PROMOTING INTEREST AND SKILLS IN STATISTICAL AND MULTIVARIABLE THINKING WITH SOCIAL JUSTICE DATA INVESTIGATIONS

Josephine Louie[1], Soma Roy[2], Beth Chance[2], Jennifer Stiles[1], and Emily Fagan[1]
[1] Education Development Center, U.S.A
[2] California Polytechnic State University – San Luis Obispo, U.S.A
jlouie@edc.org

The Strengthening Data Literacy across the Curriculum (SDLC) project has been developing and researching curriculum modules to build interest and skills in data science among U.S. high school students from historically marginalized groups. SDLC modules are centered on investigations of social justice questions using large-scale social science data and the Common Online Data Analysis Platform (CODAP). This paper examines the extent to which students show increased interest in statistics and data analysis, and stronger understanding of core statistical concepts and multivariable thinking, after completing a three-week SDLC module. This paper also discusses ways in which a social justice focus may contribute to students' interests in and understanding of data analysis.

BACKGROUND

In a world awash with data, there is increasing demand for people with robust data analysis skills and who understand the process and practices of statistical inquiry. Research analysts have sounded the alarm that schools are neither preparing students adequately nor attracting enough students to the study of statistics and data science (Henke et al., 2016; Manyika et al., 2011). Groups that have been underrepresented in mathematics and scientific fields (e.g., women, Blacks, and Latinx) are entering data science fields at disproportionately low rates (Priceonomics, 2017). Facing greater needs to diversify student populations who are prepared to enter these fields, schools today need new strategies to foster students' interests and practices in statistics and data analysis.

The *Strengthening Data Literacy across the Curriculum* (SDLC) project has been exploring the promise of drawing broader student populations to statistics and data science with high school curriculum materials that support student investigations of social justice issues using authentic social science data sets. As noted by leading educators of civic statistics (ProCivicStat Partners, 2018), the availability of rich data on topics such as employment and income, immigration, poverty, public health, and education offer prime opportunities to build students' data analysis practices as well as deeper understandings of their social and economic worlds. A driving premise of the project is that students may become more motivated to learn important statistics and data practices when lessons are centered on examining real-world questions of direct relevance to themselves, their families, or their communities (Howery & Rodriguez, 2006; Lesser 2007). When students investigate the conditions of different groups in society, "students can ask real questions about real-life situations. These in turn raise ethical and moral questions, which motivate students' learning, making the subject matter more relevant and interesting" (Rouncefield, 1995, p. 3). Examining the socioeconomic conditions of different groups may be of particular interest to students from historically marginalized racial, ethnic, or other demographic populations (Gutstein, 2003; Ladson-Billings, 1995, Nyman, 2015). To diversify and strengthen the fields of statistics and data science, it is critical to employ strategies that foster interest in these fields among underrepresented populations, given that interest is an important predictor of engagement and career choices (Linnenbrink-Garcia et al. 2013; Palmer et al., 2017).

A curriculum intervention

Drawing on these ideas, the SDLC project has been developing and testing prototype curriculum modules with a social justice focus, targeted toward U.S. high school students in non-Advanced Placement (AP) mathematics and statistics classes. Designed as applied data investigations that take up to three weeks of instructional class periods to complete, the modules offer opportunities to examine the social and economic conditions of groups in U.S. society using person-level microdata from the American Community Survey (ACS) and the U.S. decennial census. In one module, *Investigating Income Inequality in the U.S.*, students address questions such as: What is income inequality? How has U.S. income changed over time? How much income inequality exists between males and females? Does education help to explain the wage gap between males and females?

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

Students address these questions by working through the four steps of the data investigation cycle (Bargagliotti et al., 2020). They complete activities designed to strengthen conceptual understandings of measures of center and variability, the impact of outliers and skewness on these measures, sampling variability and margins of error, and comparisons of quantitative distributions. Students also learn to recognize and explain confounding and interaction effects within multivariable data – practices that are critical for working with large-scale data yet are not currently emphasized in high school curricula (Engel, Gal, & Ridgway, 2016). Students analyze data with the Common Online Data Analysis Platform (CODAP), a tool that supports data visualization and conceptual understanding of statistical ideas over calculations. In their analyses, students confront widening disparities between upper and lower-income earners over time as well as the persistence of the gender wage gap after controlling for other variables. Lessons encourage collaborative inquiry and prompt students to discuss and make sense of data with their classroom peers. At the end of the module, students conduct a data investigation of their choice, exploring questions such as: To what extent does the male-female wage gap vary by race or ethnicity, or by U.S. region? These approaches follow widespread recommendations for statistics learning (Bargagliotti et al., 2020; Garfield & Ben-Zvi, 2008; Ridgway, 2016). The curriculum also strives to support a component of Gutsein's (2003) pedagogy of social justice by raising awareness of socioeconomic inequalities. Sample lessons can be found at http://cadrek12.org/projects/strengthening-data-literacy-across-curriculum-sdlc.

METHODS

A study was conducted in fall 2019 and early 2020 to examine the following questions: RQ1) To what extent do students show a) increased interest in statistics and data analysis, as well as b) stronger understanding of core statistical concepts and multivariable thinking, at the end of the *Investigating Income Inequality in the U.S.* module compared to at the start? RQ2) What role might the focus on a social justice topic play in students' interests in and understanding of statistics and data analysis? Study participants involved seven high school mathematics teachers and their students in six schools in a northeast metropolitan U.S. region. Participating teachers implemented the module in fourteen classrooms of students in grade 12 who were taking non-AP statistics or data analysis courses. The schools' student populations were 55%-91% Black and Latinx, 35%-67% whose first language was not English, and 38%-65% economically disadvantaged. Percentages of students in each school who had met or exceeded expectations on the Grade 10 state assessment (required for graduation) ranged from 6%-58% in English language arts and from 12%-52% in mathematics.

Data sources and analysis

To address RQ1, students completed a pre- and post-module survey measuring their academic interests and self-concept in statistics. Interest has been described as a "psychological state of engaging or the predisposition to reengage with particular classes of objects, events, or ideas over time" (Hidi & Renninger, 2006, p. 112). Academic self-concept has been described as "a person's confidence in his or her competencies in [a] particular domain" (Sproesser, Engel, & Kuntze, 2016). Using instruments developed by Linnenbrink-Garcia et al. (2010) and Sproesser, Engel, & Kuntze (2016), 23 seven-point Likert-type items were adapted to measure self-concept as well as four stages of interest in statistics (Hidi & Renninger, 2006). These four stages are: triggered situational interest (often sparked by surprising or attention-grabbing stimuli in the environment), maintained situational interest (marked by more positive or meaningful connections with the content), emerging individual interest (a deeper connection to the content, often signaled by curiosity and effort to engage), and well-developed individual interest (a stable and strong connection to the content, marked in part by persistence in tasks even in the face of challenges). Increased situational interest has been found to support the development of more long-term, stable individual interest in a domain (Linnenbrink-Garcia et al., 2013; Palmer et al., 2017).

Of the 277 students who submitted the pre-survey, 210 submitted the post-survey (an attrition rate of 24%). Factor analyses were conducted on the pre- and post-survey data separately. A principal components analysis (PCA) with varimax rotation yielded four principal components that were consistent across the two sets of data, and exploratory factor analyses (principal axis factoring, promax rotation, kappa =2) yielded a similar four-factor structure. Four scales were assembled to align with the four primary components of the PCA. One scale represents a single measure of situational interest,

combining (1-SI, 8 items; see Table 2 for sample items). A second scale represents a single measure of individual interest (2-IN, 4 items). A third scale represents self-concept (3-SC, 5 items), and a fourth scale represents the value students place on data analysis (4-VC, 6 items). Cronbach's alpha for each scale ranged from 0.73 to 0.92 on the pre-survey and from 0.76 to 0.92 on the post-survey, suggesting acceptable to very good reliability for each scale (DeVellis, 2012). Pre- and post-survey score differences were examined using paired-sample *t*-tests.

Table 2. Example interest and self-concept survey items and associated principal components

Item	Component
Our statistics lessons over the past couple of weeks sparked my interest in investigating data.	1-SI
What we learned in *statistics class over the past couple of weeks* is important to me.	1-SI
Investigating data is one of my favorite activities.	2-IN
I like to think about statistics and data even when I'm outside of my statistics class.	2-IN
Understanding tasks with diagrams and statistical data is easy for me.	3-SC
I am good at solving statistical problems.	3-SC
It is important to me to be a person who can analyze data statistically.	4-VC
Statistics and working with data are useful for me to know.	4-VC

Notes: Phrases in italics were replaced with "the Income Inequality data lessons" in the post-survey. Each item had a seven-point rating scale, ranging from 1 (Strongly disagree) to 7 (Strongly agree).

In addition to the interest and self-concept survey, students were invited to complete a pre- and post-module learning assessment measuring understanding of core statistical concepts and capacities for multivariable thinking. The assessment consisted of 19 multiple-choice items drawn primarily from item pools developed by Jacobbe et al. (2014) and Garfield et al. (2006). Items were selected to measure understanding in five domains: sampling and data collection (3 items); data representation (4 items); measures of center (5 items); measures of variability (4 items), and multivariable thinking (3 items). In this last domain, two items were developed by the second and third authors of this paper (see Figure 1 for an example). Pre- and post-survey scores were calculated for each domain; pre- and post- score differences were examined using paired-sample *t*-tests.

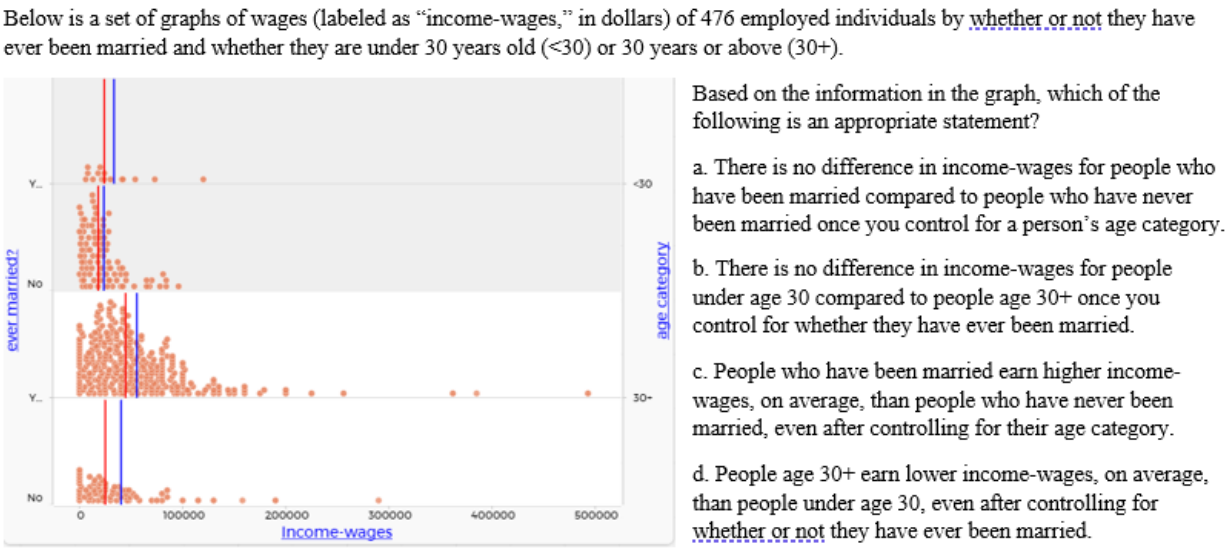

Figure 1. Sample item designed to measure students' ability to demonstrate multivariable thinking

To address RQ2, authors conducted five student focus group interviews after module completion in December 2019 and January 2020. Focus group interviews contained 4-6 students, lasted 40-50 minutes, and were audio recorded and transcribed. Two project researchers generated a

preliminary set of *a priori* codes related to ways in which the module supported (or did not support) student interest in or learning of statistics and data analysis. One researcher coded the transcripts with these codes, and both researchers examined coded statements to identify themes related to how the income inequality topic may have supported students' interests in and learning of statistics.

RESULTS

For RQ1, scores from the interest survey suggest that students' situational interest in their statistics learning (1-SI), self-concept in statistics and data analysis (3-SC), and perceptions of the value of statistical content (4-VC) grew between the start and end of the module, but the changes were not statistically significant at $p < 0.05$ (Table 3). Students' individual interest in statistics and data analysis, as measured by the second principal component (2-IN), was significantly higher at $p < 0.01$ after completion of the module, with a small effect size of Cohen's $d = 0.25$.

Table 3. Pre- and post-module interest in statistics and data analysis: Summary of score differences

	n	Pre Mean (SD)	Post Mean (SD)	Post – Pre Mean (SD)	*p*-value	Cohen's *d*
Situational interest (1-SI; 8 items)	193	4.96 (0.98)	4.98 (1.05)	0.02 (1.02)	0.799	0.02
Individual interest (2-IN; 4 items)	195	3.82 (1.20)	4.10 (1.24)	0.29 (1.18)	0.001	0.25
Self-concept (3-SC; 5 items)	196	4.70 (1.04)	4.71 (1.06)	0.01 (0.81)	0.986	0.01
Value of content (4-VC, 6 items)	190	4.97 (0.94)	5.07 (0.93)	0.10 (0.84)	0.082	0.12

Note: Scores for each scale were calculated as the average score for items in each scale. Minimum and maximum possible scores for each item and scale were 1.0 and 7.0, respectively. Sample sizes vary due to missing responses.

Based on results from the learning assessment, students' overall understanding of assessed statistical concepts grew between the start and end of the module, and the change was statistically significant at $p < 0.001$ (Table 4) with a moderate effect size ($d = 0.43$). Students demonstrated statistically significant growth in their understanding of data representation ($p < 0.01$, $d = 0.19$), measures of center ($p < 0.001$, $d = 0.38$), and multivariable thinking ($p < 0.001$, $d = 0.35$), whereas changes in understanding of sampling and data collection as well as measures of variability were not statistically significant at the 5% level.

Table 4. Pre- and post-module understanding of statistics concepts: Summary of score differences

	Pre Mean (SD)	Post Mean (SD)	Post – Pre Mean (SD)	*p*-value	Cohen's *d*
Overall (19 items)	9.66 (2.67)	10.94 (3.16)	1.28 (2.99)	<0.0001	0.43
Sampling & Data Collection (3 items)	1.53 (0.80)	1.57 (0.79)	0.04 (0.92)	0.280	0.04
Data Representation (4 items)	2.91 (0.91)	3.13 (0.97)	0.22 (1.15)	0.006	0.19
Measures of Center (5 items)	2.15 (1.08)	2.68 (1.28)	0.53 (1.41)	<0.0001	0.38
Measures of Variability (4 items)	1.57 (0.91)	1.71 (0.97)	0.13 (1.19)	0.072	0.11
Multivariable thinking (3 items)	1.50 (0.86)	1.86 (0.89)	0.36 (1.03)	<0.0001	0.35

Note: Items were multiple choice and scored as 1 = correct, 0 = incorrect or missing, with $n = 180$ students who provided a full set of pre and post responses.

For RQ2, a preliminary analysis of student focus group interview data suggests that the module's focus on income inequality in the U.S. boosted students' interests in their statistics learning. Students indicated stronger interest because the data and topic were authentic. One student explained: "It's real data. It's not…made up. It's from real people. I think that was the thing that makes it interesting, that it's real." When the data are authentic, students may be better able to relate to the data. Another student observed that "we could even be one of the people that are part of the graph. And it's really interesting how…it's different between males and females. And who makes more money depending on their race or… how much school they've been through." In addition, focusing on topics that are personally relevant and important in students' lives may be critical. Another student said: "What we're doing right now or all the time is comparing like GPA to height or different sexes, and

it's so boring. But to have the topic of income wage and comparing it to factors in our daily life – I thought that was something we could relate to, so it was more interesting."

The social justice issues raised by students' data investigations also appeared compelling and eye-opening to them. One student elaborated on how the module affected his societal understandings: "It was interesting 'cause we took a break from bookwork and we looked at real life problems and income inequality based on genders, race, and between different states…It was just very interesting being able to graph out everything and finding all these differences and like, oh, this actually is an issue. You always hear about it, like there's income inequality between different genders, and when we actually did the data it was like, oh wow, this is real. People aren't just making things up. This is a real problem, and maybe hopefully we can figure something out."

Other students appreciated opportunities to examine socioeconomic data and to discuss the issues they raise. One student shared: "I think we need to do more of [these modules] in school, to be honest with you. A lot of people are afraid to talk about issues like that… If more kids tackle it head-on, they'll have a better understanding of how our world really works." Another student concurred, explaining: "You can't really talk politics in school, but this is like the closest you can probably get where the kids can form their own opinion." He then shared how transformative such data investigations could be: "I bet you a lot of kids' opinions about this topic changed… I know myself and a few other students were on a certain side of the issue, but [in] the end that was not the case."

The authentic social issue that students explored also appears to have helped students develop a stronger understanding of statistics content. When situated in a topic that is meaningful, statistical analysis no longer becomes an abstract exercise. One student explained: "We're used to looking at graphs: here's the x-axis, the y-axis, analyze it." In contrast, "[This module] has made it easier to understand this is how you work with real problems and how you're supposed to set [graphs] up. It's not just numbers on the bottom and the top… [you have to] apply your knowledge." Another student echoed these thoughts by saying: "With normal graphs… it's just supposed to make sense. But with this, you go more with it. You learn about the graph. You don't just say, okay there's the graph, there's the equation, solve it, there you go, you're done. You just learn more."

DISCUSSION

Analyses of quantitative data from pre- and post-module survey and assessment data as well as qualitative data from student focus group interviews suggest that a three-week module focused on data investigations of income inequality in the U.S. was associated with growth in individual interest in statistics and data analysis, as well as understanding of measures of center, data representation, and multivariable thinking, among approximately 190 high school students in high school non-AP mathematics and statistics classes. In interviews, students suggested that the focus on a social issue with data drawn from real people helped increase their interest in and learning of statistics and data analysis for multiple reasons: the issue was important and relevant to their lives; they could see themselves in the data; and the social justice issues raised helped them better understand conditions in the wider world. They suggested that the topic of their data investigations made quantitative explorations less abstract and more meaningful. Coming in large proportions from historically marginalized racial and ethnic populations and from working-class households, the students in this study displayed knowledge of income wages from their own and their families' work experiences and could draw on this knowledge when trying to explain wage gaps among groups in society.

These findings are limited by the preliminary nature of the team's qualitative data analyses, the exploratory nature of the study's interest survey scales, the correlational nature of the study and lack of a control group comparison (which prevents causal claims), and the lack of adjustments for clustered data (which overestimates statistical significance levels). Further research could explore why individual interest scores grew significantly while situational interest scores did not – although student survey and interview data suggest that the length and iterative structure of the module lessons may have been a factor (Louie et al., 2021). Despite these limitations, this study suggests that data investigations with a social justice focus hold promise as a curriculum approach that may promote students' individual interest in statistics and data analysis, particularly among those from historically marginalized groups. Future research with comparison groups is needed to study the efficacy of using this approach to motivate underrepresented student populations to pursue statistics and data science careers – and potentially toward greater social and civic participation as well.

ACKNOWLEDGMENT

The *Strengthening Data Literacy across the Curriculum* project is supported by the National Science Foundation (NSF) under Grant No. DRL 1813956. Opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., Spangler, D.A. (2020). Pre-K-12 guidelines for assessment and instruction in statistics education II (GAISE II): A framework for statistics and data science education. American Statistical Association.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Engel, J., Gal, I., & Ridgway, J. (2016). Mathematical literacy and citizen engagement: The role of civic statistics. Presented at the 13th International Conference on Mathematical Education, Hamburg, Germany.

Garfield, J., delMas, B., Chance, B., Ooms, A. (2006). *Summary statistics for a national sample of undergraduates: Comprehensive Assessment of Outcomes for a first course in Statistics (CAOS).* The Web ARTIST Project.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice* (1st ed.). Springer Netherlands.

Gutstein, E. (2003). Teaching and learning mathematics for social justice in an urban, Latino school. *Journal for Research in Mathematics Education*, 37–73.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., Sethupathy, G. (2016). The age of analytics: Competing in a data-driven world. McKinsey Global Institute.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111–127.

Howery, C. B., & Rodriguez, H. (2006). Integrating data analysis (IDA): Working with sociology departments to address the quantitative literacy gap. *Teaching Sociology, 34*(1), 23–38.

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014). Establishing the validity of the LOCUS assessments through an evidenced-centered design approach. In *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9).*

Ladson-Billings, G. (1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice, 34*(3), 159–165.

Lesser, L. M. (2007). Critical values and transforming data: Teaching statistics with social justice. *Journal of Statistics Education, 15*(1), 1–21.

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S.

A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70*, 647–6

Linnenbrink-Garcia, L., Patall, E., Messersmith, E. (2013). Antecedents and consequences of situational interest. *British Journal of Educational Psychology, 83,* 591-614.

Louie, J., Stiles, J., Fagan, E., Roy, S., Chance, B. (2021). Data investigations to further social justice inside and outside of STEM. *Connected Science Learning, 3*(1).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Nyman, W. (2015). Social justice in the social studies classroom: Using culturally relevant pedagogy to promote civic engagement. *Rising Tide, 7*, 1-23.

Palmer, D., Dixon, J. Archer, J. (2017). Using situational interest to enhance individual interest and science-related behaviours. *Research in Science Education, 47,* 731-753.

Priceonomics. (2017, September 28). The data science diversity gap. *Forbes*.

Ridgway, J. (2016). Implications of the data revolution for statistics education: The data revolution and statistics education. *International Statistical Review, 84*(3), 528–549.

ProCivicStat Partners (2018). *Engaging civic statistics: A call for action and recommendations.* A product of the ProCivicStat Project.

Rouncefield, M. (1995). The statistics of poverty and inequality. *Journal of Statistics Education, 3*(2).

Sproesser, U., Engel, J., & Kuntze, S. (2016). Fostering self-concept and interest for statistics through specific learning environments. *Statistics Education Research Journal, 15*(1), 28–54.