CONTEXT, DATA, AND ASSUMPTIONS: THEIR ROLES IN MATHEMATICAL MODELING AND DATA SCIENCE

Sindura Kularajan[1], Cynthia Anhalt[2], Brynja Kohler[1]

[1]Utah State University, [2]University of Arizona

In this conceptual paper, we share the intersections of mathematical modeling and data science through examining the role of context, data, and assumptions within these mathematical processes. We argue, in modeling context is central to decontextualizing and recontextualizing, while in data science it drives creation and connection. In modeling, assumptions guide model formulation, but must be checked before choosing analytical methods in data investigations. In modeling, data motivates and validates models, whereas they permeate every stage of the data science cycle. We discuss three major contributions: (1) a theoretical perspective for distinguishing between modeling and data science; (2) implications for modeling and data science task design within a shared context, and (3) implications for teacher-education.

INTRODUCTION

Mathematical modeling (hereafter referred to as modeling) and data science represent central practices in contemporary mathematics education. Unlike traditional rote-based mathematics, both modeling and data science practices require students to apply their mathematical, statistical, and real-world knowledge to describe, predict, and address real-world phenomena. Thus, both modeling and data science have been successful in addressing the disparity between mathematics taught in school and its relevance in the real world. Despite their pedagogical power in making mathematics meaningful to learners, the interconnectedness of these mathematical processes has received little to no attention. While connections among distinct mathematical domains (e.g., geometry, algebra) have long been emphasized for students' mathematical proficiency (National Research Council, 2001), we make the case that mathematical processes such as data science and modeling should not be taught in silos if the goal is to cultivate students' competencies in mathematical processes. Specifically, when students have experience in one of these processes, instruction can intentionally draw on that knowledge base to support learning in the other. Attending to such connections may provide important implications for the teaching and learning of modeling and data science across K–16 contexts. Accordingly, this work is motivated by a close examination of the parallels and distinctions between these two processes. In this conceptual piece, we pose and address the following question: *Within a scientific inquiry, to what extent are the intersecting mathematical processes of modeling and data science similar and/or different?* We take up this question by examining the roles of data, assumptions, and context across modeling and data science.

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 875-882. PME.

HISTORICAL PERSPECTIVES ON MATHEMATICAL PROCESSES

Problem solving and reasoning have been at the core of historical curriculum recommendations and standards for teaching and learning mathematics in the United States. For example, the National Council of Teachers of Mathematics' *Principles and Standards for School Mathematics* (2000) emphasizes problem solving, reasoning and proof, representation, connections, and communication for mathematics learning. More recently, the *Common Core State Standards in Mathematics* (2010) reaffirms the importance of mathematical processes and practices with a focus on problem solving, reasoning, and modeling. The *Guidelines for Assessment and Instruction in Mathematical Modeling Education* (Garfunkel & Montgomery, 2019) makes an explicit effort in emphasizing the utility and value in demonstrating how mathematics can analyze and guide decision making for real-world problems, thereby influencing mathematics education to consider a more holistic approach to problem solving by including data, analysis of relationships and quantities, models, computation, interpretation, validation, and communication of the solution. More recently, the *Guidelines for Assessment and Instruction in Statistics Education II* (ASA, 2020) underscores the rise of big data and technological advances in computing, and societal needs in data literacy demands changes in mathematics curricula to include data science. Thus, mathematical modeling and data science have gained a prominent place in learning mathematics in K-12 and undergraduate education.

THEORETICAL PERSPECTIVES ON THE MATHEMATICAL PROCESSES

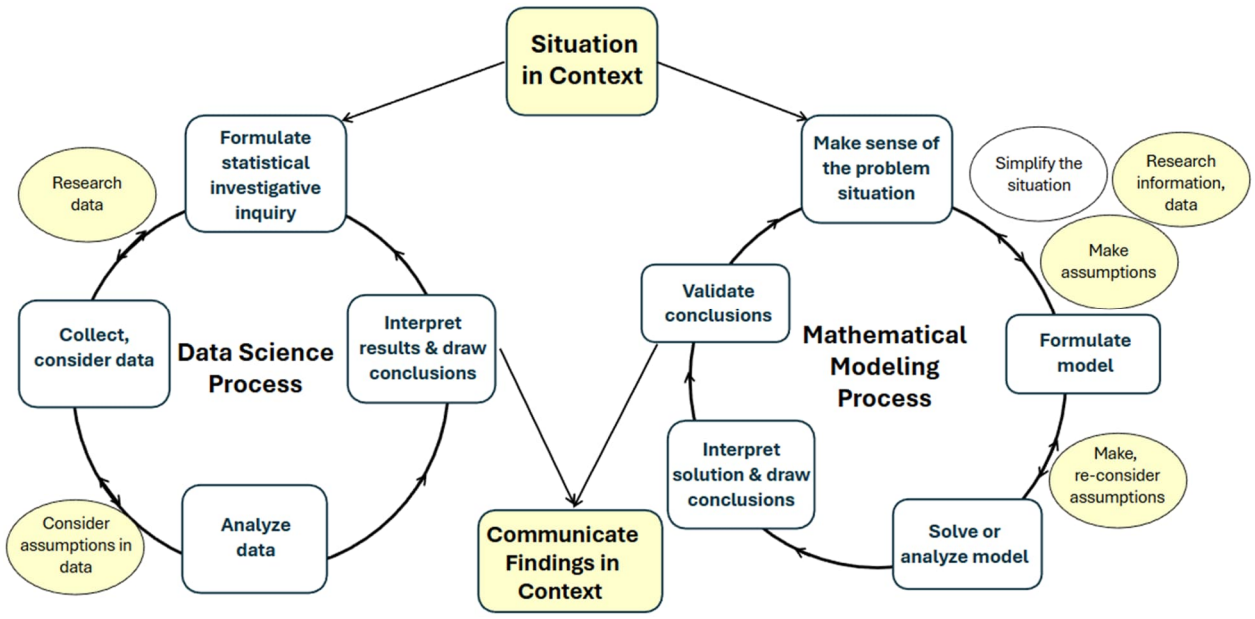
Mathematical modeling is the process of using mathematics to represent, describe, and predict real-world situations. To attain this, it is necessary to (1) make sense of the problem situation; (2) make simplifying assumptions about the situation; (3) formulate a model; (4) analyze the model; (5) interpret the solutions and draw conclusions; (6) validate the conclusions; and (7) communicate the findings to appropriate stakeholders. Thus, engaging in modeling requires frequent transitions between a modeler's mathematical knowledge and real-world knowledge, requiring the modeler to de-contextualize and re-contextualize the situation. Thus, the process of modeling can be conceptualized as a continuous feedback loop between de-contextualizing the situation and re-contextualizing the mathematical results.

Data science is an interdisciplinary field that connects mathematics and statistics, computer science, and real-world applications (e.g., social science, medicine) (Hazzan & Mike, 2023). That is, data scientists draw knowledge from these major domains to engage in data science tasks. While doing so, it is necessary to, (1) formulate statistical investigative questions about the real-world phenomenon; (2) collect new data or use existing data and engage in data cleaning, management, and storage; (3) retrieve the stored data to conduct mathematical and statistical analysis in order to make inferences about the data; and (4) interpret and draw conclusions with respect to real-world considerations (Bargagliotti et al., 2020; Baumer, 2015). In **Error! Reference source**

not found., we provide a schematic diagram that highlights the parallels and intersections of the mathematical modeling and data science processes.

In the remainder of this paper, we theorize the roles of context, assumptions, and data, positioning them as *prominent elements* across both mathematical processes. Our focus on these elements is motivated by two considerations. First, context, assumptions, and data are foundational to both processes. Second, although these elements are deeply interwoven within each process, they function in distinct, and at times overlapping, ways across the two. Thus, to advance their use as theoretical and pedagogical constructs, we examine these elements separately and articulate how each informs and constrains the modeling and data science processes.

Figure 1 Parallels and intersections in mathematical modeling and data science processes



THE PROMINENT ELEMENTS: CONTEXT, ASSUMPTIONS, AND DATA

Role of context

Real-world contexts serve as powerful motivators for student learning (Czocher et al., 2021; Dunn & Marshman, 2019; Neumann et al., 2013). In modeling, the modeler relies on context to make sense of and simplify the problem situation, research necessary information, and make assumptions that lead toward formulating a model. In other words, the *context is explicit* and has impact early in the process while determining the assumptions made, quantities conceived, parameters identified, and the mathematics that organizes them, acting as the *driver* in model development. However, the model development process entails *decontextualization*, which is the removal of contextual aspects from the objects. In essence, the real-world context is separated from the mathematics that is used to represent it. This separation of the real-world situation from mathematics is grounded in the epistemological belief that, within

modeling, one must both *decontextualize* and *recontextualize* within the process. Once the model has been developed and tested, *recontextualization* re-attaches the contextual aspects to the epistemic objects to interpret results, draw conclusions, and validate the outcomes of the model.

On the other hand, the data science process requires reasoning and making sense of the context in the presence of variability. In consideration of the data science process activities, we discuss the role of context through the three major activities in the data science process, consisting of *creating*, *connecting*, and *computing* (Hicks & Irizarry, 2018). We can argue that contextualization exists only within the connecting and creating activities, and not within the computing activities. *Creating* activities involve formulating questions and hypothesis and generating the dataset if applicable within the context of the situation inquiry. *Connecting* activities involve connecting the investigative inquiry with real-world data sets and analytical tools and interpreting the results, and further, integrating mathematics, statistics, or computer science methodologies for approaching a data science inquiry. Finally, *computing* activities include data analyses that require mathematical and statistical theory and implementation of algorithms. In computing, because data are the objects of analysis, and therefore, an abstraction of the context, *decontextualization* is a natural consequence. In summary, the context acts as a *mechanism* for evaluating, interpreting, and creating data sets. With data being context-embedded, the *context is implicit* in the data science process.

Role of data

Within modeling, data is predominantly used as a *tool for investigating* phenomena. However, the role of data in modeling can vary depending on the operationalization of data and the type of modeling. If data is operationalized as any *information* given for the investigative phenomena, the data initiates the scientific inquiry. If data is operationalized as any *artefacts* that are generated as a result of engaging in scientific inquiry, then data and modeling are inseparable. Finally, if data is operationalized as *empirical observations* researched/collected with the purpose of formulating a model, then data plays a role in the beginning phases of the modeling, such as making sense of the situation and formulating a model, and during model validation. That is, a modeler can collect or consider data to assess the reasonableness of the model they constructed.

At the same time, the role of data can vary depending on the type of modeling. Within the modeling literature, three types of modeling have been identified: theoretical, application, and data modeling (Niss et al., 2007). In theoretical modeling, a modeler constructs models from first principles, during which the modeler forms correspondence between a mathematical representation and a scientific situation (Greca & Moreira, 2002). In application modeling, the goal is to use existing, well-established models to mathematically solve a given phenomenon. Finally, in data modeling, a modeler is responsible for finding the model that fits existing data (e.g.,

curve fitting). In the latter case, data *constrain the investigation*, and parameter values emerge as findings.

In data science, data are the *objects of investigation*. Meaning, data scientists study the patterns and (in)variances exhibited by the data and use appropriate visual representations to represent them. Additionally, data scientists may also test any hypothesis they might have about the phenomena using inferential statistics. Thus, the entire scientific inquiry within data science is motivated by data that already exists or data that needs to be collected to address questions of interest. Simply said, the data science process is nonexistent without data—data is embedded within all phases of the data science process shown in Figure 1.

Role of assumptions

From a perspectivist viewpoint, both mathematical and statistical models provide partial representations of reality rather than true depictions of it. Each offers a particular way of seeing, shaped by its assumptions and purposes (Zhao, 2025). Learning to make, articulate, test, and critique assumptions is therefore a critical epistemic practice.

In modeling, assumptions are actively made as part of the *mathematization* of a real-world situation (i.e., organization of a situation into a mathematical representation). The assumptions simplify reality, and in a very real sense, define the model itself. They capture the modeler's decisions about which variables and parameters matter, what interactions are represented, and which mathematical structures—such as functions, differential equations, or interaction networks—are used to simulate known mechanisms. As models are formulated, additional implicit assumptions often arise, such as treating populations as well-mixed in space or assuming processes occur in continuous time. While such assumptions are frequently known to be false in a literal sense, they enable the translation from reality to a mathematical system that can be analyzed and interpreted back in context. Scientists knowingly accept idealized versions of reality for specific purposes.

In data science, assumptions play a different role. Rather than being freely made to enable model construction, assumptions must be checked against the data before analysis proceeds. Many statistical and machine learning methods rely on assumptions about data structure or distribution, and their conclusions are only meaningful when those assumptions are reasonably satisfied. Consequently, data cleaning, diagnostic testing, and validation are central practices in data science (Flatt & Jacobs, 2009; Osborne, 2013). Imperfect, biased, or non-representative datasets can lead to misleading results, making assumptions a potential source of error rather than a generative modeling tool.

DISCUSSION AND CONCLUSIONS

We see three primary contributions of our paper: (1) theoretical contribution regarding the nature of inquiry of modeling and data science to educational research; (2)

implications for the teaching and learning of modeling and data science; and (3) implications for teacher-education in modeling and data science.

First, we argue that differences in the nature of inquiry across modeling and data science arise from how prominent elements function within each process. In both processes, *context* acts as the primary driver of investigation—it initiates and shapes the investigation; however, it does so in distinct ways. In data science, context generates or directs attention toward data sources and analytic choices, whereas in modeling, context affords and constrains the assumptions embedded in the model. Thus, depending on which function of context is emphasized—data collection/selection or assumption-making—the inquiry follows a trajectory toward data science or toward modelling, respectively. Similarly, the role of *assumptions* can determine the trajectory of the scientific inquiry. When assumptions play a passive role, where they are merely verified against data, it aligns the inquiry with a data-driven approach, characteristic of data science. In contrast, when assumptions play an active role, requiring deliberate construction and integration into the formulation of a model, the inquiry becomes model-driven, emphasizing characteristics of modeling. This distinction underscores how the epistemic status of assumptions informs whether the focus is on extracting patterns from data or constructing explanatory models. Finally, when *data* serves as an abstraction of real-world phenomena on which the investigation is conducted, the inquiry is data-driven. In contrast, when data function primarily as a tool to formulate and validate the mathematical abstraction of the real-world phenomena, the inquiry is model-driven.

Second, our arguments have important implications for the teaching and learning of modeling and data science. As modeling and data science increasingly take precedence in K–16 curricula, implementation often outpaces the development of meaningful and time-effective task design. We argue that our findings can be leveraged to design both modeling and data science tasks within a shared context by systematically varying prominent elements such as data availability and assumptions. For example, consider an inquiry focused on CO_2 emissions (e.g., Little, 2019). Providing students with an existing dataset (e.g., from National Oceanic and Atmospheric Administration) affords the design of a data science task centered on analysis and interpretation. In contrast, withholding data or sufficient information and asking students to estimate or model CO_2 emissions supports the design of a modeling task. In this way, students develop tools for discerning which mathematical process is appropriate for a given problem situation while simultaneously building content-specific knowledge. Thus, we make the case that mathematics educators' knowledge, insight, and understanding of the roles of prominent elements in the modeling and data science processes will influence their ability to engage in effective task design.

Third, we aim to address implications for K-12 pre-service teachers (PST) and in-service teachers (IST). We posit that the intricacies of both processes used by scientists must necessarily be unpacked to understand the nature of the scientific processes so

that the mathematics education community may integrate them more effectively into the mathematical preparation and professional development of teachers. Indeed, researchers have found that PSTs and ISTs who experience the modeling and data science processes as learners themselves can simultaneously develop pedagogical knowledge in the modeling and data science processes (Anhalt et al., 2018; Lee et al., 2020). Thus, we argue that by engaging PSTs and ISTs deeply in modeling and data science activities immersed in meaningful *data*, relevant *contexts*, and practical *assumptions*, PSTs and ISTs can utilize their mathematical and data science knowledge with creativity and purpose. Such experiences enriched with pedagogical methods have the potential to advance the transformation of modeling and data science competencies into professional competencies for teaching modeling and data science.

In summary, this paper highlighted the roles of context, assumptions, and data in modeling and data science. We believe these insights extend beyond educational settings. Attending to how context frames inquiry, how assumptions inform the interpretation and analysis, and how data constrain and validate claims have implications not only for the teaching and learning of mathematical processes, but also for the broader scientific enterprise, where analytical transparency, ethics, and reproducibility are expected. We conclude that these prominent elements underscore how scientific knowledge is generated and shared through these processes.

References

American Statistical Association (ASA) (2020). Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISEII): A Framework for Statistics and Data Science Education. American Statistical Association. ISBN 978-1-7342235-1-4

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). The Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II). American Statistical Association.

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

Common Core State Standards Initiative (2010). Common Core State Standards in Mathematics. Washington D.C.: National Governors Association Center for Best Practices and the Council of Chief State School Officers. http://www.corestandards.org

Czocher, J., Melhuish, K., & Kandasamy, S. (2019). Building mathematics self-efficacy of STEM undergraduates through mathematical modelling. *International Journal of Mathematical Education in Science and Technology*, 51. 1-28. https://doi.org/10.1080/0020739X.2019.1634223.

Dunn, P., & Marshman, M. (2020). Teaching mathematical modelling: A framework to support teachers' choice of resources. *Teaching Mathematics and its Applications: An International Journal of the IMA, 39*, 127-144. https://doi.org/10.1093/teamat/hrz008.

Flatt, C., & Jacobs, R. L. (2019). Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical Reporting of Imperfect Data Sets. *Advances in Developing*

Human Resources, 21(4), 484-502. https://doi.org/10.1177/1523422319869915 (Original work published 2019)

Garfunkel, S., & Montgomery (Eds.) (2019). *GAIMME: Guidelines for assessment and instruction in mathematical modeling education, second edition.* Consortium for Mathematics and Its Applications (COMAP) and Society for Industrial and Applied Mathematics (SIAM).

Greca, I. M., & Moreira, M. A. (2002). Mental, physical, and mathematical models in the teaching and learning of physics. *Science education, 86*(1), 106-121.

Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. *The American Statistician, 72*(4), 382-391. https://doi.org/10.1080/00031305.2017.1356747

Hazzan, O., and Mike, K. (2023), "What is Data Science?" in Guide to Teaching Data Science: An Interdisciplinary Approach, pp. 19–34, Springer. https://doi.org/10.1145/3575663

Lee, H. S., Mojica, G. F., and Lovett, J. N. (2020), "Examining How Online Professional Development Impacts Teachers' Beliefs about Teaching Statistics," *Online Learning*, 24, 5–27. DOI:10.24059/olj.v24i1.1992.

Little, J. B. (2019). *Modeling and data analysis: an introduction with environmental applications* (Vol. 120). American Mathematical Society.

National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematics.* Reston, VA, USA: NCTM.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. In J.Kilpatrick, J. Swafford, and B.Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. National Academy Press.

Neumann, D., Hood, M., & Neumann, M. (2013). Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal, 12*(2), 59-70. https://doi.org/10.52041/serj.v12i2.304.

Niss, M., Blum, W., & Galbraith, P. (2007). Introduction. In W. Blum, P. L. Galbraith, H.-W. Henn & M. Niss (Eds.), *Modelling and applications in mathematics education* (pp. 1–30). Springer.

Osborne, J. W. (2013). Is Data Cleaning and The Testing of Assumptions Relevant in the 21st Century? *Frontiers in Psychology, 4*(370). https://doi.org/10.3389/fpsyg.2013.00370

Watson, J., Callingham, R., & Nathan, E. (2009). Probing teachers' pedagogical content knowledge in statistics: "How will Tom get to school tomorrow?" In R. Hunter, B. Bicknell, & T. Burgess (Eds.), *Crossing divides* (pp. 563-570). MERGA.

Zhao, K. (2025). What Are Statistical Assumptions About? An Answer From Perspectivism. *Harvard Data Science Review*, 7(1). https://doi.org/10.1162/99608f92.7fcd521d