GROUNDED RELEVANCE: HOW EXPLORING SOIL DATA CAN PROMOTE DATA LITERACY, LEARNING, AND MOTIVATION

Ian Thacker, Ph.D. UT San Antonio ian.thacker@utsa.edu	Rebecca Schroeder, Ph.D. UT San Antonio rebecca.schroeder@utsa.edu	Sara Shields-Menard, Ph.D. UT San Antonio sara.shields-menard@utsa.edu
Corina Lopez UT San Antonio corina.lopez@my.utsa.edu	Sandrina Ramirez UT San Antonio sandrina.ramirez@my.utsa.edu	Tanvir Alam UT San Antonio tanvir.alam@my.utsa.edu

The ability to analyze, interpret, and draw insights from data and data visualizations is quickly becoming a necessary skill for success across multiple disciplines and careers. However, people struggle to make meaning from data, and traditional data-science curriculum falls short of emphasizing its relevance to underrepresented students. To create opportunities for meaningful applications of data-science for diverse students, we developed and implemented an online learning module focused on engaging N = 298 undergraduate students at a Hispanic Serving Institution (HSI) in an analysis of place-based soil data. Using a pretest posttest study design, we found that student's perceptions of data-science relevance microbiology knowledge improved. We also inductively coded qualitative survey responses and used automated text analysis to explore how students framed "relevance" and how perceptions changed from pretest-to-posttest.

Keywords. Conceptual Change, Data Literacy, Integrated STEM, Place-Based Education

Data literacy—decision-making through statistical methods and techniques—is essential across multiple industries (Ben-Zvi & Garfield, 2008; Gould, 2017; Steen, 2001). With the increasing volume of data in today's world, being able to analyze, interpret, and draw insights from data and data visualizations is becoming essential for career success, likened to reading and mathematical literacy (Börner et al., 2019; Gal, 2002). Despite the global demand to improve data science education, traditional courses are not meeting the needs of those seeking training (Baumer, 2015), and statistics education, while firmly grounded in mathematics curriculum (Ben-Zvi & Garfield, 2008; NCTM, 2000), does not traditionally tap into topics that students find relevant, such as interdisciplinary and sociopolitical applications (Kokka, 2019; Weiland, 2017). This disconnect contributes to a lack of diversity with regards to race and gender in math-intensive STEM fields (NSF, 2015). Hispanic students, in particular, are underrepresented in STEM fields, particularly those related to statistics and data science (Fry et al., 2021).

Theoretical Framework

To frame how soil data collection and analysis can support science learning and perceptions of data science relevance for underrepresented students, we integrate frameworks of Conceptual Change, Data Literacy, and Place-Based Education. *Conceptual change* is a process where individuals restructure their conceptual knowledge to be more aligned with experts after engaging with novel information (Dole & Sinatra, 1998; Lombardi et al., 2016). In these models, learner characteristics (e.g., their *beliefs*, *motivation*, and *emotions*) and information characteristics (e.g., *comprehensibility*, *compellingness*, and *relevance*) interact to determine students' levels of cognitive engagement. Higher levels of engagement, along with shifts in motivation and emotion, predict more serious consideration (or reconsiderations) of whether scientific ideas are *plausible* (Lombardi et al., 2016), and higher likelihood of conceptual change (see Figure S1 in the Supplemental Materials [SM]). Of these many factors, evidence suggests that data *comprehensibility* and *compellingness* can bolster student engagement and conceptual change (Thacker, 2023; 2024; Thacker et al., 2024; 2025; Thacker & Sinatra, 2022).

Though there is no consensus on a definition, the term "data literacy" refers to statistical competencies, methods, and techniques that facilitate decision-making (Gould, 2017). It includes competencies of *understanding*, *acquiring, reading, interpreting, evaluating, managing, visualizing,* and *using* data (Börner et al., 2019; Carlson & Johnston, 2014; Kim et al., 2023; Prado & Marzal, 2013; Qiao et al., 2024; Ridsdale et al., 2015). This study focuses on promoting students' *translation* of relevant problems of interest into problems of data (Börner et al. 2019). That is, students must first link data to meaningful problems and identify measurable variables.

A useful framework for contextualizing data and creating personally relevant data experiences for students is to connect to their "sense of place" (Semken et al., 2017). Place-based learning is grounded in students' local contexts and systems of meaning such as culture, history, and community. Place-based learning is ideal for microbiology, geoscience, and agricultural education because of their direct relation to one's lived environment and can be used as a means to enhance culturally inclusive practices for diverse learners (NGSS, 2013; Semken et al., 2017). However, its effectiveness for diverse learners and the role of motivation remain understudied (Gosselin et al., 2016; Semken et al., 2017). We address these gaps in three research questions:

- RQ1. To what extent do student's perceptions of data science relevance and microbiology knowledge change after exploring place-based soil data and data visualizations?
- RQ2. How do students describe the relevance of data-science? What levels of analytical thinking are evident in those descriptions? And how do these dimensions change after exploring place-based soil data and data visualizations?
- RQ3. What soil-related variables do students identify as being relevant for data exploration? And how do these perceptions change from pre- to post-intervention?

Methods

Intervention Developed

Prior to conducting the current study, we developed and tested an interdisciplinary data-literacy/microbiology learning intervention (see Thacker et al., 2025). The intervention is an asynchronous, open-access learning module developed in SoftChalk Cloud that introduces undergraduate students enrolled in a microbiology lab course to the Tiny Earth Initiative (Hurley et al., 2021), with a particular focus on introducing students to data-science applications. Tiny Earth is a national initiative concentrated on identifying new antibiotics in soil to combat the escalating antibiotic resistance crisis by encouraging undergraduate students to collect soil from the place where they live, study the bacteria in that soil, analyze the bacteria for antibiotic activity, and add their data to an online repository. The module introduced antibiotic resistance, microbial ecology factors influencing antibiotic-producing bacteria, guided research question generation, and included a tutorial on interpreting soil data visualizations. Students explored whether a selection of variables (landform type, soil type, annual precipitation, pH, and annual air temperature) were related to the percentage of antibiotic producers in a selection of map visualizations generated from the Tiny Earth national repository data (see SM, Figure S2).

Participants and Procedures

To answer our research questions, we recruited $N = 298$ undergraduate students from an HSI in a southwestern state of the USA over the course of three semesters. Students reported their year of study (1% first year, 13% second year, 38% third year, 38% fourth year, and 10% other), gender (75% Female, 22% Male, 1% Nonbinary, 2% prefer not to say) ethnicity (55% Hispanic), race (1% American Indian/Alaska Native, 12% Asian, 7% Black/African- American, 9% Two or more races, 62% White/Caucasian, 9% Other race), and whether they were enrolled in a STEM major (79% STEM major, 15% not STEM, 1% plan to enroll in a STEM major, 4% Other).

All participants first completed a researcher-created 12 item pretest questionnaire measuring their microbiology knowledge, which was averaged and converted to percentage points for all analyses. Students also completed a three-item data literacy measure based on principles from Börner et al. (2019). Two open-ended item prompted students to "Explain why you think that data science is or is not relevant for the field of soil microbiology. Provide examples if possible." The second item prompted students to, "Please make a list of any variables that might be relevant for data scientists to investigate when exploring information about soil." The third item was a single item assessing students' perceptions of data science relevance on a scale from 1 = *Not at all relevant* to 5 = *Very relevant* (see SM, Appendix B–C).

After the pretest, learners completed the ~60 minute module and then completed an identical post-test of microbiology knowledge and data literacy. Cronbach's alpha for the microbiology knowledge scale was .65 at pretest and .71 at posttest.

Findings

RQ1: Pre-Post Improvements in Data Science Relevance and Microbiology Knowledge

To investigate whether there were significant changes in mean data relevance perceptions and microbiology knowledge, we used paired Wilcoxon signed rank tests to account for skewed distributions. Students' perceptions of data-science relevance significantly improved from pretest to posttest (M_{pre}=4.2, SD_{pre}=0.88, M_{post}=4.5, SD_{post}=0.81; W=2425, p<.001, Cohen's d=.33), as did microbiology knowledge (M_{pre}=76%, SD_{pre}=17, M_{post}=83%, SD_{post}=16; W=4326, p<.001, d=.39). For visualizations of distributions, see Figure S3 in the supplemental materials.

RQ2: Qualitative Analyses of Student Perceptions of Data-Science Relevance

To explore how students perceived the relevance of data-science relevance in soil microbiology analysis, we analyzed their open-ended responses in two ways. The first was by using traditional inductive coding processes. Two graduate student researchers openly-coded responses to the prompt: "Explain why you think that data science is or is not relevant for the field of soil microbiology…" They read each student's response at pretest and posttest and categorized them freely. We then met to discuss common codes, emerging themes related to all student responses, and collectively developed a codebook (see Appendix A in SMs) which we used to systematically recode all responses (Saldaña, 2021; Chamraz; 2015). Four dimensions emerged: data-science offers: (a) *understanding,* and illuminates, explains, or provides insight into properties of soil by offering comparing and contrasting information or showing trends; (b) *utility* to address real-world problems in specific fields, using specific analytic techniques, by visualizing problems, saving time, and other vague applications; (c) *is vaguely useful or relevant* with unclear reasoning for why, and (d) that data-science is *not relevant* in their perspective.[1] A summary of the results are presented in SM, Table S1, and examples of student responses for each code in Appendix A. Overall, we found that the most common code was utility perceptions, noted by 74% of students at pretest, 68% at posttest, then understanding (34% pre, 35% post), followed by vague explanations (15% pre, 16% post), and "not relevant" (5% pre, 3.7% post).

[1] The two coders independently applied the codebook to the full dataset. Interrater reliability analysis revealed high percent agreement (average 91%) but low kappa values due to infrequent code use and data sparsity. Coders met to reconcile discrepancies and arrived at consensus-codes for all reported analyses.

The second way in which we coded the qualitative data was using LIWC 2022 (Linguistic Inquiry and Word Count) a validated automated text analysis program (Boyd et al., 2022). Specifically, we assessed students' use of words pertaining to *analytical thinking* dimensions, which demonstrate cognitive skill and engagement (see, e.g., Markowitz, 2023; Pennebaker et al., 2014) at pretest and posttest and compared them. Across the five dimensions of analytic cognitive processes (insight, causation, discrepancy, tentativeness, certitude, and differentiation), Wilcoxon signed rank tests revealed significantly lower levels of "discrepancy" language at posttest (p=.006, d= -.22), and significantly lower levels of "tentative- ness" at posttest (p=.006, d= -.19; see SM, Table S1 for means and SDs across analytic thinking).

RQ3: Relevant Soil-Variables Identified

We also used LIWC to analyze the list of variables that students provided at pretest and posttest that they perceived would be relevant in data-exploration. See SM, Table S2 for pre-post word counts, word clouds, and differences in frequencies. Comparing pre and post, it is evident that students became more specific in their language, salience of "moisture," "temperature, "pH," and other terms at posttest compared to pretest, while "soil" remains prominent across both. In a comparative analysis between pretest-and-posttest that highlights words that are more associated with pre vs post, we see that "animals" and "organisms" are more prominently related to the pretest, whereas more chemical and landscape factors are more prominent in the posttest.

Discussion

We sought to develop a learning intervention that promotes data-literacy skills as they apply to place-based soil microbiology learning experiences. We found that students significantly improved their perceptions of data science relevance and science knowledge from pretest to posttest, consistent with prior theory and evidence (Lombardi et al., 2016; Thacker et al., 2025).

Student's open-ended explanations revealed that students tended to explain data-science relevance in terms of its utility across medical and STEM professions and its capability to help scientists better understand properties of soil-data. This suggests that students are predisposed to expect data-tools to support their understanding of multidisciplinary topics and appear to be eager to learn and apply what they know to problem-solve in multiple domains.

Automated text analysis revealed that students used less discrepant and tentative language at posttest compared to pretest when discussing data relevance, indicating improved levels of certainty. We also found that students identified lists of soil-related variables that became more specific and specialized over time. Terms such as "animals" and "organisms" became less emphasized in place of more relevant microbiological properties such as "pH", "nitrogen," and "chemical." This suggests that students improved in terms of confidence and identification of relevant variables, both of which are critical for persisting in data-literate professions.

Generally speaking, our findings demonstrate that data-literacy supports combined with place-based learning experiences can dually enhance students' knowledge, motivation, and data-literacy in terms of identifying analytical variables (Lombardi et al., 2016; Semken et al., 2017; Börner et al., 2019; Carlson & Johnston, 2014). We encourage researchers and practitioners to ground data-analysis experiences in real world and place-based scenarios to help students understand the scientific and societal factors underlying what they observe.

Acknowledgements

This work is supported by the Hispanic Serving Institution Education Grant [2024-77040-43179/1032913], from the U.S. Department of Agriculture, National Institute of Food and Agriculture.

References

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician, 69*(4), 334-342.

Ben-Zvi, D., & Garfield, J. (2008). Introducing the emerging discipline of statistics education. *School Science and Mathematics, 108*(8), 355-361.

Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(6), 1857–1864. https://doi.org/10.1073/pnas.1807180116

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin.

Carlson, J., & Johnston, L. (2015). *Data information literacy: Librarians, data, and the education of a new generation of researchers.* Purdue University Press.

Charmaz, K. (2015). *Grounded theory. Qualitative psychology: A practical guide to research methods,* 3, 53-84.

Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist, 33*(2-3), 109-128.

Fry, R., Kennedy, B., & Funk, C. (2021). *STEM jobs see uneven progress in increasing gender, racial and ethnic diversity.* Pew Research Center Science & Society.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1-25.

Gosselin, D., Burian, S., Lutz, T., & Maxson, J. (2016). Integrating geoscience into undergraduate education about environment, society, and sustainability using place-based learning: three examples. *Journal of Environmental Studies and Sciences, 6*, 531-540.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25. http://iase-web.org/Publications.php?p=SERJ

Hurley, A., Chevrette, M. G., Acharya, D. D., Lozano, G. L., Garavito, M., Heinritz, J., ... & Handelsman, J. (2021). Tiny earth: a big idea for STEM education and antibiotic discovery. *MBio, 12*(1), 10-1128.

Kim, J., Hong, L., Evans, S., Oyler-Rice, E., & Ali, I. (2023). Development and validation of a data literacy assessment scale. *Proceedings of the Association for Information Science and Technology, 60*(1), 620-624.

Kokka, K. (2022). Toward a theory of affective pedagogical goals for social justice mathematics. *Journal for Research in Mathematics Education, 53*(2), 133-153.

Lombardi, D., Nussbaum, E. M., & Sinatra, G. M. (2016). Plausibility judgments in conceptual change and epistemic cognition. *Educational Psychologist, 51*(1), 35-56.

Markowitz, D. M. (2023). Analytic thinking as revealed by function words: What does language really measure?. *Applied Cognitive Psychology, 37*(3), 643-650.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Authors.

National Science Foundation (2015). *Science and engineering degrees, by race/ethnicity of recipients: 2002-12*. Arlington, VA.

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* Washington, DC: The National Academy Press.

Pennebaker, J. W. (2001). *Linguistic inquiry and word count: LIWC 2001*. Austin, TX: Pennebaker Conglomerates.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLOS One, 9*(12), e115844. https://doi.org/10.1371/journal.pone.0115844

Prado, J., & Marzal, M. A.´ (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri, 63*(2), 123–134.

Qiao, C., Chen, Y., Guo, Q., & Yu, Y. (2024). Understanding science data literacy: a conceptual framework and assessment tool for college students majoring in STEM. *International Journal of STEM Education*, *11*(1). https://doi.org/10.1186/s40594-024-00484-5

Ridsdale, C., Rothwell, J., Smith, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D., Matwin, S., & Wuetherick, B. (2015). *Strategies and best practices for data literacy education: Knowledge synthesis report.* Dalhousie University. https://dalspace. library.dal.ca/handle/10222/64578

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

Semken, S., Ward, E. G., Moosavi, S., & Chinn, P. W. (2017). Place-based education in geoscience: Theory, research, practice, and assessment. *Journal of Geoscience Education, 65*(4), 542-562.

Steen, L. A. (2001). Mathematics and numeracy: Two literacies, one language. *The Mathematics Educator, 6*(1), 10-16.
Thacker, I. (2023). Climate change by the numbers: Leveraging mathematical skills for science learning online. *Learning & Instruction, 86*, 101782. https://doi.org/10.1016/j.learninstruc.2023.101782
Thacker, I. (2024). Supporting secondary students' climate change learning and motivation using novel data and data visualizations. *Contemporary Educational Psychology, 78,* 102285. https://doi.org/10.1016/j.cedpsych.2024.102285
Thacker, I., French, H., & Feder, S. (2024). Estimating climate change numbers: Mental computation strategies that can support science learning. *International Journal of Science Education, 47*(1) 1–22. https://doi.org/10.1080/09500693.2024.2307473
Thacker, I., Shroeder, R., Shields-Menard, S., & Goforth, N. (2025). Dirt Don't Hurt: How relevant soil data can support learning and motivation at a Hispanic Serving Institution. *International Journal of Science and Mathematics Education, 23*, 803–826. https://doi.org/10.1007/s10763-024-10491-1
Thacker, I. & Sinatra, G. M. (2022). Supporting climate change understanding with novel data, estimation instruction, and epistemic prompts. *Journal of Educational Psychology, 114*(5), 910–927. https://doi.org/10.1037/edu0000729
Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. *Educational Studies in Mathematics, 96*(1), 33-47.