# SCAFFOLDING DATA SCIENCE CONCEPTS FOR FUTURE MATHEMATICS TEACHERS

Heather Gallivan
University of Northern Iowa
heather.gallivan@uni.edu

Eric Weber
Iowa State University
esweber@iastate.edu

*Data science education is becoming more prevalent as standards and coursework make their way into PreK-12 schools. Mathematics teacher preparation programs have a responsibility to prepare future teachers to teach data science. This project aimed to iteratively develop and evaluate a 4-week data science curriculum module for secondary preservice mathematics teachers, focusing on optimization models. The first iteration did not support preservice teacher learning of optimization concepts as anticipated. A scaffolded activity was created, which better supported the development of their understanding of this important data science concept.*

Keywords: Data Analysis and Statistics, Preservice Teacher Education

Data science education is getting more attention in the mathematics education field. The *PreK-12 Guidelines for Assessment and Instruction in Statistics Education* (GAISE) I and II reports (Bargagliotti et al., 2020; Franklin et al., 2007) as well as the National Council of Teachers of Mathematics and the American Statistical Association (2022; see also Dykema, 2024) have called for more of an emphasis on statistics and data science education for PreK-12 students. Given this, at least five states now have standards that are data science specific (e.g., California Department of Education, 2023). As a result, data science curriculum is already being developed and taught in K-12 schools (e.g. Gould et al., 2016; McNamara & Hansen, 2014). The Association of Mathematics Teacher Educators (AMTE, 2017) and *The Statistical Education of Teachers* document (Franklin et al., 2015) have pushed for more of a focus on the preparation of K-12 mathematics teachers to teach statistics topics. Middle and high school mathematics teachers will need to extend their knowledge of data analysis and statistics to deliver data science concepts in the classroom. Thus, mathematics teacher educators (MTEs) have a responsibility to prepare pre-service secondary mathematics teachers (PSMTs) to teach data science (Dykema, 2024; Gallivan & Weber, 2023).

The goal of our research project is to iteratively develop and evaluate the effectiveness of a 4-week data science module for PSMTs. The learning objectives for this 4-week module include data science content knowledge (e.g. understand the classification of data points in space) as well as required mathematical/statistical knowledge for teaching (Ball et al., 2008), such as analyzing and interpreting student work in data science. The purpose of this paper is to describe some of the materials we used to support PSMTs' learning of data science content, discuss the effectiveness of those materials on PSMTs' learning, and how we subsequently revised those materials to better scaffold the learning of PSMTs.

## Theoretical Framework

To situate our study, we broadly define "data science" as "the science of learning from data…" (Donoho, 2017, p. We conceptualize data science as the reversal of the scientific method to our students. Meaning, instead of starting with a statistical question to answer, we start with data that has been collected and reverse engineer the question (i.e. what question(s) does this data answer?). More specifically, we scaffold data science concepts over prior mathematical knowledge. For example, statistical and data science models are essentially functions we use to

Zbiek, R. M., Yao, X., McCloskey, A., & Arbaugh, F. (Eds.). (2025). *Proceedings of the forty-seventh annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Pennsylvania State University. https://doi.org/ 10.51272/pmena.47.2025

describe data. We will draw upon this idea of scaffolding towards new data science knowledge through the use of already known mathematical knowledge in this paper.

### Data Science and Statistical Content Knowledge of PSMTs

Due to the newness of data science in mathematics education, there is very little, if any research on PSMTs learning of data science concepts. However, because there is significant overlap of statistics, mathematics, and data science content (Drew, 2024), we draw upon what we do know about PSMTs' developing statistical content knowledge. Research suggests that PSMTs may not have the required content knowledge they need to teach statistics (Hannigan et al., 2013; Lovett & Lee, 2017). Combined with the new and developing field of data science, it can be extrapolated that PSMTs are likely to be similarly unprepared to teach data science concepts. It is imperative as MTEs that we develop curriculum materials that will support PSMTs' learning of data science concepts as well as their mathematical and statistical knowledge for teaching data science (Dykema, 2024; Gallivan & Weber, 2023).

### Mathematical and Statistical Knowledge for Teaching

To design the module, we drew upon the mathematical knowledge for teaching framework (MKT; Ball et al., 2008). MKT involves both subject matter knowledge and pedagogical content knowledge (Ball et al., 2008). Developing PSMTs' MKT is an essential element in their preparation for future mathematics teaching (e.g. Hill et al., 2005). Since mathematics and statistics are related but different fields (Cobb & Moore, 1997), a framework for statistical knowledge for teaching (SKT) was developed (Groth, 2007, 2013). For this project, we adopted features of MKT and SKT to support PSMTs in developing what we will ultimately call data science knowledge for teaching (DSKT). Our module includes goals for both content that PSMTs will teach (common content knowledge, specialized content knowledge; Ball et al., 2008; Groth, 2007) as well as content that goes beyond the high school curriculum (horizon knowledge; Ball et al., 2008; Groth, 2007).

## Data Science Module

To develop PSMTs' knowledge of data science content, we created a 4-week self-contained, "drop-in" module for content courses taken by PSMTs. The framework of the content includes incorporating horizon knowledge (Ball et al., 2007; Groth, 2007) to better situate PSMTs in a data science context. This horizon knowledge builds upon standard mathematics and statistics concepts within mathematics teacher preparation programs, including linear regression models, projections, and elementary optimization routines. Our content learning outcomes include visualization and exploration of data (Outcome A), prediction (B), communication (C), and optimization models (D). These content learning outcomes form the foundation of the module lessons and activities. For this paper, we focus on PSMT learning of Outcome D: Optimization.

## Methods

The data science curriculum materials were implemented in two different content courses for PSMTs at two different universities in the Midwest in Spring of 2023 and 2024. One course was a statistics content course, and the other course was a general topics course. In year one, the courses had 13 and 6 PSMTs, respectively, who participated in the data science module. In year 2, the courses had 20 and 12 PSMTs. We will discuss the evaluation of our 4-week module for both iterations, specifically as it pertains to optimization (Outcome D).

### Data Collection and Analysis

We measured the PSMTs' data science content knowledge through a pre-post test. During the first iteration, the pre-post test consisted of three open response tasks that addressed each of the

four learning objectives. The first task assessed student understanding of Outcomes A, B, and C through linear regression models. To assess their understanding of Optimization (Outcome D), one task required PSMTs to classify animal pathway data using an optimization algorithm (described below). The final task required students to consider a modified traveling salesman problem. During the second iteration of the module, we revised the pre-post test based on the results from the first iteration and changes we made to the module lessons and activities. The first task on the pre-post test was unchanged. The animal pathways task was revised with more scaffolding for PSMTs to encourage students to demonstrate their knowledge of optimization algorithms. We did not include the traveling salesman task as part of the second version of the pre-post test as we incorporated it into the module lessons.

A rubric was created for each task to score the participants' tests. The animal pathway task from the first iteration did not get scored since we felt that the responses (and lack thereof) from the PSMTs were better analyzed using the constant comparative method of qualitative analysis (Merriam & Tisdell, 2016). The authors independently scored all of the pre-post tests and resolved any disagreements collaboratively. Paired samples *t*-tests were utilized to determine whether there was a significant change in the PSMTs' data science content knowledge from before the data science module was implemented and after.

## Data Science Module Iterations and Results

In this section, we will discuss details of the first and second iterations of the data science module as it pertains to Outcome D: Optimization.

### First Iteration of Data Science Lessons

The first iteration of the data science module included an introduction to data science concepts through a familiar statistics topic, linear regression models, as this is a well-known optimization problem (i.e. minimizing the sum of the squared residuals). Then, we introduced concepts around data visualization and classification of points in space (Outcome A) through visual inspection. We explored these concepts through the context of prized animals wandering in a field towards a food source. Much of the discussion in these first lessons were on the decisions PSMTs were making to classify the data points based on different attributes ($x$ and $y$ coordinates, time, direction vectors, etc.) as well as how a computer program might be coded to do the classification. PSMTs considered how to classify the data points through visual inspection. The next lesson in the sequence required PSMTs to use an Excel spreadsheet with a similar set of data points to implement an optimization routine utilizing projections (a topic familiar to PSMTs) to determine the classification. Finally, PSMTs briefly looked at "solutions" for the animal pathway classification utilizing different optimization models; interpreting and evaluating their solutions in context.

### Results from the First Iteration

Overall, while there is evidence that the PSMTs demonstrated learning some data science concepts overall by the post-test ($t = 8.320$, $df = 18$, $p < .001$), they did not perform as well on the direct measures of their understanding of optimization models (Outcome D). For example, they did not perform significantly better on the post-test on the task that asked them to consider the traveling salesman problem to find the shortest pathway between five coordinate points ($t = 1.931$, $df = 18$, $p = .069$). While we anticipated that they would suggest finding all possible pathways, none of the PSMTs could give a satisfactory answer to how they would find the shortest pathway. On the post-test for the animal pathways task, many of the PSMTs (42.1%) did not respond to the task at all. Of those that responded (11 of the 19 PSMTs), three PSMTs assigned points to animal pathways with no explanation of how they did so (27.3%). Four

PSMTs attempted to classify the animal pathways using visual inspection. While three PSMTs attempted to use an optimization algorithm for the solution, only one of them provided their calculations and a complete explanation for what they had done.

### Second Iteration of Data Science Lessons

Due to these results, we developed a new lesson and revised some of the previous lessons to scaffold student understanding. The second iteration of the module was the same as the first, with the scaffolded material incorporated prior to the lesson on using projections to classify points. Our intent was to allow PSMTs to engage with the data science concept of optimization using a more familiar mathematical topic first; namely, Euclidean distance.

The scaffolded lessons were a modified version of the traveling salesman problem. For our lesson, we gave students a set of 5 coordinate points and asked them to find the shortest possible pathway using Euclidean distance by hand and/or with a basic calculator. We had them discuss strategies they used to find the shortest pathway, hoping they would suggest finding the length of all possible pathways (i.e. a brute force algorithm) or using the idea of finding the next closest point (i.e. a "greedy" algorithm). We want PSMTs to discover: 1) through finding the total number of possible pathways for $n$ points, that finding all possible pathways by hand was going to become untenable, even for a computer, and 2) that the greedy algorithm does not always provide the shortest pathway, but it can be a good approximation (i.e. a model does not need to be perfect to be useful). Then, we move back to the animal pathway data and have PSMTs use the greedy algorithm with Excel spreadsheets to classify the points as belonging to a particular animal using: 1) Euclidean distance; and 2) projections in distinct prediction models.

### Results from the Second Iteration

Preliminary results from the second iteration suggest that PSMTs developed a stronger grasp of the mathematical concepts surrounding the use of optimization models for data classification ($t = 8.786$, $df = 30$, $p < .001$). Specifically, many PSMTs (77.4%) demonstrated some understanding of how to classify points through visual inspection by the post-test when relatively few PSMTs (25.8%) were able to do this on the pre-test. Many PSMTs developed some understanding of how to utilize optimization algorithms by the post-test ($t = 5.916$, $df = 30$, $p < .001$). Many of these PSMTs used the Excel spreadsheets provided during coursework to apply at least one optimization algorithm. While this is promising, several of the PSMTs (22.5%) claimed to still not be able to use such algorithms to classify the animal pathway data. However, this is an improvement from the 42.1% of PSMTs who didn't respond to this task in the first iteration. Overall, the PSMTs demonstrated a better understanding of optimization (Outcome D) during the second iteration of the data science module.

## Discussion and Conclusions

As data science coursework becomes more popular in K-12 schools, mathematics teacher educators need to consider how to best prepare PSMTs to teach data science. Through the development of our module, we determined that we could leverage PSMTs' mathematical knowledge as a scaffold for them to learn foundational data science content. The results of this study will be used to further improve the 4-week module over time as well as develop the module for dissemination to other mathematics teacher educators to use with their PSMTs.

## Acknowledgments

This project/material is based upon work supported by the Iowa Space Grant Consortium under NASA Award No. 80NSSC20M0107 and the National Science Foundation under awards #1830254, #2152117, and #2219959.

## References


Association of Mathematics Teacher Educators. (2017). *Standards for preparing teachers of mathematics*. Association of Mathematics Teacher Educators. amte.net/standards

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59(*5), 389-407. https://doi.org/10.1177/0022487108324554

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 guideline for precollege statistics and data science education*. National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

California Department of Education. (2023). *Mathematics framework for California Public Schools: Kindergarten through grade twelve (Mathematics Framework)*.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104(*9), 801–823.

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Drew, D. E. (2024). The limits of data science. *Journal of Humanistic Mathematics, 14(*1), 305-315.

Dykema, K. (2024, April 4). *The importance of data science* [Press release]. National Council of Teachers of Mathematics. https://www.nctm.org/News-and-Calendar/Messages-from-the-President/Archive/Kevin-Dykema/The-Importance-of-Data-Science/

Franklin, C., Bargagliotti, A., Case, C. A. Kader, G. D., Scheaffer, R. L., & Spangler, D. A. (2015). *Statistical education of teachers (SET)*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/EDU-SET.pdf

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: A pre-K-12 curriculum framework*. American Statistical Association. http://www.amstat.org/education/gaise/GAISEPreK12_Full.pdf

Gallivan, H. R., & Weber, E. (2023). Artificial intelligence, data literacy, and preservice mathematics teacher training. *Iowa Council of Teachers of Mathematics.* Retrieved from: https://iowamath.org/Articles/13285066

Groth, R. E. (2007). Towards a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education, 38(*5), 427-437.

Groth, R. E. (2013). Characterizing key developmental understandings and pedagogically powerful ideas within a statistical knowledge for teaching framework. *Mathematical Thinking and Learning*, *15*, 121–145.

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., Tangmunarunkit, H., Trusela, L., & Zanontian, L. (2016). Teaching data science to secondary students: The mobilize introduction to data science curriculum. In J. Engel (Ed.), *Promoting understanding of statistics about society. Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE)*. https://iase-web.org/documents/papers/rt2016/Gould.pdf

Hannigan, A., Gill, O., & Leavy, A. M. (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education*, *16*, 427-449.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42(*2), 371-406.

Lovett, J. N. & Lee, H. S. (2017). New standards require teaching more statistics: Are preservice secondary mathematics teachers ready? *Journal of Teacher Education, 68(*3), 299-311.

McNamara, A. & Hansen, M. (2014). Teaching data science to teenagers. In K. Makar, B. deSousa, & R. Gould (Eds.), *Sustainability in Statistics education. Proceedings of the 9th International Conference on Teaching Statistics (ICOTS9, July 2014)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

National Council of Teachers of Mathematics & American Statistical Association. (2022). *Joint NCTM-ASA position statement on preparing PK–12 teachers of statistics and data science* [Position statement]. https://www.nctm.org/Standards-and-Positions/Position-Statements/Preparing-Pre-K-12-Teachers-of-Statistics