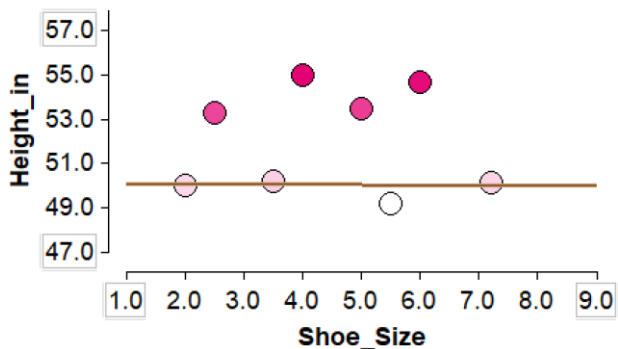
Choice goes through three collinear points, targets a localist conception of association. One of two relatively flat line choices.

Task 3: Athlete height and long jump distance

A high school track and field coach collected data on their 12 students' height and their long jump, both measured in inches. The data were organized in the plot shown below:
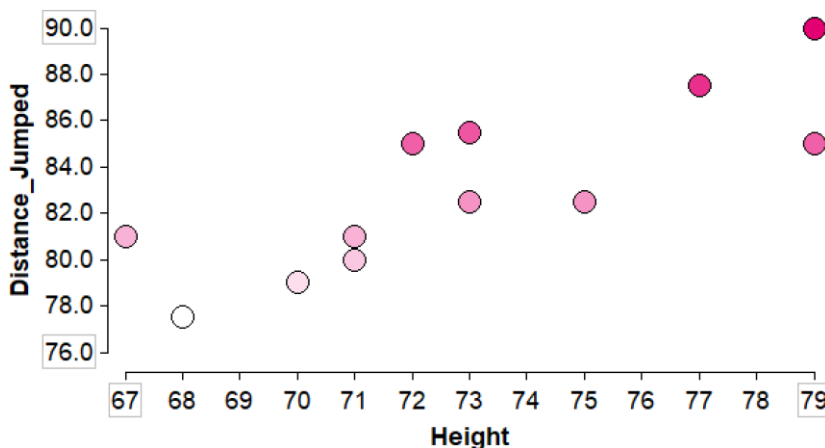
Figure 8. Height and distance scatterplot as shown in survey.

The choices below all show the same plot of these track students' height and long jump distance. Which of the six lines do you think is the line that best fits the given data?

Table 16. Answer choices for height and distance scatterplot in survey.

Line Choices	Comments
	One of three upward sloping line choices, where this choice reflects a line that is potentially too steep.
	Choice goes through three collinear points, targets a localist conception of association. The only relatively flat line choice.
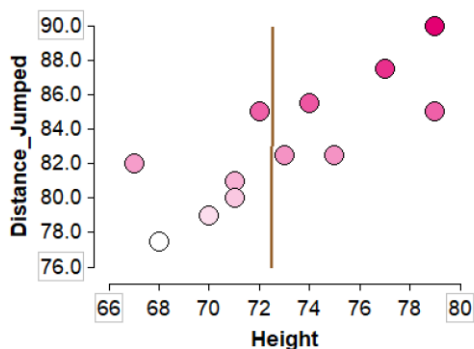
The only choice of a completely vertical line. Reflects that students may want to divide the data points into equal halves, like with the "equal above and below" code, but does so in a way that does not characterize the relationship.
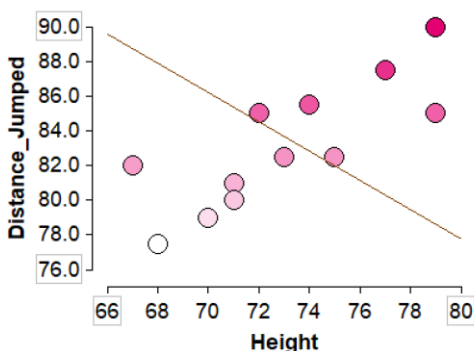
The only choice of a negatively sloped line. Reflects that students may want to divide the data points into equal halves, like with the "equal above and below" code, but does so in a way that does not characterize the relationship.
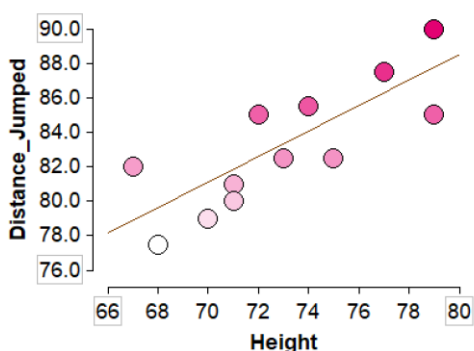
Reflects the least squares line for this data. One of three positively sloped line choices. Notable that this line does not equally divide the data above and below it.
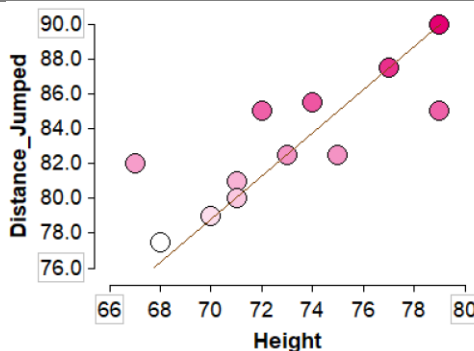

Choice goes through five near-collinear points, targets a localist conception of association. One of three positively sloped lines.

Appendix B: Interview Task Protocol

Students will be presented with scatterplots in TinkerPlots and a line tool. All students will be reminded/introduced to how to adjust the line on the plot, which is especially important for traditional students who have not used TinkerPlots before. The interviewer will open these TP files on their computer and have the student request for computer control via Zoom so they can control the line.

Four scatterplots and contexts will be presented to students:

Table 17. Details of the four line-fitting interview tasks.

Context	Scatterplot	Comments
Grades and attendance	*Scatterplot: FinalGrade vs Attendance*	Intended to be a more straightforward line placement task to gain a baseline for students understanding of how to place a line without targeting a specific conception.
Ocean temperature and salinity	*Scatterplot: Temp (°C) vs Salinity (g/kg)*	Targets a localist conception of association by having a cluster of dots separate from the rest. Students' placement of lines may aim to reflect how much they account for these dots in their line placement.
UberEATS distance and tip	*Scatterplot: Tip vs Distance*	Targets a univariate conception to determine if students recognize data with no slope. The context may also challenge students' prior beliefs if they believe a longer delivery should correspond with a higher tip.
Accidental deaths	*Scatterplot: Truck vs Bed*	Targets students prior beliefs by presenting two variables that seem completely unrelated yet display a purely spurious correlation.

For each of the four scenarios, ask students the following questions:

1. Can you explain to me why you placed your line in that location, and why you think that best fits the data?
2. Did you use any criteria for placing your line?
   o Expected criteria that students may provide:
     ▪ Through as many points as possible
     ▪ Equal number of points on both sides
     ▪ As close to all points as possible
     ▪ Reflects expected relationship based on context
     ▪ Through the first and last points (leftmost/rightmost)
3. Do you think this reflects how the line of best fit was determined in your class, or did your class use different criteria when determining the line of best fit?

Unstructured follow-up questions based on interesting features of students' responses may be asked to gain insight into their line-fitting strategies.

Additionally, for the accidents task, ask students if the plot indicates that more falling out of bed deaths causes there to be fewer deaths by truck crashing into objects.

Chapter 3: Comparing Student Outcomes on Testing for a Statistical Association

for Traditional and Simulation-Based Curricula

Abstract: Simulation-based inference has been advocated by educators and researchers for its power in helping students understand statistical inference at a deeper conceptual level. This study adds to the wealth of comparison literature by focusing on student approaches to conducting hypothesis tests for the slope of a least squares line. This study also focuses on the Change Agents for the Teaching and Learning of STatistics (CATALST) curriculum, which is unique among simulation-based curriculum for its focus on probability modeling in TinkerPlots. Students completed pre/post-survey instruments and task-based interviews to track the effectiveness of both a simulation-based curriculum and a traditional curriculum to compare their effectiveness. Results revealed that students from the simulation-based course not only showed greater progress in their learning from the classroom intervention, but were more prepared to apply inferential concepts to a novel data scenario before formally learning this content. These results have implications for teaching in emphasizing the importance of generalization in hypothesis testing and distinguishing testing from other descriptive methods in linear regression like correlation.

Introduction

For at least the past decade, the proliferation of high-powered computers has made statistics and data science more accessible. However, the introductory statistics curriculum has not caught up with the technology available, and the traditional curriculum focused on rote algebraic statistical tests still prevails as the consensus curriculum. Cobb (2007) argued that the introductory statistics course should emphasize

key inferential concepts and leveraging technology through simulation-based techniques like bootstrapping and randomization tests. Since Cobb's appeal for this shift toward simulation-based curricula, a plethora of research indicates generally positive student outcomes in these courses in comparison to the traditional curriculum (Chance et al., 2016, 2022; Hildreth et al., 2018; Tintle et al., 2012, 2014).

Statistics educators have also taken a particular focus on students' modeling techniques. Modeling itself is an essential practice of statistics and should be an important aspect of the introductory statistics class. One curriculum that gives students authentic modeling experiences is the Change Agents for Teaching and Learning STatistics (CATALST) curriculum (Garfield et al., 2012). This curriculum aims to have students explore statistics concepts by both building probability models and carrying out simulations using those models. Many simulation-based curricula focus on students working with applets that serve as prepared models to students, giving students the ability to adjust just the parameters of these models to carry out simulations. Students may gain some insight by recognizing the consequences of adjusting these model parameters, but do not get the authentic, expressive experience of building statistical models from scratch (Doerr & Pratt, 2008). This particular focus of the CATALST curriculum on modeling gives students opportunities to model real world phenomenon as statisticians do themselves, which best aligns the introductory statistics course with actual practice.

In a traditional introductory statistics course, students are typically only exposed to models through a simple linear regression model, and the typical presentation of this topic is very static in nature. The regression line or least squares line, which acts as the

model for the relationship between the two variables, is typically found through procedures or computation. There are also many other statistical concepts that accompany linear regression, including descriptive measures like the correlation or r-squared value, as well as inferential techniques such as the t-test statistics and p-values that are used to test for a significant linear relationship. Students will likely rely on some form of technology to compute these statistics associated with linear regression, both the descriptive and inferential. The processes for computing such output would be very procedural in nature, leveraging some software package like Excel, R, or SPSS.

In the CATALST classroom, there is more of a distinction in the methods used to compute descriptive measures and to carry out inferential tests. Where students in a traditional class could carry out a test for the slope of a least squares line with just a few clicks of a dialog in their respective software package, students in the CATALST classroom are engaged in a modeling and simulation process that has them engage with many different statistical aspects of the process. In this paper, I argue that the CATALST curriculum is more effective than a traditional curriculum in giving students a greater sense of the purpose and interpretation of statistical inference than students in a traditional statistics course.

When CATALST students carry out an inferential test for the slope of the least squares line, students must model the scenario appropriately under the null hypothesis in their TinkerPlots sampler. Then, they must use this sampler to produce one sample of data and determine the appropriate statistic of interest that best addresses their research question. Finally, they simulate data from their model many times to produce a sampling

distribution and find a *p*-value to draw their conclusions. Carrying out a test in this way requires students to carefully think about relevant aspects of the data, context, and statistical question they are trying to answer. This enables students to actively think like modelers rather than just follow procedures for finding results of a hypothesis test. I hypothesize that this multifaceted modeling and simulation process may help students set apart the key interpretations and conclusions drawn from significance testing and descriptive statistics like correlation. In a traditional course, calculating the correlation or carrying out a statistical test have near identical procedures: load the data, and then click the appropriate dialog, which may make distinctions in their purposes and interpretations less clear. To this effect, this study aims to investigate students from two different curricula (a CATALST-inspired introductory statistics course and a traditional introductory statistics course) and compare their approaches to carrying out a significance test for the slope of a regression line.

Background and Literature Review

This section details the relevant literature and motivation for conducting the present study. First, I will detail the importance of generalization in inference and how this should be set apart from descriptive statistics. Next, we will look at example data scenarios that highlight the relationship between correlation and inferential methods. Finally, we will highlight the importance of modeling in simulation, and how the CATALST curriculum is best posed to teach students the importance of generalization and how it conceptually differs from descriptive statistics like correlation.

Generalization

Statistical association and lines of best fit are a key component of the introductory statistics course. It is often the deepest topic in terms of both conceptual understanding and computation that students experience in their introductory course. On top of this, students must also understand the purpose of significance testing for linear relationships as a method to generalize results from a sample. Generalization of results is a key tenet of inferential reasoning, and students must be able to set apart the purposes of exploratory data analysis and significance testing (Makar & Rubin, 2009). I conjecture that this may be difficult for students in a traditional, algebra-based introductory statistics course. The procedures for both descriptive and inferential statistics in linear regression rely heavily on computers, making it more challenging to set apart their conceptual differences. The computation of p-values relies heavily on computation in any data scenario in the traditional classroom, as it relies on a calculator or computer to perform a calculus-based computation of area under a distribution function. For concepts typically taught earlier in the course like tests of means or proportions, procedures for the computations of descriptive statistics are potentially approachable to students conceptually. Students likely know how to calculate proportions or means by hand, even if software is still typically used to do this. Knowing how to calculate these descriptive statistics may set them conceptually apart from computing the p-value, which often is computed by software and may appear as a "black box" procedure to students. This idea is supported by comparison studies that show students from simulation-based courses make significantly greater improvement on test items that focus on inferential reasoning by the end of the course (Chance et al., 2022; Hildreth et al., 2018; Tintle et al., 2012, 2014).

However, when learning linear regression, both descriptive and inferential statistics act as a black box to students in a traditional class. Students typically lean heavily on software for descriptive statistics like residual standard error, correlation, and the slope/intercept for the least squares line, and are less familiar with how to calculate these measures themselves. This may lead to problematic conceptual understanding if students obscure the differences in purpose between these two methods. Correlation and p-values are often described with similar descriptors like the "strength" of results, adding to the potential conflation of their purpose. While measures like correlation may give a descriptive measure for the strength of association within a given data set by measuring how close data values are to the least squares line, examining this value alone cannot confirm a generalizable result, and may also discount potential meaningful relationships that do not have a correlation value typically seen as representing a strong relationship.

Relating correlation and p-values

To illustrate the relationship between correlation values and significance, Table 18 shows the corresponding t-test statistics at a given sample size and correlation, with the statistical significance marked by asterisks. Note that tests of correlation and tests of the slope of a line of best fit produce the same test statistic and p-value. First, values typically associated with "strong" or at least "moderately strong" correlations around 0.5 to 0.7 are not significant at the 0.05 significance level for very small samples like $n = 10$. That being said, in the age of data science and big data, most analyses are not done on very small samples. What is more notable is that statistical significance at the 0.05 level can be achieved with large sample sizes and correlation values typically thought of as a

Table 18. Values for t-test statistics for the given correlation and sample size values.

Correlation	$n = 10$	$n = 20$	$n = 30$	$n = 50$	n = 100
0.2	0.58	0.87	1.08	1.41	2.02*
0.3	0.89	1.33	1.66	2.18*	3.11**
0.4	1.23	1.85	2.31*	3.02**	4.32***
0.5	1.63	2.45*	3.06**	4.00***	5.72***
0.6	2.12	3.18**	3.97***	5.20***	7.42***
0.7	2.77*	4.16***	5.19***	6.79***	9.70***

$*p < 0.05, **p < 0.01, ***p < 0.001$

"weak" association. While large sample sizes often bring up caution about practical significance, this still highlights potential variables which could be informative to some linear model may not require correlation values normally thought of as strong. A low or "weak" correlation value may not provide an accurate prediction, but if the relationship is significant, it does inform that the typical or average value for the response variable changes with the predictor. Thus, only looking at a correlation value to determine the strength of a relationship may lead to ignoring potentially informative relationships.

To illustrate this idea, I present two example data sets to examine. The first examines the relationship between the average environmental temperature and energy used by a residence in a given summer month. Obtaining this data is not terribly difficult, especially for a utility that is already collecting usage data for billing purposes. Now consider another data scenario that examines the relationship between the weight of a hen and the weight of the eggs produced by the hen. Here, obtaining data on this requires observing and working with animals, which is more time consuming and costly, so larger samples are difficult to obtain. Thus, when examining scatterplots of each of these scenarios in Figure 9, we can see that the energy scenario has 100 observations, where the hen data only has 20. However, the correlation values are much different, due to the
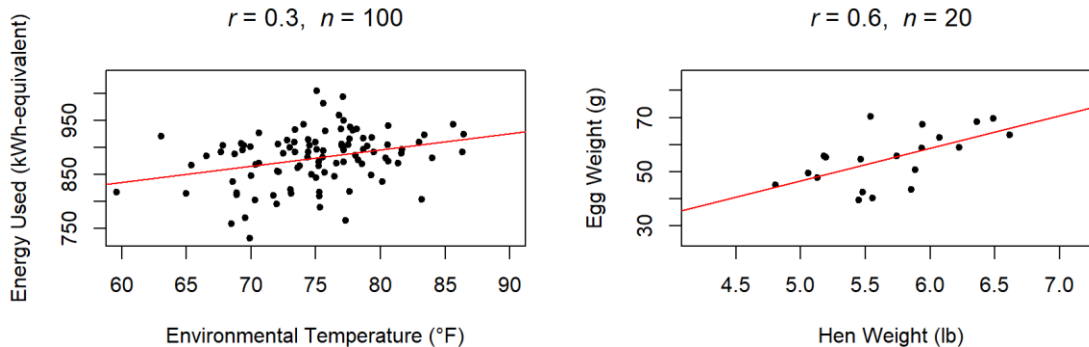
Figure 9. Example scatterplots with various sample sizes and correlations.

nature of these relationships. There are many other factors that control energy usage in a

residence, such as the size of the residence or their desired indoor temperature, so the

correlation is only 0.3. For hens, other than factors like breed which are likely already

correlated with their weight, there are not as many obvious, measurable variables that

could predict egg size, so the correlation is stronger at 0.6. When checking these

combinations of correlation and sample size in Table 18, we would find they would both

produce a *p*-value under 0.01, indicating both are significant relationships. Thus, while it

would be difficult to accurately predict one residence's bill based on the temperature

alone, as the correlation itself is low due to many outside factors, the relationship

between environmental temperature and energy usage is clearly meaningful for

determining an average energy usage based on a given temperature. The scatterplot of the

energy data set does not yield an obvious trend visually, but hypothesis testing reveals

this to be just as meaningful as the hen scenario, where the relationship is more visually

obvious and the points are tighter to the line. This also highlights how different scientific

fields may have different heuristics for what correlation values are meaningful. A

biologist may typically work with smaller data sets like this one and know heuristics for

what correlation values typically produce meaningful results for data sets of that size, but if they only look for similarly strong correlations when examining larger data sets, they may potentially miss potentially surprising links between variables. Fields that typically work with larger data sets due to the ease of collecting large amounts of data could not use these same heuristics. Having a universal scale for what determines a strong or weak correlation value is thus problematic, as it should only be presented as a way to determine the predictability of results in a given context based on how close the points are to the regression line, and not for determining the generalizability or relevance of the relationship between two variables.

Importance of Modeling

This highlights the importance of significance testing and determining generalizability of results in linear regression through inference. If students are to understand the differences in purposes and use of correlation and hypothesis testing, more distinction must be made between the two. As previously discussed, the setting in which students traditionally work with both descriptive and inferential statistics for linear regression is all procedural in nature and relies heavily on technology for computation, making all of this output indistinguishable by the setting alone. Considering these challenges students face when learning significance testing for linear regression, the ideal curriculum should emphasize conceptual understanding of these topics to help students understand the purpose of different statistical measures and methods related to linear regression. A simulation-based curriculum is one potential answer for this, as this setting has the potential to best allow students to draw connections from the study design to the

logic of hypothesis testing (G. W. Cobb, 2007; Rossman, 2008). Among simulation-based curricula, the CATALST curriculum may be the best choice to meet this purpose, as it doesn't just allow students to carry out simulations, but to create the data generation devices that carry them out using TinkerPlots software (Konold & Miller, 2018). These random generating devices are based on real-world, physical devices like lottery ball machines or spinners. This allows them to act as models for students, allowing them to deepen their understanding of the data generation process they create. Creating TinkerPlots samplers in this way is a form of expressive modeling, which best reflects the actual practice of statistical modeling (Doerr & Pratt, 2008). Students in CATALST courses are more successful in identifying the purpose of simulations than students from other traditional and even simulation-based curricula (Hildreth et al., 2018). And by using TinkerPlots models and modeling in the classroom, this also provides a rich environment for the exposition of students' statistical reasoning (Pfannkuch et al., 2018). Thus, CATALST and TinkerPlots seem to give students the ideal environment to act as modelers and understand the purpose of the simulation they are carrying out with these models. In TinkerPlots, carrying out a hypothesis test for the slope of a least squares line can be done using a randomization test, which will be discussed in the following section detailing the activities used with students.

CATALST Activities for Linear Regression

Many characteristics of the CATALST curriculum identified in the literature seem ideal for learning topics surrounding linear regression, especially with regards to hypothesis testing. However, the CATALST curriculum as originally designed does not cover this content, as it was originally designed to cover fewer topics to focus on

statistical thinking and literacy (Justice et al., 2020). The present study focuses on students from a classroom that used a curriculum based on CATALST with additional activities that cover topics traditionally taught in most introductory statistics courses, like linear regression. The activities designed for linear regression have students explore three main ideas: transitioning from analyzing univariate data distributions to bivariate data distributions, understanding how to best fit a line to data, and conducting a test on the slope of the least squares line. The following subsections will detail these activities and motivations for why these activities should best support students' learning of linear regression and the inferential techniques surrounding this topic.

From Univariate to Bivariate Data

Before students can understand bivariate data distributions like a scatterplot, they must have a solid foundation with distributions of a single variable. Zieffler and Garfield (2009) used quantitative methods to analyze students gains through testing students on items related to distributional reasoning and bivariate reasoning at several points during the course. They found that students who had made progressively larger progress on distributional reasoning test items early in the course made corresponding larger gains on the bivariate data items compared to other students, backing up the claim that fluency with univariate distributions is a precursor to understanding bivariate data.

To leverage students' knowledge of univariate data, Cobb, McClain, and Gravemeijer (2003) suggest a learning trajectory that begins with a focus on distributional reasoning with univariate data, examining the shapes of data visualized with dot plots. After developing the use of scatterplots as a visual tool for bivariate data,

this knowledge of univariate data is leveraged by "slicing" bivariate data into several

conditional univariate distributions of the dependent variable. This allows students to see

how the variable changes as the independent variable changes, leveraging the

descriptions that students used when analyzing univariate data to describe the change.

Konold (2002) suggests a similar approach, adding the use of dot plots with a color

gradient. By using colors on each of the dots to represent a second variable, students can

leverage their familiarity with univariate distributions while also grasping how it relates

to some new variable displayed with the color.

The first activity done with students integrates these ideas, with a sample of plots

made in this activity shown in Figure 10. These three plots were made in TinkerPlots

based on a data set of drivers' ages and maximum sight distance for reading road signs.

The first plot shows a univariate distribution of the ages, with a color gradient that shows

the sight distance. This visualization allows students to leverage their existing experience

with univariate distributions with a simple way to view a second variable. Building color-

coded distributions like this is possible with many software packages, but is quite simple

the relationship between two variables. The second plot reflects the "slicing" suggested
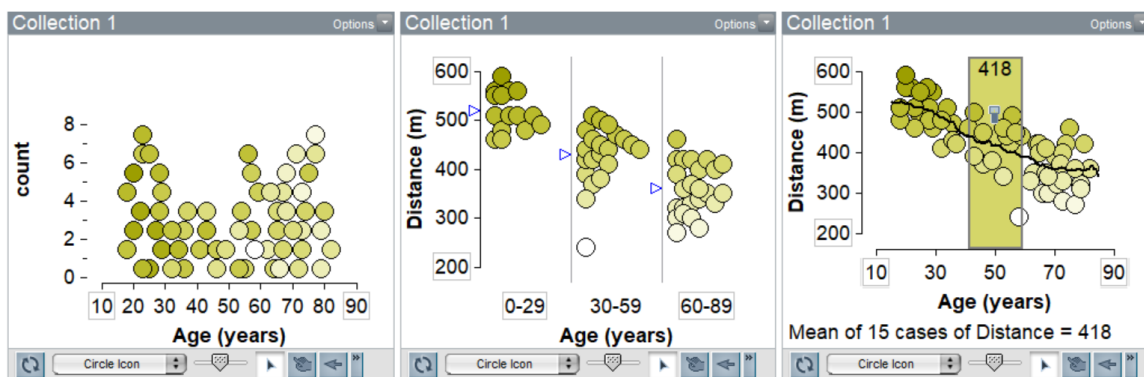
Figure 10. Plots from TinkerPlots that highlight the transition to bivariate data.

by Cobb et al. (2003), which allows students to see conditional distributions of the *y*-variable based on a range of values for the *x*-variable, while still being based on multiple univariate distributions. This allows students to see how the variable changes as the independent variable changes, leveraging the descriptions that students used when analyzing univariate data to describe the change. By seeing a plot with various conditional means, this also acts as a precursor to fitting a line to data, and understanding the line as estimating the conditional mean.

The third plot in Figure 10 is another similar precursor, which uses the "color meter" tool in TinkerPlots to trace a conditional mean line. This line is determined by the mean *y*-value of the dots within the box, tracing that value along a line as the color meter is moved. This plot can be helpful not only for the transition of the idea of center from univariate to bivariate, but to motivate the use of a straight line to characterize the shape of linearly related data, thus reinforcing a global, aggregate perspective of data.

Determining a Single Line of Best Fit

The previous activity motivated students toward summarizing data with a line, but now leaves students with the question of how to choose a line appropriately. The traditional method of fitting a line to data uses the least squares method, which minimizes the sum of the squared differences in each data point's observed *y*-value to the predicted *y*-value. This method is not simple for students to understand initially, and the motivation for why statisticians prefer squared distances over absolute distances is a nuance of calculus, which is typically not a prerequisite for introductory statistics. In fact, research indicates that using vertical distances is not an intuitive approach that students use for fitting a line informally (Sorto et al., 2011).

This activity's approach to this topic is in light of Edwards' (2005) rationale for using the median-slope algorithm, although our methodology for fitting a line differs. Edwards emphasizes the need for classroom activities to emphasize the underlying mathematical concepts behind method like linear regression and least squares, rather than hiding the results behind a black box. To this effect, these activities leverage TinkerPlots to motivate the general idea of having the line as close as possible to all points simultaneously. Students begin this activity by fitting a line informally based on where they believe it most accurately represents the relationship of the data, considering the ideas of the conditional mean and color meter tools from the previous activity. They can then measure the total distance from their line to all the points, as shown in "Sum of | Diff | of 45 cases" in Figure 11. Adjustments can then be made by the student to decrease this value. Figure 11 shows such an adjustment, with the sum of absolute deviations going from 3060.15 down to 2524.37, indicating a better fit. While TinkerPlots is restricted to measuring the absolute deviations rather than taking a typical approach of squared distances, this activity still gives students an introduction to the general idea of minimizing a criterion based on the residuals in order to produce a line of best fit.
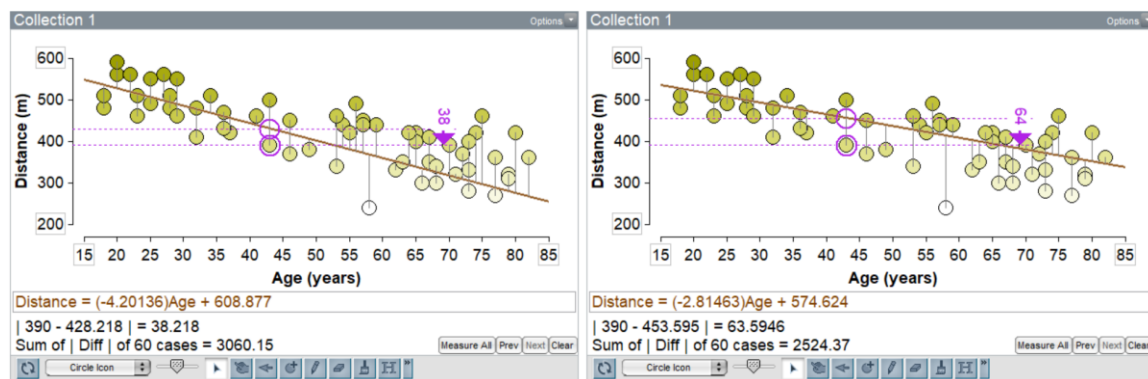
Figure 11. Plots illustrating the process for finding the line of best fit in TinkerPlots.

However, it is still best to transition these ideas to least squares, as it reflects the more widely used method for fitting lines to data, and students will use the least squares slope for testing in the next activity. To motivate students about why least squares is used in practice, one can motivate informal explanations about how squaring the distances ends up putting less emphasis on small deviations from the line. As such small deviations are expected through natural variation, a line's placement should not fixate on placing a line to further minimize distances that are already small, and are thus well explained by the relationship summarized by the line already. Additionally, showing students carefully constructed data examples that reveal that there is not always a unique line produced by minimizing the absolute deviations can help motivate the use of least squares (Lesser, 1999). These ideas parallel lessons that explore mean absolute deviation as a gateway to standard deviation, which is a commonly used activity in the CATALST classroom already. As part of this activity, students also explore other descriptive ideas surround the line of best fit, like the correlation and determination coefficients.

Testing for a Significant Linear Relationship

This final activity of the sequence will guide students in performing statistical tests on the slope of a least squares line. This requires building a sampler in TinkerPlots to conduct a simulation under the null hypothesis of no association between the two variables. To understand this null hypothesis, students will first connect the idea of no association to a line with a flat slope, and thus provide a statistical measure that can indicate relative strength of an association. This can be used as the statistic of interest for their simulation. After generating a sampling distribution of slopes, students can then

compare their slope of the least squares line from their observed data to this distribution and determine the likelihood of obtaining such a slope if the null hypothesis is true. In order to do this, students will be guided in this activity to construct sampler models in TinkerPlots that simulate data assuming that the two variables of interest are unassociated. Students will need to determine what random processes they can model in TinkerPlots if there really was no association.

Students can leverage their knowledge of inference in previous scenarios, especially those for comparing two populations or groups. In both of these scenarios, students can leverage the idea of random assignment to simulate data under the null hypothesis. The structure of the modeling process in two groups uses random assignment to re-pair values from the response or outcome variable to one of the two group or population labels. This modeling process is similar in a linear regression context, except the grouping variable becomes a quantitative explanatory variable. Thus, numerical responses are now just randomly assigned to a numerical value from the explanatory variable. Because of this similarity, students are presented with a fairly direct connection from students' past experiences with statistical inference and performing it on the slope of a regression line.

Research into students understanding of randomization tests for comparing two populations is a recent and growing area of focus. Biehler et al. (2015) investigated preservice teachers' reasoning with randomization tests, and developed a framework for the three worlds that represented how these teachers reasoned: the context world, the statistical world, and the software world. These worlds are nested within each other,

indicating that learners must be able to reason and draw connections between these three worlds in order to fully reason with these randomization tests in software like TinkerPlots, then subsequently draw conclusions in the original problem context. However, subsequent research has revealed challenges for students in properly navigating through these worlds. Noll & Kirin (2017) observed that students who created randomization models for comparing two groups moved through these worlds constantly to verify their reasoning and choices in constructing TinkerPlots models. For example, students constantly needed to re-verify that their samplers in TinkerPlots were simulating data under a hypothetical world where the null hypothesis is true, and that may not reflect what they informally observed in the sample data. This required being able to read their model appropriately in the software world, connect this to their null hypothesis in the statistical world, and realize that the real, contextual world and the data presented may not accurately reflect this world.

There are also gaps between how statisticians, students and even teachers view and understand the randomization process. Noll et al. (2021) found that students often saw a randomization process as taking a new sample of subjects rather than the process of reassigning existing subjects to new groups. This reflected their narrative views of the study, and how it would seem invalid or "unethical" to re-use the same subjects who have already participated in the study. It was also not obvious to students that a control group or taking a difference in means/proportions was necessary to properly answer the research question. Justice et al. (2018) found that preservice teachers did not have a proper view of understanding the purpose of modeling a randomization test was to understand the scope of experimental variation. These teachers also had a strong

preference for the order that their sampling devices came in, with their observed results coming before group assignments; many did not see the reverse ordering as an isomorphic model. Noll et al. (2018) observed that students had similar preferences with the ordering of devices in a probability modeling context. Despite the challenges students faced with the narrative elements of modeling, it is important to point out the CATALST curriculum at least enables students to think about these conceptual aspects of hypothesis testing, unlike traditional or even many other simulation-based curricula. There is much for students to gain in understanding the data generating process of a hypothesis test and how it is rooted in the null hypothesis assumption, which in turn can enable students to have a rich understanding of the conclusions they draw from a test.

While research on students understanding of randomization tests and bivariate data exists in isolation, there is not any work done yet on students modeling of randomization tests for the least squares line. This study aims to fill this gap while also comparing approaches for carrying out such a test to students who took a traditional, algebra-based statistics course. Studies on students understanding of bivariate data yield some mixed results when comparing student outcomes in traditional and simulation-based curricula. Students in traditional statistics courses had mostly non-significant differences gains in performance to those who took a simulation-based course on survey questions pertaining to bivariate data, with one survey item that significantly favored students from the traditional curriculum (Tintle et al., 2011). Results were more mixed when comparing retention of bivariate data topics across each curricula, with no significant differences in retention on any of the survey items (Tintle et al., 2012). However, these same comparison studies along with many others show notable gains in

performance on survey items pertaining to hypothesis testing and the purpose of inference, which makes sense due to the emphasis placed on inferential techniques and their conceptual understanding through carrying out simulations (Chance et al., 2016; Hildreth et al., 2018). These differing results emphasize the interest in the present study that combines these two content areas.

In light of the potential benefits the CATALST curriculum has in highlighting the purpose of inference for linear regression, this study aims to address the following research question: Do students from a traditional curriculum and the CATALST curriculum recognize the need to use a hypothesis test for evaluating the statistical significance of a linear relationship? How do students' approaches compare across these two curricula?

Methodology

In this study, students participated in both surveys and interviews that focused on questions about determining a significant linear relationship. The following subsections will first detail the theoretical framing for the study and give background for why individual instruments were used. Once this framing is established, I then explain the tasks students completed as part of the surveys and interviews, detail the participants in the study, and describe the method of analysis used on the data collected.

Theoretical Framing

My view on students' learning reflects the theory of social constructivism. Students are not only actively constructing their own knowledge from their experiences in the course, but by working collaboratively with their peers and being integrated into a

community of knowledge. This is especially true in the CATALST curriculum, where students work on scaffolded activities designed for students to discover statistical concepts in small groups. Knowledge in the classroom is thus constructed based on both students' statistical experiences within the course as well as personal experience outside the course that may relate to statistical concepts or the data context. These activities are rooted in various contextual settings, often with notable societal and cultural importance, so students own experiences and backgrounds add to their learning experience in the classroom. Students from the traditional classroom were also given opportunities to work in groups on practice problems with their peers in order to build their statistical knowledge collaboratively in a similar light to the CATALST classroom.

Considering this perspective and the collaborative nature of both classrooms, it may seem surprising that this study's data collection is based on individual surveys and task-based interviews. However, individual instruments can still be seen as compatible with the social constructivist perspective. Vygotsky views learning happening on two levels: the interpsychological, where ideas are shared on the social level, and the intrapsychological, where such ideas are internalized. Students come with many pre-conceived notions about reading and interpreting bivariate data and linear relationships from their own experiences, which the pre-survey aims to capture. The learning process during the course is then experienced socially with their peers, and the experiences they bring to the course individually affect their own experiences with classroom activities and the data contexts. These experiences have individual impacts on their learning, leading to this knowledge that students construct internalized once again in each individual student.

The post-survey and the interviews aim to capture what knowledge these students internalized.

For the purposes of the research I am conducting, my perspective of knowledge is more cognitive than social, as I am more focused on what knowledge an individual has constructed in each classroom. Individual achievement is very relevant to higher education institution, as for better or worse, students' knowledge is evaluated by individual grades. However, this focus on the individual still incorporates the idea that students' knowledge is not based on a totally individual experience, but on their experiences in and out of the classroom. This social perspective lends to why I believe the CATALST curriculum has an advantage over other traditional curricula. Building a TinkerPlots sampler requires a negotiation of both statistical and contextual ideas, and contextual ideas are firmly rooted in students' experiences. This gives students perspectives on identifying the most relevant aspects of context to be used in their TinkerPlots samplers while also engaging students in the statistical processes carried out by that sampler. I believe that through these classroom experiences, students' statistical knowledge that they have gained individually can be observed by the individual survey and interview instruments.

Survey

Individual surveys were administered electronically to students both before and after learning linear regression content in the course. This study focuses on responses to an open question on how students would conduct a test to determine a significant linear association with provided data. Students were asked several other questions before this to

prompt them to explore the provided data and give conjectures about the relationship, but the focus of this study will be on their responses to the question regarding determining a significant linear association. This question in particular is relevant to the research question posed, as it will reveal whether students recognize the need for inference to address whether there is a significant linear relationship. If students have gained the appropriate knowledge from their courses, they should be able to describe the correct procedure used in this scenario. Ideally, students would also convey their conceptual knowledge about inferential techniques (i.e., describe the null hypothesis, interpret how the p-value would allow them to draw conclusions), but asking more pointed survey questions regarding interpreting results from a hypothesis test may lead students toward choosing that method based on the wording of the questions rather than their knowledge alone.

To mitigate the effect of biases in the writing of the question toward a specific curriculum, the data and contextual information were pulled from the ARTIST (Assessment Resource Tools for Improving Statistical Teaching) database, with the sub-questions altered to elicit open-ended responses. Information on the data context and sub-questions presented to students in the survey can be seen in Figure 12.

Interview

Interviews were carried out with a smaller set of students from those who took both the pre and post survey. These interviews were carried out virtually via Zoom in light of the ongoing COVID-19 pandemic. Students who were interviewed were asked to revisit their survey responses and then carry out the methods they described in their

Does the outdoor temperature have a significant impact on the cost of your electric bill? A random sample of 25 electric bills across the USA was taken, and the table of data below lists the average temperature of a month and the amount of the electricity bill for that month from those 25 bills. The regression equation is: Bill Cost = 86.2 - 0.233*Temperature. Below is that table and a scatterplot of the data with the regression line superimposed.

Temp	Bill	Temp	Bill	Temp	Bill
51	$71.69	34	$77.93	45	$73.82
61	$72.64	32	$77.81	39	$74.41
74	$66.62	41	$74.43	35	$76.24
77	$70.70	43	$78.87	30	$80.80
78	$68.49	57	$69.48	49	$77.64
74	$67.88	66	$70.89	68	$68.70
59	$68.66	72	$70.89	76	$70.23
48	$76.23	72	$71.39		
44	$85.13	70	$68.31		
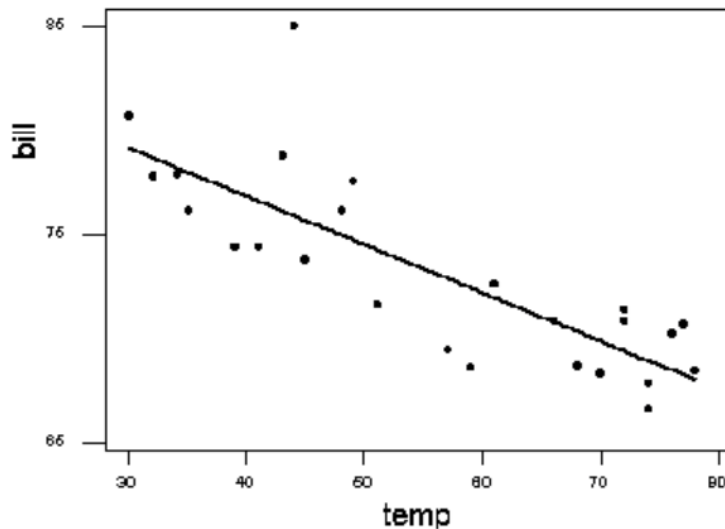
How could you test whether or not there is evidence to suggest there is a   *
significant linear relationship between the cost of an electric bill and the average temperature that month? Give details to any methods you might use to test this.

Your answer
_____

Figure 12. Background information and survey question analyzed in this study.

survey to determine if there is a significant linear relationship. The data were available for students to analyze in TinkerPlots, a virtual graphing calculator, and Excel, representing the software that students used in both the CATALST and traditional courses. Students had remote control access to the interviewer's computer through Zoom to give them access to interact with the software tools.

Revisiting the survey question aims to provide a deeper perspective of students understanding of conducting a hypothesis test for the slope of a regression line that could not be captured by the survey. To elicit ideas about their understanding of statistical inference, students were asked to explain how they would solve this problem and the underpinning concepts to someone who has never taken a statistics course. In order to determine if students recognized the distinction between the purposes of descriptive statistics and a hypothesis test, students were also asked if they could simply use the correlation value to determine if a relationship is significant. The semi-structured nature of this interview also allows for asking pointed follow-up questions in order to gain a better perspective of students' conceptual knowledge about hypothesis testing. Students who gained the appropriate knowledge in their courses should be able to carry out the test with the relevant software appropriately. They should also be able to interpret the results and p-value of their test in order to come to an appropriate conclusion based on their work.

Participants

This study examines students from two sections of a 10-week, second term introductory statistics course. This course is targeted at non-statistics majors, most of

whom come from social science backgrounds. While some students in the course may have had some experience with statistics from high school or other courses in their department, this is the students' primary exposure to statistics in college.

Of the two sections studied, one used the CATALST curriculum and had 23 students enrolled, while the other used a traditional curriculum and had 31 students enrolled. The CATALST section was taught by the author, and the traditional section was taught by another colleague who has worked with the CATALST curriculum previously. While the students are not guaranteed to be representative of students from each curricula or control for every confounding variable like instructor differences, the choice of these sections was made in order to limit these confounders. These two sections were chosen for the comparison due to the instructors holding similar teaching philosophies and leveraging in-class group work despite the differing curricula. Additionally, each section spent an equal time in-class on linear regression content, with four class sessions devoted to this topic.

A pre-survey was administered before students began this section of the course and once the topic was completed. Both sections administered the survey to students as an in-class assignment, with students receiving credit for completing the survey, not necessarily getting answers correct. Thus, all students were required to take the survey to receive in-class credit, but any non-consenting students who completed this assignment were not included in this study. For the CATALST section, 18 of the students consented to the study and participated in both surveys. For the traditional section, 17 consenting students participated in both surveys.

Students from each course were subsequently invited to participate in interviews. The selection of interviewed students was done purposefully based on their survey responses to obtain a pool of students with a wide variety of conceptions for conducting a statistical test in linear regression. For example, some students from each curricula were able to describe the appropriate test and how to interpret its results in their survey response, where others described only looking at descriptive measures like correlation. (The full range of codes assigned to student survey responses is in Table 19.) Thus, the interview sample is somewhat biased toward conceptions that were uncommon, and is intended to show the full range of possible student conceptions rather than be a representative sample. In the CATALST section, eight students were invited to participate in interviews approximately 1-2 months after the course's completion, five of whom participated. These five students are referenced in this study by the pseudonyms Dabney, Dene, Garnett, Morgan, and Riley. Eight students from the traditional section were invited to participate in interviews on the same schedule as the CATALST students, and three of those students participated. Those three students are referenced in this study by the pseudonyms Alma, Amani, and Jordan.

Analysis

Analysis of the results began with reading the survey responses and identifying common themes among student responses. Some of the notable themes that emerged included: the need to sample more data than already present, using descriptive statistics (e.g., correlation) to determine statistical significance, and the use of inferential reasoning on the pre-survey despite no introduction to this specific context. The depth of inferential

reasoning given by students, especially on the post-survey, was widely varied, with some simply specifying that a hypothesis test needs to be conducted without specifying what kind of test or how it would be carried out, and others describing a full procedure with conceptual explanations of the test conducted. Open codes designed to capture these ideas were refined into a coding scheme described in Table 19. The aim of developing this scheme was to accomplish two goals. The first is to determine the procedure they described for determining a significant linear relationship, which is reflected by the three main categories of codes: non-statistical, descriptive statistics, and inferential statistics. Within these categories, codes were further broken down to determine more specificity. For students that did not explicitly describe a statistical measure or method, it was common for students to call for more data to be collected to determine a significant relationship, so these kinds of responses were separated from students who were unsure or provided a response that was unclear or non-statistical. For students who gave a response using descriptive statistics, univariate measures like mean or proportions were separated from bivariate measures like correlation or the least squares line. If students described a hypothesis test or form of inference in their response, three different codes were used to determine their conceptual understanding about hypothesis testing they used; for example, recognizing that the hypothesis test for this scenario is about the relationship between two variables, or describing how they might interpret their results with a p-value, or describing the null hypothesis and how that is incorporated in the analysis.

Table 19. Coding scheme for survey responses.

Category	Code	Description	Example(s)
Non-statistical	Uncoded	The student is unsure of an approach or their answer is unclear, may describe non-statistical approaches to evaluate a claim.	"People are likely not running their heater as much as it gets warmer."
	Collect	The approach described by the student calls for collecting more data to properly answer the research question.	"Test more months and see if the electric bill follows the data and the regression line."
Desc. Statistics	Uni-variate	The approach described by the student involves purely descriptive measures that do not describe a bivariate relationship (e.g. mean, average, etc.) with no element of hypothesis testing or inference.	"You could test more months and see if the average temp matches what the predicted bill might be and see if the data follows the trend."
	Bivariate	The approach described by the student involves analyzing data that are of a bivariate nature, but are purely descriptive (e.g. correlation, R, slope, scatterplot) with no element of hypothesis testing or inference. (exclude terms explicitly stated in the prompts like "regression line" or "regression equation")	"Check the r squared value." "Using a line of best fit is a great way to tell whether or not there is a correlation in this scenario." "You'll use the residuals to see the average distances from the regression line. If that distance is small- we can assume strength, if not, its assumed to be weak."
Inferential Statistics	Generic	The approach described by the student mentions at least a vague idea of hypothesis testing but does not specify any characteristics of the test (e.g. does not describe hypotheses, p-value, how to draw the conclusion, etc.), or describes aspects of a test that does not apply to this data (e.g. testing for a difference in proportions). This excludes the phrase "significant linear relationship" as this phrase is used in the prompt.	"We could use a chi-square test, but this is a new concept to me so I may be wrong… Chi-square tests are often used to find direct correlation between nominal values, which is what we are trying to do in this case." "Use a hypothesis test to determine if it's sufficient enough to claim that there is a strong relationship."
	Generic+	The approach described by the student is centered around doing a hypothesis test and describes elements that indicate generic elements of the philosophy of significance testing (e.g. hypotheses, p-value) without specifying that this test is for evaluating the relationship between the two variables. This excludes the phrase "significant linear relationship" as this phrase is used in the prompt.	"You could use an Anova table, and find the p-value to decipher whether or not there is in fact evidence of a linear relationship." "You could do a random sample and collect a p-value in order to determine if its likely these results are due to random chance or to statistical significance."
	Specific	The approach described by the student describes a hypothesis test that assesses the relationship between the two variables, using a measure like the slope or correlation to test on, and derives results from the test using a p-value. Students may minimally reference the proper name of the appropriate test used in class.	"Create a sampler that represents the null hypothesis (that there isn't a correlation between the temperature and bill) and see how likely it is to get the original data if the null were true." "You could use a regression Anova test in Excel to get a p-value and compare it to your significance level."

Several students did give responses that fit multiple categories; for example, someone might see the need for more data while also using a descriptive measure like correlation to determine a significant relationship. Another student may describe a hypothesis test while also mentioning the relevance of calculating a correlation value. For these responses, students were only coded for one particular category, even if they had multiple categories of reasoning. Because the research question focuses on the ways students express their knowledge of inferential reasoning, if a student had descriptive and/or non-statistical responses in addition to inferential responses, they would only be assigned codes for the inferential reasoning category.

A limitation of using a survey is that follow-ups with students cannot be done to elucidate their response. This is especially limiting for students that appropriately cite the proper name of the test they learned in the course (e.g. test of linear association, t-test for the slope, regression ANOVA) with little further reasoning or explanation given. As terminology like this makes it clear that the students know some relevant details about the correct procedure and recognize the need to conduct a hypothesis test, these responses were still assigned the "Specific" code. This did lead to very short responses with that code, where longer, more in-depth responses that described the logic of inference but didn't specify that the test was for a relationship between two variables would only be coded as "Generic+". Given that the research question for this study focuses on determining the statistical significance of a linear relationship, and not their general knowledge of inferential reasoning, it seems most appropriate to code responses in this way. Additionally, knowing the name of a test shows that students recognize the need for conducting a hypothesis test, which also best reflects the research question. However, it is

worth recognizing that knowing the name of the test does not necessarily imply that students are capable of carrying out this test.

In order to ensure the reliability of the coding scheme used, all student responses were double coded by the author and a second coder, and all disagreements in coding were discussed until an agreement could be reached. First, a random selection of 24 responses were coded by both coders. These responses were a mix of responses from the pre-survey and post-survey. Of the 24 assigned codes, there were 5 disagreements, which were discussed until an agreement could be reached. This process led to some additional clarifications made within the code descriptions, with some initial code categories combined together. The two coders often disagreed on responses that only used language that was given by the prompts and background information, and so revisions were made so that phrases given in the prompt or background information were excluded for qualifying from particular codes. This ensures that students were not merely parroting the question given to them and required students to give reasoning in their own words to be coded appropriately. After this process, another round of double coding was performed for the remaining responses. Discussions were had again on any disagreements in coding, which were more easily resolved after revising the coding structure from the first round.

Transcripts of the interviews were generated with pseudonyms applied. The transcripts were read through for any inferential reasoning they gave, especially in regards to their response regarding the distinction between using a hypothesis test and the correlation coefficient alone to determine significance. Students' approaches were analyzed to determine if they produced an appropriate significance test for the regression

line and if they could distinguish significance testing from descriptive statistics like

correlation. This summary information aimed to provide some validation to coding for

survey responses based on this subset of students.

The summary information revealed two students from each curriculum who had

similar survey responses and initial approaches but revealed key differences in their

statistical reasoning. These key differences led to the choice of a collective case study

approach to analyzing these students (Stake, 1995). The selection of these two students is

instrumental in purpose, as the similarity in these two students' initial conceptions

provide an avenue for interpretation. Many differences in their thinking emerged as the

interview progressed, which revealed key differences in their respective curriculum

where they learned these concepts.

<div align="center">

Results

</div>

The subsections that follow will first investigate the results from the surveys

conducted in this study, then investigate the interview data. The interview data will be

presented first by summarizing the big picture of students' approaches, and then will

examine two specific students closely from each curriculum in detail.

Surveys

The survey data reveals stark differences in the responses for conducting a

significance test on the regression line across the two curricula. Table 20 shows the

counts of codes applied to students at the category level. CATALST students more

frequently gave responses that represented some form of hypothesis test than the students

from the traditional curriculum did. Traditional students were also more frequent in

Table 20. Frequency of specific code categories assigned to responses for the survey item.

	Pre		Post		Diff	
Category	**CAT**	**Trad**	**CAT**	**Trad**	**CAT**	**Trad**
Non-statistical	6	9	2	5	-5	-4
Descriptive Statistics	4	4	5	5	2	1
Inferential Statistics	8	4	11	7	3	3
Total	**18**	**17**	**18**	**17**		

giving non-statistical responses on their pre-survey and post-survey than the CATALST students. However, in terms of the intervention, students from both curricula appeared to have similar gains in advancing their response types toward one that represents a hypothesis test.

However, there is more nuance to how students' responses were coded within that hypothesis test category. Table 21 shows the full counts of codes across each curriculum and on pre-survey and post-survey. This shows that not only did many of the students in the CATALST curriculum suggest a hypothesis test of some variety, but many gave many specific details of the hypothesis test. One CATALST student with the "Specific" code gave the following response on their pre-survey:

> *"I would want to use a hypothesis test for this. I think, if there isn't a significant relationship between bill price and temperature, then the slope of the line would be closer to zero. I haven't ever done a hypothesis test for a linear relationship, but I am guessing there is probably a way to find out if this is significant (I would determine a p-value, I am guessing)."*

This student, despite no formal presentation of linear regression content, was able to anticipate the general idea of the method for determining statistical significance, identifying what the appropriate null hypothesis would be and recognizing the need to

Table 21. Frequency of specific codes assigned to students' responses for the survey item.

Category	Code	Pre CAT	Pre Trad	Post CAT	Post Trad	Diff CAT	Diff Trad
Non-statistical	Uncoded	2	6	1	4	-1	-2
	Collect	4	3	0	1	-4	-2
Descriptive Statistics	Univariate	1	0	1	0	0	0
	Bivariate	3	4	5	5	2	1
Hypothesis Testing	Generic	3	3	0	1	-3	-2
	Generic+	2	1	2	2	0	3
	Specific	3	0	9	4	6	3

determine a *p*-value. The other two CATALST students who gave responses with this code gave a similar amount of detail as well in their response.

Some students from the traditional curriculum did give responses that reflected conducting a hypothesis test on the pre-survey. However, no students were at the "Specific" level, meaning that no details about their method pertained to the use of bivariate data. The lone student who gave a response coded as "Generic+" stated "Using a hypothesis test with a 95% significance level would help to determine this theory within 5% of a doubt." This student showed that they recognized the need for a hypothesis test, but did not provide details that the previous CATALST student provided about the null hypothesis or using the slope of the line for conducting such a test. When examining the effect of the classroom intervention with learning content related to linear regression, both courses saw students generally shift toward hypothesis testing codes, as expected. It is notable, however, that CATALST had a change of 6 students using reasoning coded as "Specific," which was noticeably larger than the difference of 3 for the traditional course. Thus, while there was an equal shift toward codes in the hypothesis testing category overall across both curricula, CATALST students seemed to show a deeper

understanding of conducting an appropriate test of significance in this scenario than traditional students did, based on the survey responses.

Interview

As mentioned in the analysis section, assuming that students' survey responses are a representation of their knowledge at a specific time or gained through the intervention is suspect, especially since students were not asked to perform the test that they described in their survey response. To gain a better picture of students' conceptions regarding significance tests with linear regression, this subsection will investigate the eight interviewed students. Table 22 summarizes these interviewed students, highlighting the codes that were assigned to their survey responses as well as features of the students' statistical approaches in the interview. These features include whether they conducted an

Table 22. Summary of interviewed students' codes for survey responses and aspects of their interview responses.

		Survey		Interview	
Student	**Class**	**Pre**	**Post**	**Appropriate test**	**Corr. vs. Sig. test**
Dabney	CATALST	Non-stat: Uncoded	Desc: Bivariate	Yes, but not initially.	Initially same, then different
Dene	CATALST	Desc: Bivariate	HT: Specific	Yes	Same
Garnett	CATALST	Desc: Univariate	Non-stat: Uncoded	No	N/A
Morgan	CATALST	HT: Specific	HT: Specific	Yes	Different
Riley	CATALST	HT: Generic	HT: Specific	Yes	Different
Alma	Traditional	Non-stat: Collect	HT: Generic+	Yes	Same
Amani	Traditional	Desc: Bivariate	Desc: Bivariate	Yes, but not initially.	Same
Jordan	Traditional	Desc: Bivariate	HT: Specific	Yes	Same

appropriate significance test and their perception on the distinction between correlation and a significance test. One noticeable difference across the two curricula is that CATALST students seemed to have a better understanding of the difference between correlation and the hypothesis test for the regression line—all three students who were interviewed in the traditional classroom did not make a distinction in meaning between correlation and conducting a significance test, where three of the five CATALST students provided reasoning for how they are different. One CATALST student was not able to conduct an appropriate significance test for this scenario. As this student did not offer any approach when interviewed, they were not asked regarding the distinction between these two concepts.

To further investigate the interview data, the cases of Amani and Dabney will be investigated further. These two participants were chosen for a few reasons. First, they both gave non-hypothesis test responses on the surveys, and yet were able to conduct an appropriate test but only after discussion with the interviewer, which provided interesting episodes regarding their statistical thinking. Additionally, they had diverging opinions on the distinction between correlation and significance tests. Their different responses were fairly representative of other interviewed students in their respective curricula, despite their common initial intuition for conducting a significance test.

The following subsections review Dabney and Amani's interviews in three separate phases. The first phase is where they describe their initial approach with various descriptive statistics. The next examines whether they believe their descriptive statistics can determine statistical significance. These first two phases of the interview reveal many

similarities in their approaches. In the final phase of the interview, Dabney and Amani were prompted to explore the idea of conducting a significance test, which reveals interesting divergences in their thinking.

Descriptive Statistics. Both Amani and Dabney's survey responses indicated that they could use descriptive statistics for analyzing the significance of a linear relationship through purely descriptive methods. Amani, a student from the traditional classroom, gave this response on the post-survey: "To test whether or not there is evidence to suggest there is a significant linear relationship between the cost of an electric bill and the average temperature that month, I would do a r-squared statistical analysis to see what the relationship between the two variables are." Dabney, a CATALST student, gave this response on the post-survey: "You could test this by calculating average residual in order to determine how well the line would describe the relationship." While these are not identical statistical methods, the average residual size and r-squared both are descriptive measures of how close the data are to the regression line, with the latter being a normalized measure.

These post-survey responses influenced their initial methods they conducted during the interview. Amani created a scatterplot in Excel that showed the r-squared value of 0.6463 on it, and interpreted this value in this way:

Amani:	Because it's 0.6463, it's like, moderately -- okay, to me that's low. Usually an r-squared value of 0.8 or higher shows a better correlation between the values.
Interviewer:	What are your cutoff points for r-squared for something to be significant? It sounds like 0.8 though [is significant]?
Amani:	0.8 to 1, yeah.

| Interviewer: | Gotcha. So, you would call 0.64 then sort of moderately associated then? |
| Amani: | Yeah, moderately associated. Between, I guess 0.5 and 0.79, I think that's moderate, anything lower is weakly associated. |

Amani's method for determining significance is quite strict, with an r-squared value of 0.8 or higher being significant enough for them. For comparison, when conducting a t-test, an r-squared value of 0.45 produces a *p*-value smaller than 1 percent even with a small sample of 20 data points. In this scenario, with an r-squared of 0.64, Amani agrees with the interviewer that this is only a moderate association, but we do not get a sense of what Amani believes r-squared represents as a measure itself.

Dabney's initial method is described by taking the average size of the residual. In TinkerPlots, Dabney places the least squares line on a scatterplot and uses tools within the program to measure the average size of the residual. Dabney then begins to analyze his findings:

Dabney:	So 2.1 is the value [TinkerPlots] gives. … Having [2.1] in the [units] of what's on the graph isn't necessarily meaningful… something you could say about the graph is like, "Oh, we can predict your bill based on a temperature with a range of $5.40." Is there an equation to convert that to a value between like zero and one? Is there an equation, maybe, that can relate that to -- I'm trying to turn that into a number that's more universal? … Could you convert that into a percentage?
Interviewer:	A percentage of what though?
Dabney:	Oh, [darn] it. That's a good question. [pause] Maybe I'm talking in circles, but I know that we're trying to -- I want a number that's going to describe how far off it's likely that I'm going to be, or the range at least, if I'm trying to predict it based on … if I'm trying to predict temperature from bill or bill from temperature.

Dabney calculates the average residual size, but realizes that this is difficult to interpret generally in terms of the relationship itself. They realize the need to calculate a more

universal value like a percentage but are unsure of what that value might be. After more

discussion, Dabney realizes a potential solution:

> Interviewer: Do you recall from class any measure that described the strength of a relationship via a percentage at all?
>
> Dabney: I mean, correlation is a number between zero and one, right? Like a correlation of one is -- it's absolutely the same. And then the smaller the correlation, the less likely it predicts -- oh, so maybe it's correlation -- the less likely a predictor it is.

While correlation is not truly a percentage like r-squared is, Dabney does come up with a

normalized measure of the strength of the relationship, akin to Amani's initial method.

Dabney goes on to calculate the value for correlation in TinkerPlots, which is -0.8 for the

given data. Based on this discussion, Dabney seems to have a more conceptual

understanding of this measure as well, as Dabney connected this measure to the average

residual size and the closeness to the line while connecting these ideas to the actual

context. Dabney also places their interpretations within the context of temperature and

the electric bill. Amani's conception of *r*-squared seems purely procedural and relative:

it's a value used to assess a linear relationship, and a higher number indicates a stronger

relationship.

Using Correlation to Determine Significance. In both interviews, the

interviewer confirms with each student that these descriptive methods are enough to

determine if there is a significant linear relationship. This was Amani's response to this

question:

> Interviewer: Were there any other sort of ways that you went about testing for significant linear relationships in class? Or is this sort of the main one that you used?
>
> Amani: This is the main one I used, that I typically use, actually.

Interviewer: So is r-squared helping you evaluate actual hypotheses, like testing for hypotheses? And if so, what are those hypotheses that you're testing?

Amani: I don't remember specifically, but for r-squared, I'm just looking at the correlation specifically. Not necessarily if the data is significant. You would have to deal with different tests for that. I don't remember the name of the test.

Here, Amani confirms that this value of r-squared is enough to determine a significant linear relationship, but when followed up, Amani contradicts their previous statement saying that there are other tests that they can't currently recall that would be appropriate for that. Amani's investigation into other tests will be examined later. At this point in the interview for Dabney, they give a similar response to this question:

Dabney: Okay, so our correlation is -0.8… I do know that we would be able to say there's a negative correlation, meaning as *x* increases, *y* decreases. What I'm not sure of is, like with *p*-value, how we've [got] those… ranges where it's like strong, moderate little evidence. I don't remember what those are with correlation… But I feel like if something had an 80% accuracy, … I would say that that's acceptable for me.

Interviewer: Can you use [this correlation] to answer the question of whether there is a significant relationship or not?

Dabney: Yes, absolutely. That's how you would describe it. That's how you would say like, there is, and this is the level of that the strength of that relationship… I don't know when you would say strong, moderate or whatever. But yes, absolutely. That's the number you would use.

In addition to agreeing that correlation is enough to determine a significant relationship, Dabney also alludes to potential ranges of values for correlation that determine the strength of a given relationship that Amani referenced earlier, but is not sure what those ranges are and isn't totally sure whether -0.80 is enough evidence because of that. Thus far, both students have described very similar procedures to answer this question by discussing correlation and r-squared, despite having learned their content from different

curricula. Dabney may be demonstrating a deeper conceptual understanding of these descriptive measures by connecting correlation to the average residual size and placing interpretations within the data context, but these two students thinking regarding their statistical methods both reflect descriptive statistics so far. The next episodes will challenge students on their conceptual knowledge of hypothesis testing, which is where their overall approaches will diverge further.

Hypothesis Testing Concepts. At this point of the interview, both Dabney and Amani are asked to consider whether their correlation values make it possible to evaluate hypotheses. Amani previously recognized that this idea of hypothesis testing represents a different method, but does not recall details about how to conduct this test:

| Interviewer: | Do you remember how to carry [out a hypothesis test] in Excel or on a TI or anything like that? |
| Amani: | For TI, no, maybe with Excel. Maybe if I spend a little more time on it. |

More interview time is spent on this idea, but Amani is still unsure about how to conduct such a test in Excel. The interview then shifts to focus on conceptual ideas of hypothesis testing and the *p*-value:

Amani:	I would have like a null hypothesis and alternative hypothesis, the null being that there is no significant difference between these two values, I mean, these two variables, and then alternative being there is a significant difference, or maybe I'd reverse those two. But then I would just conduct a test, I probably would end up Googling which test to use, and then apply it for my data.
Interviewer:	What would be the end result that we would use to make a decision from that test?
Amani:	Based on the *p*-value specifically, yeah. And then see if there is any significance or not.
Interviewer:	Gotcha. Can you explain what a *p*-value exactly means?

Amani: So, with a *p*-value, it's just seeing that with a 5%, confidence, or whatever confidence that we're doing, to see that 5% of that time or something? Or, actually, I feel like I'm mixing that up now with the p value definition. … [long pause] It's basically the chance, I guess, the probability of that happening. Or I guess the *p*-value is the probability of you seeing that result happening. Or not, there's a -- I guess it's as extreme as it is, like of the observed results or something like that.... Basically it's just like observing to see like can this happen in that value.

Amani is able to define hypotheses for an appropriate test, albeit with some confusion about which is the null or alternative. They also recognize the need to compute a *p*-value. However, they are not as sure about defining what the *p*-value means and how it is used to draw an appropriate conclusion. There are some phrases in Amani's response that are normative to the definition of the *p*-value. However, Amani continues on to be challenged in finding an appropriate definition. To help have a concrete example to talk about, the interviewer conducts the appropriate test in Excel for Amani, and this leads to their interpretation of those results:

Amani: I guess the *p*-value for that, if we're doing like a 5% confidence interval, because the *p*-value is less than the confidence at 5%, then we would end up not rejecting – no, we would end up rejecting the null hypothesis for that, and then accepting the alternative for whatever that may be.

Interviewer: So why does a low *p*-value lead us to rejecting the null? You talked about the *p*-value being like this probability that we get our observed result, so I'm just trying to see if you can connect that idea to why a low *p*-value ends up rejecting our null hypothesis?

Amani: I don't think I have a strong understanding of the *p*-value specifically then, going through these questions specifically.

Amani holds appropriate procedural knowledge on how to conduct a hypothesis test and draw an appropriate conclusion from a *p*-value, yet still exhibits some confusion about what the hypotheses are and what is rejected or not. Most notably though, Amani does not seem to understand the *p*-value and why this procedure works. Thus, while Amani

may have earlier recognized that there was a difference between using the correlation or

r-squared and a hypothesis test for the slope of a regression line, there is not a great

conceptual basis for how to conduct a hypothesis test and interpret the *p*-value. While this

was not addressed directly with Amani again in the interview, it would seem that they are

unsure about how the purposes of descriptive statistics like correlation and hypothesis

testing differ.

Dabney's approach to conducting a hypothesis test began with constructing a

sampler in TinkerPlots, shown in Figure 13. This sampler takes two bootstrapped random

samples of both the electric bills and the temperatures that month, and pairs each together

at random. After constructing this, Dabney initially faced some challenges in determining

how to use this sampler, specifically on what statistic to collect. They are prompted on

what the purpose of conducting these simulations is, which sparks an idea:

Dabney:	[The purpose of simulating is] to determine how likely it is that the real data that we've sampled was just generated by chance, as opposed to there being an actual trend. So it helps us determine how accurately and how likely it is that that study or data set represents reality, represents the actual population that it was sampled from.
Interviewer:	Gotcha… here we're trying to test for if there is a significant relationship, so… what can we use to measure against the data we actually sampled?
Dabney	Oh! So here. So if we were writing out a hypothesis, the null hypothesis, our null hypothesis would be there is no relationship between temperature and bill, whereas our [alternative] hypothesis, there is a relationship. So, we are going to use that concept of the null hypothesis, there being no relationship, that's how we would develop our simulation… So let's say we do five hundred simulations, we are going to then see what the percentage of them hits our correlation or stronger, or more extreme. Oh, shoot, and
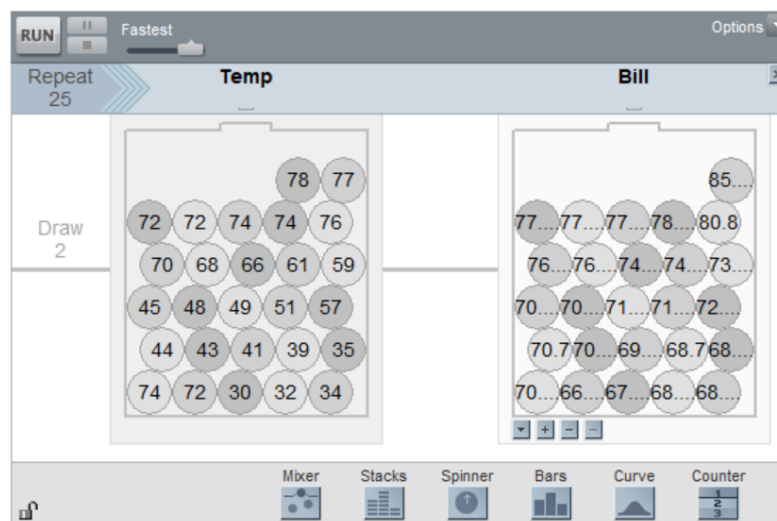
Figure 13. Dabney's TinkerPlots sampler.

we're gonna generate a *p*-value. And so then that *p*-value is going to represent exactly… how many cases out of that 500 had a correlation of -- what did ours come out to be? -0.8… So the fewer cases we have that are within that range, the stronger the evidence we have that the actual data is representative, and it's unlikely that it was generated by chance.

After being prompted to think about the purpose of simulation and what measure would be relevant, Dabney proceeds to divulge the nature and logic of hypothesis testing applied to this scenario. Dabney's statements about this hypothesis test are always grounded in the data context, which potentially alleviates the confusion Amani exhibited regarding which hypothesis is which and what is typically rejected. Dabney expresses the appropriate hypotheses to set up, that their simulation is representing the null hypothesis, and the *p*-value then represents the number of cases as extreme or more as the correlation value of -0.8. They go on to show the *p*-value in their empirical sampling distribution shown in Figure 14.

After going through this process and concluding that this is a significant linear relationship through this hypothesis test, Dabney is asked to compare this approach to
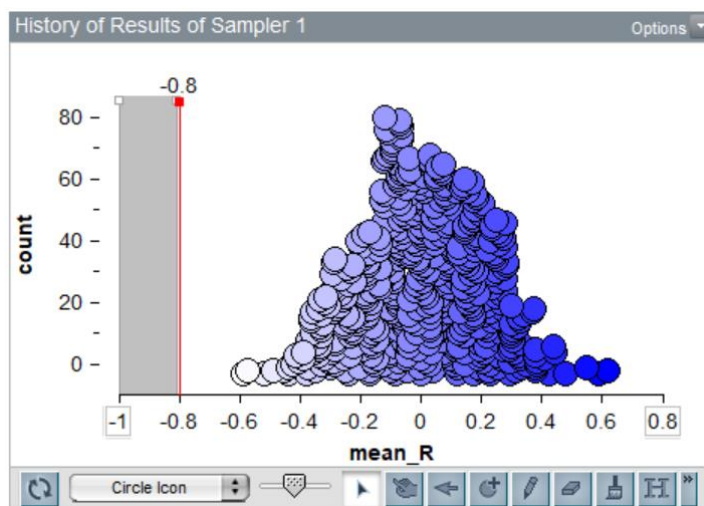
Figure 14. Dabney's sampling distribution, with shaded region representing the p-value.

their previous one using correlation, and whether they are answering the same research

question:

> Interviewer: I guess the question we were going for was is this a significant relationship? [Is this a] synonymous answer [to your previous answer using correlation]?
>
> Dabney: No, they're not synonymous. I don't think they're synonymous. I think that this is allowing us to determine that this data wasn't generated just by chance. But I think that R value is still separate…. So I guess, I see them as different concepts, I would say that *p*-value is the likelihood that the data was just generated by chance, and the R value, the correlation, is showing how strong the relationship is between our two variables. Which would also relate to its predictability.

Here, based on Dabney's understanding of the hypothesis testing process shown earlier,

they are able to make a distinction between their previous descriptive methods using

correlation and the average residual to the hypothesis test they just conducted. Amani did

not display this level of understanding in the interview, and while they recognized that

the two methods were distinct, they were unable to explain their conceptual differences as

Dabney did here.

Discussion

This study reveals many potential advantages for simulation-based curricula like the CATALST curriculum in regards to students' understanding of hypothesis tests for a regression line. These advantages will be discussed in two subsections, the first of which will discuss important aspects revealed from the survey data, and the second will investigate the interviews. Some notes about non-response in this study and the impact it may have had on the results are also given.

Surveys

Students from the CATALST curriculum were much more frequent in providing survey responses that gave aspects of a hypothesis test, even before formally learning this content in their course. While this pre-survey is primarily a baseline measure, it's important to consider that these students were in their second statistics course of a two-course sequence, and have seen hypothesis tests in various contexts before learning content on linear regression. These differences may be attributable to what students learned before taking their respective courses, but it could very well be a result of the conceptual knowledge of hypothesis testing they gained during the course before taking the pre-survey. Considering the numerous research studies reviewed that showed significant advantages for simulation curricula in learning hypothesis testing, it does not seem surprising that CATALST students would start this unit with a rich conceptual knowledge of hypothesis testing based on what they had learned thus far in the course.

An additional point of interest from the pre-survey was that the only students to be coded at the "Specific" level were from the CATALST course. This code indicates

that they were able to give specific details of a test for assessing the relationship of two numeric variables, despite no formal introduction to such a test in the course. This shows potential for students in the CATALST course as emerging statistical modelers, recognizing the important aspects of a given data scenario and what would be relevant to simulate and test. Throughout the CATALST course, students are asked to create TinkerPlots samplers with minimal instructor intervention at first, which gives students practice as statistical modelers. These three students coded as "specific" potentially demonstrate an ability to reason through relevant elements of a novel situation and encode them statistically and in TinkerPlots, shuttling through the context, statistical, and technology worlds identified by Biehler et al.'s framework (2015). Still, this was only seen in three of 18 students in the CATALST class, which confirms the challenges students have in navigating these three worlds identified by Noll & Kirin (2017). But given the limitations of a survey, it is possible that more students from the CATALST course have this foundational hypothesis testing knowledge and the ability to model and simulate these novel scenarios. While the interview came well after the end of the course, Dabney was able to reason clearly though the modeling and simulation processes in their interview despite only giving a post-survey response that described only descriptive measures.

Another interesting feature of the survey data is the shift CATALST students made toward more sophisticated expressions of their statistical knowledge on the post-survey. While both classes saw students shift toward higher level codes representing reasoning centered around a hypothesis test, there was a larger shift in the CATALST class toward the "Specific" code. Half of the CATALST class (nine out of 18) gave a

response coded as "Specific" on the post-survey, compared to 3 out of 17 in the traditional class. This seems to show that the CATALST curriculum not only prepared students for learning hypothesis testing in a deep and conceptual way, but acted as a stronger intervention in aiding students in applying those concepts to linear regression than the traditional curriculum did.

Interview

The most notable feature from the interview data as a whole was that CATALST students typically recognized the differences between descriptive and inferential statistics and their purpose. While many of the traditional students did carry out an appropriate test on the linear regression line, none of these three students saw a difference in purpose to carrying out a statistical test and examining the correlation value, whereas three CATALST students were able to explicitly explain the differences conceptually. This supports the initial conjecture of this study that CATALST students would be apt to point out such differences. Simulation and modeling seem to be powerful tools for students to develop their statistical thinking in this way.

The cases of Amani and Dabney further highlight this. Both students had very similar initial trajectories in their interviews, as they started based on survey responses that described descriptive methods only, carried out those methods, and believed them appropriate for determining a significant linear relationship. It was when they were asked to consider if their methods could evaluate statistical hypotheses that the interviews diverged. Amani was not able to carry out this test on their own with the provided software, and while there were elements of procedural knowledge for hypothesis testing

present in Amani's reasoning, they could not make distinct the conceptual differences between analyzing the correlation value and carrying out a hypothesis test on the linear regression line. On the other hand, Dabney was able to model the scenario with a TinkerPlots sampler and use it to carry out an appropriate hypothesis test while providing sound reasoning and interpretations of their test. Dabney also made appropriate distinctions in purpose and meaning between their hypothesis test and the correlation value they computed earlier in the interview. Despite the fact that a hypothesis test was not the first method that Dabney gave on the post-survey or interview, they still had a great deal of conceptual knowledge about setting up hypotheses, building a model to carry out a simulation, and interpreting the results and p-value of the test readily available. This suggests that there may be other surveyed CATALST students who did not provide answers coded within the hypothesis test category that still have deep knowledge of hypothesis testing that can be applied to linear regression.

Another interesting feature of Amani and Dabney's interview is that they both mentioned potential ranges of values for the correlation or r-squared to determine if it was a strong, moderate, or weak correlation. Dabney didn't cite specific ranges exactly, but did connect the idea of this to interpreting strength of a p-value based on it falling within different ranges of values, as well as comparing it to the significance level to draw an appropriate conclusion. This may reveal another challenge for students in keeping inferential and descriptive statistics distinct with linear regression, as both p-values and correlation are interpreted based on falling within specified ranges of values, often subjective in nature. Students could potentially conceptualize the idea that the p-value being less than a significance level 0.05 is similar to the correlation being greater than 0.6

or less than -0.6 in the scenario of linear regression based on the similarities in these procedures. This may suggest that presenting correlation in a way that assigns ranges of values to subjective qualifiers like strong, moderate, or weak may not distinguish it enough from the idea of a hypothesis test. Students in the present study seemed to have a strong conception that a correlation of 1 or -1 would result in all points falling on the line, and thus it may be suggested to have students learn the meaning of correlation by seeing various examples of scatterplots with each correlation value instead. This may lead to a more objective view of the correlation value and allow students to judge the meaning of the strength themselves. Ranges of values indicating a level of strength are quite subjective, as social sciences are often pleased with data that shows a relatively low correlation value due to the varied nature of the measures they construct, where more technical fields often have stricter qualifiers for what determines a meaningful relationship based on correlation. Given that introductory statistics courses are commonly multidisciplinary, any universal ranges learned for the strength of a correlation will be supplanted with field-specific constraints if students eventually become involved in research in their own field.

Limitations and non-response considerations

It is important to keep in mind that the results of this study are limited by the scope of this study. While the traditional classroom was chosen to have an instructor that emphasized active learning in the classroom to mitigate the effect of pedagogy, the differences in the instructor's implementation of active learning may confound the results in this study. It is also important to consider that the author of this paper is the instructor

of the CATALST course. While I attempted to remove my own personal biases in the data collection process of this study, this may have had some impacts on the research. Additionally, the sample size of this study is limiting, especially with the interviews. Only a handful of students were interviewed and were selected in a purposeful manner, and thus this study cannot provide generalizable results about the differences in the curricula. This study's results only aim to provide a view of CATALST as a practically relevant and theoretically promising curriculum for introductory statistics.

The instructor effect also appeared to have an impact on non-response in the traditional course, due to the lack of familiarity with the researcher. Nearly 80% of the students in the CATALST class consented to the study and completed both the surveys, where just over half from the traditional classroom did the same. Given that the survey instrument was a course assignment contributing to a very small amount of their final grade, responses may bias toward students that were keeping up with all assignments in the traditional class. This potentially is correlated with students who achieved higher grades in the course given their relative willingness to keep up with two minor assignments in the course, although I do not have the data available to confirm this. Still, this gives reason to believe that a higher response rate from the traditional course may have revealed an even greater advantage to CATALST students in recognizing the need for a hypothesis test.

As for interviews, only three traditional students responded to invitations to interview. These three students gave relatively high-level codes on the post-survey among their classmates, whereas the group of CATALST students was more varied.

Garnett, a CATALST student that was "Uncoded" on the post-survey, did not have another comparable student from the traditional course that interviewed due to this non-response.

Conclusions and Future Work

There are many advantages to using a simulation-based curriculum like CATALST for developing statistical thinking, especially as it pertains to the purpose and logic of hypothesis testing. Students from the CATALST curriculum were more frequent in applying an appropriate method for determining a significant linear relationship as well as distinguish these inferential methods conceptually from descriptive statistics like correlation. Regardless of the curriculum, these results have many implications for the teaching of linear regression. The purpose of conducting a hypothesis test in the context of linear regression should be emphasized as a tool to determine generalizability, that is, if the relationship present in the data could be a product of random chance of the sampling process under the null hypothesis. This may not imply that the relationship is meaningful or has a practical use or interpretation in a given field, but could be used as a tool to determine the possibility of such a relationship. Descriptive tools like correlation can aid in verifying whether a relationship is meaningful, but it should be noted that what makes a meaningful correlation is very field-specific. Textbooks often provide ranges of values for the correlation to determine what values yield a strong relationship, but this cutoff potentially further blurs the concepts of hypothesis testing and descriptive statistics, as it parallels the procedural aspects of interpreting a p-value.

This work adds to the vast literature that compares student outcomes in traditional and simulation-based curricula, and provides more evidence for the relative efficacy of simulation-based curricula. This study focused on the CATALST curriculum, which is unique in that students create their own probability models to carry out simulations, rather than be given pre-constructed simulation applets where students can only tinker with a few parameters of these models. This modeling aspect of the CATALST curriculum has revealed many advantages to advancing students' statistical thinking, and may have been a key aspect in students' ability to discern between inferential and descriptive statistics as well as provide detailed conceptual explanations of inferential techniques. Future work could consider comparing students from various simulation-based curricula, including CATALST, and identifying differences in students' reasoning.

References

Biehler, R., Frischemeier, D., & Podworny, S. (2015). *Preservice tachers' reasoning about uncertainty in the context of randomization tests* (pp. 129–162).

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centered on simulation-based inference. *Statistics Education Research Journal*, *21*(3), Article 3. https://doi.org/10.52041/serj.v21i3.6

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education*, *24*(3), 114–126. https://doi.org/10.1080/10691898.2016.1223529

Cobb, G. W. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1). https://escholarship.org/uc/item/6hb3k0nz

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction*, *21*(1), 1–78.

Doerr, H. M., & Pratt, D. (2008). The learning of mathematics and mathematical modeling. *Research on Technology and the Teaching and Learning of Mathematics: Research Syntheses*, *1*, 259–285.

Edwards, M. T. (2005). Median-slope algorithm. *Mathematics Teacher*, *98*(6).

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898. https://doi.org/10.1007/s11858-012-0447-5

Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing student success and understanding in introductory statistics under consensus and simulation-based curricula. *Statistics Education Research Journal*, *17*(1). https://iase-web.org/documents/SERJ/SERJ17(1)_Hildreth.pdf?1526347238

Justice, N., Le, L., Sabbag, A., Fry, E., Ziegler, L., & Garfield, J. (2020). The CATALST Curriculum: A Story of Change. *Journal of Statistics Education*, *28*(2), 175–186. https://doi.org/10.1080/10691898.2020.1787115

Justice, N., Zieffler, A., Huberty, M. D., & delMas, R. (2018). Every rose has its thorn: Secondary teachers' reasoning about statistical models. *ZDM*, *50*(7), 1253–1265. https://doi.org/10.1007/s11858-018-0953-1

Konold, C. (2002). Teaching concepts rather than conventions. *New England Journal of Mathematics*, *34*(2), 69–81.

Konold, C., & Miller, C. (2018). *TinkerPlots* (2.3.4). LearnTroop. http://www.tinkerplots.com

Lesser, L. (1999). The 'Ys' and 'why nots' of line of best fit. *Teaching Statistics*, *21*(2), 54–55. https://doi.org/10.1111/j.1467-9639.1999.tb00807.x

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1).

Noll, J., Clement, K., Dolor, J., Kirin, D., & Petersen, M. (2018). Students' use of narrative when constructing statistical models in TinkerPlots. *ZDM*, *50*(7), 1267–1280. https://doi.org/10.1007/s11858-018-0981-x

Noll, J., & Kirin, D. (2017). TinkerPlots model construction approachs for comparing two groups: Student perspectives. *Statistics Education Research Journal*, *16*(2). http://iase-web.org/documents/SERJ/SERJ16%282%29_Noll.pdf

Noll, J., Kirin, D., Dolor, J., & Clement, K. (2021). *Inventing and evaluating TinkerPlots models for comparing two groups: A statistical modelling story* [Manuscript in progress].

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance and context. *ZDM*, *50*(7), 1113–1123.

Rossman, A. J. (2008). Reasoning about Informal Statistical Inference: One Statistician's View. *Statistics Education Research Journal*, *7*(2).

Sorto, M. A., White, A., & Lesser, L. M. (2011). Understanding student attempts to find a line of fit. *Teaching Statistics*, *33*(2), 49–52. https://doi.org/10.1111/j.1467-9639.2010.00458.x

Stake, R. E. (1995). *The art of case study research*. SAGE Publications, Inc.

Tintle, N., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2014). *Quantitative evidence for the use of simulation and randomization in the introductory statistics course*. 9th International Conference on Teaching Statistics, Flagstaff, Arizona. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21.

Tintle, N., VanderStoep, J., Holmes, V.-L., Quisenberry, B., & Swanson, T. (2011).

Development and assessment of a preliminary randomization-based introductory

statistics curriculum. *Journal of Statistics Education*, *19*(1).

https://doi.org/10.1080/10691898.2011.11889599

Zieffler, A. S., & Garfield, J. B. (2009). Modeling the growth of students' covariational

reasoning during an introductory statistics course. *Statistics Education Research*

Journal, *8*(1).

Chapter 4: Evidence for Further Development of TinkerPlots to Support Inferential Reasoning with Linear Regression

Abstract: The Change Agents for the Teaching and Learning of STatistics (CATALST) curriculum is a popular simulation-based curricula for introductory statistics that has students create probability models to carry out simulation. However, the curriculum as originally designed does not include many topics that are typically in an introductory statistics course, such as linear regression. I detail a workaround in the TinkerPlots software that allows for conducting more advanced statistical tests, and show evidence that using this workaround with students to conduct test for the least squares line reveals many advantages to students' ability to manage the multi-leveled reasoning of inference when leveraging simulations. This setting also provides opportunities for students to engage with experiment-to-causation inference. Based on this evidence, we propose a potential technology innovation for TinkerPlots to eliminate technological hurdles associated with my workaround and better support students' inferential reasoning.

Introduction

There is a great deal of evidence supporting the use of simulation-based inference in introductory statistics classrooms. Studies that compare these curricula with more traditional, algebra-based counterparts show many positive learning outcomes for students, especially in regard to understanding the purpose of hypothesis tests and interpreting their results (Chance et al., 2016; Chance & McGaughey, 2014; Tintle et al., 2012, 2014). A primary feature of a simulation-based course is the technology used to carry out simulations. The technology innovation highlighted in this paper focuses on the Change Agents for the Teaching and Learning STatistics (CATALST) curriculum

(Garfield et al., 2012). This curriculum leverages TinkerPlots (Konold & Miller, 2018), an interactive software that allows students to build sampler devices themselves for carrying out simulations.

Rossman & Chance (2014) give three key recommendations for the choice of software in a course that uses simulation-based inference. First, the software should be easy enough to use so students can focus on learning statistics rather than technology. The appearance and functionality should also be consistent across various data scenarios. Finally, they recommend that software should use aspects like animation that connect to the real-world or physical items being simulated to avoid being a "black box" process. TinkerPlots meets these conditions in several ways, making it an ideal choice for the software to carry out simulations in an introductory statistics course. In TinkerPlots, the process of building and using these samplers is generally easy to use and does not require students to have any coding experience to build custom simulations. TinkerPlots also provides a sense of consistency: while the layout of the various plots and devices is up to the student, the functionality of these aspects is similar across various data scenarios. This process is iterative, and highlights the multi-level nature of simulations and sampling distributions. Students must first work with just one sample and identify the key statistic that will be used to produce a sampling distribution, then they carry out many repetitions of the simulation and create a sampling distribution based on their choice of statistic. Additionally, TinkerPlots leverages animations like many other simulation applets, helping to avoid the "black box" appearance of the simulation process. But TinkerPlots goes a step further by giving students the access to build the samplers themselves. This allows students to gain ownership over the simulation process, as

constructing the sampler requires students to think about exactly what aspects of the data and context are relevant to simulate. This gives students an authentic statistical modeling experience, and building a simulation through this TinkerPlots sampler from the ground up helps to connect their understanding of the simulated results to the sampler that generated them. This truly helps to prevent students from experiencing the simulation as a black box.

There are many other advantages for student learning when carrying out simulations in the TinkerPlots software environment. Comparison studies of various curricula show that CATALST students particularly excel in understanding the purpose of simulation and interpreting their results, even over other simulation-based curricula (Hildreth et al., 2018). This is potentially tied to the use of the sampler device in TinkerPlots and the ownership it provides over the simulation process. Research also shows that students become more engaged as statistical modelers in the simulation process, and are more attentive to narrative features in the presented data and context when creating these models (Noll et al., 2018, 2021). This feature of TinkerPlots makes it a wonderful tool for teachers and researchers alike in revealing and assessing student thinking in various statistical contexts (Watson & Donne, 2009).

The primary limitation of TinkerPlots and the CATALST curriculum is that it does not cover all topics typically presented in the introductory statistics curriculum. CATALST was only intended to cover a limited selection of introductory statistics topics to focus more on statistical thinking and literacy, especially in regards to inference (Justice et al., 2020). Originally designed for middle school students to engage in

exploratory data analysis, TinkerPlots was a relevant choice for CATALST with the introduction of the sampler device and simulations in a later update. However, the scope of TinkerPlots' simulation capabilities makes advanced topics usually included in an introductory statistics class like analysis of variance, Chi-square tests, and linear regression challenging to carry out. The present study is based upon CATALST-inspired activities developed for students to carry out a hypothesis test for the slope of a line of best fit in TinkerPlots. These activities require the use of clunky workarounds in TinkerPlots that enable the production of sampling distributions for a wider variety of statistics, such as the slope of a least squares line. Despite the annoyances students faced with using these workarounds, I found that students were often able to articulate many aspects of inference based upon their TinkerPlots samplers, including how their sampler reflected the null hypothesis, their choice of with or without replacement on the devices, and their interpretations of the *p*-value. Based on the relative success seen with students in this environment, I suggest a technology innovation for improving TinkerPlots to limit students' difficulty with technology.

Motivation and Research Questions

Expanding the CATALST curriculum to cover additional topics typically covered in an introductory statistics course like regression was done to meet a need of a larger study on the CATALST curriculum. Thus, while the cumbersome workarounds I present may make other simulation software alternatives more appealing for learning linear regression, this implementation allowed for students to continue using the same software throughout the course, keeping as much of the processes for conducting simulations as consistent as possible. However, given the positive impacts we observed with students

using TinkerPlots to learn hypothesis testing for the least squares line, I argue that this software may provide value to students despite the obstacles faced.

To reveal these positive impacts with students, I leverage the framework proposed by Case & Jacobbe (2018) that details the challenges students face in understanding inference through simulation. They propose three main levels that students must understand in simulations to capture the logic of inference: the true relationship, the sample, and the sampling distribution. These levels reside within two perspectives: the real world where the data was originally collected, and the hypothetical world defined by the null hypothesis of the test. Through this framework, they identified that students face challenges with distinguishing the ideas of simulation and replication that reflect each of the two perspectives of their framework. Students often saw their sampling distributions as actual real-world samples, rather than a hypothetical distribution based on repeated sampling under the null hypothesis assumption. Students also faced challenges working with the multi-level nature of inference, distinguishing samples from sampling distributions, which can result in difficulties interpreting p-values from their simulated sampling distributions and drawing conclusions. These challenges highlight two potential key areas of focus to assess the impact of these two common challenges: how students view the connection from their sampler models to the null hypothesis, and their interpretation of the p-value from a sampling distribution.

Another feature that is highlighted by the use of TinkerPlots samplers is the focus on the type of inference that is conducted. Most introductory statistics courses focus on sample-to-population inference, where a random sample is taken to make inferences

about what the larger population looks like. There is often not a focus placed on experiment-to-causation inference, where random assignment is leveraged to infer cause and effect conclusions based on the grouping variable that is controlled for. Simulation-based courses offer an opportunity for students to engage in connecting the design of the study presented to the simulation, and findings show that students are able to connect the purpose of the study to the relevant probability model used in a simulation-based course (Pfannkuch et al., 2015). In TinkerPlots, students can choose to use a bootstrap simulation method and sample their devices with replacement to emulate a random sample being taken, reflecting sample-to-population inference. If the students want to reflect an experimental design with random assignment, a randomization test can be constructed which would use sampling without replacement on each device to emulate this random assignment process. This choice gives students an opportunity to think critically about the study's design when building their sampler and connect this to their conclusions.

This previous work gives us three areas to potentially highlight in regards to students' inferential reasoning: their conceptualization of the null hypothesis and how it relates to their TinkerPlots sampler, their understanding of replacement as it relates to the study design, and their use of sampling distribution to interpret a p-value for a test. Using the TinkerPlots sampler as a probability modeling environment may provide advantages to developing these aspects of inferential reasoning in students and addresses the challenges described in previous work. Thus, this study aims to answer the following research questions: What technology challenges do students face when using TinkerPlots to carry out a test on the least squares line, and what can be done to address these

challenges? How does using TinkerPlots for conducting this hypothesis test aid their inferential reasoning and address the common challenges faced when using simulation?

I aim to address these research questions one-by-one in the sections that follow. The next section will address the TinkerPlots technology and the workaround implemented in TinkerPlots, compare this to other applications that can conduct simulations, and propose technology innovations for TinkerPlots. Next, we will look at empirical evidence to show the efficacy of TinkerPlots in addressing the three main challenges in simulation-based courses identified previously in the literature. This aims to provide an argument for why my proposed technology innovation should be considered for TinkerPlots.

TinkerPlots Background and Proposed Innovation

The methods for conducting a statistical test for the slope of a least squares line using TinkerPlots parallel the methods used for conducting a test comparing two populations or groups. Thus, to give a basis for how simulations in TinkerPlots are typically conducted, I will first examine procedures for a hypothesis test comparing two groups or populations. The next section presents workarounds for how to conduct a similar test for the slope of a least squares line, highlighting the cumbersome nature of conducting this in TinkerPlots. Finally, I propose a potential technology innovation for TinkerPlots to streamline this simulation process and avoid this clunky workaround.

Comparing Two Groups in TinkerPlots

The procedure for producing data for a hypothesis test in TinkerPlots generally follows three general steps: creating a sampler to simulate data, creating a plot of one

sample to identify a statistic of interest from the plot, and then simulating from your

sampler many times to generate a sampling distribution. This section aims to illustrate

this process and show evidence that TinkerPlots meets the three main recommendations

for simulation software proposed by Rossman and Chance (2014).

The first step, creating the sampler or probability model in TinkerPlots, requires

identifying the variables of interest to include in this model. When conducting a test

either for comparing two groups or the slope of the least squares line, you have an

explanatory and response variable, and are assessing the association between the two

variables. However, when students are presented with tests that compare two groups, it is

not often presented as a test of association initially, and instead just as a test of a

difference between two means or proportions, which somewhat obscures the explanatory

or grouping variable. However, when creating probability models in TinkerPlots, students

need to recognize each variable and model them as separate devices in the sampler. An

example TinkerPlots sampler for a problem comparing two groups is given in Figure 15.
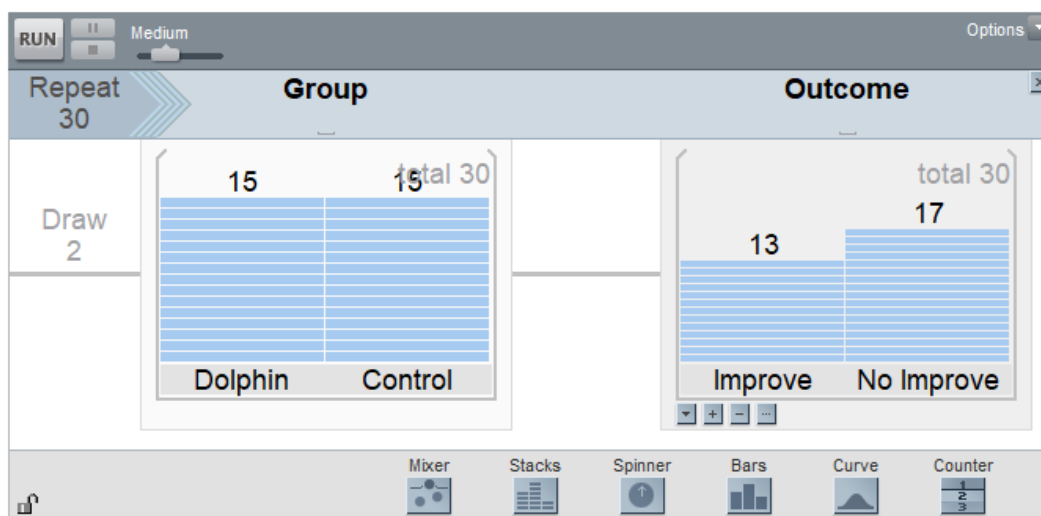
Figure 15. TinkerPlots sampler used to simulate data for the dolphin therapy problem.

This sampler represents data from a study on the efficacy of dolphin therapy for patients with depression (Antonioli & Reveley, 2005). When using this device, cases from each variable are randomly re-paired to each other, which thus represents the null hypothesis, as the two groups are treated no differently as to what cases of the response variable are assigned. This logic can be similarly applied to probability models for linear regression as well. Devices in the probability model can be used to represent the two variables, and placing all cases of each variable in these devices facilitates this re-pairing process.

The use of this device is ideal for meeting Rossman and Chance's recommendation of avoiding simulations from becoming a "black box" to students. Working with probability models like this in TinkerPlots is not only beneficial for seeing their random processes animated, but is also beneficial for giving students the ability to construct this device based on their own thinking. Giving students the freedom to build the sampler from the ground up is the novel part of the CATALST curriculum. This gives students ownership of the simulation process and avoids the sampler from being viewed as a black box. When they run the simulation in TinkerPlots, they have a more intimate perspective of how results are simulated because the students had to design the simulation process themselves.

The next step of the simulation process in TinkerPlots is to plot a single sample of results and identify the statistic of interest. When the device is run, a group card and an outcome card is sampled from each deck and paired together 30 times, producing a new set of sample data under the assumption that the null hypothesis is true. This data can then be plotted and displayed with percentages, as shown in Figure 16. Values like counts
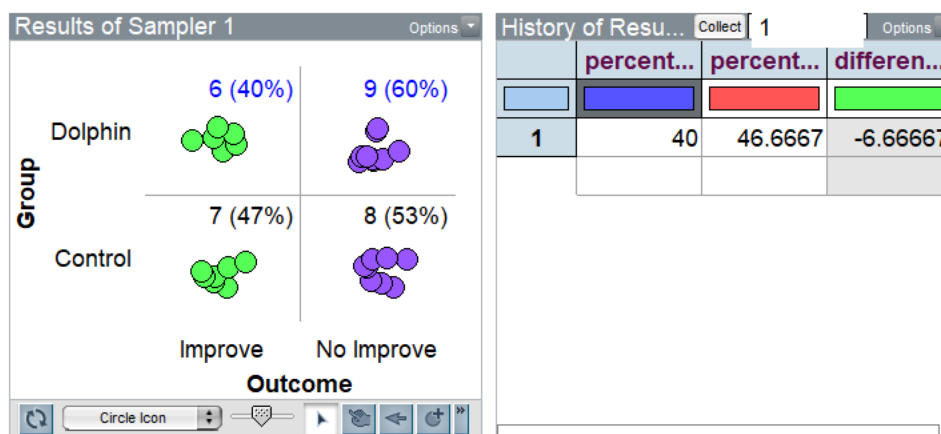
Figure 16. Plot of data and history table for the dolphin therapy problem.

and percentages displayed on plots can then be tracked over many simulations to build up

sampling distributions. In the same figure, a table of the two percentages and their

difference are presented. This design of TinkerPlots allows students to build simulations

in a more iterative way – first by running one single sample, analyzing it to determine the

statistic of interest, and then running the simulation many times. The functionality and

appearance of this process within TinkerPlots is consistent across many different data

scenarios that focus on sampling distributions for a count, percentage, or mean, which

meets another recommendation from Rossman and Chance. Additionally, this design

does not require students to pre-plan every aspect of their simulation before running their

first trial. This is quite different to the simulation process used in the popular Common

Online Data Analysis Platform (CODAP) software, where the all repetitions of a

simulation must be executed before students have the chance to explore their simulated

data further. The process of plotting data and identifying a statistic of interest for a

sampling distribution is important for students to make revisions to their initial

probability model, and this revision process is important to the probability modeling
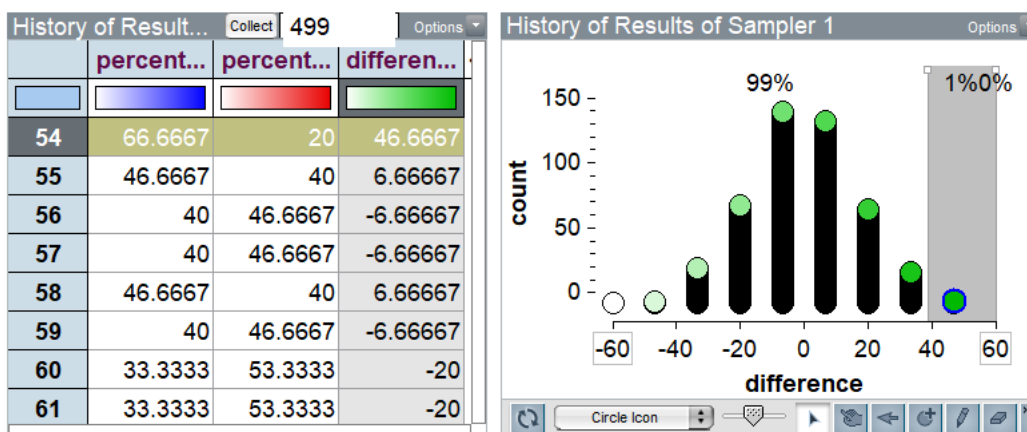
Figure 17. Table and sampling distribution for the dolphin therapy problem.

Finally, once students have identified an appropriate statistic to collect, they can run their simulation many times and create a sampling distribution, as shown in Figure 17. This allows students to then determine an empirical *p*-value based on their simulated trials, as shown by the shaded region in the figure labeled with 1%. Students often have great success with interpreting this *p*-value and drawing conclusions in this curriculum, as the null hypothesis assumption is readily visible to students through building their samplers under this assumption itself. This is in line with comparative curricula research that has shown CATALST students have a deep understanding of the purpose of simulation and interpreting the results of statistical inference (Hildreth et al., 2018).

Rossman and Chance recommend that software should not interfere with the learning of statistics. Overall, the process of building and carrying out a simulation in this data context of comparing two groups or populations is relatively straightforward for students to conduct themselves in TinkerPlots. One could argue that this process may take additional software knowledge compared other simulation software like applets with pre-built simulations. However, modeling itself is an important skill and should be emphasized when learning inference. Students must make connections between their null

hypothesis when constructing their model, which enhances their perspective on the simulations they conduct. The time students engage with the context and statistical assumptions and how they integrate that into their TinkerPlots samplers has the potential to enhance their statistical reasoning with these simulations. Thus, the process of building the TinkerPlots sampler is not simply an additional burden placed on students that may distract from their learning of statistics; in fact, it should enhance their learning of both inference and statistical modeling.

Testing the Slope of a Least Squares Line in TinkerPlots

This process of creating a sampler and producing a sampling distribution is structurally similar across various statistical scenarios in the CATALST curriculum. For the scenario of the slope of a least squares line, this process is initially familiar in the first step when creating the sampler. However, conducting this simulation further requires

Caffeine is a widely used stimulant and psychoactive drug found in many drinks that we consume, and has various effects on your body and health. Researchers collected data to attempt to quantify the effects caffeine has on resting heart rate. The researchers recruited individuals who drink a daily cup of coffee and were able to secure 50 volunteers. Each of these coffee drinkers were randomly assigned an amount of caffeine to be put into their drink. Their heart rate was recorded once before they were given their coffee and once again 1 hour after the drink was first consumed. Using this data, the researchers would like to answer the following research question: **Is there a significant relationship between the amount of caffeine someone drinks and their heart rate an hour after drinking it?** Data on the amount of caffeine given to the patients (mg) and the change in heart rate (bpm) can be found on the *caffeine.tp* data file. (A preview of the data in TinkerPlots is given to the right.)

Collection 1	Caffeine	Heart_R...
1	5	-1.2
2	10	8.4
3	15	-13.4
4	20	4.6
5	25	-6.8
6	30	3.2
7	35	4.2
8	40	7.1

Figure 18. The caffeine and heart rate problem context and data preview.

workarounds that break this familiar structure for students. In this section, we will

investigate a study on caffeine and heart rate, which is given in Figure 18. To build a

sampler that generates under the null hypothesis, the randomization or card shuffling

approach still works for this data scenario, but the categorical variables are now replaced

by numerical variables. Thus, a sampler that utilizes two lottery ball mixers with all of

the various numerical outcomes for each variable would be appropriate to carry out this

randomization, as displayed in Figure 19.

Next, students would plot the data in order to determine a statistic of interest.

TinkerPlots can plot the simulated data in a scatterplot; however, there is not a direct way

to visualize the least squares line in TinkerPlots. The best way to do this is to calculate

the values for the slope and intercept of the least squares line using the formula editor,

and a diagonal line can then be manually adjusted on the line to match these values

approximately. Figure 20 displays this plot with the superimposed diagonal line that is

close to the least squares line calculated in the table. This visualization, however, is time-
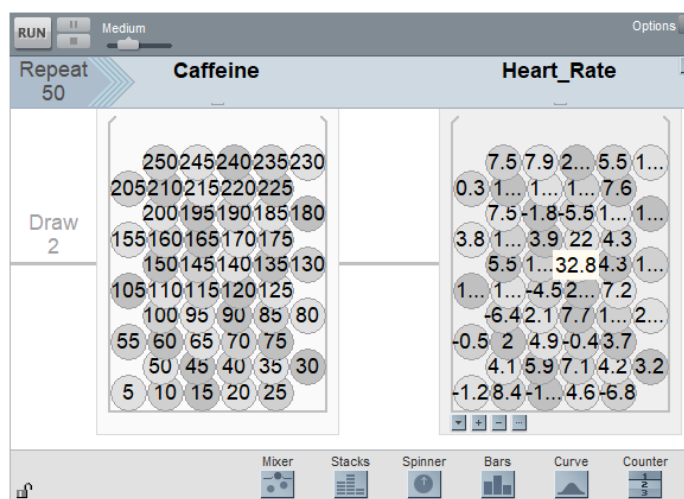
Figure 19. TinkerPlots sampler used to simulate data for the caffeine problem.
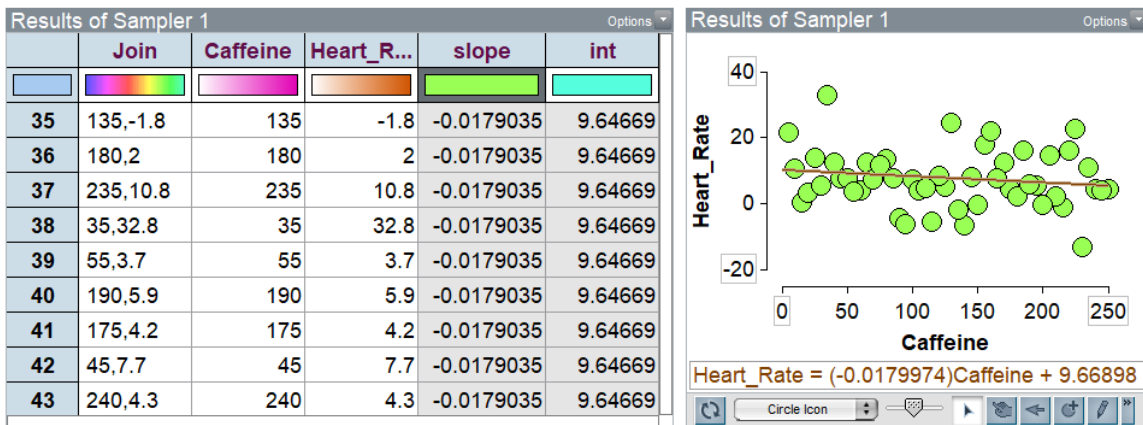
Figure 20. Data table and plot of a simulated trial for the caffeine problem.

consuming to construct, and diverts attention from learning statistical concepts to

wrangling with technology. Additionally, to set up the collection of a statistic over many

trials in TinkerPlots, the plot must display the numeric value to collect on, and this figure

must be set to update when new trials are run. While the slope of the diagonal line

appears on the plot in Figure 21, there is no way to collect on this number, as the line is

placed manually and will not update when a new trial is run. Due to the lack of support

for plotting a least squares line in TinkerPlots, there is no way to easily collect statistics

on the slope of this line.
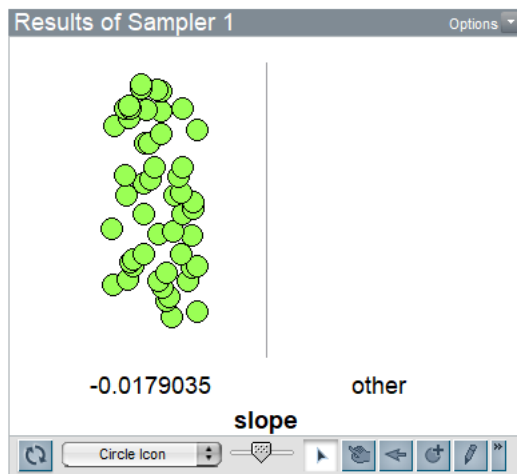
Figure 21. Plot of the slope variable, using an erroneous categorical axis.

To escape this issue, a workaround can be conducted with TinkerPlots. This workaround is not elegant, but will facilitate the building of a sampling distribution for the slope. It should be first noted in Figure 20 that the calculation of the slope of a least squares line creates a column in the table which repeats that slope value across the entire column. Thus, one could imagine plotting these values on a distribution, display the mean of the distribution, and collect statistics on the mean. This mean value is identical to the slope, as it is just the mean of many copies of the slope value itself. However, in the stable release of TinkerPlots version 2.3, TinkerPlots identifies the variable as categorical rather than numerical, seemingly due to the nature of the variable having identical values across all entries. Thus, as a result, there is no numerical axis that is plotted, preventing a mean from appearing on the plot, as shown in Figure 21. To circumvent this issue, you can follow the steps outlined in Table 23, which performs a "bait and switch" method with the equations in TinkerPlots to ensure that a numerical axis is used. While this method is functional, it is often confusing to students, especially CATALST students who are accustomed to the functionality of TinkerPlots directly connecting to the statistical concepts they are learning. While performing the workaround in Table 23 and manipulating the formula editor in TinkerPlots, one student noted "I kind of don't understand why we're doing this... I don't know what the narrative purposes are with these columns [that are calculating the slope]." This workaround is a strictly procedural method to produce the sampling distribution, which was clearly uncomfortable with this student who wanted to better see the meaning of this slope value and why this procedure was being carried out in TinkerPlots. However, once these procedures were set up with support from the instructor, students thrived being able to carry out this test in a familiar

Table 23. Instructions for setting up collecting statistics on a slope.

Description	Image
Create a new column for the slope in the data table. Right click on the header for the column and select "Edit Formula."	
Rather than use the formula for the slope (LinRegrSlope) in TinkerPlots right away, simply type in the name of one of the variables into the formula editor. For this problem, one could use "Caffeine" as this variable.	
Create a plot of this new slope column in TinkerPlots. Be sure to fully separate all the dots by dragging a dot all the way to the right so that the bins disappear and an x-axis appears.	
Right click on the header for the slope column again and select "Edit Formula." Enter in the formula for the slope (LinRegrSlope, can be found under Functions, Statistical, Two Attributes) and click OK.	
The plot of the slope will now reflect the slope value and the numerical x-axis persists. Enable the mean tool (triangle button) on the plot. Right click on the triangle and select "Collect Statistics" to begin tracking the slope value.	

modeling environment, which will be explored with the data presented later in this chapter.

If using the TinkerPlots version 3 release, this method described in Table 23 is unnecessary, as it will use a numerical axis by default. For this reason, I highly suggest using TinkerPlots 3 with students if you plan on implementing this workaround with students to give them the easiest software experience possible with TinkerPlots. However, it is worth noting that version 3 is an ongoing re-build of the original TinkerPlots 2.3, and so there may be bugs or with this version. Additionally, there are certain features that you may require in TinkerPlots 2.3 (e.g. robust copy/paste features, scrolling workspace, etc.). Currently, both versions are offered on the TinkerPlots download page due to the differences in stability and features. This method can be used on a variety of measures that TinkerPlots can calculate with the formula editor as well. There are built-in functions for many other statistical measures that could be used to build up a variety of tests using TinkerPlots software.

Proposed Technology Innovation

To best leverage the potential CATALST and TinkerPlots have in engaging students in probability modeling for linear regression, I propose a future technology innovation for the TinkerPlots software to better facilitate the building of an empirical sampling distribution for the slope. To address the complications of collecting statistics on the slope in TinkerPlots, the diagonal reference line tool should have the ability to lock-in at the position of the least squares line. There is currently a feature that locks the line at the origin in TinkerPlots, allowing for one remaining degree of freedom when

manually adjusting the line. This feature should work similarly to that, but would not allow further adjustments to the diagonal reference line when enabled. Figure 22 shows the proposed functionality for the diagonal reference line tool. This functionality is proposed in such a way that it does not replace the diagonal line tool, and simply gives the option to visualize a least squares line on a plot. The ability to informally place lines and adjust them is valuable to students' development on the concept of the least squares line. Activities that have students place lines informally and establish criteria for how well a line fits the data is an important process in building up the motivation for the least squares line and how that line is determined.

Using this functionality to place a line of best fit can be done quickly to create a plot of the observed data with a least squares line, allowing for a brief exploration of the data. Currently, the only way to plot a least squares line in TinkerPlots is to separately calculate the slope and intercept, and then place a diagonal line on their plot to approximately match these values. This feature should also be useful for plotting
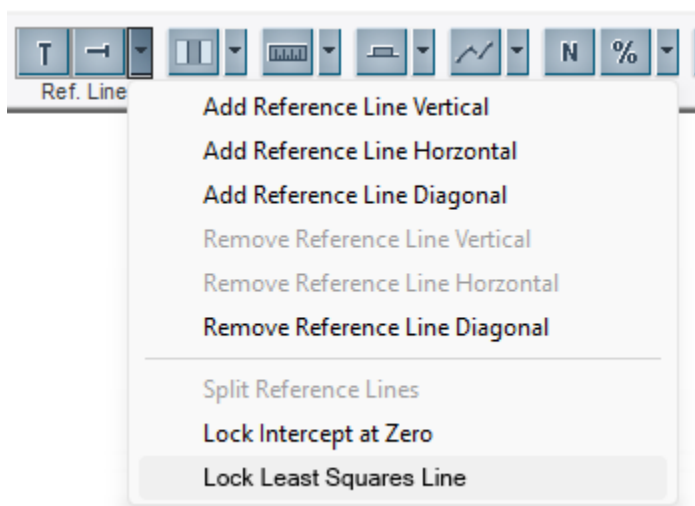
Figure 22. Proposed functionality for the least squares line in TinkerPlots.

simulated data, determining a statistic to collect, and building an empirical sampling

distribution. This proposed functionality is shown in Figure 23. When this new locking

feature is enabled, the equation of the least squares that appears in the plot should enable

interactions with the slope and intercept for collecting statistics in a history table. When a

statistic is enabled for tracking in a history table, it can be clicked on such that a grey box

appears around that statistic, and right-clicking on it will allow for the "Collect Statistic"

option to be selected. Functionality should also be added for collecting statistics on the

intercept value as well. While statistical tests for the intercept are often not the primary

focus when teaching linear regression, having the option available to students forces a

choice to be made. If the technology only allowed for the slope to be collected, students

would simply begin creating a distribution based on the only available option, rather than
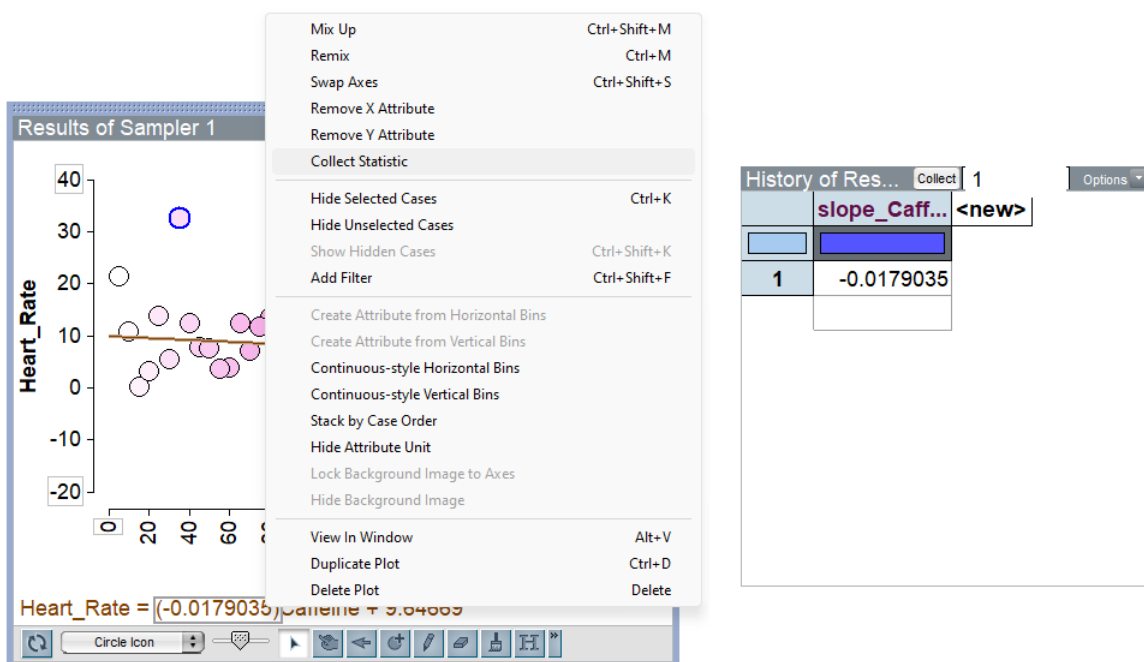
Figure 23. Proposed functionality for collecting statistics on the least squares line slope.

think critically about the equation of the line and decide which statistic will assess the relationship between the two variables.

Considering Software Alternatives

Given the cumbersome nature of conducting a test like this in TinkerPlots, one may ask why other software options are not considered for the purpose of conducting a simulation-based test for the line of best fit. The previous section identified two features of this analysis that are cumbersome with TinkerPlots software: creating the least squares line on the plot of observed data, and setting up the collection of statistics on a slope. Many existing software packages exist currently that have these features. However, we argue that the benefit of students creating the sampler device and simulation process in TinkerPlots is a worthwhile, unique feature of this software. In this section, we will highlight two other popular choices for simulation-based inference: CODAP and web-based applets.

As mentioned previously, CODAP is a very popular statistics software package due to its focus on data science through its use of nested data structures. However, its potential for simulation is currently lacking. The sampler plugin available in CODAP currently only supports simulating from one device. The TinkerPlots samplers for comparing two groups or linear regression as shown in Figures 15 and 19 each have two devices, so that values from the two variables can be re-paired with each other randomly. While multiple variables from a data set can be placed into this single device, the outcomes of each variable for a single case are inherently linked, thus preventing any simulation under the null hypothesis through randomization or bootstrapping.

Additionally, as previously mentioned, the hierarchical nature of CODAP's data structures make it difficult to emulate the iterative nature of simulation in TinkerPlots. Students must determine the number of repetitions up front before a statistic has been determined, which can create an overwhelming amount of information to process immediately. This can potentially interfere with the cyclic nature of probability modeling, where testing an initial model and reading the output of single samples often informs revisions to that probability model.

There are also many web-based applets available to conduct simulations for the slope of the least squares line, many of which share very similar features. In this paper, I consider the applets developed by Rossman and Chance (2014), as they have the most features for exploratory data analysis. An example simulation using this applet can be
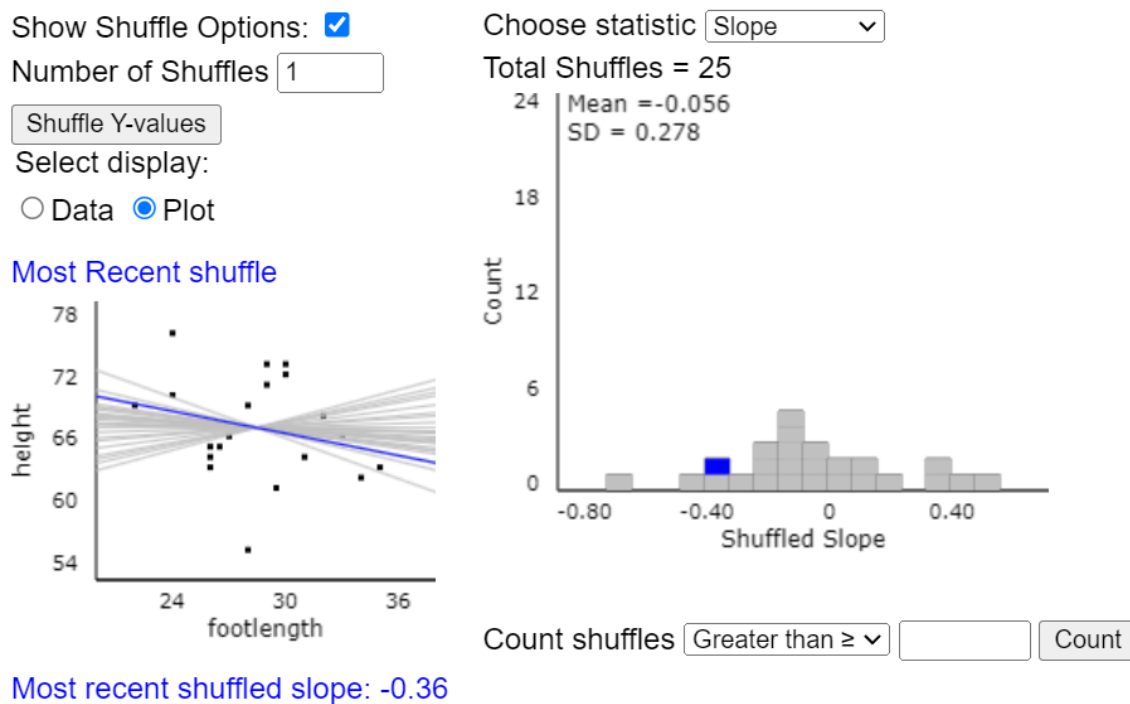
Figure 24. Simulation of least squares slope in Rossman and Chance applet.

seen in Figure 24. While this applet addresses many feature gaps in TinkerPlots, like directly plotting a least squares line on a scatterplot and easily simulating a sampling distribution for a slope, students do not have any authority in building the simulation process in these applets. After loading the relevant data into the applet, the probability model used for simulations is pre-constructed and obscured to students. While the visualizations of individual samples and the sampling distribution are well-connected by highlighting the current blue least squares line and blue observation in the sampling distribution, there is little connecting the observed sample to the data generating process. Students simply click the "shuffle y-values" button that is given to generate their sampling distribution. Instruction can assist in this connection to the simulation process, but students do not get to actively think about this process for themselves or consider the type of inference done in the study and how that may connect to the choice of simulation. Additionally, there is no opportunity for students to engage in modeling with this applet, which is an important goal of learning statistics itself.

To summarize the options with available software, Table 24 organizes these features of TinkerPlots, CODAP, and the Rossman and Chance web applet. Based on the options presented in this table, one might propose that an alternative technology innovation could be to further develop CODAP's sampler plugin so that the sampler's

Table 24. Summary of software and features for testing the slope of a least squares line.

Software	Plot LS Line	Simulation	Collect on Slope
TinkerPlots	Manually	Yes, fully customizable with sampler	Yes, but cumbersome
CODAP	Yes	No, cannot simulate multiple variables	Yes
Web Applet	Yes	Yes, pre-built	Yes

features are similar to TinkerPlots. This would not require any further work outside of the

applet, as CODAP already supports calculating the slope of each sample and creating a

sampling distribution within its existing data structure. However, the previously

identified features in how the hierarchical nature of CODAP's data structures and

requiring the number of repetitions up-front may interfere with the iterative and cyclic

nature of the probability modeling process. The web applets are capable of carrying out

all of these aspects currently, but are limited by the model being obscured to students and

thus making it difficult for students to know what is being simulated. TinkerPlots can

perform all three of these aspects of plotting the least squares line, carrying out a

simulation and collecting on the slope, but is not as easy to use as the alternatives. If

implemented, the technology innovation presented would make TinkerPlots an ideal

software tool for teaching linear regression, including its powerful and unique sampler

tool. My aim in the rest of this paper is to provide evidence that TinkerPlots can

strengthen students inferential reasoning with linear regression, even with using the

current cumbersome workarounds.

Methodology

This section details the perspective on learning I take in analyzing data from this

study, the data collection instruments used, participants in the study, and the analysis

done on the data.

Learning Perspective

The CATALST curriculum leverages carefully scaffolded activities that have

students work in groups to discover new statistical concepts in TinkerPlots. Within the

classroom, I take a social constructivist view to learning. Students come with many pre-conceived notions about statistical and probabilistic concepts, and this impacts how they might model a problem in the TinkerPlots sampler. In a group context, the experiences students bring to the course individually affect their own experiences with the activities and the data contexts, and this results in a sociocultural-based discussion of relevant contextual elements that are important to model. These activities also leverage students' zone of proximal development to allow for students to experience novel concepts for themselves rather than be directly instructed on how to approach them.

However, for the purposes of assessing students' learning, I am more concerned about individual knowledge, representing a cognitive approach to learning. These two approaches can be reconciled, as social constructivism involves students shuffling between interpsychological and intrapsychological levels, where students bring their individual experiences to a social setting and center learning within a group of students. Knowledge from this group setting is then internalized once again through this experience. Thus, to assess what knowledge students have internalized from their classroom experiences, the primary source of data we will examine is student responses to final exam questions. This will be supplemented with select episodes from group discussions in class.

Data Collection Instruments

This study focuses on an introductory statistics course using the CATALST curriculum. This study focuses on two data sources, in-class screen recordings, and

student responses to the course assessment. The following subsections detail these two instruments.

In-Class Recording. Students participated in the caffeine activity described previously in Figure 18. Students were placed into groups of 3 or 4 students to work through the activity and write up their responses in a shared google document. This activity, like many other CATALST activities, is constructed to allow students to explore ideas for themselves, discuss among their peers, and then bring these ideas to a full class discussion. Questions on the activity ask them to pose conjectures about data, build their samplers and justify the choices they make in creating them, and explain the results they generate from the samplers and their relevance in answering the statistical questions at hand. Groups of students that consented to be recorded had their screen sharing and audio recorded via Zoom. Students' written work from the activity was also collected.

Assessment. The assessment of focus for this study is the final exam, which occurred just a week after completing the linear regression content in the course. The final exam had students complete a full investigation that leveraged linear regression techniques similar to the activity. Students were instructed to complete this assessment with no other outside resources besides TinkerPlots, but given that this assessment was conducted remotely through Zoom, it is impossible to know for sure if students followed these directions. Unlike the activity's context that leveraged experimental design and random assignment, the context posed to students had students assess a random sample of diamonds from Singapore for the relationship between their carat and price (Chu, 1996). This gave students an opportunity to model a scenario with a different study design in

linear regression, which will challenge students to model the choice of replacement in their sampling process to best reflect the study design.

To best reflect the motivations for this study, this study focuses on three questions from the assessment. The first question asks students to determine how their sampler device reflects the null hypothesis, in order to gauge their understanding of whether they recognize the sampler generating data under the hypothetical perspective that there is no association between price and carat. The next question of focus asks students to justify their choice of replacement for the sampler, to capture the idea of study design as just discussed. Finally, we focus on their interpretation of the *p*-value, which will aim to help determine students' ability to decipher their results from their simulated sampling distribution. The relevant details of this assessment can be found in Appendix A.

Participants

This study focuses on one CATALST classroom of 23 students in the second 10-week course of an undergraduate introductory statistics sequence. Of those 23 students, 21 consented to participate in the study. This course is targeted at non-statistics majors, most of which come from a social science background. Some students in the course may have had some prior statistics knowledge from high school or other courses in their own departments, but for the most, this course is their primary exposure to statistics in college.

In the caffeine activity conducted in class, the 23 students were divided up into seven groups. Six of these groups were composed of all consenting students who, and had their screen and audio recorded via Zoom. Due to the nature of this course being taught remotely during the COVID-19 pandemic, group discussion was often not as vocal as an

in-person classroom, with many off-topic discussions or silent writing in their google document. As a result, the presentation of results will focus on just one of these six groups that was consistently vocal about their thinking through the activity. The students in this group are referred to by the pseudonyms Micaiah, Dabney, and Riley. All 21 consenting students in the course took the final exam, and their responses to the three questions of focus were analyzed in this study.

Analysis

Analysis of the survey responses began with the development of coding structures for student justifications in each of three assessment questions: how their TinkerPlots sampler reflects the null hypothesis, their choice of replacement and justification, and their interpretation of the p-value. The choice of these three assessment questions was made in light of the Case and Jacobbe (2018) framework that highlighted common challenges in simulation-based courses, as well as the experiment-to-causation inference

Table 25. Connection between the research literature and the assessment questions.

Research Literature	Challenge Identified	Assessment Question
Case and Jacobbe (2018)	Recognizing the simulation as a simulation under a null hypothesis assumption rather than a replication of real-world data.	Connecting a TinkerPlots sampler to the null hypothesis
	Discerning between a sample of data and the sampling distribution, and interpreting the difference between these.	Interpreting a p-value
Pfannkuch et al. (2015)	Recognize the difference in meaning and interpretation between experiment-to-causation inference and sample-to-population inference.	Justifying the choice of replacement for a TinkerPlots sampler

aspects highlighted by Pfannkuch et al. (2015). Table 25 highlights the aspects of each of these pieces and how they connect to the asessment questions I focused on.

For all three questions, responses were read and an open coding process was conducted to highlight interesting themes for each of the three assessment questions. Based on these themes, three coding schemes were developed. The following subsections detail these coding schemes for the three concepts presented.

Null Hypothesis. To assess students' understanding of how their sampler device represents the null hypothesis, students were asked the following question: "Describe how your [sampler] model reflects the null hypothesis. Be sure to give a detailed explanation in the context of the problem." An initial reading of student responses was conducted to determine overall themes to their explanations. An overwhelming majority of students gave some explanation that describes how the model paired values from each variable at random with no biases toward certain types of pairings, thus assuming no relationship upon the data. Those students who didn't give this type of explanation often claimed that the model assumed no association without the background of this re-pairing process. Another alternate explanation students gave was attributing the choice of replacement to the nature of how it reflected the null hypothesis. As a result, these three types of responses were the basis for coding, which is detailed in Table 26. Codes assigned to student responses for this structure are unique, so all responses are assigned exactly one of these three codes. Given that the "re-pairing" code was applied the overwhelming majority of student responses and that only 21 students were studied, this

Table 26. Coding scheme for responses to null hypothesis question.

Code	Description	Example Response
Re-pairing	The student conveys that their model pairs values from each of the two devices in a way that assumes no relationship.	My model will simulate randomly sampling diamonds (their sizes and carats and prices) from retailers, randomly pairing carats and prices in such a way as to simulate there being no particular relationship between the two. This is in order to assess whether such a seemingly strong positive correlation could be explained away by random chance.
Assumption	The student states that their model assumes no relationship, but does not explain anything pertaining to how the model ensures this.	This model reflects the null hypothesis because we are assuming that under the null there is no correlation between the variables being tested of carat and price. Creating this model, I am going under the assumption from the information of the study that these diamonds given are a random sample. So with using the random sample method it allows me to test under the null hypothesis at random.
Replacement	The student justifies that the model reflects the null hypothesis due to their choice of replacement.	My model reflects the null hypothesis because with both devices being set to "without replacement", it assumes that there is no correlation between the carat of a diamond and its price.

surely does not reflect the complete scope of possible student responses, but it does help

to provide a view of students' thinking in this course.

Replacement. Students were asked the following question on their assessment:

"Describe how you chose replacement for the [sampler] model, and how you think this

best reflects the context of this problem." To justify the choice of replacement in this

scenario, students provided a variety of explanations in their responses, especially

depending upon the choice of replacement that they initially made. Students can justify

the use of with replacement by referencing the study design as reflecting an observational

study using random sampling, which is best emulated through bootstrap sampling with

Table 27. Coding scheme for responses to replacement question.

Sampling	Code	Description	Example
With	Sampling	The student references that their model emulates the idea of taking a random sample or re-sampling.	I chose to do replacement for both, because the study says that we should assume random selection of diamonds.
	New data	The student desires their device to emulate the aspect of finding new observations (diamonds) in their samples.	When re-randomly sampling 48 new diamonds from population (in newspaper from retailer) per trial, they will have NEW overall data, so they are not FIXED from trial to trial.
	Obs. Study	The student references that their model represents an observational study (not simply just "study"), or references that their choice best represents the original study's design.	I chose to put both devices with replacement so that the original conduction of the study is represented.
Without	Random Assignment	The student references that their model represents the process of random assignment.	I chose "without replacement" for both of my devices because I wanted to simulate prices being randomly assigned to carats
	No duplicates	The student desires to have all data points represented once and only once in their simulated data.	Given we are comparing two variables, repeating values would not be ideal as the comparison may be thrown off if the same values were being seen over and over again.
	Experiment	The student references that their model represents an experiment, or references that their choice best represents the original study's design. Be sure they say experiment to mean study design, and not as a term synonymous with terms like "study" or "research."	The sampler was set to without replacement because it is an experiment with random assignment.

replacement. Students also can reference the idea that the re-simulation process under the null would require some idea of re-sampling new subjects. If a student chose without replacement, they could also similarly reference the study design, which would reflect an experimental design leveraging random assignment. They may also reflect a desire to have all values selected with no duplicates, as done in a typical randomization test. To encompass these ideas that emerged in the data, we coded student responses with the coding scheme presented in Table 27. After the type of replacement that the student used is identified, reasoning codes from this table were applied to their response. Unlike the codes for the null hypothesis, responses may be assigned multiple reasoning codes, as students often gave multiple justifications represented by these codes for their replacement choice.

p-value. Students were asked the following question on their assessment: "Explain what the p-value means in the context of the problem. Do not simply state how strong or weak the evidence is, give an interpretation of the percentage or probability that you found." To evaluate responses to this question, I identified and analyzed three main components of their statement: a reference to the observed slope, an inequality statement (e.g. the observed slope value *or something larger*), and a reference to assuming the null hypothesis is true. For each of these three components, a separate coding scheme was created to assess students' responses. These coding schemes for each of these three components are detailed in Table 28. Each response was assigned one code from each of the three components to characterize each student's interpretation of the p-value for a total of three codes per response.

Table 28. Coding scheme for responses to p-value question.

Component	Code	Description	Example
Observed Slope	Present	References the actual slope value in their study or calls back to the slope from the original study.	"There is a 0% chance that I would be able to obtain the slope from the original data…"
	Generic	References the *p*-value being about some undefined value or other statistic from original study.	"Because the p-value was 0%, this means that it is impossible for the results of the study to have been by random chance."
	Missing	Does not reference a value from original study.	"We can conclude that the p-value of 91% indicates strong evidence against the null hypothesis."
Inequality	Present	Explicitly gives detail that p-value looks for the slope value in the original study or more than what was provided.	"There is a 0% chance that we would get a slope value as steep as 3721.02 or more under the null."
	Implied	Inequality is potentially implied, as observed slope was not close to simulated slope statistics. Students reference that their simulation did not produce slopes anything near the original or observed value.	"Our original slope of about 3,700 didn't show up a single time, meaning that the samples with no relationship were far from close to what the original data shows."
	Missing	No inequality referenced.	"There is a 29% chance that the price of diamonds is dependent on the carat size."
Null	Detailed	Gives a detailed statement of the null hypothesis in their interpretation.	"…assuming the carat and price of a diamond have no association (under the null hypothesis)."
	Contextless	References that the *p*-value assumes the null to be true without stating what that null is.	"Under the null hypothesis results is valid, that we wouldn't get the original value of the slope being 3721.02 and above from the data I have collected."
	Random Chance	References that the p-value is the result of what happens randomly or by random chance without explaining the random chance process.	"…a zero percent chance of the data's original slope occurring randomly."
	Missing	No reference to the null or random chance.	"we have around a 0% chance that we would get results of the same as the original study or more extreme."

Analyzing student responses for these three components not only gives an opportunity to assess students' conceptions of the *p*-value itself, but it also reveals how students recognized the various perspectives and levels of simulation-based inference identified in Case and Jacobbe's framework. A recognition of their *p*-value as being a probability about the slopes of the least squares line gives evidence that the student recognizes the sampling distribution as representing this statistic rather than the observed data itself. And recognizing the null hypothesis in their statement also shows evidence that if students hold the appropriate hypothetical view of the null hypothesis, then these students view the data in the sampling distribution as representing a hypothetical perspective. While the latter point is assessed in their responses to the null hypothesis question, this gives some additional verification that this hypothetical perspective held throughout the simulation and their analysis, and that they could manage these perspectives throughout the analysis.

Results

On the whole, students were able to procedurally carry out a test for a least squares line with great success on their assessment. Of the 21 students who took the assessment, all 21 were able to create an appropriate sampling distribution based on their TinkerPlots samplers, with 19 of them identifying the correct *p*-value and drawing an appropriate conclusion. For the two students that did not reach an appropriate conclusion, one made a reading error on the slope of the least squares line (used 370 instead of 3700) which affected their observed statistic that their *p*-value was based on, and the other used their sampling distribution to make an empirical confidence interval of their simulated results. While a confidence interval approach could be a valid method in this scenario,

this students' distribution was still generated using a sampler that reflected the null hypothesis rather than bootstrapping on the observed data.

However, this does not necessarily reflect students' understanding and interpretations of the procedures they conducted. To understand this, the following subsections will investigate the results of student responses to the three coding schemes developed for the null hypothesis, replacement, and *p*-value. I will then follow-up to discuss technology difficulties students faced with TinkerPlots when carrying out the test for the least squares line.

Null Hypothesis

Overall, students were mostly successful in identifying how their device represented the null hypothesis that a diamond's carat was not correlated to the price. The table of codes for student responses is given in Table 29, which shows that 81% of students were able to recognize that their sampler device in TinkerPlots was randomly re-pairing values from each variable, thus representing that there is no association. Many students were quite explicit about defining the process not just as "random," but stating that this re-pairing process is independent, and does not bias certain pairs of values that may show an association. One student provided the TinkerPlots sampler shown in Figure 25 and gave the following response:

Table 29. Counts of codes for responses to the null hypothesis question.

Code	Count
Re-pairing	17 (81.0%)
Assumption	2 (9.5%)
Replacement	2 (9.5%)

Regardless of the carat size randomly selected in the right device, [this will have] no effect [or] have an impact on the price of the diamond (can randomly sample any price from very low and freely pair it with any carat size). That way the diamond price is determined randomly, in an independent way from the carat, showing no linear association.

Another student explained this idea of independence by stating that "This pairing is not weighted in any way (i.e. higher carat being paired with higher price)." While these responses show great promise in students' probability modeling techniques in managing assumptions, these statements do not guarantee that students are considering their null hypothesis assumption throughout the analysis. An examination of students' *p*-value interpretations will potentially triangulate this data and reveal if they potentially carried this assumption through the analysis.
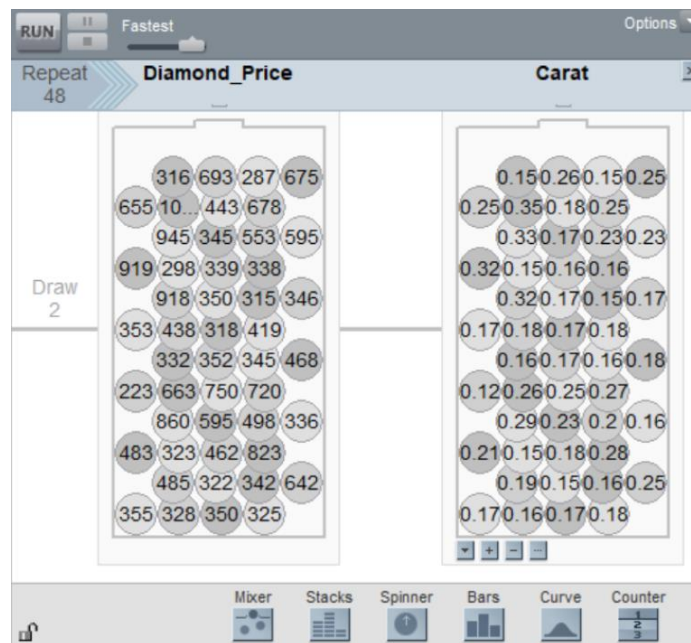

Figure 25. TinkerPlots sampler for diamonds and carat provided by student

Replacement

Students were more divided on the choice of replacement for this particular

scenario, as shown by the counts of codes assigned in Table 30. A majority of students

chose to use with replacement on their TinkerPlots devices, but many still chose without

replacement. As the in-class activity on caffeine and heart rate where this simulation

method was originally introduced leveraged sampling without replacement due to the

experimental design, students may have received the impression that it was ideal to use

random assignment and have no duplicated values in their samples to best simulate the

data. This seems evident in this student's response:

> *I chose without replacement for both categories as I want the outcomes to simply
> be randomly paired up without double dipping… if we were to do with
> replacement, it would not be adequate to use the [original study's results] and the
> new [simulated] information if the info is not kept the same… Also, I think we are
> using a Linear Association deal and that is done without replacement.*

This student gives a justification that reflects that the simulated samples best reflect the

original study's results if no duplicates are used in the simulated sample data, and follows

up with the idea that testing for a linear association is done without replacement. While

there were two students who did mistake the diamond scenario for being an experimental

Table 30. Counts of codes for responses to the replacement question.

Sampling	Code	Count
With (13/21)	Sampling	13/13 (100%)
	New data	5/13 (38.5%)
	Obs. Study	4/13 (30.8%)
Without (8/21)	Random Assignment	7/8 (87.5%)
	No duplicates	7/8 (87.5%)
	Experiment	2/8 (25%)

design, most students who chose without replacement seemed to think that this choice was inherently linked to testing for linear associations. This may be due to the only in-class activity on testing for a linear association was the caffeine and heart rate study, which was experimental in nature and leveraged without replacement in the TinkerPlots sampler.

For students who chose with replacement, every student attributed their choice of replacement to the idea that their sampler should emulate random sampling. While not many students explicitly stated the study design (observational study) or made it clear that they wanted this to represent a new set of data from a larger population, every student who made this choice appeared to identify the random process associated with the choice of replacement, like this student:

> *I chose with replacement on both explanatory and response variables in order to simulate a random sampling from retailers such that any combination of carat and price is equally likely rather than simply simulating resampling from this same data set where there are a fixed number of diamonds of particular carat or price.*

This student recognized that with replacement allowed for samples within each variable that could have varied sets of data within each variable, which is akin to bootstrap sampling. While they did not reference the exact study design or see the need for samples to reflect a new set of diamonds, this student shows evidence that they understand the purpose of their choice of replacement relevant to the key random aspect that they are trying to emulate, which is random sampling. The following student's response gives an example of a student who did give this level of detail:

> *Both devices are with replacement, to represent how the original study was conducted (by randomly sampling 48 diamonds from the newspaper (random*

sample sold from retailer which we can later generalize to our findings to)), and record their carat size and price. Both are with replacement hence, because when re-randomly sampling 48 new diamonds from population (in newspaper from retailer) per trial, they will have NEW overall data, so they are not FIXED from trial to trial, and we have some variability (this is what is being represented in this model too). Also, even though this is an observational type study…

This student's response shows reasoning about how the choice of with replacement allows the model to best represent that each sample they generate represents a new set of diamonds, and cites that they want to best represent the original study design, which they define later as an observational study. This student also recognizes the need for their sampler to produce "new" data, in that it represents a new random sample of diamonds, rather than the existing diamonds having the existing prices randomly assigned to them. This was only made explicit in 5 of the 13 student responses that chose with replacement, but reveals another aspect to students' thinking about the impact of the choice of replacement and how this relates to the study design.

Another important feature of replacement that was not fully accounted for by the coding scheme is that students are often willing to reference previous activities and the study designs they used and relate that as confirmation for their choice. As seen by many of the example responses already, students' reasoning is firmly rooted within the data context, and these contextual elements help students draw comparisons across various study designs and the TinkerPlots samplers they use to analyze them. One student who chose with replacement on their assessment said that "I am not trying to shuffle cards, a la Dolphin Therapy. This is similar to the hybrid car mpg/price model we previously worked on." This student not only referenced to a previous activity on linear regression,

but compared also to study designs for comparing two populations like the dolphin therapy study.

Students drawing connections to previous activities and data contexts was also observed in the classroom as students were first learning this concept in their activity on caffeine and heart rate. This episode here now shifts from the assessment data to the classroom activity on caffeine and heart rate that came earlier in the course, which was experimental in nature rather than an observational study like the diamonds context. When this group began the activity and read through the context of this caffeine study, Riley initially made the connection to the dolphin therapy study that they previously examined in class:

Riley:	[The subjects] were put into different groups, so basically, they are groups, right? It's like the dolphin study again. They're put into different groups because they weren't given like some random numbers of caffeine but they were put into, you know, specifically assigned an amount of caffeine.
Dabney:	Was it? Or you because you can see the [data] file already. You looked at the file?
Riley:	No, I'm looking at [the context written in the activity]. It's like they were randomly assigned some amount of caffeine but I'm assuming -- oh, maybe it's not groups. Yeah.
Dabney:	Yeah, I think it's the opposite, at least based on this limited thing we've read.
Micaiah:	You're saying it's one group?
Riley:	Oh, yes. Yeah, it is. It is a large number of numbers. Yes, it's not groups.

While Riley's initial assumption that the subjects were going to be placed into groups was discovered to be incorrect after looking through the data file, they did recognize the random assignment aspect of this study, and connected that to the dolphin study they did

previously in class. This conversation about the random assignment aspect of this study

continued when they began constructing their TinkerPlots sampler:

Riley:	Basically, we're going to shuffle the heart rate and caffeine. Is that all we're doing?
Dabney:	I think so. Wait, shuffle heart rate and caffeine. Like the dolphin therapy?
Riley:	Yeah, basically.
Dabney:	Because it's an experiment?
Riley:	Yeah.
Dabney:	So we're gonna do without replacement for both, right?
Riley:	Why?
Dabney:	Because it's an experiment.
Riley:	Seems to me like you want to use all the ones that are there just once each, right? Because you're disassociating them from each other.
Micaiah:	Yeah. So we would say without replacement.
Dabney:	Without replacement. Right? Isn't that what I said initially?
Riley:	Oh, I thought you said with replacement.
Micaiah:	That's what I thought too, I must have misheard you.
Dabney:	Without replacement. Maybe I did. But what I meant was, we are not adding any new -- it's like that we're just tearing the cards up and mixing them. We're not pulling a new [sample]. Okay. Maybe I did say it the wrong way, but that is conceptually what I meant.
Riley:	There you go. (Riley begins constructing a TinkerPlots sampler as shown previously in Figure 19.)
Dabney:	Okay, and we're doing that because this is an experiment [rather than] an observational study. And so we want to work with the data points we already have.
Riley:	Yes. Yes. Because we want to work with the data points we already have we don't want to -- we're not like simulating sampling

> a million times, right? [We're not using] different samples. We're
> still using the same people.

These students here correctly identified the experimental nature of the caffeine study

where subjects are randomly assigned to varying levels of caffeine. This was relevant to

them as they built their sampler, as they wanted it to reflect this random assignment

process, which requires sampling without replacement on both devices to ensure all data

values from both variables are used once and only once.

P-value

As mentioned previously, 19 of the 21 students were able to procedurally find the

correct *p*-value in TinkerPlots based on the samplers they built. However, their

interpretations of this as a probability may have lacked some precision. The counts of the

codes assigned to the students' *p*-value interpretations are given in Table 31. A majority

of the class did seem to include some statement about the observed slope value from the

original data in their response, with some other students referencing some value vaguely

that wasn't clearly defined as the slope. Not as many students clearly identified the

inequality of the *p*-value, but given that the observed slope was clearly outside of the

Table 31. Counts of codes for responses to the p-value interpretation question.

Component	Code	Count
Observed Slope	Present	14 (66.7%)
	Generic	4 (19.0%)
	Missing	3 (14.3%)
Inequality	Present	9 (42.9%)
	Implied	7 (33.3%)
	Missing	5 (23.8%)
Null	Detailed	7 (33.3%)
	Contextless	3 (14.3%)
	Random Chance	5 (23.8%)
	Missing	6 (28.6%)

range of students' simulated slope values under the null hypothesis, many students stated that their p-value was 0% or close to 0% because of this. It is difficult to know if students would have considered an inequality in their statement if there were multiple extreme cases to consider in their sampling distribution, but these statements seem to imply that they may be considering this aspect.

With regards to the null hypothesis in their p-value interpretations, most students referenced some idea of the null, but again, many lacked precision in their response. While seven of the students explicitly referenced their detailed null hypothesis in their statement, an additional eight referenced the idea of the null hypothesis without any further context, or referenced the idea of "random chance" without explaining the random chance process exactly. Regardless of the level of precision, these students show evidence that they may be reasoning about this sampling distribution as an artifact of the hypothetical world by citing this assumption in their interpretation. Still, six students did not reference the null in their interpretation. Of those six, four were students that did successfully connect their TinkerPlots sampler to the null hypothesis through recognizing that it is re-pairing values randomly and independently. This could show evidence that some students may lose this hypothetical perspective as they move toward using their samplers and interpreting the results, but could also simply be attributed to their response being an incomplete picture of their interpretation of these results.

Considering students statements as a whole, 13 of the 21 students provided a p-value interpretation with all three components present, that is, no "Missing" codes were applied. Four of those students gave a p-value where the slope and inequality were both
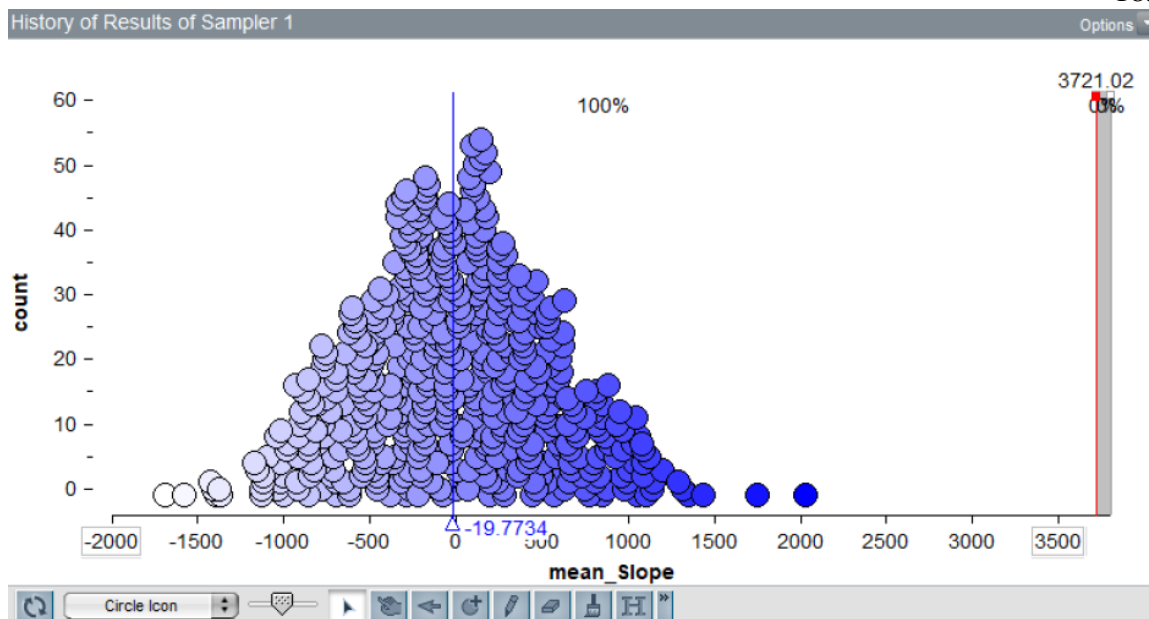
Figure 26. TinkerPlots sampling distribution and p-value provided by student

present and the assumption of the null hypothesis was detailed. One such example is

given below:

> *There is a close to 0% chance we'd get our observed positive 3721.02 slope from the data collected in original study, or more extreme (more steep/significant= away from 0), under the null model assuming no correlation between the carat size of the diamond and its price.*

This student gives a full picture of the *p*-value, including citing the exact slope value, a

relevant inequality statement, and a re-statement of their null hypothesis in context.

While this response is ideal and demonstrates coordination between the output they are

interpreting with their constructed sampler and the null assumption that went into it, other

students also gave interpretations that may demonstrate this as well. Consider this

student's *p*-value displayed in Figure 26 and their interpretation:

> *The p-value of 0 means that there is strong evidence against the null, which means that it is EXTREMELY unlikely that the researchers came to their data due to chance… [They have] evidence that carat size and price are positively correlated because reaching a slope of 3721.02 did not occur once in the null*

simulation. So, we can say that they have a significant correlative relationship based on the data when compared to the null.

This student was coded as having the slope present in their interpretation, an implied inequality, and the null as "random chance." Despite that they were not precise about this statement, they did reference the idea of the null hypothesis in their response, which was articulated clearly in other responses given on the assessment. While not required for the prompt given to the student on this question, their interpretation is inherently linked to the conclusion drawn from the study, showing they understand the implications of how they interpreted their *p*-value. This interpretation is also firmly rooted in the details of the data context as well, showing further support for CATALST students placing importance on the data contexts in their modeling and interpretation. Thus, other codes that indicate a lack of precision in their *p*-value interpretation may not obstruct their inferential reasoning overall.

Discussion and Conclusions

When only considering the limitations and workarounds needed to leverage TinkerPlots for teaching linear regression, one might argue that other software alternatives that provide fewer barriers would be better for students' learning. However, the evidence seen with students' inferential reasoning while using TinkerPlots to carry out tests for linear regression provides reason to be hopeful about the use of CATALST and TinkerPlots for learning simulation-based methods. It is important to consider that this study only looked at one single classroom, and that it is difficult to tease apart the impact of the curriculum and instructor on these students. Still, given the challenges students have with simulation-based inference that have been identified in the literature,

the results of this study show that CATALST is a potentially promising avenue for addressing these concerns. Even with the difficulties experienced with the technology, I argue that the benefits students appeared to experience through modeling their simulations in TinkerPlots gives reason to consider the proposed software improvements to further student learning in this setting.

A novel feature of TinkerPlots is the sampler device itself. Being able to create probability models and have students work creatively and collaboratively to build these samplers is the key feature that sets curricula like CATALST apart from other simulation-based curricula. Understanding the contextual and narrative aspects of statistical problems is a key feature of student reasoning in this course, as research studies previously discussed have shown that students have a strong desire to understand narrative aspects of the data and problem context (Noll et al., 2018, 2021). This was evident with the student attempting to determine the "narrative purpose" of using the formula editor in the workaround described. The results also described many episodes of students leveraging previously explored data contexts and applying them to new scenarios to aid their modeling and interpretation of statistical results. Contextual details of the data are key to learning and understanding statistics, and sets it apart from the more abstract nature of mathematics, or even traditional statistics courses that focus on procedures and formulas.

These sampler devices also allow students to draw connections between different statistical methods and connect these aspects of designing their sampler to the study design. This was seen in the transcripts of the activity work, where students referenced

the Dolphin Therapy study, and connected similar features in the types of variables used and the experimental design using random assignment. While more focus on the choice of replacement and how it connects to the study design may need more attention, students still often gave reasonable explanations for their choice of sampling, regardless if it was with or without replacement. Statisticians commonly leverage randomization tests that sample without replacement for the test of a slope, even if the study design does not reflect an experiment with random assignment. This is due to the minimal impact the choice of replacement has on the simulated results and associated p-values. However, focusing on the random process leveraged in a given context is important to stress for students for interpreting the results. While the choice of replacement and how it related to their study design and use of randomness was a feature of this study, this assessment did not focus on students' conclusions relative to the type of inference conducted. This may be a potential area of research focus for CATALST students specifically in assessing their thinking in this area as well as developing a learning trajectory that focuses on types of inference with TinkerPlots.

Students did seem to have relative success with the two challenges previously identified by Case and Jacobbe's (2018) framework: recognizing their simulation as a product of a hypothetical assumption and not a product of replicating real-world data, and distinguishing their simulated sampling distribution from observed samples. Over 80% of the class understood that their sampler in TinkerPlots represented the null hypothesis through independently re-pairing values from each of the two variables together randomly. While four students did not carry this assumption to their p-value interpretation, a majority of the class showed evidence that they viewed their simulation

as residing in the hypothetical world throughout the analysis. A strong majority of students also interpreted the *p*-value as a probability related to the value of a slope, indicating students seem to recognize their sampling distribution as a distribution of slope values rather than a single sample. This may be due to the nature of linear regression analysis, where samples of data are visualized via scatterplot rather than a univariate distribution. Still, students still must actively create these visualizations in TinkerPlots themselves, which may aid in helping students navigate the perspectives and levels identified by Case and Jacobbe.

Because this was a summative assessment for students, one may suggest that students prepared responses that may not reflect a deep understanding of inferential reasoning. This is often true for many introductory statistics courses, especially *p*-value interpretations, where students are often given a fill-in-the-blank style structure to follow for their *p*-value interpretations. The example responses provided in the results section show that this is not the case with this course. The wording of interpretations is up to the students themselves, and students do not have templates to follow for interpretations. Students build up their understanding and interpretations of statistical results through group and full class discussions, and refine them based on feedback from their peers and instructors. This is done to ensure that students are articulating their own thinking and work through statistical problems not just procedurally but conceptually. No two students emulated any other students' responses word-for-word throughout the assessment, reflecting how students think critically about the statistical simulations conducted in the course.

Adding the proposed functionality to TinkerPlots would bring the software in line with recommendations proposed by Rossman & Chance (2014) while capturing the benefits of using the TinkerPlots sampler. As evidenced by the students' thinking about random assignment in the caffeine scenario, TinkerPlots already enables students to think critically about their simulation process, greatly mitigating the risk of students perceiving the TinkerPlots sampler as a black box device. However, these innovations would not only make using TinkerPlots for linear regression easier on students, it would bring the simulation procedures in-line functionally and visually with other data scenarios like the dolphin therapy problem. After students build their samplers and generate one sample of data, students would work to plot their data in both scenarios. They would then use various TinkerPlots tools to create statistical measures on their plot that can be tracked over multiple simulations. The proposed innovation uses many similar visual cues as dolphin therapy, like the statistics being highlighted by the gray box when they can be collected on. Integrating these innovations into a future update of TinkerPlots would allow students to engage in discussions about the purpose and rationale for building their simulation in TinkerPlots and carry out that simulation with relative ease.

Future work should consider how to expand TinkerPlots to cover more concepts typically covered in an introductory statistics course. While the CATALST curriculum was intentionally designed to cover fewer topics so that more focus can be placed on students' statistical thinking and reasoning (Justice et al., 2020), a limited list of topics covered greatly hinders its adaptation. Course outlines for introductory statistics courses that are determined by statistics departments frequently contain topics not covered by CATALST like linear regression, analysis of variance, and chi-square tests. There are

numerous benefits to using CATALST over other simulation-based curricula, as seen by

the way students engaged with the simulation process in the caffeine scenario. Students

leveraging their narrative understanding of statistical processes in this curriculum has

been documented in various other statistical scenarios (Noll et al., 2018, 2021). In order

to achieve a greater adaptation of CATALST, providing more activities and technology

advancements to TinkerPlots for topics typical to the introductory statistics curriculum is

necessary. Activities for chi-square tests have been implemented with relative success

using TinkerPlots software (Dolor & Noll, 2015), but could consider further

improvements to streamline the technology experience. Focus on future work should

focus on how to implement analysis of variance in the CATALST curriculum, both on

creating activities and suggesting potential innovations for the TinkerPlots software to

accommodate these activities.

References

Antonioli, C., & Reveley, M. A. (2005). Randomised controlled trial of animal facilitated

    therapy with dolphins in the treatment of depression. *Bmj*, *331*(7527), 1231.

Case, C., & Jacobbe, T. (2018). A framework to characterize student difficulties in

    learning inference from a simulation-based approach. *Statistics Education*

    Research Journal, *17*(2), 9–29.

Chance, B., & McGaughey, K. (2014). Impact of a simulation/randomization-based

    curriculum on student understanding of p-values and confidence intervals.

    Proceedings of the 9th International Conference on Teaching Statistics, *9*.

    https://icots.info/9/proceedings/pdfs/ICOTS9_6B1_CHANCE.pdf

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on

    simulation-based inference: A preliminary report. *Journal of Statistics Education*,

    24(3), 114–126. https://doi.org/10.1080/10691898.2016.1223529

Chu, S. (1996). Diamond ring pricing using linear regression. *Journal of Statistics*

    Education, *4*(3).

Dolor, J., & Noll, J. (2015). Using guided reinvention to develop teachers' understanding

    of hypothesis test concepts. *Statistics Education Research Journal*, *14*(1), 60–89.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and

    thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898.

    https://doi.org/10.1007/s11858-012-0447-5

Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing student success and

    understanding in introductory statistics under consensus and simulation-based

curricula. *Statistics Education Research Journal*, *17*(1). https://iase-web.org/documents/SERJ/SERJ17(1)_Hildreth.pdf?1526347238

Justice, N., Le, L., Sabbag, A., Fry, E., Ziegler, L., & Garfield, J. (2020). The CATALST Curriculum: A Story of Change. *Journal of Statistics Education*, *28*(2), 175–186. https://doi.org/10.1080/10691898.2020.1787115

Konold, C., & Miller, C. (2018). *TinkerPlots* (2.3.4). LearnTroop. http://www.tinkerplots.com

Noll, J., Clement, K., Dolor, J., Kirin, D., & Petersen, M. (2018). Students' use of narrative when constructing statistical models in TinkerPlots. *ZDM*, *50*(7), 1267–1280. https://doi.org/10.1007/s11858-018-0981-x

Noll, J., Kirin, D., Clement, K., & Dolor, J. (2021). Revealing students' stories as they construct and use a statistical model in TinkerPlots to conduct a randomization test for comparing two groups. *Mathematical Thinking and Learning*, 1–20. https://doi.org/10.1080/10986065.2021.1922858

Pfannkuch, M., Budgett, S., & Arnold, P. (2015). Experiment-to-causation inference: Understanding causality in a probabilistic setting. *Reasoning about Uncertainty: Learning and Teaching Informal Inferential Reasoning*, 95–128.

Pfannkuch, M., Budgett, S., Fewster, R., Fitch, M., Pattenwise, S., Wild, C., & Ziedins, I. (2016). Probability modeling and thinking: What can we learn from practice? *Statistics Education Research Journal*, *15*(2). https://iase-web.org/documents/SERJ/SERJ15(2)_Pfannkuch.pdf

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning

introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*,

6(4), 211–221. https://doi.org/10.1002/wics.1302

Tintle, N., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., &

VanderStoep, J. (2014). *Quantitative evidence for the use of simulation and*

randomization in the introductory statistics course. 9th International Conference

on Teaching Statistics, Flagstaff, Arizona. https://iase-

web.org/icots/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention

of statistical concepts in a preliminary randomization-based introductory statistics

curriculum. *Statistics Education Research Journal*, *11*(1), 21.

Watson, J., & Donne, J. (2009). TinkerPlots as a Research Tool to Explore Student

Understanding. *Technology Innovations in Statistics Education*, *3*(1).

https://doi.org/10.5070/T531000034

Appendix C: Assessment Questions

Table 32.
Diamonds data set.

Carat	Price
0.17	355
0.16	328
0.17	350
0.18	325
0.25	642
0.16	342
0.15	322
0.19	485
0.21	483
0.15	323
0.18	462
0.28	823
0.16	336
0.2	498
0.23	595
0.29	860
0.12	223
0.26	663
0.25	750
0.27	720
0.18	468
0.16	345
0.17	352
0.16	332
0.17	353
0.18	438
0.17	318
0.18	419
0.17	346
0.15	315
0.17	350
0.32	918
0.32	919
0.15	298
0.16	339
0.16	338
0.23	595
0.23	553
0.17	345
0.33	945
0.25	655
0.35	1086
0.18	443
0.25	678
0.25	675
0.15	287
0.26	693
0.15	316

This appendix provides the data and context (Chu, 1996) with only the questions from the assessment that are analyzed in this paper.

A Singapore-based newspaper, *The Straits Times*, published an advertisement from a diamond retailer. The advertisement contained 48 different pictures of diamonds, with the carat (size) of the diamond listed along with its price in Singapore dollars. We will assume that this selection of diamonds in the newspaper is a random sample of diamonds sold from retailers. Based on this data, we would like to determine how to predict the price of a diamond based on its carat.

The data from this study is presented in Table 32. (Students were provided this data via a TinkerPlots file.)

Null Hypothesis Question: Describe how your model [that you provided] reflects the null hypothesis. Be sure to give a detailed explanation in the context of the problem.

Replacement Question: Describe how you chose replacement for the model above, and how you think this best reflects the context of this problem.

***p*-value Question:** Explain what the *p*-value means in the context of the problem. Do not simply state how strong/weak the evidence is, give an interpretation of the percentage/probability that you found.

Chapter 5: Conclusion

With the increased efforts to promote simulation-based curricula in the introductory statistics classroom as a result of Cobb's call to action (2007), it is important to reflect upon how these curricula achieve his primary goal: an emphasis on conceptual understanding of statistical techniques. The modeling focus of the Change Agents for the Learning And Teaching of STatistics (CATALST) curriculum is especially relevant for bringing conceptual understanding to the forefront. Its use of TinkerPlots software makes CATALST especially powerful in revealing and assessing student thinking in various statistical contexts (Watson & Donne, 2009), and past work has shown that CATALST students engage in full statistical investigations by constructing narratives within their model and engages students in the full statistical investigation cycle, reinforcing their understanding of inferential techniques (Noll et al., 2018, 2021). This dissertation aimed to investigate statistical association and linear regression in a CATALST course, which is a topic not previously covered in the original curriculum. Given the relevance of statistical association in the goals of the introductory statistics course outlined in standards (Carver et al., 2016) and its relevance in statistical literacy (Crocker, 1981; McKenzie & Mikkelsen, 2007; Schield, 2017), the goals of this dissertation address a relevant gap in statistics education research and teaching. In the following sections, I summarize the main findings of the previous three chapters and highlight their main contributions to the field. After this, I conclude with some final remarks and future directions.

Contributions from Previous Chapters

Students' Knowledge about Lines of Best Fit in a Modeling and Simulation Introductory Statistics Curriculum. My first paper investigated CATALST students' conceptions and strategies for informally fitting lines of best fit to scatterplots. Research identifies many various conceptions that students hold about statistical association. These include the univariate conception, which is a bias toward identifying upward sloping associations; the localist conception, which is characterized by placing a line based only on a small subset of points that are often collinear, and prior beliefs, where students place their line based on their own beliefs about the context despite the data presented. These conceptions have been observed in previous studies on informal line fitting that examined teachers and middle school students (Casey, 2015; Casey & Wasserman, 2015). To further this research on the population of CATALST students, I asked the following research questions: What are CATALST students' intuitive strategies for placing lines of best fit before and after formally learning about least squares criterion?

Students completed line-fitting tasks via survey instruments administered both before and after learning this content in their class. These questions allowed students to give a brief justification for their choice. Select students were invited to semi-structured task-based interviews to complete additional line fitting tasks, where a greater perspective on students' strategies could be obtained from follow-up questions. Results from both the survey and interviews indicated that many of these existing conceptions still persisted among CATALST students. In many scenarios where the data exhibited no association, the students in this study justified their choice by indicating there was a positive

association present, indicating a univariate conception. While justification with prior beliefs was overall rare on surveys and interviews, it most often occurred with uncorrelated data, suggesting that students may often seek out associations based on their own beliefs even when not present in the data. This reflects patterns of social prejudice, where people make judgments based on their own beliefs even in the face of data that tell a different story. Statistics content should be structured in a way to have students challenge these biases so students can recognize when they are not making appropriate claims.

One promising feature that emerged from interviews is the use of offsetting distances as a criterion for determining the line of best fit. This criterion often emerged when students were presented with data that contained outliers, and made placing a line using strategies that divided the data above and below it were not as effective. Students would try to determine groups of data points whose residuals summed to zero to determine if their line was accurately placed. This is a promising strategy to focus on for future curriculum materials on informal line fitting, as this aligns with the property of a line of best fit having all residuals sum to zero, which is a necessary condition of the least squares line. Students who used this strategy seemed to use this intuitively, and provided many analogies to justify their use of the criterion, such as a "weight" that a data point is pulling on the line. Future curricula that focus on statistical association and linear regression should consider this criteria as a point of focus with students, as this may help strengthen their informal line fitting skill in a way that is agreeable with the least squares line.

Another finding from this study was that students placed different emphasis on outliers depending on their placement along the *y*-axis. If an outlier was in the corner of the graph, along the extremes in the *y* direction, students would often successfully account for this outlier's impact on the line of best fit when placing their line. However, for outliers in the middle of the graph along the *y*-axis, students would often not fully account for their impact on the line of best fit appropriately, often not labelling the data points as outliers. This study was not originally designed to test for this, and so confounding variables exist between the data contexts that presented these outliers, such as the number of outliers, the sample size, and the slope of the line. Future research could examine students' perceptions of corner and middle outliers with more emphasis placed upon these confounding factors.

Comparing Student Outcomes on Testing for a Statistical Association for Traditional and Simulation-Based Curricula. This paper compared students across two classes, one which used the CATALST curriculum, and another that used a traditional curriculum. The focus of the comparison was with the approaches they described for determining whether two variables exhibit a significant linear relationship. Students from simulation-based courses often excel in the purpose and concepts of inferential techniques (Chance et al., 2018, 2022; Hildreth et al., 2018; Tintle et al., 2012, 2014), and this paper aims to add to this comparative literature by taking a particular focus on tests for the least squares line. Given the amount of descriptive statistics that surround linear regression and the complexity of the calculations behind these statistics, traditional students need to leverage software to generate output for any of these computations,

including conducting a hypothesis test. This may lead to a difficulty in separating the purpose and interpretation of descriptive and inferential statistics in this context.

Results from this study showed that CATALST students were far more likely to use inferential techniques to determine a significant linear relationship, with many more traditional students only suggesting examining a correlation value. After formally learning linear regression, CATALST students made larger gains than traditional students in using hypothesis tests to determine significant linear relationships. But CATALST students were also more likely to describe a hypothesis test for this scenario even before learning linear regression content in the course, suggesting that the CATALST curriculum is generally better at preparing students to understand the purpose of statistical inference. A case study of two students from each of these curricula who both described purely descriptive methods on their survey responses revealed gaps in their conceptual understanding of linear regression and hypothesis testing. After the interviewer hinted to both students at the idea of testing hypotheses, the CATALST student was able to produce and interpret a hypothesis test and distinguish it from descriptive methods like correlation, where the traditional student recognized these two methods as distinct but could not conceptually explain their differences. These results have implications for teaching content of linear regression. Correlation is often described with ranges of values that indicate levels of "strength" or "weakness" of a relationship, just as p-values indicate the strength of a result in light of a null hypothesis. Regardless of the curriculum used, instruction should be careful to make a distinction between these two methods, as correlation is purely an indicator of how close data is to a line, where the

p-value of a hypothesis test indicates if the relationship that exists within the data can be generalized in some manner.

Evidence for Further Development of TinkerPlots to Support Inferential Reasoning with Linear Regression. This third paper of the dissertation served two purposes: to demonstrate CATALST students inferential reasoning for hypothesis tests of the least squares line, and to suggest a technology innovation in TinkerPlots to streamline the technology experience with linear regression content so that learning statistical content is the main focus. CATALST has shown great potential in unlocking students' narrative reasoning (Noll et al., 2018, 2021) and exposing their conceptual reasoning with modeling and inference (Watson & Donne, 2009). However, the CATALST curriculum does not cover all topics typically taught in an introductory statistics course, such as linear regression. As part of this study, I designed activities leveraging TinkerPlots to introduce students to linear regression and conducting a hypothesis test for a slope. These activities required workarounds in TinkerPlots that are cumbersome, and do not meet recommendations for software in a simulation-based course (Rossman & Chance, 2014). Still, there is great potential in using this software for teaching students modeling techniques and unlocking their conceptual understanding of inference. I propose a potential technology innovation that would bring TinkerPlots in line with Rossman and Chance's recommendations while leveraging the modeling capabilities of TinkerPlots. To support the proposal of this technology innovation, I investigated students' inferential reasoning on an assessment at the conclusion of the course. This investigation was based upon the following research question: How does using TinkerPlots for conducting a

hypothesis test for the least squares line aid students' inferential reasoning and address common challenges faced when using simulation?

Analysis of student assessments was based on a combination of Case and Jacobbe's (2018) framework on challenges students face in simulation-based inference and the relevance of experiment-to-causation inference and connection to the study design and model design (Pfannkuch et al., 2015). This highlighted three areas of focus for analysis: connecting the null hypothesis to the TinkerPlots sampler, the choice of sampling with or without replacement and how that impacts study design, and the interpretation of the sampling distribution and p-value. The findings of this study generally showed that students were successful on the whole in connecting their TinkerPlots samplers to the null hypothesis as well as interpreting their p-value and drawing conclusions, although some language surrounding the p-value could have been more precise. Student responses regarding their choice of replacement were more mixed, and revealed that students did not always connect their choice to the study design and the type of inference constructed; however, students did provide reasoning consistent with their choice of replacement. Given that statisticians would normally conduct randomization tests without replacement regardless of the study design in this scenario, this procedure is not out of the ordinary. However, students should be able to appreciate the differences between experiment-to-causation and sample-to-population inferences and when each type of inference is appropriate. This is a potential avenue for future study in developing materials that strengthen students inferential reasoning and assessing their effectiveness. Overall, these results show promise in the use of CATALST and

TinkerPlots for more advanced introductory statistics topics like linear regression for supporting students' modeling techniques and inferential reasoning.

Concluding Remarks

This dissertation project aimed to investigate students' reasoning with statistical association and linear regression in the CATALST curriculum in order to support students' statistical literacy. Originally, the activities on linear regression were created out of a need to meet the demands of both institutional requirements for the introductory course and a larger NSF project. However, the ways that students interacted with these activities and the TinkerPlots technology in fitting lines to scatterplots and conducting randomization tests with TinkerPlots sampler models provided motivation for studying student outcomes based on the activities.

Statistical association is a topic necessary for statistical literacy. Visualizations like scatterplots and trend lines are ever present in the media, and statistics curricula should prepare students to read, analyze and critique statistical claims that are based in this data. Results from the first paper of this project focused on how students informally fit a line to scatterplots, which reflects how students read and summarize a linear relationship in a scatterplot. While the current activity sequence in CATALST showed that students' line fitting strategies were not consistent with statistical practice, there was promise in finding strategies that are intuitive to students. This may provide future avenues for bridging the gap to the least squares criterion, which is conceptually challenging to grasp and leverage informally.

The final two papers focused on statistical inference with linear regression. When students recognize trends and linear relationships in scatterplots, inferential techniques are crucial for understanding the importance and relevance of that trend, and whether that trend is generalizable in some way. If students are to analyze a claim like the one discussed at the beginning of the introduction chapter about the link between autism and proximity to freeways, they need to know how the inference was made. Does the data show a strong, meaningful trend? Can a causal link be made based on how the data was collected? Inferential reasoning is necessary for statistical literacy to evaluate and critique claims like this, as the news article does not always quote the researchers who challenge the causal links made in the headline. These two studies provided more evidence for the improved student outcomes with inferential reasoning in simulation-based curricula, but also gives a theoretical basis for the advantages CATALST has over other simulation-based curricula with its focus on modeling. Modeling itself is a goal of the introductory statistics course (Carver et al., 2016), and has the capability to enhance student learning of statistics concepts (Justice et al., 2018; Noll et al., 2016, 2018, 2021).

Studies that compare statistics curricula have long focused on comparing various simulation-based curricula to the traditional curriculum, but future studies should take a focus on comparing various simulation-based curricula with a special focus on the technology students use. How do various applets or applications support students learning? What limitations do pre-constructed models in applets have on student outcomes? Does the modeling environment of TinkerPlots enhance students inferential thinking? What other benefits are there to engaging students in modeling, and what other tools and enhancements should a simulation offer to students? As statistical practice

continues to shift from frequentist to Bayesian, research is beginning to investigate integrating this philosophy in the introductory statistics course (Paul, 2017). Students modeling capabilities need not be limited to the capabilities of the TinkerPlots sampler, and future work should support the development of both software and curricula that evolves with the practice of statistics.

References

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Roswell, G., Velleman, P., Witmer, J., & Wood, B. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association.

Case, C., & Jacobbe, T. (2018). A framework to characterize student difficulties in learning inference from a simulation-based approach. *Statistics Education Research Journal*, *17*(2), 9–29.

Casey, S. A. (2015). Examining student conceptions of covariation: A focus on the line of best fit. *Journal of Statistics Education*, *23*(1).

Casey, S. A., & Wasserman, N. H. (2015). Teachers' knowledge about informal line of best fit. *Statistics Education Research Journal*, *14*(1).

Chance, B., Mendoza, S., & Tintle, N. (2018). Student gains in conceptual understanding in introductory statistics with and without a curriculum focused on simulation-based inference. *Proceedings of the 10th International Conference on Teaching Statistics*. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3B2.pdf

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centered on simulation-based inference. *Statistics Education Research Journal*, *21*(3), Article 3. https://doi.org/10.52041/serj.v21i3.6

Cobb, G. W. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1). https://escholarship.org/uc/item/6hb3k0nz

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin*, *90*(2), 272–292. http://dx.doi.org/10.1037/0033-2909.90.2.272

Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing student success and understanding in introductory statistics under consensus and simulation-based curricula. *Statistics Education Research Journal*, *17*(1). https://iase-web.org/documents/SERJ/SERJ17(1)_Hildreth.pdf?1526347238

Justice, N., Zieffler, A., Huberty, M. D., & delMas, R. (2018). Every rose has its thorn: Secondary teachers' reasoning about statistical models. *ZDM*, *50*(7), 1253–1265. https://doi.org/10.1007/s11858-018-0953-1

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33–61. https://doi.org/10.1016/j.cogpsych.2006.04.004

Noll, J., Clement, K., Dolor, J., Kirin, D., & Petersen, M. (2018). Students' use of narrative when constructing statistical models in TinkerPlots. *ZDM*, *50*(7), 1267–1280. https://doi.org/10.1007/s11858-018-0981-x

Noll, J., Gebresenbet, M., & Glover, E. D. (2016). A modeling and simulation approach to informal inference: Successes and challenges. In *The teaching and learning of statistics* (pp. 139–150). Springer.

Noll, J., Kirin, D., Clement, K., & Dolor, J. (2021). Revealing students' stories as they construct and use a statistical model in TinkerPlots to conduct a randomization test for comparing two groups. *Mathematical Thinking and Learning*, 1–20. https://doi.org/10.1080/10986065.2021.1922858

Paul, W. (2017). An exploration of student attitudes and satisfaction in a GAISE-influence introductory statistics course. *Statistics Education Research Journal*, *16*(2), Article 2. https://doi.org/10.52041/serj.v16i2.203

Pfannkuch, M., Budgett, S., & Arnold, P. (2015). Experiment-to-causation inference: Understanding causality in a probabilistic setting. *Reasoning about Uncertainty: Learning and Teaching Informal Inferential Reasoning*, 95–128.

Rossman, A. J., & Chance, B. L. (2014). Using simulation-based inference for learning introductory statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(4), 211–221. https://doi.org/10.1002/wics.1302

Schield, M. (2017). GAISE 2016 promotes statistical literacy. *Statistics Education Research Journal*, *16*(1), 50–54.

Tintle, N., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2014). *Quantitative evidence for the use of simulation and randomization in the introductory statistics course*. 9th International Conference on Teaching Statistics, Flagstaff, Arizona. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8A3_TINTLE.pdf

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21.

Watson, J., & Donne, J. (2009). TinkerPlots as a Research Tool to Explore Student Understanding. *Technology Innovations in Statistics Education*, *3*(1). https://doi.org/10.5070/T531000034